



Universiteit
Leiden
The Netherlands

Open-world continual learning via knowledge transfer

Li, Y.

Citation

Li, Y. (2026, January 27). *Open-world continual learning via knowledge transfer*. Retrieved from <https://hdl.handle.net/1887/4287955>

Version: Publisher's Version

License: [Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden](#)

Downloaded from: <https://hdl.handle.net/1887/4287955>

Note: To cite this publication please use the final published version (if applicable).

Open-world Continual Learning via Knowledge Transfer

Proefschrift

ter verkrijging van
de graad van doctor aan de Universiteit Leiden,
op gezag van waarnemend rector magnificus,
volgens besluit van het college voor promoties
te verdedigen op dinsdag 27 januari 2026
klokke 11:30 uur

door

Yujie Li

geboren te Ganzhou, China

in 1996

Promotores:

Prof.dr. M.M. Bonsangue

Prof.dr. X. Yang (Southwestern University of Finance and Economics)

Promotiecommissie:

Prof.dr. M.S. Lew

Prof.dr.ir. F.J. Verbeek

Dr. D.M. Pelt

Prof.dr. T. Zhong (University of Electronic Science and Technology of China)

Dr. L. Huang (Southwestern University of Finance and Economics)



**Universiteit
Leiden**
The Netherlands



This thesis was performed at the Theory Group, Leiden Institute of Advanced Computer Science, Leiden University, The Netherlands.

Yujie Li was partially supported through by China Scholarship Council (CSC) to participate in the PhD program of Leiden University with grant number 202306980040.

Copyright © 2025 Yujie Li. All rights reserved.

ISBN 978-94-6522-966-9

To my grandmother, Peng Chenglan,
who taught me the meaning of love, forever in my heart.

Abstract

This thesis presents a systematic investigation into Open-world Continual Learning (OWCL), a newly emerging paradigm that fundamentally redefines how intelligent systems learn in non-stationary and uncertain environments. Rather than being a mere combination of continual learning and open-set recognition, OWCL establishes a new research problem that addresses the simultaneous needs for continuous adaptation and open-world awareness. It envisions learning agents that not only retain and transfer knowledge across evolving tasks, but also autonomously identify, characterize, and incorporate novel information beyond previously known categories. In doing so, OWCL departs from the traditional closed-world assumption underlying most deep learning frameworks and moves toward a truly dynamic, self-evolving learning paradigm. The overarching goal of this thesis is to construct the theoretical, methodological, and practical foundations of OWCL, thereby advancing the development of lifelong, adaptive, and trustworthy artificial intelligence.

The thesis begins by formalizing OWCL through a rigorous mathematical definition and identifying four representative real-world scenarios that capture its essential characteristics. This formalization provides a principled framework for understanding the inherent openness and non-stationarity of real-world data, distinguishing OWCL from traditional continual learning paradigms. Through extensive empirical analyses, the study reveals that OWCL is defined not merely by sequential learning but by its continuous exposure to the unknown—posing challenges related to open-set boundary expansion, knowledge retention, and adaptive generalization. These findings lay the groundwork for addressing the stability–plasticity dilemma within an open and evolving context.

Building upon this conceptual foundation, the thesis advances a knowledge transfer-centric perspective as a unifying principle for solving the challenges of OWCL. Within this framework, three novel models are introduced to address complementary aspects of continual adaptation. The first, Pro-KT, develops a prompt-based mechanism for representing and transferring both task-specific and task-agnostic knowledge. By dynamically encoding and reusing knowledge across tasks, Pro-KT enhances the system’s ability to maintain adaptability while preserving previously learned information, achieving a balanced trade-off between stability and plasticity. Extending this direction, PEAK explores the few-shot open-world continual learning

setting, where labeled data are scarce and task boundaries are ambiguous. Through a joint architecture designed to mitigate representation collapse and boundary uncertainty, PEAK demonstrates strong resilience in learning under limited supervision while maintaining effective open-set detection. (Chapter 4 & Chapter 5)

To provide a unified theoretical and architectural foundation for OWCL, the thesis introduces HoliTrans, a holistic framework that integrates open-set recognition with continual learning through a shared decision space. By analyzing the intrinsic coupling between open-set dynamics and task progression, HoliTrans derives optimization principles that govern effective learning in open environments and embeds them into a transferable model design. This framework not only achieves state-of-the-art empirical performance but also offers provable guarantees on decision boundary stability and knowledge preservation, bridging the gap between theoretical understanding and algorithmic implementation. (Chapter 6)

Beyond theoretical and algorithmic contributions, this thesis demonstrates the practical value of OWCL through an applied study in cross-regional fraud detection. Real-world financial systems often face the dual challenges of distributional shifts and evolving fraudulent behavior across regions. To address these, the CCL framework is proposed, formulating fraud detection as a continual graph learning problem. CCL combines prototype-based knowledge replay, parameter smoothing, and heterogeneous graph modeling to adapt to region-specific fraud patterns while retaining critical knowledge from previous distributions. Experimental results show that CCL not only mitigates catastrophic forgetting but also enhances cross-regional adaptability, thus validating the real-world relevance and scalability of OWCL principles. (Chapter 7)

Collectively, the four research questions explored in this thesis chart a coherent trajectory from theoretical formulation to holistic methodology and real-world validation. The contributions of Pro-KT, PEAK, HoliTrans, and CCL together form an evolving narrative of how knowledge can be accumulated, transferred, and generalized in open, uncertain, and continually changing environments. This work establishes open-world continual learning as a distinct and necessary paradigm for next-generation artificial intelligence—one that is capable of embracing novelty, uncertainty, and continual evolution. The findings of this thesis not only deepen our theoretical understanding of open learning systems but also open new avenues for future research in lifelong, knowledge-centric, and robust AI.

Contents

1	Introduction	1
1.1	Background	1
1.2	Research Questions and Contributions	7
1.3	Thesis Outline	12
2	Literature Review	15
2.1	Continual Learning	15
2.2	Open-Set Learning	17
2.3	Out-of-Distribution Detection	17
2.4	Open-world Continual Learning	18
2.5	Concluding Remarks	20
3	Preliminaries	23
3.1	General Definitions	23
3.1.1	Continual Learning	23
3.1.2	Open-Set Learning and Out-of-Distribution Detection	25
3.2	Problem Definition	27
3.3	Four OWCL Scenarios	28
3.4	Pretrained Models in OWCL: A Brief Overview	30
3.4.1	Attention Mechanism and Transformer Architecture	31
3.4.2	Prompt-based CL Methods with PEFT Techniques	32
3.4.3	Other PTMs-based CL Methods with PEFT Techniques	35
4	Learning to Prompt Knowledge Transfer for Open-world Continual Learning	37
4.1	Objectives of OWCL	40
4.2	Methodology	41
4.2.1	Prompt Bank	42
4.2.2	Task-Aware Open-Set Boundary	44
4.2.3	The Overall Objective and Algorithms	45
4.3	Evaluation	48
4.3.1	Experimental Results and Analysis	51
4.3.2	Additional Results	53
4.4	Summary	56

5	Improving Open-world Continual Learning under the Constraints of Scarce Labeled Data	59
5.1	OFCL Problem Definition	62
5.2	Methodology	63
5.2.1	Instance-wise Token Augmentation	64
5.2.2	Margin-based Open Boundary	66
5.2.3	Adaptive Knowledge Space	68
5.2.4	Overall Objective and Process	69
5.3	Evaluation	71
5.3.1	Experimental Setup	71
5.3.2	Main Results	73
5.3.3	Ablation Studies	75
5.3.4	Additional Results	77
5.3.5	Visualization and Analysis	79
5.4	Summary	79
6	Exploring Open-world Continual Learning with Knowns-Unknowns Knowledge Transfer	81
6.1	Toward Holistic Knowledge Transfer in OWCL	82
6.2	Rethinking OWCL: A Scenario-Based Framework	86
6.2.1	Problem Formulation	86
6.2.2	Four OWCL Scenarios	88
6.2.3	Experiments on All OWCL Scenarios	90
6.2.3.1	Experimental Findings Summary	92
6.3	Theoretical Bases: Open Risk and Incremental Prediction Error	93
6.4	Methodology	99
6.4.1	Initialization: Fine-tuning on the First Task	100
6.4.2	The Proposed Method	101
6.4.3	Overall Optimization and Complexity Analysis	102
6.4.4	Why NRP performs effectively in Knowledge Transfer for OWCL	104
6.5	Experiments	105
6.5.1	Experimental Settings	105
6.5.2	Experimental Results and Analysis	108
6.6	Summary	112
7	Cross-regional Fraud Detection via Continual Learning with Knowledge Transfer	115
7.1	Continual Learning for Cross-Regional Fraud Detection	116
7.2	A Brief Recall on Fraud Detection	119
7.3	Methodology	121
7.3.1	Overview and Problem Formulation	121
7.3.2	Construct Heterogeneous Trade Graphs	122
7.3.3	The Proposed CCL Model	125

7.3.3.1	Prototype-Based Knowledge Replay	126
7.3.3.2	Regularization-based Parameter Smoothing	128
7.3.4	Graph Representation Learning Backbone	129
7.3.5	Overall Framework and Algorithm	130
7.4	Experiments	132
7.4.1	Experimental Settings	132
7.4.1.1	Implementation Details and Datasets	132
7.4.1.2	Baselines	133
7.4.1.3	Metrics	134
7.4.2	Main Results	134
7.4.3	Ablation Study	136
7.4.4	Sensitivity Analysis of Hyperparameters	138
7.4.5	Interpretability Analysis	139
7.5	Summary	141
8	Conclusions and Future Work	143
8.1	Summary of Contributions	144
8.2	Future Work	147
	Bibliography	153
	List of Figures	163
	List of Tables	169
	English Summary	171
	Nederlandse Samenvatting	173
	Acknowledgements	177
	Curriculum Vitae	179

Chapter 1

Introduction

1.1 Background

It is not the most intellectual of the species that survives; it is not the strongest that survives; but the species that survives is the one that is able best to adapt and adjust to the changing environment in which it finds itself.

— Charles Darwin ¹

Learning serves as the fundamental mechanism enabling intelligent systems to adapt to dynamic environments. Throughout biological evolution, humans and other organisms have developed exceptional adaptive capabilities, allowing them to continuously acquire, refine, and apply knowledge in response to environmental changes [1, 2, 3, 4, 5]. Inspired by these natural learning processes, researchers aim to enrich artificial intelligence (AI) systems with analogous continual learning abilities. This pursuit has led to the development of continual learning (CL), a paradigm focused on training models to incrementally process sequences of tasks while maintaining performance on previously learned ones (see Figure 1.1). Continual learning enables knowledge acquisition across diverse domains, including new skill learning, skill refinement, environmental adaptation, and contextual understanding [6, 7, 8].

In the literature, continual learning is often synonymous with incremental learning and lifelong learning [4, 8, 9, 10]. While these terms are frequently used interchangeably, they collectively embody the overarching objective of developing artificial intelligence systems capable of dynamic evolution, preserving acquired knowledge while avoiding strict dependence on static data distributions.

Unlike traditional machine learning models that assume static data distributions, continual learning specifically addresses evolving, dynamic distributions. A cen-

¹This quote is widely attributed to Charles Darwin, but it does not appear verbatim in his writings. The phrasing is believed to originate from Professor Leon C. Megginson, who paraphrased Darwin’s ideas. Despite its frequent misattribute, the quote effectively captures the essence of Darwinian evolution and has since been popularized in both scientific and managerial literature.

tral challenge in continual learning is catastrophic forgetting (CF) [11, 12], where learning new information severely degrades a system’s ability to retain and utilize previously acquired knowledge. This phenomenon reveals the fundamental tension between two competing objectives: **learning plasticity** (the capacity to acquire new knowledge rapidly) and **memory stability** (the ability to preserve existing knowledge). Achieving an optimal balance proves challenging, as excessive plasticity compromises stability, and vice versa [2, 12, 13].

Furthermore, effective continual learning systems must demonstrate strong **generalizability** to handle distributional shifts both within individual tasks and across multiple task domains (see Figure 1.1 **b**). While one potential solution involves re-training models on all previous data, this approach presents substantial limitations, including prohibitive computational costs, extensive storage requirements, and potential privacy issues [14].

The core objective of continual learning is consequently to enable efficient model updates that learn new data incrementally while preserving knowledge from past experiences, maintaining computational and storage efficiency, and scaling effectively with continuous changes.

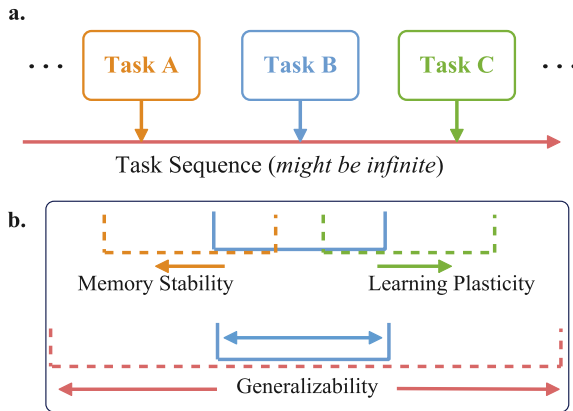


Figure 1.1: A conceptual framework of continual learning. **a.** Continual learning requires adapting to incremental tasks with dynamic data distributions. **b.** A desirable solution should ensure an appropriate trade-off between stability (orange arrow) and plasticity (green arrow), as well as a generalizability to intra-task (blue arrow) and inter-task (red arrow) distribution differences.

Despite significant advancements in continual learning, most current approaches operate under a **closed-world assumption**, limiting systems to recognizing and processing only those object classes explicitly encountered during training [8, 15, 16]. This assumption implicitly excludes the possibilities of unexpected or novel entities emerging during testing or deployment, thus falling short in real-world scenarios, as the environment is inherently open and constantly presents unknown objects.

Human cognition, by contrast, treats novelty as a powerful *intrinsic motivator* for learning. The human brain responds to new stimuli with distinct neural activity and behavioral changes, driving exploration and enabling flexible encoding of unfamiliar information [8]. Over time, however, this novelty response attenuates as the brain rapidly adapts to repeated exposure to the same stimuli [17, 18], demonstrating an adaptive balance between exploration and habituation.

To operate effectively in an **open-world environment**, artificial intelligence systems must simulate these human capabilities. Specifically, an ideal system would autonomously identify and respond to novel entities, incrementally integrate new knowledge while maintaining existing competencies, and adapt dynamically without requiring a periodic offline retraining. Such capabilities would enable artificial intelligence systems to continuously evolve through a sequence of new tasks, allowing real-time adaptation to novel scenarios in practical applications.

In this context, Open-world Continual Learning (**OWCL**) [8, 19] has emerged as both a highly practical yet profoundly challenging machine learning paradigm. This framework requires artificial intelligence systems to continuously adapt to unbounded task sequences in a dynamic open environment [15, 20], where novel elements might appear unpredictably during testing phases [5, 21, 22].

The core objective of the open-world continual learning framework includes operating under the open-world assumption, recognizing unseen/open samples, and incrementally acquiring knowledge from new tasks while preventing catastrophic forgetting [23, 24, 25]. These requirements introduce two critical challenges. Due to the potential occurrence of novelties in continual learning [14], open-world continual learning requires detecting unknown samples to prevent misclassification into known categories. At the same time, open-world continual learning requires the model to retain previously learned knowledge without forgetting during continuous open detection. Consequently, the open-world continual learning paradigm faces a dual complexity. On the one hand, the presence of unknown samples complicates the fundamental tension between *knowledge stability* [1, 14] and *knowledge plasticity* [26, 27]. On the other hand, the continuous expansion of the knowledge space through incremental learning of new tasks complicates open detection in the embedding space.

Although existing continual learning methods have made progress in addressing catastrophic forgetting, they fundamentally operate under three limiting assumptions [28, 29]: (1) task boundaries are always clearly defined, (2) all test-time samples belong to known classes, and (3) data distributions remain relatively stable [24]. These assumptions break down in open real-world environments where models encounter novel categories, shifting contexts, and ambiguous task transitions, precisely the challenges our work addresses.

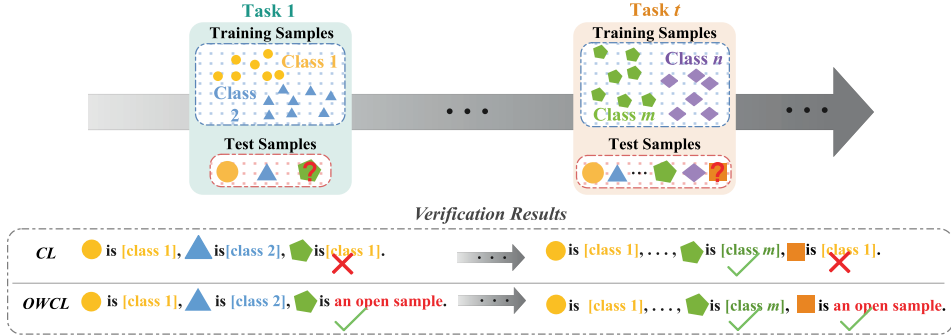


Figure 1.2: An illustration of the open-world continual learning problem and comparison between continual learning. In open-world continual learning, the test set of each task may include samples from classes unseen during training, deviating from the closed-world assumption of traditional continual learning. As shown in the figure, this setting imposes two key challenges: (1) preserving knowledge from previously learned tasks to mitigate catastrophic forgetting, and (2) performing robust open-set recognition to correctly identify and reject instances from unknown classes. While continual learning models tend to misclassify open samples as one of the known classes (e.g., the green pentagon and orange square), open-world continual learning models are explicitly designed to distinguish between known and unknown classes, thereby enabling more reliable and adaptive decision-making in dynamic environments.

To further illustrate the unique challenges of open-world continual learning, we refer to the scenario depicted in Figure 1.2. Unlike continual learning, where all test-time samples are assumed to belong to previously seen classes, open-world continual learning confronts a more realistic setting where unknown classes may emerge during testing. This introduces the first key challenge: open-set detection, i.e., accurately identifying and rejecting samples from unseen categories rather than misclassifying them into existing known classes. As shown in Figure 1.2, while continual learning models tend to misclassify such unknowns (e.g., the green pentagon and orange square) into existing classes, open-world continual learning explicitly distinguishes them as open samples.

Another challenge lies in the need for stronger generalization and boundary stability. Open-world continual learning models must not only acquire new knowledge incrementally and retain prior learning—addressing the classical stability–plasticity dilemma—but also maintain a consistent open-set recognition capability across tasks. This requires the ability to form and update well-defined task or class boundaries over time, ensuring that the decision boundaries between known and unknown categories remain reliable despite ongoing changes in the data distribution.

Lastly, and most critically, open-world continual learning raises a fundamental question that goes beyond mere classification: how can models exploit knowledge from unknown samples? While traditional continual learning gains knowledge by learn-

ing from labeled data across successive tasks, the presence of unlabeled and unpredictable unknowns in open-world continual learning complicates this process. Effectively mining and leveraging knowledge from open-set instances, without ground truth supervision, becomes essential for sustained learning in dynamic and open environments.

To provide a more detailed description of the diverse scenarios that may arise in open-world continual learning, we construct four representative open-world continual learning settings based on the three classical continual learning paradigms [6]: task-incremental learning, domain-incremental learning, and class-incremental learning. It is important to note, however, that a critical distinction in open-world continual learning lies in the absence of task identity information during testing, as unknown samples can appear unpredictably at any time [30]. This fundamental characteristic renders conventional task-incremental and domain-incremental settings insufficient for fully capturing the complexity of open-world continual learning scenarios.

Therefore, we propose a novel scenario designed for open-world continual learning, called knowledge-incremental learning (**KIL**), characterized by the ability to solve tasks previously seen without knowing which task is being performed (no task identifiers available), non-strictly disjoint training sets across different tasks, and potential distribution shifts among known categories.

Knowledge-incremental learning both subsumes class-incremental learning and accommodates category-specific distribution shifts, thereby providing a more realistic and consequently more challenging framework for open-world continual learning. Based on this foundation, we then introduce four distinct scenarios for open-world continual learning of increasing difficulty: class-incremental with non-repetitive open samples (**CINO**), class-incremental learning with repetitive open samples (**CIRO**), knowledge-incremental learning with non-repetitive open samples (**KINO**), knowledge-incremental learning with repetitive open samples (**KIRO**).

Table 1.1: Overview of different OWCL scenarios.

<i>Scenario</i>	CINO	CIRO	KINO	KIRO
Known Classes Repeatedly Appear	✗	✗	✓	✓
Unknown Samples Repeatedly Appear	✗	✓	✗	✓

Having established four representative open-world continual learning scenarios based on classical continual learning settings, we now turn to the practical significance of developing robust open-world continual learning systems. Recent advances in deep representation learning for computer vision and natural language processing have catalyzed growing interest in open-world continual learning. However, many current approaches [25, 31] still treat open-world continual learning as a loose combination of open-set recognition and continual learning, rather than embracing it as a unified and

principled learning paradigm. This fragmented view leads to two core limitations: (1) knowledge transfer is typically confined to labeled data, overlooking the rich semantic information that may be embedded in unknown samples; and (2) the lack of standardized problem formulations and evaluation protocols impedes consistent benchmarking and method selection.

In response, this thesis takes a knowledge transfer perspective to investigate the open-world continual learning problem. Rather than treating unknown-class detection and known-class classification as isolated objectives, we aim to understand and enhance the underlying transfer mechanisms across both labeled and unlabeled samples in dynamic, open environments. Throughout the thesis, we explore multiple model architectures, training strategies, and evaluation protocols to advance the understanding and practical utility of knowledge transfer in open-world continual learning, which contributes not only to methodological innovation but also to enabling practical open-world continual learning systems that exhibit robust adaptability and generalization across a wide range of real-world scenarios.

For instance, in financial security, open-world continual learning models can support real-time detection of emerging fraud patterns, such as AI-generated deepfake scams, while maintaining accuracy on previously encountered attack types, thus reducing reliance on costly periodic retraining. In healthcare, open-world continual learning systems could incrementally incorporate novel disease markers (e.g., viral mutations) without compromising diagnostic performance for known conditions. These examples highlight open-world continual learning’s potential to enhance operational robustness in dynamic environments, while also addressing pressing societal challenges.

Beyond immediate applications, the broader societal implications of open-world continual learning are equally profound. By mimicking human-like continual learning, open-world continual learning systems can reduce the environmental footprint of artificial intelligence by eliminating repeated retraining cycles, lower the barrier for adoption by organizations with limited access to large-scale labeled data, and foster more trustworthy artificial intelligence systems through transparent novelty detection. Nonetheless, these benefits come with ethical considerations, such as ensuring privacy in continuous surveillance settings or maintaining accountability in medical diagnostics—issues we explicitly address through built-in uncertainty quantification mechanisms in our framework.

While the empirical studies in this thesis primarily focus on image-based data, the proposed open-world continual learning framework is designed to be modality-agnostic. Its core principles—adaptive knowledge transfer, novelty awareness, and dynamic representation updating—can naturally extend to other data modalities such as video, text, speech, and multimodal sensor streams. In particular, applying this framework to sequential modalities (e.g., video surveillance or dialogue systems) could reveal new insights into temporal transfer and incremental context modeling

under a realistic open-world environment. Future research may further explore how the same open-world principles can be instantiated across heterogeneous domains, paving the way toward a unified and generalizable paradigm of continual learning beyond visual tasks.

1.2 Research Questions and Contributions

Our approach advances open-world continual learning beyond current limitations through three key innovations: (a) a dynamic knowledge representation that evolves with both known and unknown samples, (b) a unified optimization framework that jointly preserves stability and enables plasticity, and (c) task-agnostic detection mechanisms that operate without explicit task identifiers. This combination theoretically enables infinite learning capacity while maintaining practical deployability. Note that all subsequent investigations of the problems addressed in this thesis are conducted within the context of image tasks, and all experimental validations in subsequent chapters are carried out using image benchmark datasets.

Research Question 1 (RQ 1) *Can we develop an effective open-world continual learning model to accumulate the knowledge gained in the past and to use the knowledge to help divide the test data into knowns and unknowns/opens, especially for enlarging the open-set boundary between knowns and unknowns?*

Existing methods fail to exploit previous unknowns from former tasks to help new tasks, leading to performance degradation on new tasks. Hence, the main motivation of RQ 1 here is to learn knowledge for both knowns and unknowns from previous tasks and later use the knowledge to formulate an open-set boundary between knowns and unknowns for new tasks. This question targets the foundational challenge of open-world continual learning: maintaining previously acquired knowledge while remaining adaptable to new, unseen classes. A robust open-world continual learning system must not only classify known samples accurately but also detect and separate unknowns with high confidence.

The key challenge of **RQ 1** lies in addressing two critical issues: (1) *Knowledge Stability*, where newly acquired knowledge from a task that is significantly different from previous ones might conflict with existing knowledge; and (2) *Knowledge Plasticity*, which requires the model to update its understanding by reclassifying unknown samples as knowns once their ground-truth data becomes available in subsequent tasks. Our primary contribution to addressing this research question is the development of a novel **Prompt-enhanced Knowledge Transfer** approach, referred to as **Pro-KT**. Pro-KT is introduced in Chapter 4 and based on the following publication [30]:

- **Li, Y.**, Yang, X., Wang, H., Wang, X., & Li, T.: Learning to Prompt Knowledge Transfer for Open-World Continual Learning. In: The 38th Annual AAAI Conference on Artificial Intelligence. AAAI '24 (2024) 38(12), 13700-13708.

Our main ideas are *knowledge transfer* and *prompt learning* [32, 33, 34]. Pro-KT delineates an innovation of prompt learning for open-world continual learning, a novel plug-and-play prompt bank for knowledge transfer, and two adaptive threshold selection strategies for determining the open-set boundary. Specifically, to address *knowledge stability*, we create a prompt bank designed to encode knowledge through the use of prompts. These prompts serve as instructions for directing the model in task execution. By flexibly selecting prompts from the proposed prompt bank, Pro-KT can facilitate effective knowledge transfer with both task-generic and task-specific knowledge across diverse tasks. To address *knowledge plasticity*, we design two adaptive threshold-selection strategies for determining the open-set detection boundary. Through these strategies, the open-set decision boundary will be updated according to the newly learned knowledge continually, so as to handle knowledge plasticity.

While RQ1 addresses the stability-plasticity trade-off, its success heavily depends on access to labeled data. In practice, such labels are often scarce. This motivates the investigation in RQ2, which explores how to achieve strong open-world continual learning performance under severe label constraints by leveraging transfer learning and few-shot techniques.

Research Question 2 (RQ 2) *How can we enhance the performance of existing open-world continual learning models under the constraint of limited labeled data by employing an improved knowledge transfer approach?*

Existing open-world continual learning still requires a large amount of labeled data for training, which is often impractical in real-world applications. Given that new categories/entities typically come with limited annotations and are in small quantities, a more realistic scenario is open-world continual learning with scarce labeled data, i.e., few-shot training samples, which addresses our second research question **RQ 2**.

In RQ 2, we investigate the problem of open-world few-shot continual learning (OFCL), challenging in (i) learning unbounded tasks without forgetting previous knowledge and avoiding overfitting, (ii) constructing compact decision boundaries for open detection with limited labeled data, and (iii) transferring knowledge about knowns and unknowns and even update the unknowns to knowns once the labels of open samples are learned. To address RQ 2, we introduce a novel framework, also referred to as *PEAK*, in Chapter 5 based on the following publication [35], that, to the best of our knowledge, is the first work to study the open-world few-shot continual learning problem:

- **Li, Y.**, Wang, X., Yang, X., Bonsangue, M., Zhang, J., Li, T.: Improving Open-World Continual Learning under the Constraints of Scarce Labeled

Data. In: Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining. KDD '25 (2025).

Specifically, we first propose an instance-wise token augmentation (ITA) aimed at acquiring additional ‘knowledge’ to mitigate the inadequate representation caused by scarce labeled data. Moreover, instance-wise token augmentation can facilitate knowledge transfer by matching learnable tokens to each sample embedding. Additionally, given the scarce labeled data where certain exemplar points (hubs) appear among the nearest neighbors of many other points, test samples will be assigned to it regardless of their true label, resulting in low accuracy [36]. To mitigate this, inspired by the embedding representations on a hypersphere [37], we introduce a novel and compact margin-based open boundary (MOB) and an adaptive knowledge space (AKS) consisting of learnable hyperspheres, where each hypersphere is characterized by a class centroid and an associated radius. In particular, enables the formulation of compact decision boundaries between known and unknown samples, thereby enhancing open detection. Simultaneously, the adaptive knowledge space encourages the model to learn incrementally from unknowns and classify previously encountered unknown samples in new tasks, effectively transforming unknowns into knowns over time.

Even with enhanced few-shot learning (RQ2), existing open-world continual learning methods often disregard the latent information embedded in unknown samples. RQ3 challenges this orthogonality assumption by asking whether knowledge from the open world—especially unlabeled or outlier data—can be actively incorporated into future tasks through holistic knowledge transfer.

Research Question 3 (RQ 3) *How can we leverage knowledge from both known and unknown samples to enable knowledge transfer in open-world continual learning, effectively improving open-set detection with incremental classification at both theoretical and architectural levels?*

Unlike traditional learning models that operate in a closed and fixed set of classes, open-world continual learning aims at learning on the job in an open-world assumption with the goal of recognizing unseen/open samples and incrementally acquiring knowledge from new tasks without forgetting [23, 24, 25]. Due to the potential occurrence of novelties in continual learning [14], open-world continual learning models need to accurately detect unknowns to prevent unknown samples from being incorrectly classified into known categories. At the same time, open-world continual learning requires the model to retain previously learned knowledge without forgetting while continually performing open detection. In summary, the open detection for unknowns and classification for knowns are *interdependent* in open-world continual learning: the incremental learning of new tasks makes open detection in an embedding space more challenging with the expanding knowledge space, and vice versa.

However, current approaches [25, 31] still treat open-world continual learning as a simple combination of open-set recognition and continual learning, rather than as an integrated paradigm, making it only effective in knowledge transfer related to known samples, while neglecting the knowledge derived from unknown samples. Therefore, a promising open-world continual learning model must be capable of **knowns-unknowns knowledge transfer**, i.e., effectively transferring knowledge both for known categories and unknown samples. Besides, there is a lack of problem formulation and thorough empirical explorations of potential issues in open-world continual learning, making it difficult to compare the performance of existing methods and making it unclear how to choose one method over another. In Chapter 6, we explore the issues arising in open-world continual learning and adopt an integrative perspective to deal jointly with unknown samples’ detection and known samples’ classification, particularly the knowledge transfer for both unknowns and knowns, based on the following paper [38]:

- **Li, Y.**, Lai, G., Yang, Y., Li, Y., Bonsangue, M., & Li, T.: Exploring Open-World Continual Learning with Knowns-Unknowns Knowledge Transfer. arXiv preprint arXiv:2502.20124.

Specifically, we provide a formal and unified formulation of the open-world continual learning problem, delineating four distinct yet interrelated scenarios that capture the multifaceted nature of open-world dynamics. Our empirical investigations reveal a strong and non-trivial interaction between the tasks of open-set detection and incremental prediction, challenging the prevailing assumption that these components can be independently optimized. This insight calls for a more holistic approach to open-world continual learning, one that integrates both discovery and adaptation within a unified learning process.

From a theoretical standpoint, we rigorously define the decision space of open-world continual learning and uncover the core optimization objectives that govern learning under open-world conditions. Building upon this foundation, we propose a novel framework, **HoliTrans**, which effectively transfers knowledge from both known and previously unseen (unknown) instances. HoliTrans unifies open-set recognition and continual adaptation through a single principled mechanism, thereby enabling robust generalization in the presence of evolving data distributions and task boundaries.

Extensive empirical evaluations across all four open-world continual learning scenarios corroborate our theoretical insights. HoliTrans consistently outperforms state-of-the-art baselines, demonstrating not only superior performance in isolated metrics but also stable adaptation across open and incremental learning tasks. These results affirm the necessity of treating open-world continual learning as an inherently integrated problem and highlight HoliTrans as a promising step toward a unified, theoretically based open-world continual learning framework.

Having established methods for knowledge retention, transfer, and open-sample utilization, RQ4 shifts focus to practice. It examines whether continual learning and open-world continual learning frameworks can go beyond benchmarks and offer tangible value in complex, real-world tasks characterized by data drift, concept emergence, and geographical heterogeneity.

Research Question 4 (RQ 4) *How can we improve the model performance by formulating practical problems with an open-world continual learning framework and developing effective models for challenging real-world applications?*

Building upon the aforementioned works, we further investigate the potential of continual learning and open-world continual learning to address real-world challenges. As emerging and increasingly influential paradigms in machine learning, continual learning and open-world continual learning offer the promise of reshaping how practical systems adapt to dynamic environments. In contrast to conventional static models, these frameworks inherently accommodate both the continuity of data streams and the openness of task boundaries, making them particularly well-suited for applications characterized by evolving knowledge spaces.

For instance, in the domain of dynamic financial transactions, OWCL-enabled systems can incrementally learn from newly observed fraud types (e.g., deepfake-based voice scams) while simultaneously preserving detection capabilities for previously encountered fraud patterns (e.g., credit card cloning or identity theft). Similarly, in smart healthcare, continual learning has been leveraged to enhance pathological diagnosis systems, empowering them to incrementally incorporate rare or emerging disease categories—such as novel variants of infectious diseases—without requiring complete retraining or catastrophic memory loss.

Motivated by these promising applications, we aim to extend the frontier of open-world continual learning by formulating a novel cross-regional fraud detection problem under the continual learning paradigm. Unlike conventional fraud detection approaches that are confined to static datasets within isolated geographical regions, our proposed setting reflects a more realistic scenario wherein financial service providers operate across multiple cities or countries, each exhibiting unique transaction patterns and fraud behaviors.

The inherent challenges in such settings include high retraining costs, data privacy constraints, and the risk of catastrophic forgetting when adapting models to new regional data. In light of these challenges, we propose in Chapter 7 a unified framework that treats cross-regional fraud detection as an incremental learning task. This formulation enables the development of robust, scalable fraud detection systems that can generalize across diverse regions while effectively retaining prior knowledge and adapting to novel threats. This chapter is based on the following publications [39, 40]:

- **Li, Y.**, Yang, Y., Gao, Q., & Yang, X: Cross-Regional Fraud Detection via Continual Learning (Special Program). In: Proceedings of the AAAI Conference on Artificial Intelligence, 37(13), 16260-16261.
- **Li, Y.**, Yang, X., Gao, Q., Wang, H., Zhang, J., & Li, T.: Cross-Regional Fraud Detection via Continual Learning With Knowledge Transfer. In: IEEE Transactions on Knowledge and Data Engineering. TKDE'24 (2024) 38(12), 7865-7877.

1.3 Thesis Outline

The overall organization of this thesis is as follows. In this chapter, we give a brief introduction to the thesis’s background and motivation, the main research questions, and contributions. Chapter 2 provides a literature review for existing methods in continual learning and open-set recognition, and focuses more on open-world continual learning methods. Chapter 3 presents some general definitions, which are used throughout the whole dissertation, problem definitions of open-world continual learning, and different scenarios, respectively.

Subsequently, we present investigations corresponding to the four research questions, detailed respectively in Chapters 4 through 7. Specifically, in Chapter 4, we address **RQ 1** by introducing *Pro-KT*, a prompt-based knowledge transfer method for open-world continual learning. We design a plug-and-play prompt bank to balance knowledge stability and plasticity by enabling flexible transfer of task-specific and task-generic knowledge.

Chapter 5 addresses **RQ 2** and introduces a novel framework, called *PEAK*, which is designed to tackle the challenges of open-world continual learning under few-shot learning scenarios. Within this framework, we propose three core modules that collectively enable robust performance when labeled data is severely limited, offering an exploration of the unique difficulties posed by limited supervision in open-world continual learning.

Chapter 6 addresses **RQ 3** and introduces *HoliTrans*, a unified framework for open-world continual learning. We provide a formal theoretical construction of the open-world continual learning decision space and identify four interrelated scenarios that reflect the complexity of open-world dynamics. Our findings reveal a strong interaction between open-set detection and incremental prediction, challenging the assumption of their separability and motivating a holistic approach that unifies discovery and adaptation within a holistic paradigm.

Chapter 7 addresses **RQ 4** by introducing *CCL*, a continual learning framework for cross-regional fraud detection. Motivated by real-world deployments, we extend open-world continual learning to a more practical setting where financial institutions operate across diverse regions with varying transaction patterns and fraud behaviors.

Unlike traditional methods limited to static, region-specific datasets, CCL enables adaptive learning across regions, capturing both shared knowledge and local nuances in a unified model.

At last, Chapter 8 concludes the contributions of the thesis and discusses possible future work.

Chapter 2

Literature Review

First, we examine continual learning (CL) methods, categorizing them into three principal families according to their mechanisms for preserving and utilizing task-specific knowledge across sequential learning episodes: (1) replay-based methods, which retain and revisit past samples; (2) regularization-based methods, which constrain parameter updates; and (3) structure-based methods, which dynamically expand or isolate model components. For each category, we provide a critical analysis of its respective advantages and fundamental limitations.

Second, we investigate open-set learning (OSL) and out-of-distribution (OOD) detection, beginning with formal definitions to then moving to an evaluation of their core methodologies. This reveals significant gaps in current approaches, particularly regarding their ability to handle incremental knowledge integration and practical deployment scenarios.

Finally, we integrate these perspectives into the emerging paradigm of open-world continual learning (OWCL), which combines the sequential adaptation of continual learning with the novelty detection capabilities of open-set learning. While recent open-world continual learning works demonstrate promising results, our analysis identifies critical shortcomings in the knowledge transfer mechanism and provides insights into the importance and uniqueness of knowledge transfer in open-world continual learning.

2.1 Continual Learning

Driven by the resurgence of neural network research, continual learning and catastrophic forgetting have received considerable attention. Common approaches to addressing these challenges include reducing representation overlap through parameter regularization [21], memory-based approaches that replay either real or synthetic historical samples [41, 42], and architectural solutions employing dual-network configurations [43].

Early continual learning research was often constrained by limited computational resources, restricting experiments to shallow architectures and modestly-sized datasets [44]. Recent empirical analyses have examined how dropout [45] and various activation functions influence forgetting in sequential tasks [46]. Theoretical perspectives have also been explored, especially in task-incremental settings [19]. Recently, there has been a shift towards practical scaling to longer task sequences and larger datasets, utilizing large-scale pre-trained foundational models as backbones with fine-tuning strategies. This evolution reflects the field’s progression from constrained laboratory settings to more realistic, large-scale applications.

Next, we categorize recent continual learning methods according to their mechanisms for preserving and utilizing both task-specific and transferable knowledge during incremental learning processes. This taxonomy organizes approaches into three principal categories based on their core operational principles.

Replay-based methods. Replay-based methods address forgetting by retaining and revisiting representative examples from past tasks. Classical approaches, such as iCaRL [42], retrain models on a limited set of past samples alongside new task data. Other works incorporate constrained optimization into the replay process [47, 48]. When actual past samples are unavailable, pseudo-rehearsal via generative models serves as an alternative, though it introduces complexity, including potential mode collapse and difficulties in balancing old and new samples. Replay methods also inherently face data privacy and security challenges, significantly limiting their practical deployment when previous data is inaccessible or restricted by privacy regulations.

Regularization-based methods. Regularization-based methods apply constraints on model updates to retain knowledge across tasks. For instance, Elastic Weight Consolidation (EWC) [21] mitigates forgetting by penalizing significant changes in crucial model weights. Another classical approach, Learning without Forgetting (LwF) [49], utilizes knowledge distillation techniques to preserve previous task knowledge. Recent work further explores knowledge distillation strategies [50], though these methods remain highly dependent on task similarity. Consequently, a growing body of research has begun addressing the robustness of continual learning models against varying task distributions [51].

Structure-based methods. Structure-based methods address continual learning by expanding the model’s architecture to accommodate new tasks, preserving previously learned knowledge [52, 53]. These approaches typically involve either dynamically growing the model structure for new tasks or employing task-specific masks to isolate parameters of previous tasks [54]. However, without careful management, these methods risk saturating model capacity, hindering future task learning. Additionally, given recent advancements in large-scale pre-trained models, strategies

relying heavily on model parameter reuse and storage face significant limitations in scalability, particularly when applied to extensive downstream tasks.

2.2 Open-Set Learning

The conventional closed-world assumption is inadequate for real-world applications, where models must handle unknown samples during deployment while facing restricted access to historical training data. This challenge motivates open-set learning, which addresses two core objectives: (1) accurate classification of known classes and (2) robust rejection of unknown samples during inference [15, 55]. Open-set learning explicitly models the uncertainty inherent in dynamic environments where novel objects frequently appear [8].

Formally, given a training set $D_{tr} = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$, where each y_i belongs to the set C_B of *known classes*, a set u denoting the *unified unknown category*, and an evaluation set $D_{ev} = \{(x'_1, y'_1), (x'_2, y'_2), \dots, (x'_m, y'_m)\}$, with each $y'_i \in (C_B \cup \{u\})$, open-set learning requires:

- Precise discrimination between known C_B and unknown $\{u\}$ samples;
- No further granularity within the unknown category.

This formulation captures the essential challenge of open-set recognition while maintaining practical applicability.

Recent advances in open-set learning have produced several innovative approaches to handle unknown samples [56, 57, 58]. For example, ORE [55] combines contrastive clustering with energy-based classification for novel object detection. Additionally, [59] develops an open-set classifier using an extreme value theory-based classifier. PointCLIP [60] integrates CLIP and GPT-3 multimodal foundation models for zero-shot open-set tasks, and CEC [61] introduces an open detection framework featuring a dual-component framework with proposal advising and class-specific expulsion. These methods demonstrate the field’s progression from basic thresholding to sophisticated architectural solutions.

2.3 Out-of-Distribution Detection

Parallel developments in out-of-distribution detection address the related challenge of identifying samples from non-training distributions. Current algorithms predominantly rely on the independent and identically distributed (IID) assumption, despite prevalent distribution shifts in practical settings. Formally, given training data $D_{tr} = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$, drawn from a distribution $P_{tr}(X, Y)$ and test samples $x'_i \sim P_{OOD}(X')$, out-of-distribution detection evaluates model confidence to

flag distributional outliers during testing phase. Metrics commonly used for out-of-distribution detection include the model’s prediction confidence or certainty, typically measured via softmax probability scores, assessing the correct identification of out-of-distribution samples.

The field has evolved from early post-hoc techniques like ODIN [62], employing temperature scaling and input perturbations as post-hoc adjustments to enhance differentiation between ID and out-of-distribution data, to modern approaches using energy-based scoring [63, 64, 65] with theoretical guarantees [66], enhanced frameworks like Generalized ODIN [67] for covariate shift robustness, extension of the ODIN’s framework [67] by adopting specialized training objectives and optimizing hyperparameters, like perturbation magnitude, specifically on ID data, and hierarchical semantic organization of known classes [68] with specialized classifiers such as top-down classification [69] and group softmax training [70] that have been proven very effective within out-of-distribution.

Notably, recent work [19] emphasizes the importance of novelty detection and integrates out-of-distribution detection with continual learning, as exemplified by SOLA [8], a framework that presents a strategy to enhance novelty detection, facilitate on-the-fly task adaptation, and support incremental learning.

Despite the significant progress, existing methods still fall short in adequately addressing the representation and integration of knowledge for effective transfer to future tasks. The presence of open samples further exacerbates these challenges, posing substantial difficulties in knowledge representation and accumulation. Crucially, both open-set learning and out-of-distribution methods classify unknown samples into an *unknown class*, thus preventing models from fully utilizing the informative content within these samples, leading to incomplete exploitation of available knowledge.

2.4 Open-world Continual Learning

Continual learning [21, 49] typically operates under a closed-world assumption, meaning the model presumes that all samples encountered during testing or deployment belong to predefined classes previously seen during training [15, 16, 71]. However, such an assumption is unrealistic for dynamic, real-world environments, where systems frequently encounter novel or previously unseen data [8, 30].

While continual learning and open-set learning address complementary challenges, preserving knowledge across tasks and detecting unknowns, respectively, their integration introduces fundamental tensions. Continual learning methods prioritize stability to mitigate forgetting but assume a closed-world setting, while open-set learning emphasizes plasticity to recognize novelty at the risk of disrupting learned representations. This creates a three-way trade-off: (1) retaining past knowledge

(stability), (2) acquiring new knowledge (plasticity), and (3) dynamically detecting open-set samples without prior exposure. For instance, replay-based continual learning methods may erroneously classify unknowns as known if trained only on historical data [42], while open-set learning detectors struggle when task distributions shift incrementally [20].

Practically, continual learners need to detect, adapt to, and incrementally learn from these new, unknown classes or samples while retaining previously learned information [72, 73]. These conflicting objectives explain why open-world continual learning cannot be trivially reduced to combining existing continual learning and open-set learning techniques, but rather requires novel frameworks to unify stability, plasticity, and open-set robustness. Therefore, effectively detecting and incrementally learning novelties while preserving prior knowledge becomes essential for realistic deployment.

Open-world continual learning extends continual learning to accommodate open-world scenarios. Recently, it has gained significant attention, also because it presents substantial challenges [8, 14]. To enable existing continual learning frameworks to recognize unknown or open samples effectively, initial research in open-world continual learning has integrated methods from open-set recognition and out-of-distribution detection into standard continual learning models. [59] proposed an OSR framework utilizing extreme value theory to manage incremental tasks within dynamic environments. Similarly, [55] introduced a model employing contrastive clustering coupled with energy-based identification techniques [74], enabling the continual learner to identify and integrate novel data.

Building upon these approaches, recent open-world continual learning studies have increasingly emphasized incorporating out-of-distribution detection into continual learning frameworks. For instance, [19, 25, 30] underscored novelty detection as a critical component of open-set learning, suggesting that existing out-of-distribution detection strategies could be effectively integrated within continual learning settings. Furthermore, recent frameworks such as SOLA [8, 75] combine out-of-distribution detection methods with incremental task adaptation, enabling improved novelty detection and task-specific learning in open-world environments.

Nevertheless, despite these advances, current open-world continual learning research remains constrained by three fundamental limitations. First, the absence of standardized problem formulations and evaluation protocols [72, 76] hinders fair comparison across methods. Second, most approaches rely on simplistic integrations of continual learning with out-of-distribution detection techniques [25, 30, 31, 77], failing to address the critical challenge of asymmetric knowledge transfer between known and unknown samples. While existing methods effectively transfer knowledge for known classes (via replay or regularization), they largely ignore the informational value of unknowns, creating both experimental and theoretical gaps: Experimentally, evaluations disproportionately focus on known-class accuracy while lacking benchmarks

to assess whether detected unknowns improve future task learning [73]. Unknown samples are typically discarded rather than analyzed for transferable features (e.g., hierarchical relationships to known classes [5]). Theoretically, no framework exists to quantify how unknowns should contribute to knowledge transfer. Unlike known classes, where distillation or replay provides clear mechanisms, unknowns present unique challenges: (i) preserving latent representations without destabilizing known classes, (ii) refining decision boundaries for future tasks, and (iii) balancing plasticity for new unknowns with stability for knowns. This ambiguity forces ad-hoc solutions like threshold-based rejection [30] or heuristic clustering [25] without theoretical guarantees. Ultimately, the lack of robust theoretical foundations reduces open-world continual learning to a superficial combination of continual learning and open-set techniques, rather than a principled framework for unified knowledge transfer and updating.

2.5 Concluding Remarks

Table 2.1 summarizes the core distinctions between continual learning, open-set learning, and open-world continual learning across four dimensions: objectives, strengths, limitations, and representative works. While continual learning excels at sequential knowledge retention and open-set learning specializes in novelty detection, open-world continual learning must reconcile its competing demands—preserving stability for known classes while maintaining plasticity for unknowns, all within dynamically expanding task distributions. This reveals open-world continual learning’s critical gap: current methods [8, 30] lack mechanisms to transfer knowledge from unknown samples while balancing the stability-plasticity trade-off.

This chapter shows critical gaps in current open-world continual learning research that our work addresses: (1) asymmetric knowledge transfer because of the prevailing neglect of unknown samples’ informational value (solved through *HoliTrans*’ unified knowledge transfer in Chapter 6), (2) dynamic evaluation because of lack of benchmarks for unknown reuse (tackled via few-shot protocols in Chapter 5), and (3) theoretical guarantees because of the absence of formal bounds for open-set continual learning (resolved by *Pro-KT*’s stability-plasticity analysis in Chapter 4). These advancements are designed with practical deployment constraints in mind, as demonstrated through our real-world fraud detection applications in Chapter 7.

Table 2.1: Comparison of continual learning, open-set learning, and open-world continual learning.

Aspect	Continual Learning	Open-Set Learning	Open-world Continual Learning
Primary Goal	Avoid forgetting known tasks	Detect unknowns in static settings	Learn incrementally and adapt to unknowns
Key Strength	Task-specific stability [21]	Strong open-set detection [15]	Unified openness and continuity [8]
Limitation	Assumes a closed world	No incremental learning	Lack of theoretical framework for unknown utilization
Representative Work	EWC [21]	ORE [55]	SOLA [8]

Chapter 3

Preliminaries

In this chapter, we establish the foundational definitions and formal problem statements for Open-world Continual Learning (OWCL) that underpin our research. We also introduce key technical concepts that will be needed throughout this thesis.

To begin with, we present widely-adopted definitions of three core paradigms: Continual Learning (CL), Open-Set Learning (OSL), and Out-of-Distribution (OOD) detection. Building upon these, we then formalize the basic definitions and notation specific to open-world continual learning, which are used throughout all subsequent chapters. Additional chapter-specific definitions will be introduced in their respective contexts.

3.1 General Definitions

Before formalizing the core components of open-world continual learning, we first present the definitions of its two basic paradigms: continual learning and open-set recognition. These concepts, while often studied independently, are combined in a complex interaction in open-world continual learning systems.

3.1.1 Continual Learning

Continual Learning is formally defined as a learning paradigm where models adapt to evolving data distributions across sequential tasks. In this framework, the training samples arrive in temporal order, with each task t potentially drawn from a distinct distribution $\mathbb{D}_t$. A continual learning model $f_\theta(\cdot)$, parameterized by θ , must learn new tasks with minimal access to previous training data and maintain performance on all previously encountered tasks.

We adopt the following notation: tr and te denote training and testing phases, respectively. In addition, $\mathcal{D}_t^{tr} = \{\mathcal{X}_t^{tr}, \mathcal{Y}_t^{tr}\}$ represents the training set of task t , where $\mathcal{X}_t^{tr}$ are the input samples, $\mathcal{Y}_t^{tr}$ their corresponding labels, and $t \in \{1, \dots, T\}$

the task identifiers. Finally, $\mathbb{D}_t^{tr} := P_t^{tr}(\mathcal{X}, \mathcal{Y})$ denotes the true data distribution for task t .

Under the standard continual learning assumption, the testing distribution $\mathbb{D}_t^{te}$ matches $\mathbb{D}_t^{tr}$ for each task t . This leads to our formal definition:

Definition 1 *Continual Learning.* Let $\mathcal{T} = \{\mathcal{T}_1, \dots, \mathcal{T}_t, \dots, \mathcal{T}_N\}$ be a sequence of learning tasks, where each task $\mathcal{T}_t$ comes with its own training data $\mathcal{D}_t^{tr}$ consisting of input-output pairs $(\mathbf{x}_i, y_i)$.

A continual learning model learns a parameterized function $f_{\theta_t} : \mathcal{X} \rightarrow \mathcal{Y}$ by minimizing the expected loss:

$$\mathcal{L}(\theta_t) = \mathbb{E}_{(\mathbf{x}_i, y_i) \sim \mathcal{D}_t^{tr}} [\ell(f_{\theta_t}(x_i), y_i)],$$

where ℓ is a task-specific loss (e.g., cross-entropy). To prevent forgetting knowledge from previous tasks $\mathcal{T}_1, \dots, \mathcal{T}_{t-1}$ learning is additionally constrained by

$$\mathcal{L}_t^{CL}(\theta) = \underbrace{\mathcal{L}_t(\theta)}_{\text{Current task}} + \lambda \sum_{k=1}^{t-1} \underbrace{\Omega_k(\theta, \theta_k^*)}_{\text{Memory of past tasks}}.$$

Here, Ω_k acts as a "memory anchor" [21], encoding knowledge from the task $\mathcal{T}_k$, and λ is a trade-off parameter that regulates the balance between plasticity (learning new tasks) and stability (retaining previous knowledge). Empirically, appropriate values of λ (e.g., in the range $[0.3, 0.7]$) have been shown to effectively mitigate catastrophic forgetting while preserving learning flexibility, as we further demonstrate in the experimental analysis presented in Chapters 4 to 7.

The final goal is to learn one single function f_{θ^*} that works well in all tasks:

$$\theta^* = \arg \min_{\theta} \sum_{t=1}^N \mathcal{L}_t^{CL}(\theta).$$

Continual learning aims to acquire new tasks sequentially without catastrophic forgetting of previously learned knowledge. To achieve this, typical "memory anchors" Ω_k are used, including regularization, replay (store old examples), or architectural (separate storage areas per task).

In real-world applications, two critical pieces of information may be unavailable during training: the task labels $\mathcal{Y}_{tr}^t$ and the task identities t . This practical constraint leads to three fundamental continual learning scenarios, categorized by data arrival patterns and task identifier availability [6]:

Task-Incremental Learning (TIL): In this setting, the model receives explicit task identifiers during both training and inference phases. This task-awareness enables the use of dedicated output modules for each task, typically implemented through a shared backbone network with multiple task-specific classifier heads. For

Table 3.1: Overview of the three CL scenarios.

Scenario	Required at test time
Task-IL	Solve tasks so far, task-ID provided
Domain-IL	Solve tasks so far, task-ID not provided
Class-IL	Solve tasks so far <i>and</i> infer task-ID

instance, a visual recognition system might maintain separate output layers for different domains (e.g., one head for medical images, another for satellite images) while sharing convolutional feature extractors.

Domain-Incremental Learning (DIL): In this scenario, task identity is not provided at test time, so that the model only needs to perform the correct task without identifying which one it is. The assumption is that the task structure remains fixed across domains, while the input distributions vary. A practical example would be an agent operating across multiple environments, where the task/environment identification is not required, while data distributions change as they depend on the specific environment.

Class-Incremental Learning (CIL): As the most demanding scenario, class-incremental learning requires models to not only solve all previously seen tasks but also infer which task a given input belongs to. The system must gradually incorporate new categories while preserving performance on previously learned ones, without access to task identifiers. This mirrors real-world applications like retail product recognition, where new items are continuously added to inventory while maintaining accurate identification of existing products.

3.1.2 Open-Set Learning and Out-of-Distribution Detection

Traditional supervised learning operates under a closed-world assumption, where the label spaces for training ($\mathcal{Y}_{\text{train}}$) and testing are identical. Formally, a model learns a mapping $f : \mathcal{X} \rightarrow \mathcal{Y}_{\text{train}}$ with the guarantee that all test samples belong to $\mathcal{Y}_{\text{train}}$. While effective for constrained environments, this assumption fails in realistic scenarios where systems encounter new categories during deployment.

To address this limitation, various strategies have been proposed in recent years. Two prominent and increasingly studied directions are **Open-Set Learning** that extends classification to include rejection of unknown categories, and **Out-of-Distribution Detection** that focuses on distinguishing in-distribution from out-of-distribution samples. Though both approaches handle distributional shifts, they differ fundamentally in their objectives and formulations.

Definition 2 Open-Set Learning. *In open-set learning, the test-time label space is extended to include unknown classes:*

$$\mathcal{Y}_{\text{test}} = \mathcal{Y}_{\text{train}} \cup \mathcal{Y}_{\text{unknown}}, \quad \mathcal{Y}_{\text{unknown}} \cap \mathcal{Y}_{\text{train}} = \emptyset.$$

The system must simultaneously achieve two objectives: (1) accurate classification of samples $x \in \mathcal{X}$ from a known class $x \in \mathcal{Y}_{\text{train}}$, and (2) reliable detection or rejection of samples from unknown classes $y \in \mathcal{Y}_{\text{unknown}}$.

The above dual requirement implies that an open-set learning model not only maintains discriminative power for established categories but also identifies novel inputs that deviate from the training distribution. OOD detection instead creates binary in/out decision boundaries without explicit class differentiation.

Definition 3 Out-of-Distribution Detection. Let $\mathcal{D}_{\text{in}}$ be the in-distribution data (training distribution), and $\mathcal{D}_{\text{out}}$ be any distribution such that $\text{supp}(\mathcal{D}_{\text{out}}) \cap \text{supp}(\mathcal{D}_{\text{in}}) = \emptyset$. The goal of OOD detection is to learn a function $g : \mathcal{X} \rightarrow \{0, 1\}$ such that:

$$g(x) = \begin{cases} 1 & \text{if } x \sim \mathcal{D}_{\text{in}} \text{ (in-distribution)} \\ 0 & \text{if } x \sim \mathcal{D}_{\text{out}} \text{ (out-of-distribution)} \end{cases}$$

Optionally, if x is classified as in-distribution, it may be passed to a classifier f for label prediction.

Following the two definitions above, Table 3.2 contrasts open-set learning and out-of-distribution detection across several dimensions. Although both paradigms address the limitations of closed-set assumptions, they diverge fundamentally in their objectives and implementations.

Open-set learning methods like OpenMax [15, 57] and C2AE [78] focus on discriminative classification of known classes while rejecting unknowns within an expanded label space. In contrast, out-of-distribution detection techniques such as ODIN [62], Mahalanobis distance [79], and MSP [80] establish binary decision boundaries to separate in-distribution from out-of-distribution samples, without explicit class differentiation.

These foundational concepts share a common structural framework, where data is partitioned into base known classes available during training and unknown classes encountered only during deployment, with no overlap between them by definition. This difference between known and unknown categories forms the base of our investigation into open-world continual learning.

Our work extends these principles by developing novel knowledge transfer mechanisms that bridge continual learning and open-set recognition. At its core, the framework addresses three interconnected challenges: the preservation of base-class knowledge during incremental updates, the transfer of discriminative features between known and unknown categories, and the maintenance of an optimal stability-plasticity balance as the class space evolves dynamically. These objectives are pursued while respecting the inherent constraints of real-world deployment scenarios.

The **open-world environment** we consider significantly generalizes conventional open-set learning and out-of-distribution settings through its temporal dimension,

Table 3.2: Comparison between OSL and OOD Detection.

Aspect	OSL	OOD
Objective	Classify known classes and reject unknowns	Detect whether a sample belongs to the training distribution
Label Space at Test	$\mathcal{Y}_{\text{train}} \cup \mathcal{Y}_{\text{unknown}}$	No specific label set for out-of-distribution samples; only binary in/out decision
Output	Class label or rejection	Binary label: in-distribution vs. out-of-distribution
Assumption	Unknown classes exist and are not seen during training	Outliers come from a distribution disjoint from training data
Application Focus	Open-world classification, e.g., novel class discovery	Model safety, anomaly detection, robust deployment
Representative Methods	OpenMax [57], C2AE [78]	ODIN [62], MSP [80], Mahalanobis [79]

where both base and unknown classes may evolve across learning episodes, and its emphasis on the dynamic interplay between familiar and novel categories. This perspective, which formalizes the interdependence of incremental learning and open-set recognition, is formally defined in the next section.

3.2 Problem Definition

To formally characterize the open-world continual learning challenge, Definition 4 gives the learning objective as a double optimization problem. This formulation fundamentally extends traditional continual learning in two critical aspects: first, by explicitly accounting for open-class samples during testing through an open-set risk term; and second, by requiring simultaneous minimization of both this detection risk and the standard incremental learning error across all tasks. The resulting framework captures the essential tension between maintaining existing knowledge while remaining responsive to novel, previously unseen categories.

Definition 4 (OWCL Problem Formulation.) *Consider a task t characterized by a training set $\mathcal{D}_t^{\text{tr}} = \{\mathbf{x}_i \in \mathcal{X}, y_i \in \mathcal{Y}_t\}_{i=1}^{N_t}$ with M_t classes, and a test set $\mathcal{D}_t^{\text{te}} = \{\mathbf{x}_j \in \mathcal{X}, y_j \in \mathcal{Y}'_t\}_{j=1}^{N'_t}$ where $\mathcal{Y}'_t \supset \mathcal{Y}_t$ (i.e., it contains additional unknown classes). During training, an open-world continual learning model only accesses $\mathcal{D}_t^{\text{tr}}$, while during testing it must handle both known ($\mathcal{Y}_t$) and unknown ($\mathcal{Y}'_t \setminus \mathcal{Y}_t$) classes. Let $h(\cdot)$ be a feature extractor mapping inputs to a latent space, and define $\mu(h(\mathbf{x}), h(\mathcal{D}_t^{\text{tr}}))$ to be the open-set risk measuring detection capability for unknowns, and $\epsilon_{\mathcal{D}_t^{\text{tr}}}(h)$ to be the prediction error on known classes. The open-world continual learning objective*

learns an optimal $h^* \in \mathcal{H}$ that minimizes these two factors across all tasks:

$$h^* = \arg \min_{h \in \mathcal{H}} \left\{ (1 - \lambda) \sum_{t=1}^T \underbrace{\mu(h(\mathbf{x}), h(\mathcal{D}_t^{tr}))}_{\text{Open-set risk}} + \lambda \sum_{t=1}^T \underbrace{\epsilon_{\mathcal{D}_t^{tr}}(h)}_{\text{Incremental error}} \right\} \quad (3.1)$$

where $\mathcal{H}$ is the hypothesis space of possible models, $\lambda \in [0, 1]$ is a trade-off parameter between old and new knowledge, $\epsilon_{\mathcal{D}_t^{tr}}(h)$ is the generally suggested prediction error term on task t [15], μ can be defined on any scoring function for out-of-distribution detection, and T is the total number of tasks.

In practice, μ can be instantiated using a variety of out-of-distribution detection scores, such as energy-based scores, Mahalanobis distance, or margin-based confidence [63, 79]. The choice of μ depends on the specific nature of the feature space and the distribution shift between known and unknown classes.

As formalized in Definition 4, the open-world continual learning paradigm considers a (potentially infinite) sequence of tasks, each associated with a training set containing only known-class examples. The model is required to incrementally learn from these tasks while maintaining the ability to recognize both previously seen and unseen (novel) classes at test time. Therefore, a promising open-world continual learning model must not only accurately classify samples from known classes, but also reliably detect inputs belonging to previously unseen (unknown) classes.

To do this, the model learns a feature extractor that maps inputs into a useful representation space. The learning objective then tries to find the best such feature extractor by balancing two goals: keeping the prediction error on known classes low (called incremental error), and being good at recognizing unknown or out-of-distribution inputs (called open-set risk). A parameter λ adjusts how much the model focuses on remembering what it has learned versus being cautious about unfamiliar data, and this balancing act happens continually as new tasks arrive.

To fix our formal problem formulation and facilitate discussion of the proposed methods, Table 3.3 provides a summary of the key mathematical notations used throughout this thesis.

3.3 Four OWCL Scenarios

Traditional continual learning is typically divided into three scenarios based on whether task identifiers are available at test time: *TIL*, *DIL*, and *CIL*.

However, as outlined earlier in Chapter 1, in the context of open-world continual learning, such assumptions no longer hold. During testing, the model cannot rely on task identifiers, as unknown or open-category samples may appear arbitrarily alongside known ones. Moreover, data distributions in open-world continual learning evolve dynamically through both novel class emergence and recurring open samples.

Table 3.3: Notations used in the OWCL problem formulation.

Symbol	Description
t	Index of the current task
T	Total number of tasks encountered so far
θ_k or (θ_k^*)	Parameters optimized after learning task t
$\mathcal{L}_t(\theta)$	Expected loss on task t
$\Omega_t(\theta, \theta_t)$	Constraint/regularization term to retain knowledge from task t
$\mathcal{D}_t^{tr}, \mathcal{D}_t^{te}$	Training, test set for task t
N_t	Number of training samples for task t
N'_t	Number of test samples for task t
M_t	Number of classes in the training set of task t
M'_t	Number of classes in the test set of task t
$\mathcal{Y}_{i=1}^{M_t}$	Label set for training samples in task t
$\mathcal{Y}_{i=1}^{M'_t}$	Label set for test samples in task t
$\mathbf{x}$	A sample input from the input space $\mathcal{X}$
$h_\theta(\cdot)$ or $f_\theta(\cdot)$	Latent feature extractor or encoder function
h^*	The optimal function to be learned
$\mu(\cdot, \cdot)$	Open risk function, e.g., based on out-of-distribution scoring
$\epsilon_{\mathcal{D}_t^{tr}}(h)$	Prediction error on task t 's training set
$\ell(\cdot, \cdot)$	Task-specific loss function (e.g., cross-entropy)
$\mathcal{H}$	Hypothesis space (function space)
λ	Trade-off hyperparameter balancing open risk and prediction error

This renders conventional continual learning scenarios inadequate for open-world continual learning settings.

Hence, we formulate four progressively challenging open-world continual learning scenarios to reflect different patterns of task evolution and unknown-class recurrence in open-world environments:

- CINO: Class-Incremental Learning with Non-Repetitive Open Samples.** In CINO, each task introduces a disjoint set of new training classes, and the open-category samples encountered during testing do not reappear in future tasks. The model learns each class only once, without access to past data or recurring unknowns. This scenario is particularly suited for one-pass learning applications where data cannot be revisited, such as species identification in biodiversity monitoring or new material classification in industrial inspection.
- CIRO: Class-Incremental Learning with Repetitive Open Samples.** In CIRO, training classes remain disjoint across tasks as in CINO, but open-category samples that appear during testing can recur in future tasks. This setting requires the model to incrementally accumulate knowledge about recur-

ring unknowns without revisiting known class labels. CIRO is representative of real-world settings such as autonomous driving or robotic exploration, where unknown objects (e.g., road signs, dynamic obstacles) may resurface over time, and adaptive open-set handling is essential.

- **KINO: Knowledge-Incremental Learning with Non-Repetitive Open Samples.** KINO relaxes the disjoint-class assumption, allowing training classes to repeat across tasks, and includes distributional shifts over time. However, open-category samples remain non-repetitive—they appear only once during testing and do not recur. This scenario models data streams where user preferences or content characteristics evolve, such as in personalized recommendation systems or adaptive spam filtering, where the model must cope with shifting distributions while handling unseen content once.
- **KIRO: Knowledge-Incremental Learning with Repetitive Open Samples.** KIRO represents the most complex and realistic open-world continual learning scenario, where both known-class distributions and open-category samples can change and recur over time. *While KINO handles shifting distributions of known classes, KIRO additionally requires models to recognize and consolidate knowledge about recurring unknowns.* This setting reflects the demands of long-term, real-world systems such as intelligent surveillance, financial fraud detection, or user behavior modeling, where evolving patterns—whether known or unknown—must be continually tracked, recognized, and adapted to.

Remarks. Compared to task-, domain-, and class-incremental learning settings, knowledge-incremental learning introduces significantly greater complexity. It requires the model to learn new classes incrementally without access to task identifiers, while also maintaining robustness to shifts in data distributions across tasks. Moreover, the unpredictable presence of open-category data means the model must not only recognize unfamiliar classes but also transfer knowledge across both known and unknown examples. These combined challenges make Open World Continual Learning fundamentally more demanding than traditional continual learning paradigms.

3.4 Pretrained Models in OWCL: A Brief Overview

As outlined above, this thesis aims to push continual learning into open-world scenarios, where data distributions shift over time and the model must continually handle previously unseen instances. Within this context, we explore how knowledge can be effectively transferred under such open and dynamic conditions.

The rise of large-scale pretrained models (PTMs) has brought a paradigm shift to both the continual learning and open-set recognition communities, including in

tasks such as out-of-distribution detection. These models are increasingly used as foundational backbones, followed by task-specific fine-tuning. This shift has underscored the critical influence of the backbone architecture on the learning behavior of open-world continual learning models. Furthermore, several open-world continual learning methods are tightly coupled to specific backbone designs, which limits their generalizability across architectures.

To establish the foundations for the open-world continual learning framework proposed in this thesis, this section provides a concise overview of recent PTM-based methods in continual learning. This review both contextualizes our approach within the current research landscape and motivates the architectural choices made in the design of our proposed method.

3.4.1 Attention Mechanism and Transformer Architecture

In recent years, pretrained models have shown strong performance in continual learning, attracting significant attention from the research community. However, these approaches still suffer from catastrophic forgetting, primarily because importance weights associated with previous tasks cannot be recovered once overwritten—even when using a pretrained backbone. To mitigate this issue, many studies have explored Parameter-Efficient Fine-Tuning (PEFT) techniques, which enable model adaptation to new tasks through minimal parameter updates, thereby preserving the knowledge stored in the frozen backbone.

In the vision domain, current methods commonly employ a pretrained Vision Transformer (ViT) [81] as a fixed feature extractor f_θ . The ViT architecture is built upon a sequence of multi-head self-attention layers [82], where attention scores are computed via a scaled dot-product mechanism, allowing the model to capture complex relationships across image patches.

Definition 5 Scaled Dot-Product Attention. Let $K \in \mathbb{R}^{N \times d_k}$ be a key matrix with N key vectors, and $V \in \mathbb{R}^{N \times d_v}$ be a value matrix with N corresponding value vectors. Given a query matrix $Q \in \mathbb{R}^{M \times d_k}$, attention over (K, V) is defined as

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V, \quad (3.2)$$

where the softmax function acts on the rows of matrix $QK^\top \in \mathbb{R}^{M \times N}$.

The Scaled Dot-Product Attention mechanism forms the core computational unit of transformer architectures, enabling models to weigh the relevance of different input elements when producing contextualized representations. However, relying on a single attention function may limit the model’s capacity to capture diverse relationships. To address this, transformers, including (ViTs), use Multi-head Self-Attention (MSA), which extends the basic attention operation by computing multiple attention

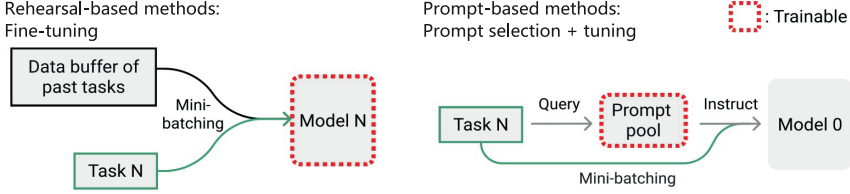


Figure 3.1: In contrast to rehearsal-based continual learning approaches that sequentially adapt full or partial model parameters to new tasks and rely on a rehearsal buffer to alleviate catastrophic forgetting, prompt-based methods adopt a unified backbone model augmented by learned prompts [32, 83]. Task-specific information is encapsulated within the prompts, thereby eliminating the necessity of a rehearsal buffer. Furthermore, different methods have different ways of selecting and updating prompts dynamically, usually enabling task-agnostic inference at test time without explicit task identity.

functions in parallel. Each head attends to different subspaces of the input representation, allowing the model to capture richer and more nuanced dependencies. The formal definition of the MSA layer is given below.

Definition 6 Multi-head Self-Attention (MSA) Layer. Let X^Q, X^K, X^V denote the input query, key, and value matrix, respectively, where $X^Q = X^K = X^V = [x_1, \dots, x_N]^T \in \mathbb{R}^{N \times d}$, and N is the length of the input sequence. The output is expressed as

$$MSA(X^Q, X^K, X^V) := \text{Concat}(h_1, \dots, h_m)W^O \in \mathbb{R}^{N \times d}, \quad (3.3)$$

$$h_i := \text{Attention}(X^Q W_i^Q, X^K W_i^K, X^V W_i^V), \quad i \in [m] \quad (3.4)$$

where $W^O \in \mathbb{R}^{md_v \times d}$, $W_i^Q \in \mathbb{R}^{d \times d_k}$, $W_i^K \in \mathbb{R}^{d \times d_k}$, and $W_i^V \in \mathbb{R}^{d \times d_v}$ are projection matrices, and m is the number of heads in the MSA layer. In ViTs, they use $d_k = d_v = d/m$.

3.4.2 Prompt-based CL Methods with PEFT Techniques

Building on the transformer-based backbone discussed above, prompt-based tuning has emerged as a flexible, lightweight, and rehearsal-free approach for adapting frozen models to new tasks in continual learning settings [32, 83, 84]. Instead of modifying the entire network, these methods introduce a small number of trainable parameters, known as prompts, that augment the model’s input representations. These prompts serve to encode task-specific knowledge and interact with the attention computations in the PTM-based backbone, enabling efficient adaptation while preserving previously learned capabilities. This design enables task-specific knowledge adaptation efficiently, making it particularly suitable for continual learning with limited memory or parameter sharing.

As illustrated in Figure 3.1, prompt-based approaches [83, 84] enhance model flex-

ibility by integrating learnable prompt vectors into the Multi-Head Self-Attention mechanism. These prompt tokens are injected into the query, key, and value matrices at each attention layer, effectively conditioning the model on new tasks. Let the prompt parameters be denoted by $p \in \mathbb{R}^{L_p \times d}$, where L_p is the prompt length and d denotes the hidden dimensionality.

According to prior work [83], two main formulations have been widely adopted: Prompt Tuning (ProT) [85] and Prefix Tuning (PreT) [86]. In Prompt Tuning, the same prompt vectors are concatenated to the inputs of the query, key, and value projections simultaneously. In contrast, Prefix Tuning splits the prompt into two components, $p^K, p^V \in \mathbb{R}^{\frac{L_p}{2} \times d}$, which are prepended to the key and value matrices only, leaving the query stream unchanged. This decoupled design enables finer control over attention flow and has demonstrated improved performance across various downstream tasks:

$$\begin{aligned} f_{\text{prompt}}^{\text{Pre-T}}(p, \mathbf{X}^Q, \mathbf{X}^K, \mathbf{X}^V) &:= \text{MSA} \left(\mathbf{X}^Q, \begin{bmatrix} p^K \\ \mathbf{X}^K \end{bmatrix}, \begin{bmatrix} p^V \\ \mathbf{X}^V \end{bmatrix} \right) \\ &= \text{Concat}(\tilde{h}_1, \dots, \tilde{h}_m) W^O. \end{aligned} \quad (3.5)$$

A broad spectrum of prompt-based continual learning methods has been developed to mitigate catastrophic forgetting by dynamically introducing new prompts as learning progresses. In this paradigm, prompts act as task-specific adapters: for each new task, a set of adaptive prompts is either generated or selected to guide the model's behavior. This modular design enables the model to retain task-relevant knowledge and reuse it effectively when encountering related inputs. At inference time, appropriate prompt configurations can be retrieved to enable accurate prediction on previously learned tasks [32].

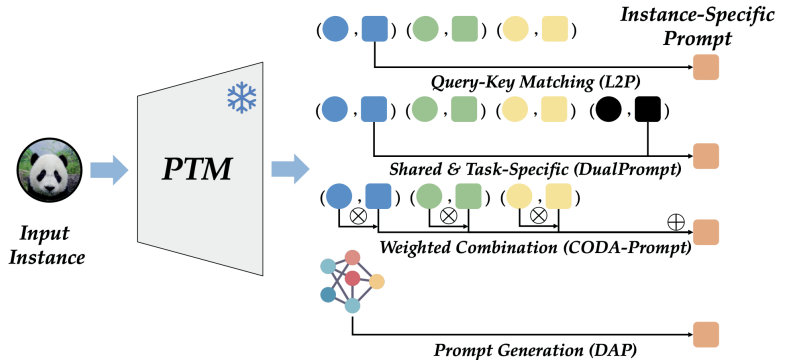


Figure 3.2: Different kinds of prompt selection, including key-value matching, shared and task-specific retrieval, attention-based combination, and instance-specific prompt generation.

An overview of prompt-based continual learning methods is provided in [87], and visually summarized in Figure 3.2. This figure categorizes a range of prompt selection strategies, including hard prompt retrieval (L2P) [32], structured prompt grouping (DualPrompt) [83], attention-based soft aggregation (CODA Prompt) [88], and meta-generated prompts (DAP) [89].

One of the earliest contributions in this line of work, L2P [32], incorporates prompt learning into the continual learning setting by introducing a shared pool of learnable prompts. A key-query matching mechanism dynamically selects a subset of prompts from the pool based on the current input, enabling input-dependent adaptation. Building upon this, DualPrompt [83] introduces a structured decomposition of the prompt space into General Prompts (G-Prompts), which encode task-agnostic knowledge, and Expert Prompts (E-Prompts), which specialize in capturing task-specific features. This separation encourages generalization while preserving task-level specificity.

Similarly, S-Prompt [90] adopts a task-specific prompt learning strategy, maintaining a separate prompt for each task, akin to L2P but without retrieval. In contrast, CODA-Prompt [88] enhances flexibility by growing the prompt pool over time and using an attention-based mechanism to softly aggregate prompts. Each prompt is weighted by task-specific attention scores, allowing the model to form context-aware combinations instead of relying on hard selection.

Moving beyond explicit retrieval mechanisms, DAP [89] introduces a meta-network, typically a multilayer perceptron (MLP), that generates prompts conditioned on input representations. This instance-level prompt generation enables fine-grained adaptation and removes the need for a fixed prompt pool.

Recent studies have also explored alternative strategies, such as appending all available prompts to the input or adopting *visual prompts*, i.e., low-level pixel-space modifications to the image itself. Furthermore, the rise of multimodal pre-trained models has led to approaches that leverage textual information to guide prompt selection or generation, particularly within vision-language pretraining frameworks.

A notable advancement in this space is HiDe-Prompt [84], which redefines the continual learning objective through a hierarchical decomposition. By optimizing each sub-objective independently, HiDe-Prompt enhances both performance and modularity, establishing a new benchmark for continual image classification.

Therefore, inspired by the aforementioned prompt-based continual learning frameworks, we further investigate their applicability and extensibility under the open-world continual learning setting. In open-world continual learning, models are expected to make inferences without access to task identifiers and to distinguish between known and unknown classes. Prompt-based methods naturally align with the demands of open-world continual learning, providing a task-agnostic solution to this challenge. For instance, CODA-Prompt’s soft selection strategy enables inference

without relying on explicit task labels, aligning well with open-world continual learning’s open-set and label-free nature. Similarly, dynamic prompt generation methods like DAP offer instance-level adaptability, which is essential when encountering unknown or evolving data distributions. These properties position prompt-based continual learning methods as promising candidates for scalable, open-world scenarios, particularly in settings with data privacy constraints where revisiting past samples is infeasible and task-agnostic.

3.4.3 Other PTMs-based CL Methods with PEFT Techniques

Recent advances in parameter-efficient continual learning studies extend beyond prompt-based methods, with adapter-based approaches demonstrating particular success. Works like SimpleCIL [91], ADAM [92], and EASE [93] employ lightweight modular architectures [94] that attach task-specific adapter layers to frozen pre-trained backbones. These adapters achieve dual objectives: encoding new task knowledge through localized parameter updates, while preserving existing representations via regularization that anchors new features near previously learned embeddings. This dual mechanism effectively balances plasticity and stability, minimizing interference during sequential learning.

Complementing these methods, vision-language models like CLIP [95] provide inherent open-world capabilities through their contrastive image-text pretraining. Their zero-shot classification ability, enabled by measuring alignment between visual features and textual prompts (e.g., "a photo of a dog"), offers natural open-set detection through confidence thresholding. However, CLIP’s static architecture faces two key limitations in open-world continual learning settings: (i) fixed decision boundaries that cannot adapt to evolving task distributions, and (ii) domain gaps when applied to specialized datasets. Our proposed Pro-KT and MOB frameworks address these gaps through dynamic prompt tuning and adaptive hypersphere boundaries, respectively.

These developments reflect a broader paradigm shift toward modular, efficient adaptation of foundation models. Contemporary approaches increasingly favor lightweight task-specific components (prompts/adapters) over full model retraining, explicit mechanisms for knowledge preservation during incremental updates, and unified architectures that support both closed-set and open-set recognition.

Our work contributes to this direction by introducing preference-conditioned adaptation techniques into the PTM backbone. Specifically, our method dynamically adjusts to task requirements through learned gating mechanisms, maintains backward compatibility with previously learned representations, and jointly optimizes for both known class accuracy and unknown detection. This approach demonstrates superior adaptability, robustness, and generalization capabilities compared to static

architectures, as evidenced by our experiments simulating dynamic, unpredictable, and open-ended data environments.

Chapter 4

Learning to Prompt Knowledge Transfer for Open-world Continual Learning

As discussed in Chapter 3, open-world continual learning (OWCL) extends traditional continual learning by embracing the open-world assumption, an appealing yet highly challenging paradigm. In this chapter, we begin by presenting Figure 4.1 and Figure 4.2 to visually illustrate the open-world continual learning setting and its key distinctions. We then further introduce the core challenges associated with open-world continual learning, setting the stage for the methods and solutions explored throughout the chapter.

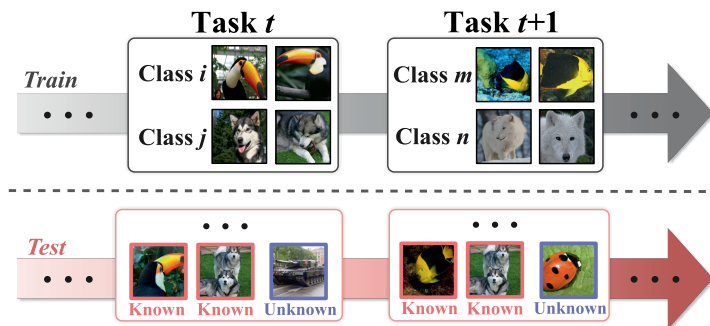


Figure 4.1: A simple illustration of OWCL.

As illustrated in Figure 4.1, each new task incrementally introduces new classes (e.g., *class m* and *class n* in task $t + 1$), while the test sets for both task t and task $t + 1$ contain open classes (highlighted in blue) that were not observed during training. This reflects the fundamental challenge of open-world continual learning, where models must cope with both continual task acquisition and the presence of unknown categories unseen during training.

To further highlight the limitations of existing continual learning methods under open-world continual learning settings, we conducted a series of preliminary experiments. These experiments aim to uncover the performance bottlenecks that arise when conventional continual learning models are extended to handle open-world scenarios. As shown in Figure 4.2, we evaluate the evolving model performance on the test sets of all tasks after sequentially learning each new task. This evaluation compares a representative continual learning baseline with our proposed approach.

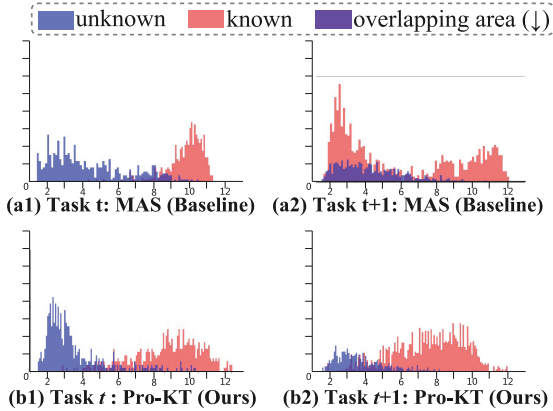


Figure 4.2: Distributions of the outputted logits scores of test samples, using MAS (a typical continual learning baseline) and the proposed *Pro-KT* on Task 1 and Task 2, respectively. The x -axis is the unscaled class-maximum logits of samples, and the y -axis shows the number of samples.

In Figure 4.2 (a1) and (b1), both MAS [96] and our proposed method demonstrate an ability to distinguish known from unknown classes in the previous task (t). However, in the current task ($t + 1$), MAS fails to maintain this separation. As seen in Figure 4.2 (a2), the logit distributions for unknown and known samples significantly overlap, indicating MAS’s inability to detect open-set entities, resulting in degraded performance. In contrast, our method *Pro-KT*, depicted in Figure 4.2 (b2), exhibits better separation between knowns and unknowns, with reduced overlap in the logit distributions.

These observations form the basis of our main motivation: to address the research question **RQ1** introduced in chapter 1 by developing an open-world continual learning model that not only accumulates knowledge over time but also uses it to accurately distinguish between known and unknown categories. Specifically, our goal is to expand the open-set boundary to improve recognition of unfamiliar instances. After learning each new task, the previously unseen classes are incorporated into the set of knowns, progressively refining the model’s understanding of the environment [97].

Challenges and Our Approach: Open-world continual learning presents several core challenges. (1) The first is the issue of *knowledge stability*, which arises when new tasks diverge significantly from previously encountered ones. In such cases, newly acquired knowledge may conflict with earlier representations, risking degradation of past learning. (2) The second challenge is *knowledge plasticity*, which requires the model not only to incrementally learn new tasks but also to reliably detect open or unknown classes during testing, and ideally, to integrate those unknowns into the set of knowns. The tension between stability and plasticity is fundamental to open-world continual learning, as it governs the model’s ability to retain prior knowledge while adapting to novel information. Effectively balancing these forces is crucial for ensuring both robustness and adaptability in evolving environments.

To address these challenges, we propose a new framework: **Prompt-enhanced Knowledge Transfer (*Pro-KT*)**. *Pro-KT* represents a significant advancement in prompt-based continual learning. It features a modular prompt bank to facilitate structured knowledge transfer and introduces two adaptive strategies for open-set boundary modeling to manage the dynamic nature of open-world learning.

To mitigate Challenge (1) (knowledge stability), *Pro-KT* constructs a prompt bank that encapsulates both general and task-specific knowledge. These prompts act as modular, instructional units that guide the model’s behavior across tasks. By selectively activating relevant prompts, the model achieves effective, scalable transfer without overwriting prior knowledge.

To address Challenge (2) (knowledge plasticity) and answer research question RQ1 from Chapter 1.3, *Pro-KT* introduces two adaptive thresholding mechanisms for open-set recognition. These strategies allow the model to dynamically refine decision boundaries as it accumulates new knowledge, supporting continual integration of novel classes while preserving discrimination capabilities.

Our main contributions and advantages of *Pro-KT* are summarized as follows:

- *Prompt-based, rehearsal-free knowledge transfer paradigm.* We introduce a novel prompt bank designed for open-world continual learning, enabling structured representation and transfer of knowledge without relying on data replay. This decouples learning from memory-intensive rehearsal and provides a scalable alternative for continual learners.
- *Balanced handling of stability and plasticity via task-generic and task-specific prompts.* The prompt bank encodes both general and specialized task knowledge, enabling flexible and efficient adaptation. Prompts are selected dynamically based on the task context, allowing the model to generalize across tasks while remaining sensitive to new information.
- *Adaptive open-set recognition through dynamic boundary modeling.* We develop two strategies for task-aware thresholding, allowing the model to distinguish known from

unknown classes throughout continual learning. These adaptive boundaries evolve with the learning process, supporting seamless incorporation of new concepts.

- *Empirical effectiveness and robustness across real-world benchmarks.* Extensive experiments on two real-world open-world continual learning benchmark suites demonstrate that *Pro-KT* consistently outperforms strong baselines across a wide range of tasks. Qualitative visualizations and case studies confirm the model’s practical utility in handling open-set and dynamic learning conditions.

This chapter is based on the following publication [30]:

- **Li, Y.**, Yang, X., Wang, H., Wang, X., and Li, T.: Learning to Prompt Knowledge Transfer for Open-World Continual Learning. In: The 38th Annual AAAI Conference on Artificial Intelligence. AAAI ’24 (2024) 38(12), 13700-13708.

4.1 Objectives of OWCL

Recall from Definition 4 in Chapter 3 that the *Open-world Continual Learning* problem arises when a learner is exposed to a (potentially infinite) sequence of tasks and must handle the emergence of novel, previously unseen objects during testing. To prevent misclassification, such novel/open instances should be identified as unknowns rather than being erroneously assigned to known categories. Subsequently, when ground-truth labels for these unknowns become available in future tasks’ training sets, the learner should evolve them into known classes and integrate them into its knowledge base accordingly.

Furthermore, the learner must learn continuously and incrementally on the job without access to training data from previous tasks. In other words, it must perform *incremental learning without forgetting* [30], thereby minimizing the incremental prediction error while retaining knowledge from earlier tasks. Notably, retraining the model from scratch is not an option.

In this chapter, we formally refine the definition and objectives of open-world continual learning. The notation used throughout this chapter is summarized in Table 4.1. It is worth noting that, unless otherwise specified, all superscripts in this chapter denote task identifiers.

Definition 7 Objectives of OWCL. *open-world continual learning learns from a sequence of tasks, $1, \dots, t, \dots$. Each task t provides a training dataset $\mathbf{D}_{tr}^t = \{(x_i^t, y_i^t)_{i=1}^n\}$, where n is the number of training samples in task t , $x_i^t \in \mathbf{X}^t$ is an input instance, and $y_i^t \in \mathbf{Y}^t$ is its corresponding class label. Here, $\mathbf{Y}^t$ denotes the set of all classes in task t . The label sets are mutually disjoint, i.e., $(\mathbf{Y}^t \cap \mathbf{Y}^{\hat{t}} = \emptyset, \forall t \neq \hat{t})$, and the cumulative set of all training labels is given by $\bigcup_{t=1}^T \mathbf{Y}^t = \mathbf{Y}_{tr}$. Similarly, we denote by $\mathbf{Y}_{te}$ the cumulative set of all labels at the testing phase.*

Table 4.1: Notation used in this chapter.

Notation	Explanation
$(\mathbf{P}, \mathbf{K})$	Prompt bank
N	Total # of tasks
M	Total # of prompts learned from each task
$(\mathbf{k}^t, \mathbf{p}^t)$	The set of key-query pair of t
(k_i^j, p_i^j)	The i key-query pair in task j
L_P	The token length of each prompt
D_e	The embedding dimension
$\mathbf{Q}_x$	The query function
$\mathbf{h}_i^t$	The embedding of x_i^t
$\mathbf{P}_{S_K}$	A subset of the prompt bank
$\mathbf{h}_i^{\prime t}$	The input embedding extended by $\mathbf{P}_{S_K}$
f_{pr}	Pre-trained backbone
f_{θ^t}	Trainable classifier for t
$prob_i^t$	Maximum unscaled softmax entropy of x_i^t
$ \mathbf{X}^t $	Total # of task t 's training samples
r	A deviation degree
λ	A trade-off degree
$[\mu^1, ..., \mu^t, ..., \mu^N]$	All learned open detection thresholds.

Open-world continual learning accommodates the emergence of previously unseen or open classes during testing, meaning that we do not require $\mathbf{Y}_{te}$ to be a subset of $\mathbf{Y}_{tr}$. Therefore, the goal of open-world continual learning is to learn a single predictive function $f_{\theta_t} : \mathbf{X} \rightarrow \mathbf{Y}$ after learning task t that can: 1) detect and reject open/unknown test instances, and 2) correctly classify known test instances—while retaining knowledge across tasks and mitigating catastrophic forgetting.

4.2 Methodology

Figure 4.3 presents the overall architecture of the proposed *Pro-KT* framework, which operates in two stages: a *training* phase and a *testing* phase.

As previously discussed, traditional continual learning approaches are typically based on the closed-world assumption, an unrealistic constraint in practical settings, where previously unseen and unknown classes frequently emerge. Although recent efforts have explored this problem under the open-world continual learning paradigm, the task remains inherently challenging. Existing methods often misclassify novel inputs, accumulate fragmented or incoherent knowledge, or even reinforce incorrect patterns, while largely neglecting the potential of explicit knowledge transfer mechanisms.

To overcome these challenges, *Pro-KT* introduces a prompt-based framework designed to encode, transfer, and evolve task-specific and task-generic knowledge in an open-world setting. Unlike rehearsal-based methods, *Pro-KT* avoids storing past

4. LEARNING TO PROMPT KNOWLEDGE TRANSFER FOR OPEN-WORLD CONTINUAL LEARNING

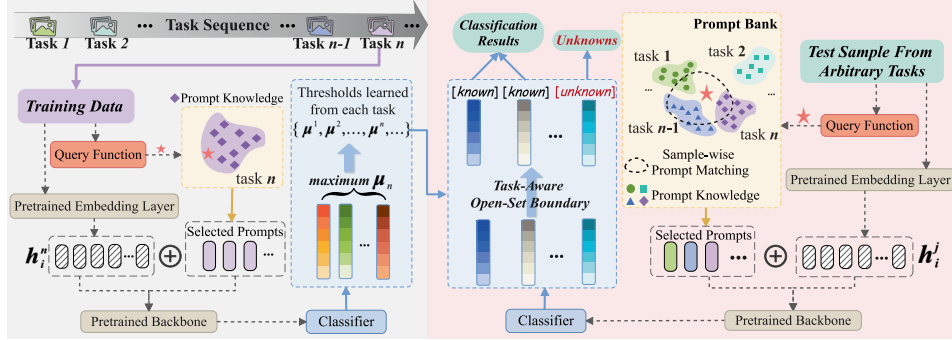


Figure 4.3: Overview of the proposed *Pro-KT* framework. Left: Training phase. Right: Testing phase. During training, input samples are first projected using a query function. *Pro-KT* learns a new set of prompts and stores them in a prompt bank. Each sample is then enriched with selected prompts, and the resulting prompt-enhanced embeddings are processed by a frozen pre-trained backbone. The final representations are passed to a trainable classifier. Additionally, an adaptive threshold for a task-aware open-set boundary is learned based on the softmax entropy of these enhanced representations. In the *testing phase*, test samples are similarly projected and matched with prompts from the prompt bank using a sample-wise selection strategy, enabling transfer of both task-specific and generic knowledge. The enriched embeddings are fed into the unified classifier to obtain unscaled logit scores. Finally, *Pro-KT* classifies each sample as either *[known]* or *[unknown]*, using the learned adaptive threshold.

data by maintaining a structured prompt bank that supports modular, efficient, and scalable knowledge transfer across tasks.

With this motivation and problem setting in mind, we now detail the core components of the *Pro-KT* framework, including its prompting strategy, training objectives, and overall learning procedure.

4.2.1 Prompt Bank

We propose leveraging a prompt bank as the central mechanism for knowledge transfer in open-world continual learning. Specifically, prompts are used to encode both task-generic and task-specific knowledge, which are retained in a dedicated prompt bank. These stored prompts are later transferred to new tasks, enabling effective reuse of prior knowledge and supporting long-term knowledge stability across the continual learning process.

Given the current task n with its associated training data $D_{tr}^n = (\mathbf{X}^n, \mathbf{Y}^n)$, the learner derives a set of prompts $(p_1^n, \dots, p_i^n, \dots)$ tailored to this task. These task-specific prompts are then stored alongside prompts from previous tasks in the prompt

bank:

$$\begin{aligned}\mathbf{P} &= \bigcap_{n=1}^N (\mathbf{p}^1, \dots, \mathbf{p}^n, \dots, \mathbf{p}^N) \\ &= \{(p_1^1, p_2^1, \dots, p_M^1), \dots, (p_1^n, p_2^n, \dots, p_M^n), \dots, (p_1^N, p_2^N, \dots, p_M^N)\},\end{aligned}\quad (4.1)$$

where N is the total number of tasks and M denotes the number of prompts learned per task. Each prompt $p_i^j \in \mathbb{R}^{L_p \times D_e \times NM}$ represents the i -th prompt from task j , where L_p is the prompt token length and D_e is the embedding dimension. This structured prompt bank promotes knowledge transfer by capturing both specific and generalizable representations from each task.

A key challenge, however, lies in how to encode and selectively retrieve the most relevant prompting knowledge. Inspired by the key-value query mechanism introduced in L2P [32], we associate each prompt with a learnable key. For a given task n , we construct key-value pairs to support sample-wise prompt matching as follows:

$$(\mathbf{k}^n, \mathbf{p}^n) = (k_1^n, p_1^n), \dots, (k_M^n, p_M^n), \quad (4.2)$$

where $\mathbf{p}^n = p_m^n M$ is the set of prompts and $\mathbf{k}^n = k_m^n M$ is the corresponding set of keys. Upon learning task n , the prompt bank maintains the full set of learned key-prompt pairs as:

$$(\mathbf{K}, \mathbf{P}) = \{(\mathbf{k}^1, \mathbf{p}^1), (\mathbf{k}^2, \mathbf{p}^2), \dots, (\mathbf{k}^N, \mathbf{p}^N)\}. \quad (4.3)$$

To facilitate more flexible and automatic knowledge transfer, we introduce a sample-wise prompt matching mechanism, where each input sample autonomously selects the most relevant prompts from the prompt bank to serve as auxiliary knowledge. Given an input x_i^n from task n , we first project it into a feature space $\mathbb{R}^{D_e}$ using a pre-trained projection layer $\mathbf{Q}_x$:

$$\mathbf{h}_i^n = \mathbf{Q}_x \cdot x_i^n. \quad (4.4)$$

This projected embedding $\mathbf{h}_i^n$ then acts as a query to retrieve the top- K most relevant prompts by matching against all stored keys in the prompt bank. Specifically, we solve the following optimization problem to identify the K keys with the smallest distance (measured using cosine similarity in this work):

$$\underset{[1, K]}{\operatorname{argmin}} \sum_{j=1}^N \operatorname{dis}(\mathbf{h}_i^n, \mathbf{k}^j), \text{ with } \mathbf{k}^j = \{k_m^j\}_M. \quad (4.5)$$

This procedure enables each input to retrieve the most relevant prompts purely based on semantic similarity, without requiring access to any task identity. Once selected, the input embedding $\mathbf{h}_i^n$ is augmented with the retrieved subset of prompts

$\mathbf{P}S_K \subset \mathbf{P}$ from the prompt bank:

$$\mathbf{h}_i^m = \mathbf{h}_i^n \oplus \mathbf{P}_{S_K}, \quad (4.6)$$

where $\oplus$ denotes dimension-wise concatenation, and the subset size K is smaller than the total number of prompts M per task.

To handle the presence of unknown classes during testing, these enriched embeddings are subsequently used to learn an adaptive detection boundary, which allows the model to distinguish between known and unknown samples on a per-task basis.

4.2.2 Task-Aware Open-Set Boundary

In this section, we describe how to establish an adaptive boundary between known and unknown samples during the open-world continual learning process. To this end, we propose two adaptive thresholding schemes for open-set detection, aimed at determining this boundary in a task-aware manner.

During training, the prompt-based knowledge stored in the prompt bank is derived from the labeled training data of each task. Crucially, we leverage the information learned from known classes to help detect unknown samples. Specifically, we propose learning a threshold based on the softmax entropy of prompt-enhanced embeddings, using this known-class information to estimate a decision boundary that separates knowns from unknowns.

After completing the training for task n , each prompt-enhanced training sample $\mathbf{h}'_i{}^n$ is passed through a pre-trained backbone f_{pr} , and the resulting representation is fed into a trainable classifier f_{θ^n} :

$$prob_i^n = f_{\theta^n}(f_{pr}(\mathbf{h}_i^m)), \quad (4.7)$$

where $prob_i^n$ represents the maximum value of the unscaled softmax entropy corresponding to the training sample $\mathbf{h}_i^m$.

To derive the open-set detection threshold μ , we compute the average softmax entropy across all training samples $\{prob_1^n, \dots, prob_i^n, \dots\}$ from task n and scale it using a deviation factor r :

$$\mu^n = r \cdot \frac{1}{|\mathbf{X}^n|} \sum_{i=1}^{|\mathbf{X}^n|} prob_i^n, \quad (4.8)$$

where $|\mathbf{X}^n|$ denotes the number of training samples, and r is a hyperparameter that controls the sensitivity of the threshold.

Once computed, the threshold μ^n for task n is stored alongside thresholds from previous tasks. This collection $[\mu^1, \dots, \mu^n]$ allows the model to dynamically retrieve a suitable threshold during inference in the testing phase.

In the testing phase, when evaluating whether a test sample x_i^j from any task T^j belongs to a known or unknown class, we use two simple yet effective strategies to select the most appropriate threshold from the stored set. These strategies guide the open-set detection mechanism and help maintain robust performance across tasks.

We consider two testing scenarios for open-set detection. In the first scenario, task identifiers (task IDs) are available at test time. This allows the model to directly use the threshold associated with the given task ID j for open-set classification:

$$\begin{aligned} x_i^j : f_{\theta^n}(\mathbf{h}_i^j) \leq \mu^j &\rightarrow [unknown], \\ \text{otherwise, } &\rightarrow [known]. \end{aligned} \quad (4.9)$$

In the second scenario, task IDs are unavailable during testing—a more realistic and challenging setting for open-world continual learning. As the model learns tasks incrementally, the prompt bank accumulates knowledge from all previously seen tasks. In this case, the model selects the most recently learned threshold μ^n learned from the current task n for the open-set detection boundary:

$$\begin{aligned} x_i^j : f_{\theta^n}(\mathbf{h}_i^j) \leq \mu^n &\rightarrow [unknown], \\ \text{otherwise, } &\rightarrow [known]. \end{aligned} \quad (4.10)$$

After applying this decision rule, if a sample is classified as an unknown object, it is labeled as $[unknown]$ and set aside for potential inclusion in future tasks. However, if the sample is classified as belonging to one of the known classes, the model proceeds to assign it the appropriate class label from the set of previously learned categories.

4.2.3 The Overall Objective and Algorithms

During the training phase of task n , the prompt bank is updated by learning M new prompts along with their associated keys. The prompt bank $(\mathbf{K}, \mathbf{P})$ retains all the prompt knowledge accumulated from tasks T^1 through n . For each known training sample, the model retrieves the top- K most relevant prompts using the key-query matching described in Equation 4.5, and extends the corresponding embedding with auxiliary prompt knowledge as defined in Equation 4.6. At the same time, the open-set detection boundary is refined through the adaptive threshold-setting strategies proposed earlier.

The overall training objective for task n can be formally defined as:

$$\min_{\theta^n, (\mathbf{K}^n, \mathbf{P}^n)} \{ \mathcal{L}(f_{\theta^n}(f_{pr}(\mathbf{h}^n)), \mathbf{y}^n) + \lambda \sum_{i=0}^M dis(\mathbf{h}^n, k_i^n) \}, \quad (4.11)$$

4. LEARNING TO PROMPT KNOWLEDGE TRANSFER FOR OPEN-WORLD CONTINUAL LEARNING

where $\mathbf{h}^n$ denotes the projected embeddings of the training samples. The first term represents the standard softmax cross-entropy loss, promoting accurate classification of known samples. The second term is a regularization loss that encourages the selected prompt keys to remain close to the query embeddings, thus reinforcing meaningful key-prompt associations. The hyperparameter λ controls the trade-off between classification performance and prompt alignment.

As previously emphasized, during the training of task n , only the prompt set p^n is updated. In contrast, during testing, the model selects prompts from the entire prompt bank $\mathbf{P}$ to perform open-world prediction.

To provide a clearer understanding of the proposed *Pro-KT* framework, we present the training procedure in algorithm 1 and the testing procedure in algorithm 2, respectively.

Algorithm 1: The Training Process of *Pro-KT*

Input: training set $D_{tr}^n = \{(\mathbf{X}^n, \mathbf{Y}^n)\}$, a prompt bank $(\mathbf{K}, \mathbf{P}) = \{(\mathbf{k}^n, \mathbf{p}^n)\}_{n=1}^N$, trainable classification head of the n -th task f_{θ^n} , a pre-trained Backbone f_{pr} .

Initialize: $(\mathbf{K}, \mathbf{P})$, θ^n , μ^n .

```

1: for  $n = 1, 2, \dots, N$  do
2:   for each epoch do
3:     for  $(x_i^n, y_i^n)$  in  $(\mathbf{X}^n, \mathbf{Y}^n)$  do
4:       Map  $x_i^n$  to  $\mathbf{h}_i^n$  via Equation 4.4;
5:       Calculate the cosine similarity of  $x_i^n$  and per key from  $(\mathbf{k}^n, \mathbf{p}^n)$ ;
6:       Select  $(\mathbf{k}_{S_K}^n, \mathbf{p}_{S_K}^n)$  from  $(\mathbf{k}^n, \mathbf{p}^n)$  by top-K algorithm;
7:       Extend  $\mathbf{h}_i^n$  to  $\mathbf{h}_i^m$  via Equation 4.6;
8:       Calculate the prompted feature by  $f_{pr}(\mathbf{h}_i^m)$ ;
9:       Calculate the sample loss via Equation 4.11;
10:    end for
11:    Update  $\theta^n$ ,  $(\mathbf{k}^n, \mathbf{p}^n)$  by backpropagation;
12:    Calculate the maximum unscaled probability of each training sample via Equation 4.7;
13:    Calculate the open-set threshold  $\mu^n$  via Equation 4.8;
14:    updat the set of selectable threshold by appending  $\mu^n$  to  $\{\mu^1, \dots, \mu^{(n-1)}\}$ ;
15:  end for
16:  Update  $(\mathbf{K}, \mathbf{P})$ .
17: end for

```

The complete training procedure of *Pro-KT* is presented in algorithm 1. For each task $n \in 1, \dots, N$, the model performs incremental learning using the task-specific labeled dataset $D_{tr}^n = (\mathbf{X}^n, \mathbf{Y}^n)$, in conjunction with the evolving prompt bank $(\mathbf{K}, \mathbf{P})$ and a trainable classification head f_{θ^n} .

During each training epoch, for every input-label pair (x_i^n, y_i^n) from task n , the model begins by projecting the input into an initial embedding $\mathbf{h}_i^n$ (Line 4). It

then computes cosine similarities between this embedding and all keys in the task-specific prompt set $(\mathbf{k}^n, \mathbf{p}^n)$, retrieving the top- K most relevant prompt pairs via a similarity-based matching mechanism (Lines 5–6). These selected prompts are used to enhance the input representation, yielding the prompt-augmented feature vector $\mathbf{h}_i^n$ (Line 7), which is subsequently processed by the frozen pre-trained backbone f_{pr} to extract the final prompted representation (Line 8).

The model computes the classification loss for each training sample using the predicted logits and the corresponding ground-truth label. This loss is then used to jointly update the classification head θ^n and the task-specific prompt parameters $(\mathbf{k}^n, \mathbf{p}^n)$ via backpropagation (Line 9).

At the end of each epoch, the model estimates a task-specific open-set threshold μ^n by evaluating the maximum unnormalized probability for each training sample. This threshold is then stored and added to the cumulative set of thresholds across tasks to support task-aware open-set recognition during inference at the testing phase (Lines 10–11).

Once all tasks have been trained, the prompt bank $(\mathbf{K}, \mathbf{P})$ is updated to include the newly acquired prompts, thereby consolidating transferable knowledge. This updated prompt bank acts as a transferable memory module that enhances representation learning and supports knowledge transfer in subsequent tasks (Line 13).

Algorithm 2: The Test Process of *Pro-KT*

Given: the prompt bank $\{(\mathbf{k}^n, \mathbf{p}^n)\}_{n=1}^N$ trained by current task n , a set of selectable thresholds, trained classification head f_{θ^n} , pre-trained backbone f_{pr} .

Input: a test example x_i^j from T^j .

- 1: Map x_i^j to $\mathbf{h}_i^j$ via Equation 4.4 and select $(\mathbf{K}_{S_K}, \mathbf{P}_{S_K})$ from $(\mathbf{K}, \mathbf{P})$ via Equation 4.5;
 - 2: Extend $\mathbf{h}$ to $\mathbf{h}_i^j$ via Equation 4.6 and feed into pre-trained backbone f_{pr} ;
 - 3: Get the $prob_i^j$ via Equation 4.7;
 - 4: **if** task IDs are available **then**
 - 5: Open detection via Equation 4.9;
 - 6: **else**
 - 7: Open detection: via Equation 4.10;
 - 8: **end if**
 - 9: **if** x_i^j detected as $[known]$ **then**
 - 10: output with known classes
 - 11: **end if**
-

During inference, *Pro-KT* utilizes the learned prompt-based representations to enable accurate and open-aware predictions. The full testing procedure is outlined in algorithm 2.

Given a test sample x_i^j from task T^j , the model performs inference using the learned prompt bank $(\mathbf{k}^n, \mathbf{p}^n)_{n=1}^N$, the trained task-specific classification heads f_{θ^n} , and a shared pre-trained feature extractor f_{pr} .

The input x_i^j is first projected into its initial embedding $\mathbf{h}_i^j$. Based on cosine similarity with the stored prompt keys, the model selects a top- K subset $(\mathbf{K}_{S_K}, \mathbf{P}_{S_K})$ from the prompt bank to enrich the input representation, resulting in an enhanced embedding $\mathbf{h}_i^j$ (Line 1). This augmented embedding is then passed through the frozen backbone f_{pr} to obtain the final feature representation (Line 2), which is used to compute the prediction probability vector $prob_i^j$ (Line 3).

To determine whether the test sample corresponds to a known or unknown class, *Pro-KT* performs open-set detection. If task identity is available, the model uses the task-specific threshold strategy defined in Equation 4.9; otherwise, it applies the general strategy described in Equation 4.10 (Lines 4–6).

If the sample is classified as belonging to a known class, *Pro-KT* outputs the final class prediction using the corresponding classifier head f_{θ_j} (Line 7). If the sample is detected as unknown, it is withheld from classification, thus enabling reliable decision-making under the open-world assumption.

4.3 Evaluation

In this section, we present the details of our experimental evaluation to assess the effectiveness of the proposed *Pro-KT* framework in comparison with competitive continual learning and open-set detection baselines. Our experiments are designed to answer the following key questions:

- **Q1:** *Does the proposed Pro-KT outperform state-of-the-art continual learning and open-set detection baselines in the open-world continual learning setting?*
- **Q2:** *How do the core components of Pro-KT contribute to its overall performance?*
- **Q3:** *How sensitive is Pro-KT to different hyperparameter configurations?*

Beyond addressing these primary questions, we also conduct additional experiments to analyze performance across different task sequences and visualize the behavior of the proposed prompt-based knowledge transfer mechanism.

Implementation Details. In our experiments, we simulate the open-world continual learning setting by removing the labels of the classes in the upcoming task (i.e., $T^{(n+1)}$) and treating them as unknowns during testing for the current task (i.e., n). Following standard practice in the continual learning literature, we randomly shuffle the task order five times for both the Split CIFAR-100 benchmark (comprising 10 tasks) and the 5-Datasets benchmark (comprising 5 tasks). All reported results represent the mean over these five randomized task orders.

We implemented both the proposed *Pro-KT* framework and all baseline methods using the PyTorch library, running on a single NVIDIA RTX 3090 GPU. For *Pro-KT*, we adopt the Vision Transformer (ViT) architecture as the backbone network. Input images are resized to 224×224 and normalized to the $[0, 1]$ range. We use the Adam optimizer with a dynamic learning rate schedule, and all models are trained using a batch size of 16 and an initial learning rate of 0.03.

For fair comparison, the hyperparameter configurations for each baseline method are adopted from the optimal settings reported in their respective original papers. On the Split CIFAR-100 benchmark, each task is trained for 20 epochs under the class-incremental learning setting. For the 5-Datasets benchmark, each task is trained for 10 epochs.

As part of our contribution, we also provide faithful PyTorch implementations of both the proposed method and a set of baseline models, all evaluated under the open-world continual learning setting.

Datasets. We evaluate our method on two widely-used benchmark datasets: **Split CIFAR-100** [98] and **5-Datasets** [99], both of which are standard in the continual learning and open-world recognition literature.

- **Split CIFAR-100.** This dataset divides the original CIFAR-100 dataset into 10 distinct tasks, with each task comprising 10 specific classes. The original CIFAR-100 dataset contains a total of 60,000 samples, of which 50,000 samples are designated for the training set, and the remaining 10,000 samples are reserved for the test set. Consequently, each single task is allocated 5,000 training samples and 1,000 test samples. In the context of open detection, the test set of the current task is merged with the test set of the subsequent task to form the open test set for the current task. As a result, except for the final task, each task consists of 2,000 open test samples. Therefore, our setting allows for the evaluation and assessment of the model’s performance on open classes during the training process.
- **5-Datasets.** We utilize a challenging dataset comprising TinyImagenet, NotMNIST, CUB200, Fashion-MNIST, and CIFAR10. Each dataset is treated as an individual task. Compared to Split CIFAR-100, all the datasets in the 5-Datasets exhibit wider variations, resulting in a greater significant forgetting issue. TinyImagenet consists of 200 classes with a total of 110,000 images, including 100,000 for training and 10,000 for testing. NotMNIST is an uncleaned dataset with 10 classes and comprises 18,256 training samples and 459 test samples. CUB200 is a widely used dataset for fine-grained visual classification tasks. It contains 11,788 images divided into 200 subcategories of birds, with 5,994 images used for training and 5,794 for testing. Fashion-MNIST includes 70,000 images from 10 categories: t-shirt, trousers, pullover,

4. LEARNING TO PROMPT KNOWLEDGE TRANSFER FOR OPEN-WORLD CONTINUAL LEARNING

dress, coat, sandal, shirt, sneaker, bag, and ankle boot. The dataset consists of 60,000 training samples and 10,000 test samples. CIFAR10 comprises RGB color images of 10 categories with 50,000 training images and 10,000 test images.

For open detection, we adopt the same setting as Split CIFAR-100, with a balanced number of known and unknown samples. Therefore, the number of open test sets for the top 4 tasks is as follows: 918, 918, 11,588, and 20,000.

Baselines. Since Open-world Continual Learning combines continual learning with novelty detection [25], we evaluate two key tasks: (1) *Unknown Detection* (identifying novel classes), and (2) *Known Classification* (learning sequentially without forgetting).

Unknown Detection. For detecting unknown classes, we compare several state-of-the-art methods. *OpenMax* [15] calibrates softmax probabilities to estimate openness, while *MSP* [80] uses the maximum softmax probability as an out-of-distribution (OOD) score. *MCD* [100] measures model disagreement on unlabeled data to quantify openness, and *EnergyBased* [63] uses energy scores to separate in-distribution (ID) and OOD data. *Entropy* [74] employs softmax entropy for uncertainty-based detection, whereas *MaxLogits* [101] improves OOD detection by relying on pre-softmax logits. Finally, *ODIN* [62] enhances MSP through temperature scaling and input perturbations to sharpen softmax boundaries.

Known Classification. For known-class classification, we benchmark against prominent continual learning approaches. Regularization-based methods include *EWC* [21], which penalizes changes to critical weights from past tasks, and *MAS* [96], which autonomously estimates parameter importance. Replay-based strategies like *DER++* [102] combine knowledge distillation with rehearsal, while *GEM* [44] enforces gradient constraints to prevent forgetting. Distillation-focused methods such as *LwF* [49] transfer knowledge across tasks via pseudo-rehearsal, and *HAT* [43] uses task-specific attention masks to isolate learned features. Bayesian approaches like *UCL* [46] incorporate variational inference for uncertainty-aware updates, and prompt-based techniques like *L2P* [32] and *DualPrompt* [83] use dynamic prompt tuning to adapt a frozen pre-trained backbone to sequential tasks.

Metrics. For unknown detection, we report the average area under the ROC curve (AUC_N), which gives us a single-number summary of how well the model separates known from unknown classes across all confidence thresholds, along with the average false positive rate (FPR_N) [74], where AUC_N measures open-class detection performance across all tasks and FPR_N specifically quantifies how often unknown samples are mistakenly accepted as known classes.

For known classification, we evaluate average final accuracy (A_N) across all learned classes and average forgetting rate (F_N) [32], with A_N reflecting overall retention

capability and F_N capturing the stability-plasticity trade-off by measuring performance degradation over sequential tasks.

4.3.1 Experimental Results and Analysis

Table 4.2: Evaluation results of unknown detection performance (%) across 10 sequential tasks (10 classes each) on Split CIFAR-100 and 5 tasks on 5-Datasets.

Dataset	Methods	AUC_N	FPR_N
Split CIFAR-100 (10 tasks)	OpenMax	49.99	50.06
	MSP	50.29	52.18
	MCD	50.17	70.37
	EnergyBased	49.66	71.86
	Entropy	49.56	51.17
	MaxLogits	52.63	74.99
	ODIN	50.81	52.64
	<i>Pro-KT</i> w/o task IDs	<u>91.01</u>	<u>41.31</u>
	<i>Pro-KT</i>	92.69	39.71
5-Datasets (5 tasks)	OpenMax	60.09	84.23
	MSP	51.22	87.96
	MCD	50.60	98.62
	EnergyBased	49.38	87.99
	Entropy	39.49	96.34
	MaxLogits	66.78	73.12
	ODIN	49.63	97.91
	<i>Pro-KT</i> w/o task IDs	<u>82.76</u>	<u>50.05</u>
	<i>Pro-KT</i>	88.60	45.70

Results on Unknown Detection (Q1). Table 4.2 presents the unknown detection performance across all evaluated methods. The results reveal two key findings: (1) existing baselines exhibit significantly degraded performance in the open-world continual learning setting, demonstrating their inability to maintain robust open-class detection capabilities across sequential tasks, and (2) our proposed *Pro-KT* achieves consistent and substantial improvements over all competitors, as evidenced by both the increased AUC_N and reduced FPR_N metrics across all experimental configurations.

Further analysis of threshold learning strategies highlights the importance of task-adaptive detection. When evaluating *Pro-KT* without task identifiers, we observe a notable performance decline in AUC_N , confirming that task-specific knowledge plays a crucial role in maintaining detection accuracy. This finding suggests that explicit task information helps refine decision boundaries for open-class detection in continual learning scenarios.

The superior performance of *Pro-KT* can be attributed to two main factors: (1) its effective knowledge transfer mechanism for open-class data across tasks, and

4. LEARNING TO PROMPT KNOWLEDGE TRANSFER FOR OPEN-WORLD CONTINUAL LEARNING

Table 4.3: Classification accuracy (%) on known classes for Split CIFAR-100 and 5-Datasets benchmarks. Experiments evaluate continual learning performance across 10 sequential tasks (10 classes each) on Split CIFAR-100, and 5 sequential tasks for 5 datasets. We used the ResNet32 and ViT backbones in the experiments.

Backbone	Method	Split CIFAR-100 (10 tasks)			5-Datasets (5 tasks)		
		$A_N(\uparrow)$	$Diff(\downarrow)$	$F_N(\downarrow)$	$A_N(\uparrow)$	$Diff(\downarrow)$	$F_N(\downarrow)$
ResNet32	<i>Upper-bound</i>	72.99	-	-	93.59	-	-
	EWC	34.16	38.83	36.20	51.25	42.34	42.06
	GEM	24.93	48.06	42.04	39.06	54.53	56.47
	LwF	16.73	56.26	51.62	36.18	57.41	59.19
	DER++	31.03	41.96	22.63	45.59	48.00	47.82
	HAT	30.43	42.56	36.73	37.00	56.59	51.35
	UCL	29.77	43.22	31.99	31.71	61.88	58.89
	MAS	29.58	43.41	22.14	49.59	44.00	33.90
ViT	<i>Upper-bound</i>	86.07	-	-	80.59	-	-
	LwF	71.93	14.14	<u>6.89</u>	19.02	61.57	44.23
	MAS	76.71	9.36	13.02	65.13	15.46	13.5
	L2P	80.63	5.44	10.93	69.06	11.59	<u>9.53</u>
	DualPrompt	83.18	<u>2.89</u>	8.42	<u>70.14</u>	<u>10.45</u>	10.11
	<i>Pro-KT</i>	84.07	2.00	5.43	71.70	8.89	5.19

(2) the innovative prompt bank design that progressively accumulates and organizes task-relevant knowledge. Notably, our method not only improves detection accuracy but also addresses the critical challenge of adaptive threshold selection through its dynamic strategy. These advantages collectively enable *Pro-KT* to set a new state-of-the-art for open-world continual learning.

Results on Known Samples Classification (Q1). We evaluate classification accuracy on known classes under continual learning settings. Notably, *Pro-KT* operates without requiring task identifiers during inference, making it suitable for task-agnostic class-incremental scenarios. For reference, we include an *Upper-bound* performance (offline joint training of all tasks) [32], regarded as the theoretical performance limit.

Table 4.3 presents comparative results on Split CIFAR-100 and 5-Datasets. Our analysis reveals three key findings: (1) *Pro-KT* consistently surpasses all baselines, establishing new state-of-the-art performance across both benchmarks; (2) When using ViT as backbone, *Pro-KT* achieves superior A_N compared to ResNet32, benefiting from both the strong representation capacity of large-scale pre-trained models and our method’s effective prompt-based knowledge transfer mechanism; and (3) While baseline methods show non-trivial performance gaps relative to the upper-bound, *Pro-KT* demonstrates the smallest discrepancy, along with the lowest forgetting rate (F_N) across sequential tasks.

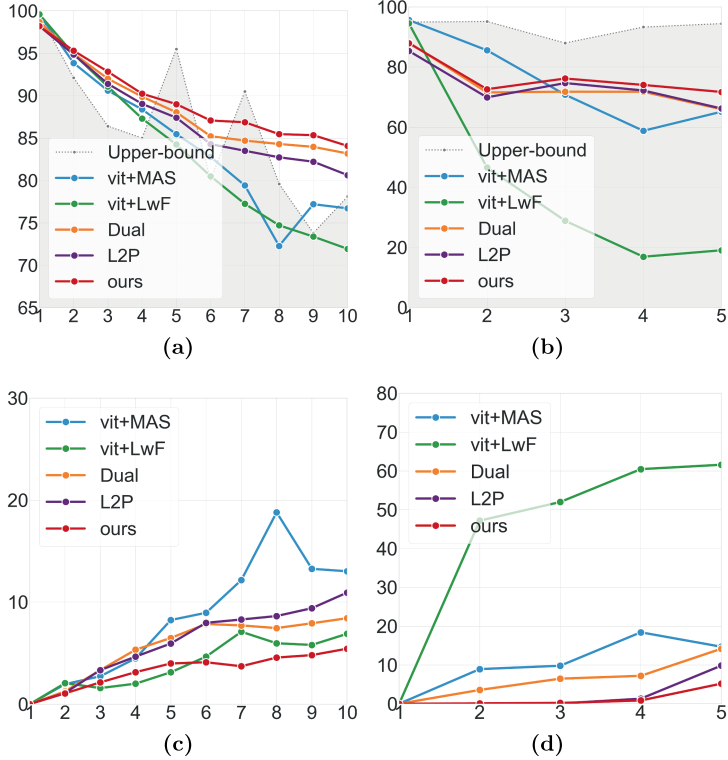


Figure 4.4: (a) and (b): Variation curve of A_N on Split CIFAR-100 and 5-Datasets, the x -axis indicates the task IDs. (c) and (d): Variation curve of F_N on Split CIFAR-100 and 5-Datasets, the x -axis indicates the task IDs.

4.3.2 Additional Results

Task Effect (Q1). Figure 4.4 presents the evolution of A_N and F_N metrics across sequential tasks. The results demonstrate *Pro-KT*'s consistent high performance across tasks and its superiority over baselines as the number of tasks increases. In fact, our method maintains consistently high accuracy (A_N) as new tasks are introduced, and it exhibits significantly lower forgetting (F_N) compared to baseline approaches. The performance gap widens progressively with increasing task complexity. These results highlight *Pro-KT*'s dual capability: effectively preventing catastrophic forgetting while sustaining high classification accuracy. The demonstrated performance superiority underscores the critical importance of our knowledge transfer mechanism in solving open-world continual learning challenges, particularly in scenarios with expanding task sequences.

Ablation Study (Q2). To demonstrate the importance of each key component in *Pro-KT*, we perform ablation studies focusing on three critical aspects: task identifiers for threshold determination, the complete task-aware open-set bound-

4. LEARNING TO PROMPT KNOWLEDGE TRANSFER FOR OPEN-WORLD CONTINUAL LEARNING

ary mechanism, and the prompt bank architecture. The experimental results are presented in Table 4.4.

First, we examine the impact of task identifiers by removing them during threshold determination. Compared to the full *Pro-KT* implementation, this results in a 1.8% drop in AUC_N on Split CIFAR-100 (from 92.69 to 91.01) and a more pronounced 6.6% drop on 5-Datasets (from 88.60 to 82.76). These results indicate that task identifiers contribute to more effective knowledge transfer by providing disambiguation across task boundaries, which becomes increasingly important when dealing with heterogeneous task distributions. Furthermore, we observe a slight increase in FPR_N under both settings, suggesting that removing task IDs can also exacerbate confusion in unknown detection.

Second, we investigate the open-set boundary by completely removing this component. The consequences are severe: the model loses its open detection capability entirely, as shown by the AUC_N score dropping to approximately 50 (equivalent to random guessing) and the FPR_N reaching its highest observed value. This demonstrates the fundamental importance of the task-aware boundary mechanism for maintaining open-world detection performance.

Finally, we assess the prompt bank’s contribution by replacing it with direct fine-tuning of the pre-trained backbone and task-specific classifiers. This ablation leads to a dramatic performance decline, revealing that naive fine-tuning approaches are fundamentally unsuited for open-world continual learning scenarios due to catastrophic forgetting. The significant performance gap underscores the prompt bank’s crucial role in enabling effective, continual learning while preserving previously acquired knowledge.

Visualization (Q2). To investigate the knowledge transfer mechanism and characterize the different knowledge types learned by our prompt bank, we visualize all prompts using both t-SNE [103] and UMAP [104] dimensionality reduction techniques on Split CIFAR-100.

Table 4.4: Ablation study on the variants of *Pro-KT*.

Unknown Detection	Split CIFAR-100		5-Datasets	
	AUC_N	FPR_N	AUC_N	FPR_N
w/o task IDs	91.01	41.31	82.76	50.05
w/o AODB	50.02	87.63	46.78	88.90
<i>Pro-KT</i>	92.69	39.71	88.60	45.70
Known Classification	Split CIFAR-100		5-Datasets	
	A_N	F_N	A_N	F_N
w/o prompt bank	40.49	17.21	15.42	40.19
<i>Pro-KT</i>	84.07	5.43	71.70	5.19

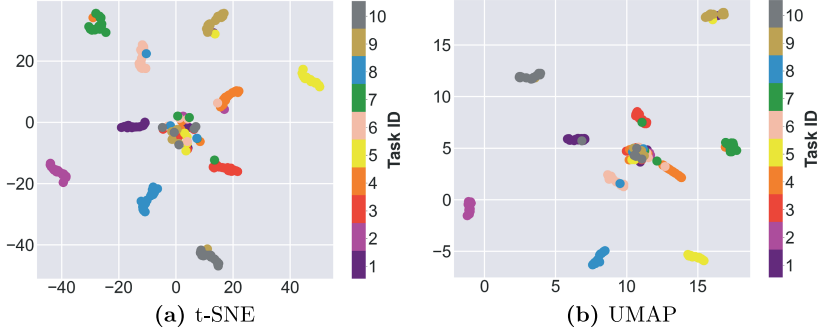


Figure 4.5: Visualization of the prompt bank via T-SNE and UMAP after training on all tasks (Split CIFAR-100 dataset).

We analyze the prompt bank from the final model after sequential training on all tasks $[1, \dots, n, \dots, N]$. Figure 4.5 reveals two distinct patterns in the prompt embedding space. First, we observe clear clustering of prompts by task (indicated by color separation), demonstrating successful acquisition of task-specific knowledge. Second, a subset of prompts forms a central cluster containing representations learned across multiple tasks, providing direct evidence of task-agnostic knowledge capture.

These visualization results confirm that our prompt bank simultaneously maintains both specialized task knowledge and transferable general knowledge. The spatial organization of prompts in the embedding space validates the model’s ability to preserve task-specific features while developing shared representations that facilitate effective knowledge transfer across the entire task sequence.

Parameter Sensitivity Analysis (Q3). Recall that there are three key parameters in our *Pro-KT*, namely the total number M of prompts learned from each task, the length L_p of a single prompt, and the selection size K used to prepend the input. Hence, $M \times N$ gives the total size of the prompt bank. Figure 4.6 shows the results on two datasets.

We now analyze the impact of different parameter settings that could affect the proposed *Pro-KT*. From the results in Figure 4.6 (left and middle), we observe that an oversized selection size K may lead to knowledge under-fitting. This phenomenon could be associated with prompt redundancy, where selecting too many prompts may introduce conflicting or irrelevant information, thereby diluting the useful task-relevant knowledge. While we find that the number of training steps L_P has little effect on performance, suggesting that prompt tuning itself is stable across moderate training durations.

Regarding the total number of prompts per task, we find that a reasonable capacity of the prompt bank is essential for encoding both task-common and task-specific

knowledge. From the results in Figure 4.6 (right), it is observed that in Split CIFAR-100, where the 10 tasks are similar, the performance remains stable across different values of M . However, in the 5-Datasets benchmark, where the tasks are highly dissimilar, increasing M introduces irrelevant prompts and exacerbates interference, resulting in degraded performance. We thus conclude that for tasks with greater dissimilarity, the hyperparameter M becomes more sensitive and requires careful tuning.

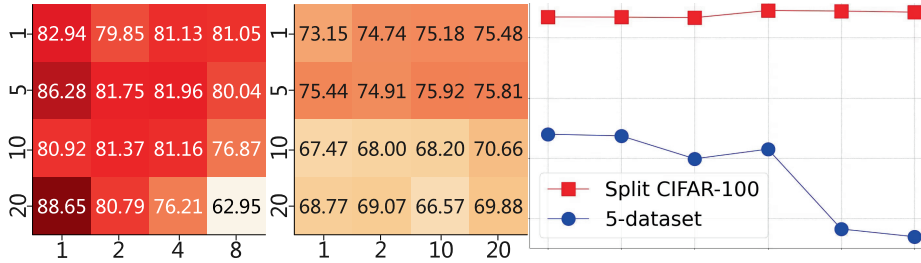


Figure 4.6: Left and middle are performance heatmaps (A_N) showing interaction effects between prompt length (L_P , y-axis) and selection size (K , x-axis) on both datasets, with fixed $M = 25$. Right is A_N variation across prompt counts (M) with optimal $L_P = 5$ and $K = 3$.

4.4 Summary

In this chapter, we introduced a prompt-based method named *Pro-KT*, designed to facilitate effective knowledge transfer through a structured prompt bank tailored for open-world continual learning. Our approach enables encoding of both task-specific and task-generic knowledge by associating learned prompts with a frozen pre-trained backbone. To support task-aware open-set decision-making, we further proposed two adaptive thresholding strategies that leverage prompt-derived representations. Extensive experiments across multiple benchmarks validate the effectiveness of *Pro-KT*, demonstrating robust performance in diverse task sequences and robustness under open-world conditions.

To better understand the limitations of our approach and guide future research directions, we conducted an in-depth analysis of representative failure cases. One notable limitation is the architectural dependence on prompt-compatible backbones. While *Pro-KT* performs strongly with transformer-based backbones (e.g., ViT), its AUC_N drops by 15.8% on Split CIFAR-100 and 21.9% on 5-Datasets when applied to ResNet-based backbones. This performance gap stems from the lack of pre-trained prompt compatibility in convolutional architectures, which impedes effective prompt activation and knowledge transfer. Addressing this challenge calls for more generalizable prompt integration strategies beyond transformers.

Another limitation arises from assumptions in the experimental setup. Specifically, we evaluated *Pro-KT* under class-balanced conditions, while real-world open-world continual learning tasks often exhibit severe imbalance, with unknown classes being rare or episodic. Our results (Section 4.4) show that under such conditions, the AUC_N of *Pro-KT* declines by up to 12.4%, and the false positive rate increases significantly, suggesting reduced model sensitivity to rare unknowns. Tackling this issue will require methods that can better exploit sparse unknown signals and incorporate fine-grained uncertainty modeling, especially in the low-resource regime.

In light of these findings, the next chapter explores a more challenging yet practically relevant setting—open-world continual learning under limited supervision. Specifically, we investigate how the core mechanisms of *Pro-KT* can be adapted to few-shot scenarios, where both known and unknown samples are scarce. This setting demands more expressive prompt representations, stronger generalization across tasks, and robust adaptation under data scarcity—all of which we address through principled model extensions and targeted empirical analysis.

Chapter 5

Improving Open-world Continual Learning under the Constraints of Scarce Labeled Data

In the previous chapter, we introduced *Pro-KT*, a prompt-based framework for open-world continual learning that employs a structured prompt bank to facilitate effective knowledge transfer and robust open-set decision-making. Although *Pro-KT* achieves strong performance in scenarios with abundant labeled data, its dependence on extensive supervision restricts its applicability in real-world settings where labeled samples are scarce—a common challenge in dynamic, open-ended environments. Similarly, prior studies [8, 25, 30] face a critical limitation: their reliance on large-scale training data for each new task.

However, this requirement for abundant labeled samples often conflicts with the inherent nature of open and dynamic data, where new tasks (or classes) typically emerge with only limited supervision—or even just a few labeled examples [105, 106]. For example, on social media platforms, user-uploaded content (e.g., photos or posts) frequently contains novel objects or scenes without corresponding labeled data. Acquiring sufficient ground-truth annotations for every new task or class is often impractical, necessitating open-world continual learning models that can adapt to extreme data scarcity.

This limitation motivates the focus of this chapter, where we extend our investigation to a more challenging yet practical setting: **Open-world Few-shot Continual Learning (OFCL)**. In this setting, the model must not only manage incremental tasks amid open-set unknowns but also operate under severe data scarcity, with both known and unknown classes represented by only a few labeled examples.

As illustrated in Figure 5.1, the scarcity of labeled data increases the two main problems in open-world continual learning, namely the misclassification of open-set samples as known classes and catastrophic forgetting due to inadequate knowledge repre-

5. IMPROVING OPEN-WORLD CONTINUAL LEARNING UNDER THE CONSTRAINTS OF SCARCE LABELED DATA

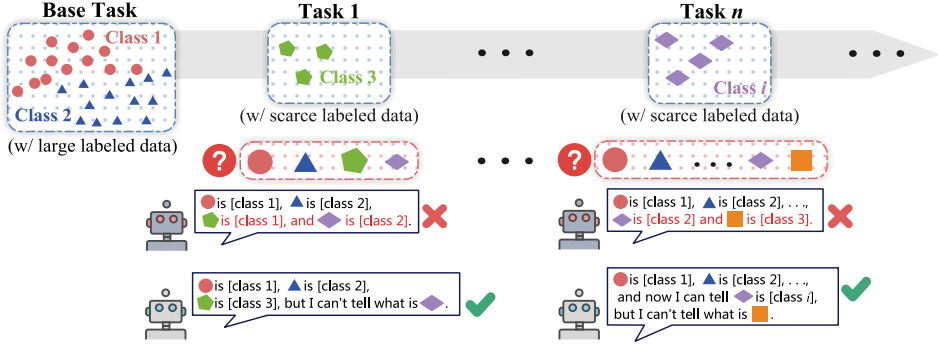


Figure 5.1: Open-world few-shot continual learning framework overview. The model incrementally learns new tasks (task 1, ..., task n , etc.) with few-shot class examples (class 3, ..., class i , etc.) while simultaneously detecting unknown samples during testing and overcoming severe forgetting and overfitting challenges inherent in data-scarce continual learning scenarios.

sensation. The limited training examples further impair the model’s ability to learn precise decision boundaries between known and unknown classes, degrading performance in open-world scenarios. Additionally, data scarcity restricts the model’s capacity to extract and generalize useful knowledge, hindering effective cross-task knowledge transfer. Consequently, the few-shot learning constraint introduces the following three challenges:

Challenge (1): Representation Collapse. With insufficient training samples, the model tends to learn shallow or overly narrow features that fail to generalize across tasks. This results in weak representations, impairing the distinction between seen and unseen classes and worsening both forgetting and misclassification.

Challenge (2): Open Detection Boundary Ambiguity. The lack of representative examples near decision boundaries makes it difficult for the model to accurately separate known and open-set classes. Consequently, open-set samples are more frequently misclassified into known categories, reducing reliability in open-world settings.

Challenge (3): Knowledge Transfer Degradation. In few-shot learning, the scarcity of task-specific data severely limits the amount of transferable knowledge. This diminishes the model’s ability to leverage prior experience for new tasks, making continual adaptation less efficient and more prone to errors.

These challenges underscore the necessity for an open-world continual learning framework capable of operating under limited supervision while preserving strong generalization, robust boundary modeling, and efficient cross-task knowledge transfer. To address these requirements, we propose a **P**rompt-**E**nriched **A**daptive **K**nowledge

transfer (*PEAK*) model, a novel framework specifically designed for open-world continual learning in data-scarce scenarios. Our approach simultaneously enhances three key aspects: (1) knowledge representation quality, (2) decision boundary reliability, and (3) cross-task transferability under extreme label scarcity.

First, we introduce a learnable token injection mechanism that enriches instance-level representations. By implicitly capturing auxiliary semantics through augmented tokens, our *Instance-wise Token Augmentation* (ITA) mitigates representational sparsity while promoting feature alignment across tasks through token-level embedding synchronization.

Secondly, by addressing the few-shot "hubness problem", where certain data points become overly frequent nearest neighbors in high-dimensional spaces, thus biasing predictions [36], we develop a hyperspherical solution inspired by [37]. We propose *Margin-based Open Boundary* (MOB) to establish well-separated decision margins through learnable class centroids with adaptive radii, effectively distinguishing known from unknown classes.

Finally, our dynamic hyperspherical space enables incremental absorption of novel class knowledge, converting "unknown" samples to "knowns". Through continuous refinement of the semantic space with new tasks, we introduce *Adaptive Knowledge Space* (AKS) to enhance both generalization capability and long-term continuity in open-world environments.

Extensive experiments across multiple benchmarks demonstrate that our *PEAK* consistently outperforms state-of-the-art approaches, establishing its effectiveness for few-shot open-world continual learning scenarios.

Our key contributions in this chapter are summarized as follows:

- **Problem Formulation:** We formalize the challenging yet practical setting of open-world continual learning with scarce labeled data, addressing the orthogonal challenges of open-set detection and knowledge transfer under extreme data limitations.
- **Technical Innovations:** We introduce Instance-wise Token Augmentation for enhancing embedding semantics and mitigating forgetting through learnable token injections. We propose a Margin-based Open Boundary to enable robust open-set detection via hyperspherical class representations, and use a novel Adaptive Knowledge Space to dynamically incorporate and update knowledge from both known and unknown classes.
- **Comprehensive Evaluation:** Our framework demonstrates superior performance against state-of-the-art open-world continual learning methods, few-shot incremental learners, and open-set detection baselines. Ablation studies further validate the robustness of each component. The source code is publicly available at <https://github.com/liyj1201/OFCL>.

This chapter is based on our publication [35]:

- **Li, Y., Wang, X., Yang, X., Bonsangue, M., Zhang, J., Li, T.:** Improving Open-World Continual Learning under the Constraints of Scarce Labeled Data. In: Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining. KDD '25 (2025).

5.1 OFCL Problem Definition

In this section, we introduce open-world few-shot continual learning, a realistic but challenging problem that aims to incrementally learn under the open-world assumption, where unseen or open samples may appear during the test phase, and to generalize effectively to new tasks with scarce labeled training data. It is important to note that we assume only the base task provides ample training samples (as shown in Figure 5.1), while each subsequent task ($t > 0$) involves only limited labeled data.

Few-shot learning inherently suffers from the scarcity of labeled data, often leading to under-trained models and sparse knowledge representations. To address this challenge, our approach integrates prompt-based learning with a novel Instance-wise Token Augmentation mechanism, enabling the model to adaptively enrich feature representations even under limited annotation. In addition, the Margin-based Open Boundary module constructs compact and discriminative decision boundaries by leveraging hyperspherical embeddings, thereby enhancing the model’s robustness in open-set detection. Complementing this, the Adaptive Knowledge Space introduces a dynamic structure for knowledge accumulation, transfer, and replay, allowing the model to incrementally refine its understanding of both known and unknown classes over time.

Together, these components enable our proposed method to overcome key limitations of previous open-world continual learning approaches, such as the reliance on fixed logits-based thresholds (e.g., *Pro-KT* [30]), and render it well-suited for real-world deployment under limited supervision.

Definition 8 *OFCL Problem Definition.* *Given a sequence of tasks $\{1, \dots, t, \dots\}$, each training set is organized in an N -way K -shot format: $D_{tr}^t = \{(x^t, C) \mid x^t \in X_{tr}^t(C), C \in CLS_{tr}^t\}$, where $X_{tr}^t(C)$ denotes the set of K training samples for class C , and CLS_{tr}^t represents the set of N classes in task t . Similarly, the test set for task t is defined as $D_{te}^t = \{(x^t, C) \mid x^t \in X_{te}^t(C), C \in CLS_{te}^t\}$, where CLS_{te}^t is the set of classes in task t .*

Under the open-world assumption, we do not assume any predefined relationships between classes in the test and training sets. Specifically, some test samples may be unknown or open, i.e., belonging to the set $(CLS_{te}^t \setminus CLS_{tr}^t)$. During evaluation, we assess performance on the cumulative test set $D_{te}^1 \cup \dots \cup D_{te}^t$ in a continual manner.

Accordingly, the objective of open-world few-shot continual learning is to learn a unified classifier f_θ , parameterized by θ , that can (1) detect unknown/open samples, (2) classify known samples accurately, and (3) incrementally transform unknowns into knowns over time.

To support a clear understanding of the above and subsequent formulations and model components, we summarize the key notations used throughout this chapter in Table 5.1.

Table 5.1: Notations and explanations.

Notation	Explanation
t	The session t
CLS_{tr}^t	The set of training classes from t
D_{tr}^t, D_{te}^t	Training and test sets of t
N	# of classes in each training set
K	# of training samples of each class
C	Class C
$\mathbf{x}^+, \mathbf{x}^-$	Positive and negative sets of class C
α, β	Scaling factors for $\mathbf{x}^+$ and $\mathbf{x}^-$
$q_\sigma(\cdot)$	A quantile function
$\mathbf{p}^t$	The token set of t
$\mathbf{p}^*$	The set of matched tokens
L_P	The length of each token
D_e	The embedding dimension
$\mathbf{h}^t = \mathbf{h}^t \oplus \mathbf{p}^*$	The augmented embedding
f_{pr}	The pre-trained backbone
f_θ	The trainable classifier
r_C	The learnable radius of class C
$\mathbf{c}_C$	The learnable centroid of class C
m	The margin
$(CLS_{tr}^t, \mathbf{C}^t, \mathbf{R}^t)$	Hyperspheres learned from D_{tr}^t
$(CLS_{open}^t, \mathbf{C}^t, \mathbf{R}^t)$	Hyperspheres learned from unknowns
λ	The trade-off parameter

5.2 Methodology

The challenges outlined above, namely, representation collapse, boundary ambiguity, and knowledge transfer degradation, make clear the need for an approach to Open-world Few-shot Continual Learning that can simultaneously address data scarcity, openness, and catastrophic forgetting. As we have discussed, existing open-world continual learning methods assume enough labeled data, and few-shot learning techniques often disregard continual adaptation. In this section, we close this gap by introducing a unified framework specifically designed for open-world few-shot continual learning. Our solution is schematically presented in Figure 5.2 and consists

5. IMPROVING OPEN-WORLD CONTINUAL LEARNING UNDER THE CONSTRAINTS OF SCARCE LABELED DATA

of three components discussed in more details below: (1) Instance-wise Token Augmentation to deal with data sparsity through dynamic prompt-based enrichment, (2) a Margin-based Open Boundary mechanism to refine decision frontiers in few-shot conditions, and (3) an Adaptive Knowledge Space to organize and transfer knowledge across tasks.

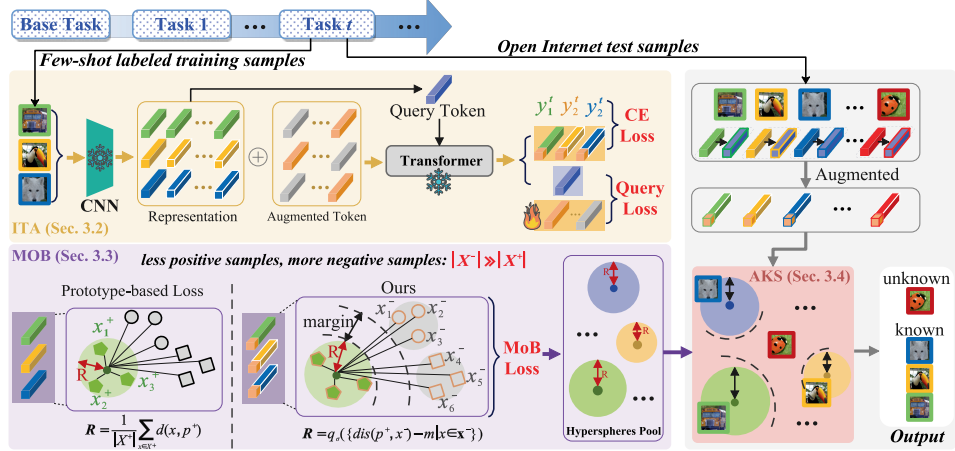


Figure 5.2: The overall *PEAK* model consists of three key components: (1) ITA (Yellow): enhancing the samples by matching them with appropriate additional tokens; (2) MOB (Purple): constructing compact decision boundaries of knowns by margin-based loss function for open detection; (3) AKS (Red): incorporating knowledge learned from both knowns and unknowns, and facilitating the transition from unknowns to knowns.

5.2.1 Instance-wise Token Augmentation

Inspired by the strong few-shot learning ability of prompt tuning methods [32, 83, 107], we propose Instance-wise Token Augmentation, a knowledge augmentation paradigm for *PEAK* designed to address the challenges of sparse supervision in open-world few-shot continual learning. This approach builds directly on the foundation presented in Chapter 4, where we demonstrated how prompt-based methods such as *Pro-KT* can enrich sample representations by injecting additional knowledge. This enriched representation plays a critical role in few-shot settings. Here, the scarcity of labeled data amplifies the risk of knowledge sparsity, undermining both representation quality and transfer efficiency. To alleviate this problem, we extend the prompt-based framework of *Pro-KT* by introducing Instance-wise Token Augmentation, a dynamic mechanism for instance-level knowledge augmentation, enabling more robust representation learning, knowledge accumulation, and cross-task transfer under limited supervision.

Following the architectural intuition of *Pro-KT*, the instance-wise token augmentation module operates through three key steps: token initialization, instance-wise token matching, and embedding augmentation.

Token Initialization. For a task t with training set D_{tr}^t , we initialize a set of learnable tokens $\mathbf{p}^t$ as:

$$\mathbf{p}^t = (p_1^t, \dots, p_l^t, \dots, p_l^t), \text{ with } p_i^t \in \mathbb{R}^{L_p \times D_e}, \quad (5.1)$$

where l denotes the number of tokens, L_p their length, and D_e the embedding dimension. During training, we maintain a frequency set $\nu^t = (\nu_1, \dots, \nu_k, \dots)$ to track how often each token p_k^t is selected for augmentation.

Instance-wise Token Matching. To identify the most relevant tokens for a given sample, we employ a top- K selection mechanism based on cosine similarity between the sample’s embedding and token-associated keys:

$$(\mathbf{k}^*, \mathbf{p}^*) = \underset{K}{\operatorname{argmin}} \sum_{i=1}^l \operatorname{sim}(\mathbf{h}^t, k_i^t) \cdot \nu_i, \text{ with } \mathbf{h}^t = \mathbf{Q}(x^t), \quad (5.2)$$

where k_i^t is the key for token p_i^t , $\operatorname{sim}(\cdot, \cdot)$ measures cosine similarity, and $\mathbf{Q}$ is a query function projecting the input x^t into the keys space.

Embedding Augmentation. The selected tokens $\mathbf{p}^*$ are concatenated with the original embedding $\mathbf{h}^t$ to produce an augmented representation:

$$\mathbf{h}^t = \mathbf{h}^t \oplus \mathbf{p}^*, \quad (5.3)$$

where $\oplus$ denotes dimension-wise concatenation.

The training objective combines classification loss with a token alignment term to ensure selected keys remain semantically relevant:

$$\mathcal{L}_{aug} = \mathcal{L}(f_\theta(f_{pr}(\mathbf{h}^t), \mathbf{y}^t)) + \lambda \cdot \sum_{i=1}^l \operatorname{sim}(\mathbf{h}^t, k_i^t), \quad (5.4)$$

where f_{pr} is the pre-trained backbone, $\mathbf{y}^t$ are task labels, and λ is a parameter to balance classification accuracy and token-feature alignment. Here, the first addend denotes the classification loss, and the second is a surrogate loss for pulling selected keys closer to corresponding query features.

Unlike prior prompt-based continual learning methods, the instance-wise token augmentation module stores all task-specific tokens in a unified token bank $(\mathbf{K}, \mathbf{P})$, treating them as reusable knowledge. During inference, the model retrieves and applies the most relevant tokens from this bank for each test sample to mitigate forgetting by preserving prior task knowledge, reduce overfitting through dynamic, instance-specific augmentation, and enhance generalization by leveraging cross-task token sharing as demonstrated empirically in Section 5.3.

5.2.2 Margin-based Open Boundary

Our approach introduces augmented training samples to facilitate the formation of precise, compact decision boundaries for open sample detection, a capability in *PEAK* where traditional methods often fail. Unlike existing approaches that simply combine continual learning with open-set recognition, we propose a Margin-based Open Boundary strategy implemented in hyperspherical space to construct compact boundaries from known samples or training classes. Our solution addresses the two challenges in open-world few-shot continual learning: the hubness problem [37], where certain points disproportionately dominate nearest-neighbor lists, leading to misclassifications, and the need for robust adaptive representations in dynamic, low-sample environments.

For each class $C \in CLS_{tr}^t$, we organize the training samples into two distinct sets following contrastive learning principles where training samples belonging to C serve as positive samples, while all other training samples are treated as negative instances (illustrated in Figure 4.3): the positive set $\mathbf{x}^+ = \{x^t | (x^t, C) \in D_{tr}^t\}$ contains all samples belonging to class C at task t , and the negative set $\mathbf{x}^- = \{x^t | (x^t, C') \in D_{tr}^t, C' \neq C\}$ comprises all other samples from the same task. This division enables our margin-based open boundary framework to learn discriminative hyperspherical decision boundaries through the following loss function:

$$\begin{aligned} \mathcal{L}_{margin} = & \frac{1}{N} \cdot \sum_{C \in CLS_{tr}} (\lambda \cdot r_C^2 \\ & + \frac{1}{\alpha} \cdot \log[1 + \sum_{x \in \mathbf{x}^+} e^{\alpha \cdot (dis(\mathbf{c}_C, f_{pr}(x)) - r_C)}] \\ & + \frac{1}{\beta} \cdot \log[1 + \sum_{x \in \mathbf{x}^-} e^{-\beta \cdot (dis(\mathbf{c}_C, f_{pr}(x)) - (r_C + m))}]), \end{aligned} \quad (5.5)$$

where N is the number of classes in the training set, m is the margin enforcing separation between classes, r_C is the hypersphere radius for class C around centroid $\mathbf{c}_C$, $dis(\cdot, \cdot)$ is the distance metric between samples, λ is a trade-off parameter for radius regularization, and α, β are the scaling factors for positive and negative set contributions, respectively. Note that the pre-trained backbone f_{pr} processes raw samples directly here, unlike in Equation 5.4 where it operates on embeddings. This design choice preserves the original feature relationships for boundary learning.

Our margin-based open boundary framework operates through three key components: (1) *known classes centroids* for establishing hypersphere centers for each known class, (2) *margin and radius learning* for optimizing class boundaries through contrastive principles, and (3) *open boundary detection* for implementing robust unknown identification.

Known Classes Centroids. Hyperspherical embeddings offer clear advantages for open-set scenarios by naturally distributing data more uniformly and avoiding

central clustering. Following approaches like noHub [37], our method maintains both class structure preservation and uniform feature distribution, a balance that allows continuous adaptation in dynamic, low-sample environments.

We construct this embedding space by representing each class as a hypersphere with a learned centroid. For any class C from task t , its centroid $\mathbf{c}_C$ is computed as:

$$\mathbf{c}_C = \frac{1}{|\mathbf{x}^+|} \sum x^t \in \mathbf{x}^+ f_{pr}(x^t). \quad (5.6)$$

The model incrementally stores these centroids throughout the continual learning process $1, \dots, t, \dots$. Each centroid serves as the center for its corresponding class hypersphere, with an optimal radius to be determined in the next step to establish well-separated decision boundaries between known and unknown samples.

Margin and Radius Learning. In our few-shot learning scenario, a given class C typically has significantly fewer positive samples $\mathbf{x}^+$ than negative samples $\mathbf{x}^-$. To maintain a clear separation between classes while accounting for this imbalance, we enforce two key geometric constraints. First, we require that all positive samples must lie within the class hypersphere:

$$dis(\mathbf{c}_C, f_{pr}(x)) < r_C, \quad (5.7)$$

where r_C is learned through optimization using Equation 5.5. The second constraint imposes that all negative samples reside outside an expanded boundary:

$$dis(\mathbf{c}_C, f_{pr}(x)) > r_C + m, \quad (5.8)$$

with margin m ensuring sufficient separation.

The optimal radius r_C is determined through:

$$r_C = q_\sigma(\{dis(\mathbf{c}_C, f_{pr}(x)) - m \mid x \in \mathbf{x}^-\}), \quad (5.9)$$

where $q_\sigma(\cdot)$ is a quantile function controlling boundary tightness via parameter σ governing deviations. For each task t , we maintain a complete hypersphere representation:

$$(CLS_{tr}^t, \mathbf{C}^t, \mathbf{R}^t) = \{(C_1^t, \mathbf{c}_{C_1^t}, r_{C_1^t}), \dots, (C_N^t, \mathbf{c}_{C_N^t}, r_{C_N^t})\}, \quad (5.10)$$

where N is the number of classes in CLS_{tr}^t . Throughout continual learning $\{1, \dots, t, \dots\}$, the model aggregates all learned hyperspheres together:

$$(CLS_{tr}, \mathbf{C}, \mathbf{R}) = \{(CLS_{tr}^1, \mathbf{C}^1, \mathbf{R}^1), \dots, (CLS_{tr}^t, \mathbf{C}^t, \mathbf{R}^t), \dots\}. \quad (5.11)$$

Open Boundary Detection. During testing, our open-set detection procedure evaluates each test sample x against all learned hyperspheres through hypersphere projection and membership verification. The first projects x into the hyperspherical

embedding space containing all learned class representations, while the latter checks if x resides within any class hypersphere by evaluating if there is a class C such that x falls within its hypersphere and thus classifies x as known in class C , and otherwise identifies x as unknown.

More precisely, the open detection process at task t consists of three steps:

- Step 1:** Identify the nearest centroid $\mathbf{c}_C^*$ to x using $\mathbf{c}_C^* = \underset{\mathbf{c}_C}{\operatorname{argmin}} \operatorname{dis}(\mathbf{c}_C, f_{pr}(x))$
- Step 2:** Compute the distance metric $\operatorname{dis}(\mathbf{c}_C^*, f_{pr}(x))$ between the centroid $\mathbf{c}_C^*$ and the test sample x ;
- Step 3:** Decide if x can be directly classified into a known class C if $d^* \leq r^*$, based on the output logits score, and otherwise, classify x as **[unknown]**.

The margin-based open boundary framework treats hypersphere centers as dynamically learnable parameters that evolve through continual learning, updating their positions based on deep feature representations. This approach extends conventional metric learning by combining margin-based optimization with hyperspherical geometry to address hubness while maintaining discriminative power. The centers serve as adaptive prototypes that progressively refine within our Adaptive Knowledge Space, enabling stable knowledge accumulation while accommodating new classes through geometric constraints. This dynamic formulation naturally handles distribution shifts and maintains robust decision boundaries throughout the learning process.

5.2.3 Adaptive Knowledge Space

Considering $(CLS_{tr}, \mathbf{C}, \mathbf{R})$ as the knowledge learned from knowns, we present an adaptive knowledge space, designed not only to store and transfer knowledge about unknowns acquired from previously learned tasks, but also to facilitate updating unknown knowledge into known knowledge for *PEAK*.

Unknown Knowledge Representation. After identifying unknowns, we then employ clustering on the identified unknown samples as follows: Given a detected unknown sample x let $\Gamma_\epsilon(x)$ represent the ϵ -neighborhood of x , *MinPts* denotes the minimum number of objects in $\Gamma_\epsilon(x)$, and $\rho(x) = |\Gamma_\epsilon(x)|$ be its density value. We define the boolean function $T(x) = 1$ if $\rho(x_j) \geq \text{MinPts}$ and $T(x) = 0$ otherwise. If $T(x) = 1$, the unknown sample x is clustered with density-reachable points, and the entire space is partitioned into M groups $\{G_1, G_2, \dots, G_M\}$. Otherwise, x is treated as noise. Each group is then assigned a group centroid, which is determined using Equation 5.6. The corresponding radius is obtained by Equation 5.9.

Since there are no *ground-truth labels* for unknowns, we generate a set of pseudo-labels and assign them to each group. Similar to hyperspheres learned from known

samples, we have :

$$(CLS_{open}^t, \overline{\mathbf{C}}^t, \overline{\mathbf{R}}^t) = \{(\overline{C}_1^t, \overline{\mathbf{c}}_{C_1^t}, \overline{r}_{C_1^t}), \dots, (\overline{C}_M^t, \overline{\mathbf{c}}_{C_M^t}, \overline{r}_{C_M^t})\}. \quad (5.12)$$

Updating Unknowns to Knowns. At task t the integrated adaptive knowledge space consists of all the learned hyperspheres during training $(CLS_{tr}, \mathbf{C}, \mathbf{R})$ together with all the new $(CLS_{open}^t, \overline{\mathbf{C}}^t, \overline{\mathbf{R}}^t)$ associated with the unknowns at all task until t . Because of this incremental growth, if previously encountered unknowns reappear, they are classified with corresponding pseudo-labels. However, if a new hypersphere learned during a subsequent task’s training overlaps with an existing hypersphere with a pseudo-label, then we integrate the corresponding pseudo-category into the newly trained categories, thus facilitating a transition from unknowns to knowns.

5.2.4 Overall Objective and Process

At any task t , *PEAK* first obtains L additional tokens via $\mathcal{L}_{aug}$ and stores all the newly learned additional tokens together with those previously learned. Subsequently, each sample is augmented by additional knowledge, and an adaptive knowledge space containing knowledge learned from knowns and unknowns is constructed. Hence, the overall objective is:

$$\min_{(\mathbf{C}^t, \mathbf{R}^t), \theta, (\mathbf{k}^t, \mathbf{p}^t)} \gamma \cdot \mathcal{L}_{margin} + (1 - \gamma) \cdot \mathcal{L}_{aug}, \quad (5.13)$$

where the γ is a balance between $\mathcal{L}_{margin}$ and $\mathcal{L}_{aug}$.

To better illustrate the proposed *PEAK* framework, we detail the training process in algorithm 3. First, the token bank and classifier are initialized. Then, for each time step t , multiple training epochs are conducted, where each sample $(\mathbf{x}^t, \mathbf{y}^t)$ undergoes feature mapping, similarity computation with stored keys, augmentation via token selection, and enhancement through the backbone network. Class-wise centroids, positive and negative sets, margins, and radii are computed to derive the overall loss, which is minimized via backpropagation. Finally, the AKS is updated with hyperspheres representing both known and unknown knowledge, enabling continual adaptation to evolving data distributions.

Accordingly, after learning task t , the inference procedure unfolds through four sequential stages:

Step 1: The model integrates newly acquired task knowledge—specifically, the information derived from labeled samples—into its existing knowledge space. In this process, previously unknown instances are transitioned into the known category space once their true labels have been observed.

Algorithm 3: : Training Process of the *PEAK* framework

Input: the training set $D_{tr}^t = \{(\mathbf{X}^t, \mathbf{Y}^t)\}$, the set of training classes CLS_{tr}^t , the token bank $(\mathbf{K}, \mathbf{P}) = \{(\mathbf{k}^t, \mathbf{p}^t)\}_{t=1}^T$, a trainable classifier f_θ , a backbone f_{pr} and the hyperparameter balance γ .

Initialize: $(\mathbf{K}, \mathbf{P})$, f_θ .

- 1: **for** $t = 1, 2, \dots, T$ **do**
- 2: **for** each epoch **do**
- 3: **for** $(\mathbf{x}^t, \mathbf{y}^t)$ in each training set D_{tr}^t with class CLS_{tr}^t **do**
- 4: Map $\mathbf{x}^n$ to $\mathbf{h}^t$;
- 5: Get the similarity of $\mathbf{h}^t$ and each key of $(\mathbf{k}^t, \mathbf{p}^t)$;
- 6: Select augmentation tokens with $(\mathbf{k}^*, \mathbf{p}^*)$ Equation 5.2;
- 7: Enhance $\mathbf{h}^t$ to $\mathbf{h}^{t*}$ with $\mathbf{p}^*$ via Equation 5.3;
- 8: Calculate the augmented feature by $f_{pr}(\mathbf{h}^{t*})$;
- 9: Calculate $\mathbf{c}_C$ for each class C in CLS_{tr}^t via Equation 5.6;
- 10: Obtain positive and negative set of class $C : \mathbf{x}^+, \mathbf{x}^-$;
- 11: Obtain margin and radius via Equation 5.9
- 12: Calculate the overall loss via Equation 5.13;
- 13: **end for**
- 14: Update f_θ , $(\mathbf{K}, \mathbf{P})$, the radius r_C and margin m by backpropagation;
- 15: **end for**
- 16: Update the overall AKS with 1) Hyperspheres learned from $D_{tr}^t : (CLS_{tr}^t, \mathbf{C}^t, \mathbf{R}^t)$ and 2) the Hyperspheres learned from unknowns: $(CLS_{open}^t, \overline{\mathbf{C}}^t, \overline{\mathbf{R}}^t)$.
- 17: **end for**

Step 2: During the inference phase, the model performs open-set recognition by actively identifying and isolating open (i.e., previously unseen) samples, thereby mitigating the risk of erroneous assignment to existing known classes.

Step 3: For samples confidently recognized as belonging to known classes, the model conducts accurate classification while simultaneously preserving knowledge acquired from previous tasks, thus alleviating catastrophic forgetting.

Step 4: The knowledge space is continuously and adaptively updated as the model incrementally incorporates novel information extracted from open samples, thereby refining its decision boundaries and expanding its representational capacity.

In summary, the proposed *PEAK* framework not only enhances the discriminative representations of known classes but also enforces tighter and more robust decision boundaries to improve open-set detection performance. This integrated approach reflects a principled extension of existing continual learning paradigms, effectively addressing the dual challenges of knowledge retention and open-world generalization intrinsic to open-world continual learning scenarios.

5.3 Evaluation

Having established the foundations of our *PEAK* framework, we now turn to empirical validation to assess its effectiveness in realistic open-world continual learning scenarios. Through experiments on standard benchmarks, we evaluate the framework’s ability to detect unknown samples while maintaining stability on known classification under few-shot conditions, as well as the practical efficiency of our adaptive knowledge mechanisms. These evaluations not only quantify *PEAK*’s advantages over existing approaches but also validate our design choices through ablation studies and sensitivity analyses.

5.3.1 Experimental Setup

Datasets. Our evaluation follows the few-shot continual learning benchmark configuration from TOPIC [108], employing two standard datasets: (1) the *CUB200 dataset* [109] (11,788 images, 200 species) partitioned into 100 base classes for task 0 followed by 10 incremental tasks (10 classes each, 5 samples per class), and (2) the *MiniImageNet dataset* [110] (60,000 images, 100 classes) divided into 60 base classes plus 8 incremental tasks (5 classes each, 5 samples per class).

Implementation Protocol. Consistent with few-shot continual learning standards, we randomly shuffle session orders (10 for CUB200, 8 for MiniImageNet) and treat subsequent session training classes as unknowns during the current session testing. All reported results average three random shuffles. Experiments are conducted on NVIDIA GeForce RTX 3090 GPUs using the PyTorch library. Input images are resized to 224×224 , normalized to $[0,1]$ range, with models trained via Adam optimizer (batch size=25, learning rate=0.03). Baseline hyperparameters follow their original optimal configurations. Our implementation includes fully reproducible PyTorch code for both *PEAK* and comparative baselines in open-world continual learning scenarios.

Baselines. To empirically validate *PEAK*’s effectiveness, we conduct comparisons against 23 state-of-the-art baselines across two standard datasets, establishing new performance benchmarks. Our evaluation includes both *Unknowns Detection* (8 representative for OOD and OSR methods) and *Knowns Classification* (15 leading few-shot continual learning approaches), providing complete coverage of the problem dimensions.

i) Unknowns Detection Baselines:

- *MSP* [80] uses the maximum softmax probability of classifier predictions as the OOD score.
- *KL* [111] proposes a scalable framework for OOD detection, resulting in large-scale datasets and improved evaluation protocols to enhance robustness and generalization.

5. IMPROVING OPEN-WORLD CONTINUAL LEARNING UNDER THE CONSTRAINTS OF SCARCE LABELED DATA

- *SSD* [112] is a unified self-supervised method that integrates diverse pretext tasks to improve representations.
- *MaxLogits* [101] enhances OOD detection performance by utilizing the maximum logit score.
- *Energy* [63] assigns lower energies to ID data and higher energies to OOD data to create an energy gap for OOD detection.
- *ViM* [113] is an OOD detection method with matching virtual logits, improving the separation between in-distribution and out-of-distribution samples through a robust scoring mechanism.
- *KNN-based OOD* [114] presents a deep nearest neighbor-based method for OOD detection, utilizing the local structure of feature spaces to identify anomalies.
- *NNGuide* [115] is a nearest neighbor guidance strategy for OOD detection, enhancing the discriminative power of feature representations by leveraging proximity-based metrics.

ii) *Knowns Classification Baselines:*

- *iCaRL* [42] is an incremental classifier and representation learning method designed for FSCIL.
- *TOPIC* [108] is a model tailored for few-shot class-incremental learning (FS-CIL), emphasizing effective adaptation to new classes with limited samples.
- *CEC* [116] accommodates new classes while preserving previous knowledge for incremental learning.
- *Meta-FSCIL* [117] is a meta-learning-based approach, leveraging meta-learning principles to facilitate fast adaptation to new sessions.
- *ALICE* [118] employs an attention-based approach to efficiently learn new classes with limited training samples.
- *FeSSSS* [105] addresses the challenge of FSCIL by utilizing feature space segmentation for improved adaptation.
- *Pro-KT* [30] Our *Pro-KT* framework (introduced in the previous chapter) that integrates task-generic and task-specific prompt libraries to encode and transfer knowledge, while leveraging task-aware open-set boundaries to effectively identify unknowns.
- *LIMIT* [119] focuses on efficient knowledge transfer and adaptation to new classes while mitigating the impact on the existing knowledge.
- *NC-FSCIL* [120] is a neural clustering-based model, utilizing clustering techniques to enhance adaptation.

- *MCNet* [121] implements memory-augmented neural networks for effective knowledge retention.
- *SoftNet* [106] utilizes the soft parameter sharing to facilitate efficient learning of new classes.
- *CoSR* [122] is an FSCIL model that employs a collaborative self-regularization mechanism for effective adaptation while preserving previous knowledge.
- *M-FSCIL* [123] proposes a memory-based label-text tuning method that leverages textual label embeddings and a memory bank to enhance generalization in few-shot class-incremental learning.
- *FSIL-GAN* [124] introduces a semantics-driven generative replay strategy, where a text-to-image model is used to synthesize class-representative samples for mitigating forgetting in few-shot incremental learning.
- *CPE-CLIP* [125] presents a multimodal parameter-efficient approach that integrates visual and textual information while freezing most model weights, enabling effective few-shot class-incremental learning with minimal overhead.

Metrics. We employ three complementary metrics: (1) For unknown detection we use the average area under the ROC (AUC_N) and false positive rate (FPR_N) across all N tasks [74]; (2) for known classification, the averaged accuracy ACC_N over N new tasks; and (3) for catastrophic forgetting, the performance dropping rate $PD = ACC_0 - ACC_N$, where ACC_0 denotes base task accuracy [30].

5.3.2 Main Results

Unknown Detection Performance. As shown in Table 5.2, conventional baselines struggle significantly in the open-world few-shot continual learning setting, revealing their limitations for real-world deployment. Our proposed *PEAK* framework demonstrates consistent superiority across both datasets, achieving state-of-the-art performance with 69.72% AUC_N and 69.96% FPR_N on CUB200 (10-way 5-shot), along with 76.20% AUC_N and 71.30% FPR_N on MiniImageNet (5-way 5-shot). These results represent substantial improvements over all baseline categories: confidence-based (MSP, MaxLogits), energy-based (Energy, ViM), distance-based (KNN-OOD), and guidance-based (NNGuide) approaches.

PEAK’s advantages stem from the integrations of three novel components: (1) Dynamic modeling of evolving knowledge spaces maintains discriminative boundaries for both closed-set classification and open-set detection; (2) Adaptive knowledge space updates enable incremental incorporation of novel open samples while preserving stability; (3) Joint optimization strategy provides principled uncertainty handling that reduces both false positives and catastrophic forgetting. The framework’s particularly strong FPR_N performance (critical for safety-sensitive applications) demonstrates the ability to avoid misclassifying unknown samples, while its

5. IMPROVING OPEN-WORLD CONTINUAL LEARNING UNDER THE CONSTRAINTS OF SCARCE LABELED DATA

Table 5.2: Results (%) regarding unknown detection. We report the results over 10 tasks for CUB200 (10-way 5-shot) and 8 tasks for MiniImageNet (5-way 5-shot).

Dataset	Methods	$AUC_N \uparrow$	$FPR_N \downarrow$
CUB200 (10-way 5-shot)	MSP	61.34	92.95
	KL	60.81	93.10
	SSD	37.70	99.52
	MaxLogits	<u>61.31</u>	<u>92.43</u>
	Energy	60.81	93.10
	ViM	54.16	97.91
	KNN-based OOD	37.20	99.16
	NNGuide	50.26	96.97
	<i>PEAK</i>	69.72	69.96
MiniImageNet (5-way 5-shot)	MSP	49.78	94.91
	KL	46.61	95.10
	SSD	50.64	95.18
	MaxLogits	46.78	94.65
	Energy	46.61	95.10
	ViM	<u>59.56</u>	<u>89.87</u>
	KNN-based OOD	52.88	93.78
	NNGuide	52.47	90.95
	<i>PEAK</i>	76.20	71.30

high AUC_N reflects a robust balance between known-class recognition and unknown rejection during continual expansion.

Overall, these results validate *PEAK* as a good solution to the fundamental challenge of maintaining robust open-set detection in continually evolving learning systems, addressing an important bottleneck in real-world deployment scenarios.

Results on Known Classification. Next, we evaluate our model’s ability to acquire novel classes while preserving prior knowledge under the challenging few-shot class-incremental learning setting. Tables Table 5.3 and Table 5.4 present the averaged classification accuracies ACC_N by different methods across incremental tasks on CUB200 and MiniImageNet datasets, respectively.

PEAK demonstrates resilience against limited data, class imbalance, and catastrophic forgetting (the three main challenges of few-shot class-incremental learning), showing state-of-the-art performance on both benchmarks. The *PEAK* framework’s final accuracy and low performance degradation (PD) validate the effectiveness of our instance-wise token augmentation and adaptive knowledge space mechanism, which jointly contribute to alleviating catastrophic forgetting while preventing overfitting under the few-shot context.

On the *MiniImageNet* benchmark, *PEAK* achieves a strong final accuracy of 87.12% with a performance degradation (PD) of 10.88, outperforming both *Pro-KT* (PD = 28.86) and M-FSCIL (which, however, exhibits a better PD of 8.08). Notably,

Table 5.3: Averaged classification accuracy (%) on CUB200 (10-way 5-shot).

Methods	ACC _N in each task (%) $\uparrow$											PD $\downarrow$
	0	1	2	3	4	5	6	7	8	9	10	
iCaRL	68.68	52.65	48.61	44.16	36.62	29.52	27.83	26.26	24.01	23.89	21.16	47.52
TOPIC	68.68	62.49	54.81	49.99	45.25	41.40	38.35	35.36	32.22	28.31	26.28	42.40
CEC	75.85	71.94	68.50	63.50	62.43	58.27	57.73	55.81	54.83	53.52	52.28	23.57
Meta-FC	75.90	72.41	68.78	64.78	62.96	59.99	58.30	56.85	54.78	53.82	52.64	23.26
ALICE	77.40	72.70	70.60	67.20	65.90	63.40	62.90	61.90	60.50	60.60	60.10	17.30
FeSSSS	79.60	73.46	70.32	66.38	63.97	59.63	58.19	57.56	55.01	54.31	52.98	26.62
<i>Pro-KT</i>	82.90	74.15	72.82	62.46	59.69	54.88	51.56	48.57	47.32	46.86	47.68	35.22
LIMIT	76.32	74.18	72.68	69.19	68.79	65.64	63.57	62.69	61.47	60.44	58.45	17.87
NC-FSCIL	80.45	75.98	72.30	70.28	68.17	65.16	64.43	63.25	60.66	60.01	59.44	21.01
MCNet	77.57	73.96	70.47	65.81	66.16	63.81	62.09	61.82	60.41	60.09	59.08	18.49
SoftNet	78.07	74.58	71.37	67.54	65.37	62.60	61.07	59.37	57.53	57.21	56.75	21.32
CoSR	74.87	73.15	68.23	63.50	62.72	59.10	57.46	55.73	53.28	52.31	51.75	23.12
M-FSCIL	81.04	<u>79.73</u>	76.62	73.30	71.22	68.90	66.87	65.02	63.90	62.49	60.40	20.64
FSIL-GAN	81.27	78.03	75.61	<u>73.72</u>	70.49	68.19	66.58	65.63	63.21	60.92	59.43	21.84
CPE-CLIP	<u>81.58</u>	78.52	<u>76.68</u>	71.86	<u>71.52</u>	<u>70.23</u>	<u>67.66</u>	<u>66.52</u>	<u>65.09</u>	<u>64.47</u>	<u>64.60</u>	<u>16.98</u>
<i>PEAK</i>	76.67	81.09	76.80	76.20	74.80	72.44	71.52	69.17	70.33	68.17	68.92	7.74

although CPE-CLIP achieves the lowest PD (7.46), it does so at the cost of lower accuracy compared to M-FSCIL. This advantage is likely attributed to the overlap between CLIP’s pretraining data and the MiniImageNet dataset. In contrast, *PEAK* achieves a superior base session accuracy ($ACC_N = 90.23\%$), further validating our model’s stronger initialization and representation capabilities—even without relying on large-scale external pretraining as required by CLIP.

A similar trend is observed on the *CUB200* dataset, where *PEAK* achieves a final accuracy of 68.92% (PD = 7.74), surpassing all major baselines including Meta-FSCIL, FeSSSS, NC-FSCIL, and SoftNet. This consistent superiority across diverse datasets demonstrates the strong generalization capability of *PEAK* under varying domain complexities and task granularities.

Long-term Stability Analysis. Figure 5.3 shows that *PEAK* performance advantage over baseline methods increases consistently as more tasks are learned. This growing performance gap demonstrates two key strengths: superior long-term stability in maintaining knowledge across extended task sequences and exceptional adaptability to evolving knowledge boundaries, the capabilities needed for real-world continual few-shot learning scenarios. These findings further confirm the robustness and scalability of our framework in addressing the challenges of long-term incremental learning.

5.3.3 Ablation Studies

In this sub-section, we evaluate the three main components of our framework: (1) the token selection mechanism in ITA, (2) the complete instance-wise token augmentation module, and (3) the full adaptive knowledge space AKS. Performance is

5. IMPROVING OPEN-WORLD CONTINUAL LEARNING UNDER THE CONSTRAINTS OF SCARCE LABELED DATA

Table 5.4: Averaged classification accuracy (%) on MiniImageNet (5-way 5-shot).

Methods	ACC_N in each task (%) $\uparrow$										PD $\downarrow$
	0	1	2	3	4	5	6	7	8		
iCaRL	61.31	46.32	42.94	37.63	30.49	24.00	20.89	18.80	17.21	44.10	
TOPIC	61.31	50.09	45.17	41.16	37.48	35.52	32.19	29.46	24.42	36.89	
CEC	72.00	66.83	62.97	59.43	56.70	53.73	51.19	49.24	47.63	24.37	
Meta-FSCIL	72.04	67.94	63.77	60.29	57.58	55.16	52.9	50.79	49.19	24.41	
ALICE	80.60	70.60	67.40	64.50	62.50	60.00	57.80	56.80	55.70	24.90	
FeSSSS	81.50	77.04	72.92	69.56	67.27	64.34	62.07	60.55	58.87	22.63	
<i>Pro-KT</i>	98.59	79.82	77.85	71.82	70.34	69.18	70.13	69.27	69.73	28.86	
LIMIT	72.32	68.47	64.30	60.78	57.95	55.07	52.70	50.72	49.19	23.13	
NC-FSCIL	84.02	76.80	72.00	67.83	66.35	64.04	61.46	59.54	58.31	25.71	
MCNet	72.33	67.70	63.50	60.34	57.59	54.70	52.13	50.41	49.08	23.25	
SoftNet	76.63	70.13	65.92	62.52	59.49	56.56	53.71	51.72	50.48	26.15	
CoSR	71.92	66.91	62.71	59.59	56.63	53.78	51.01	49.23	47.93	23.99	
M-FSCIL	93.45	91.82	87.09	88.07	86.75	<u>87.15</u>	<u>85.68</u>	<u>84.80</u>	<u>85.37</u>	8.08	
FSIL-GAN	69.87	62.91	59.81	58.86	57.12	54.07	50.64	48.14	46.14	23.73	
CPE-CLIP	90.23	89.56	<u>87.42</u>	<u>86.80</u>	<u>86.51</u>	85.08	83.43	83.38	82.77	7.46	
<i>PEAK</i>	<u>96.00</u>	<u>91.65</u>	89.17	84.90	86.46	87.18	87.57	85.71	87.12	10.88	

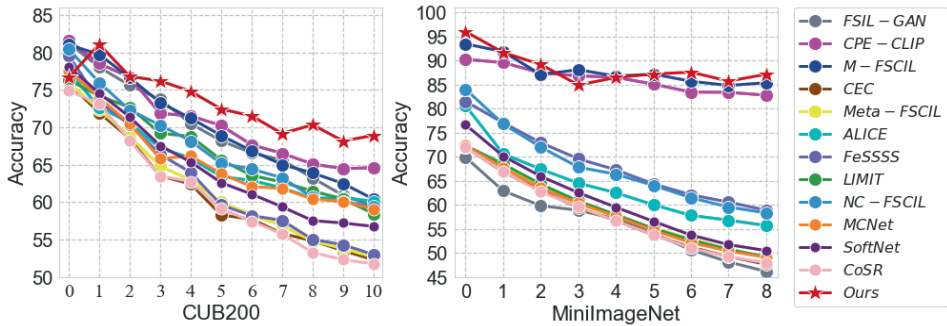


Figure 5.3: Visualizations of model performance on known classification.

measured by the average final accuracy ACC_N across all N incremental tasks, with results presented in Table 5.5.

Table 5.5 shows three key insights about *PEAK*'s architecture. First, the token selection mechanism in ITA proves essential for both classification and detection tasks, with its removal causing notable performance drops, particularly in maintaining knowledge boundaries and open-set detection. Second, the complete ITA module demonstrates the most substantial impact on classification accuracy, confirming its crucial role in preventing overfitting and catastrophic forgetting. Third, the AKS module emerges as fundamental for open-set capability, as its ablation completely disables unknown detection while also degrading known-class classification (Table 5.5), exposing the limitations of softmax-only approaches.

The *PEAK* framework's design, operating at input/output levels rather than an intermediate layer, ensures interesting backbone-agnostic properties. This is vali-

Table 5.5: Ablation studies.

Unknowns Detection	CUB200		MiniImageNet	
	AUC_N	FPR_N	AUC_N	FPR_N
w/o selection mechanism	66.68	69.96	75.73	80.54
w/o ITA	67.06	75.83	70.61	86.71
w/o AKS	-	-	-	-
<i>PEAK</i>	69.72	69.85	76.20	71.30
Knowns Classification	CUB200		MiniImageNet	
	ACC_N	PD	ACC_N	PD
w/o selection mechanism	54.52	12.48	61.34	29.66
w/o ITA	50.33	31.14	51.42	34.58
w/o AKS	55.24	28.42	63.02	34.78
<i>PEAK</i>	63.80	8.20	65.24	25.56

dated through ViT experiments: On CUB200, *PEAK* achieves 63.80% classification accuracy versus just 21.19% for fine-tuned ViT alone. This $3\times$ performance gap demonstrates that raw representational power is insufficient without our structured knowledge transfer mechanisms for continual few-shot adaptation.

5.3.4 Additional Results

Time and Space Complexity. The Adaptive Knowledge Space shows linear scaling complexity $\mathcal{O}(n)$ as new tasks emerge, demonstrating efficient practical deployment with a compact memory footprint of only 0.048 MB for CUB200 and 0.024 MB for MiniImageNet after complete training. The framework achieves real-time performance with 0.001s token matching latency and under 0.01s classification decisions, while training requires just 0.4284s per epoch on CUB200 and 0.8279s on MiniImageNet, all within a constrained peak memory budget of 28.87 MB (i.e., 28869632B), confirming its suitability for continual learning scenarios.

Parameter Sensitivity Analysis. The framework’s performance depends on several hyperparameters, including the separation margin m and the deviation constraint σ (governing boundary tightness in Equation 5.9), along with prompt configuration parameters L_P , K , and l for token length, token size K , and the number of tokens per task, respectively.

Experiments show that excessively large K or L_P values cause knowledge underfitting (Figure 5.4a), while the token quantity l maintains stable performance within $5 \leq l \leq 30$ (Figure 5.4b). The deviation constraint σ shows an interesting interaction with m : at $m = 0.5$, performance degrades with increasing σ , whereas larger margins benefit from higher σ values (Figure 5.5), highlighting the margin’s pivotal role in open-detection robustness.

The loss weighting factor γ used for balancing classification, ITA similarity, and MOB distance terms in Equation 5.13, demonstrates consistent performance across

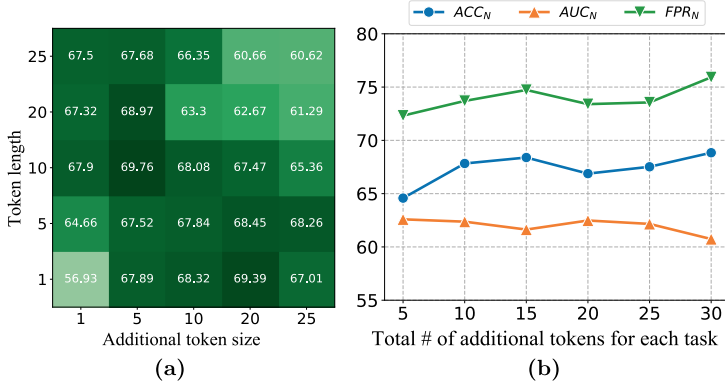


Figure 5.4: (a): ACC_N w.r.t token length L_P and additional token size K , given $l = 25$. (b): ACC_N w.r.t. l (i.e., the total number of additional tokens of each task) with $L_P = 5$ and $K = 5$ (take CUB200 for illustration).

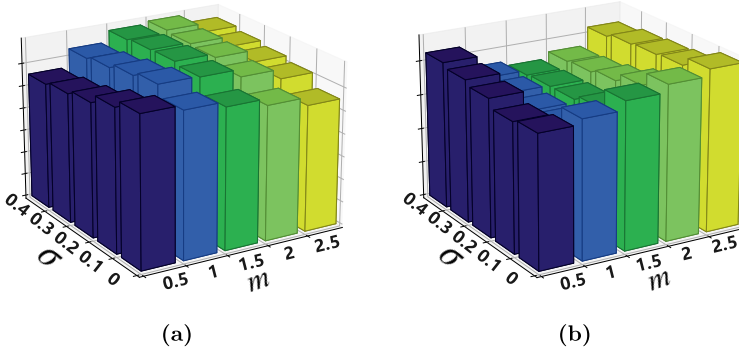


Figure 5.5: (a) and (b): AUC_N and FPR_N w.r.t constraint governing deviations σ and margin m (CUB200 dataset).

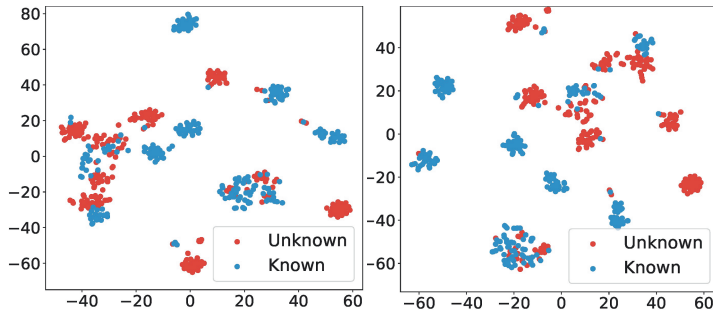
varying configurations (Table 5.6), demonstrating the *PEAK* framework’s stability to loss-weight adjustments.

In summary, the *PEAK* framework demonstrates consistent performance across diverse realistic hyperparameter configurations, confirming its inherent robustness to parameter variations. This stability suggests practical advantages for real-world deployment, where exhaustive hyperparameter tuning may be impractical. While optimal performance requires tuning 4–5 key parameters, a reasonable engineering trade-off, our open-sourced implementation achieves state-of-the-art results even with commonly used default settings. These characteristics position the proposed *PEAK* as both a robust and readily adoptable solution for open-world continual learning scenarios.

Table 5.6: Performance under different γ configurations on the CUB200 dataset.

γ	ACC_N	AUC_N	FPR_N
1:00:01	57.18	62.12	73.81
1:0.1:0.9	57.78	62.37	72.85
1:0.3:0.7	57.77	62.47	73.25
1:0.5:0.5	58.05	62.46	73.35
1:0.7:0.3	57.42	62.50	72.30
1:0.9:0.1	57.53	62.08	73.75
1:01:00	57.94	62.29	73.03
0.5:0.01:1	59.25	62.14	73.61
0.1:0.01:1	61.24	61.94	73.54

5.3.5 Visualization and Analysis

**Figure 5.6:** Visual comparison of test sample representations: (left) baseline Transformer backbone versus (right) our proposed *PEAK* framework.

To show the effectiveness of *PEAK* in learning the open decision boundary, we compare the visualizations of the final embeddings of all test samples in an arbitrary session using the vanilla backbone and our proposed *PEAK* framework.

As illustrated from the left part of Figure 5.6, we observe that there is no clear boundary between known points and unknown points, making it difficult to distinguish between them effectively. In the right part of Figure 5.6, we can observe distinct boundaries between the known points and unknown points, along with a more balanced distribution of clusters. These visualizations provide compelling evidence that our proposed *PEAK* significantly improves the learning of discriminative and compact decision boundaries for open detection.

5.4 Summary

In this chapter, we introduced *PEAK*, a novel and versatile framework designed to tackle the challenges of open-world continual learning under limited supervision.

5. IMPROVING OPEN-WORLD CONTINUAL LEARNING UNDER THE CONSTRAINTS OF SCARCE LABELED DATA

At its core, *PEAK* dynamically models the evolving knowledge space, enabling effective extraction, representation, and progressive integration of information from both known and previously unseen classes. By jointly optimizing instance-wise token augmentation and adaptive knowledge accumulation, *PEAK* not only enriches the semantic representation of known categories but also enforces tighter, more discriminative decision boundaries—significantly boosting open-set recognition performance.

A major advantage of *PEAK* is its inherently *backbone-agnostic* design, allowing seamless integration with diverse feature extractors and embedding architectures. This plug-and-play flexibility ensures broad applicability across domains and datasets while maintaining strong generalization, even under varying data distributions and task complexities.

Looking ahead, several promising research directions emerge for future research. Integrating large language models (LLMs) could enhance cross-modal reasoning and provide richer semantic priors, further improving robustness in open-world scenarios. Also, improving the transparency of learned representations by interpretable knowledge structures may yield deeper insights into model decision-making, aiding both debugging and real-world trustworthiness. Finally, developing efficient strategies for rapid generalization and dynamic adaptation in the testing phase could further strengthen *PEAK*'s ability to handle evolving class spaces in truly open-ended dynamic environments.

Chapter 6

Exploring Open-world Continual Learning with Knowns-Unknowns Knowledge Transfer

In the previous two chapters, we provided a concise problem formulation for the open-world continual learning setting, focusing on the different challenges that arise in large-sample and few-shot supervision scenarios, respectively. We also proposed effective models to facilitate knowledge transfer within open-world continual learning. However, these prior efforts were inspired by the perspectives of continual learning and open-set recognition, and thus fell short of treating open-world continual learning as an integrated problem. Despite uncovering distinct challenges in open-world continual learning, a unified theoretical framework has yet to be established..

In this chapter, we explore the open-world continual learning problem from both theoretical and empirical perspectives. We define and formalize four open-world continual learning scenarios, thereby refining our previous problem definition and scenario setting. Most importantly, based on our theoretical analysis of the open-world continual learning problem, we propose a unified technical framework, called ***HoliTrans*** (Holistic Knowns-Unknowns Knowledge Transfer), which aims to address the full spectrum of open-world continual learning scenarios and to bridge the gap between theoretical formulation and practical implementation. Finally, from a practical standpoint, this chapter presents extensive re-implementation and benchmarking of a variety of representative and competitive models. The released code-base and experimental protocols serve as a solid foundation for further research and practical exploration in open-world continual learning.

6.1 Toward Holistic Knowledge Transfer in OWCL

Unlike traditional learning models that operate under a closed and fixed set of classes, we have seen that the key distinction of open-world continual learning lies in its ability to learn continuously under an open-world assumption. This includes recognizing unseen or open-category samples while incrementally acquiring knowledge from new tasks without catastrophic forgetting [23, 24, 25].

Due to the potential occurrence of novelties in continual learning [14], open-world continual learning models must accurately detect unknowns to prevent misclassification into known categories. At the same time, they must retain previously acquired knowledge while continually performing open-set detection. In the last two chapters, we have discussed that the fundamental challenge in open-world continual learning: detection of unknowns and classification of knowns are *interdependent*. On one hand, the presence of unknowns complicates the trade-off between *knowledge stability* [1, 14] and *knowledge plasticity* [26, 27]. On the other hand, the incremental acquisition of new tasks makes open detection more difficult as the knowledge space expands.

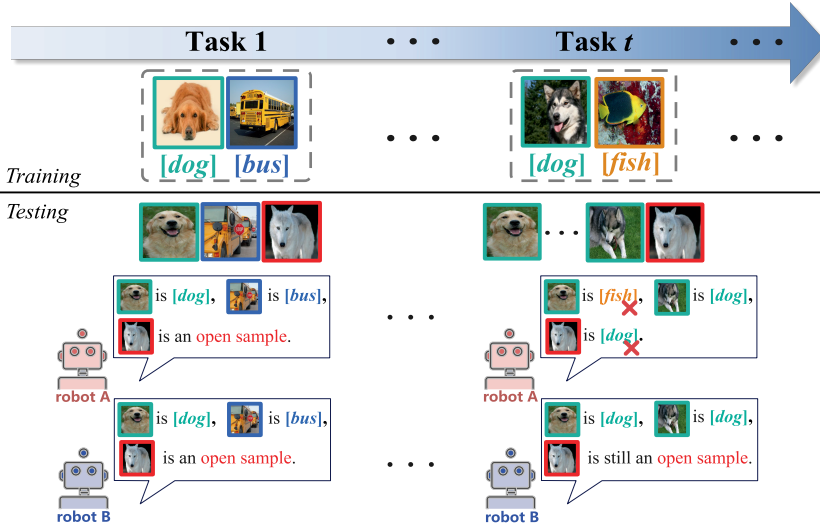


Figure 6.1: A toy example of the KIRO scenario in OWCL. During training Task 1 and Task t, Robots A and B learn *Golden Retrievers* and *Labradors*, both labeled as [dog]. In testing Task 1, both robots correctly classified [dog] and [bus] samples while identifying the unknown wolf without errors. However, in testing Task t, due to the emergence of *Labrador*, Robot A experienced distributional shifts when recognizing *Golden Retrievers*. This also led to the misclassification of *Wolves* as *Labradors*, likely due to visual similarity. In contrast, Robot B accurately classifies all known samples and consistently detects the wolf as an open sample via knowledge transfer.

Despite growing attention to open-world continual learning in recent years, current approaches [25, 31] often treat open-world continual learning as a simple combination of open-set recognition and continual learning, rather than as an integrated paradigm. As a result, these methods tend to be effective only in transferring knowledge related to known samples, while neglecting knowledge that can be derived from unknowns.

To address this limitation, we argue that a promising open-world continual learning model must support **knowns-unknowns knowledge transfer**, i.e., the ability to transfer knowledge both from known categories and from previously unseen samples. However, a thorough formulation of the open-world continual learning problem and an empirical exploration of its main challenges are lacking. This makes it difficult to fairly compare existing methods or choose an appropriate solution for specific scenarios.

As shown in Figure 6.1, in real-world settings, labeled data for a new category is rarely acquired all at once or from a single task [24]. Instead, such data typically arrives incrementally across multiple tasks [28, 29]. Consequently, a specific category may recur across tasks with shifting data distributions. Furthermore, as tasks accumulate, the boundaries between known categories become increasingly ambiguous. The recurring presence of open/unknown samples complicates the model’s ability to distinguish between known and unknown instances.

Accordingly, as outlined in Table 6.1, we identify two principles for refining open-world continual learning scenarios: (1) *whether a known class with changing data distribution appears in different tasks during training*, and (2) *whether an unknown sample reappears in different tasks during the testing phase*.

Table 6.1: Overview of the four OWCL scenarios.

<i>Scenario</i>	CINO	CIRO	KINO	KIRO
Known Classes Repeatedly Appear	✗	✗	✓	✓
Unknown Samples Repeatedly Appear	✗	✓	✗	✓

Building on these principles, we extend traditional continual learning scenarios, namely, task-incremental, domain-incremental, and class-incremental learning, into the open-world setting. However, unlike conventional continual learning, open-world continual learning disallows task identity information during testing, as unknown samples may randomly emerge [30]. This renders task-incremental and domain-incremental paradigms insufficient.

To overcome this, we introduce a novel scenario definition: **knowledge-incremental learning (KIL)**. In knowledge-incremental learning, models must solve each previously encountered task without knowing the task identity, and training sets across

6. EXPLORING OPEN-WORLD CONTINUAL LEARNING WITH KNOWN-UNKNOWN KNOWLEDGE TRANSFER

tasks are allowed to overlap. This permits shifts in the data distribution of known classes. Knowledge-incremental learning unifies class-incremental learning with distributional variation, providing a more realistic and challenging setting for open-world continual learning.

Based on this definition, we propose four open-world continual learning scenarios of increasing difficulty: **CINO** (Class-Incremental with Non-repetitive Open samples), **CIRO** (Class-Incremental with Repetitive Open samples), **KINO** (Knowledge-Incremental with Non-repetitive Open samples), and **KIRO** (Knowledge-Incremental with Repetitive Open samples).

Recent work [76, 126, 127] has emphasized the importance of both inter-task and intra-task open samples in open-set recognition. We follow this direction and propose an open-world continual learning model that explicitly distinguishes: *inter-task open samples*, which appear across multiple tasks but were not part of earlier training sets, and *intra-task open samples*, which resemble known classes within a single task while also preserving accurate classification of known categories.

To evaluate these challenges, we conduct experiments under the most difficult open-world continual learning scenario, KIRO.

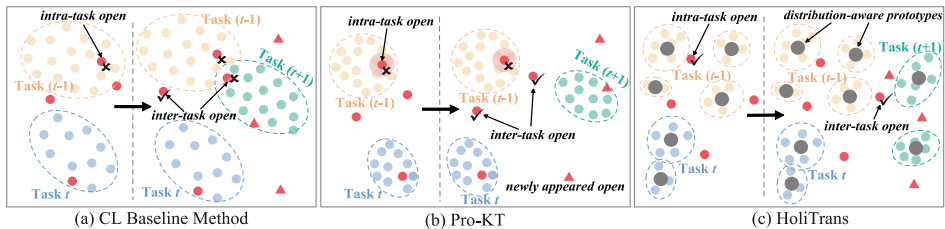


Figure 6.2: From task t to task $t + 1$ in the KIRO scenario, we illustrate the embedding space using different baselines. Red circles and red triangles denote unknown samples encountered in tasks t and $t - 1$, respectively. Orange, blue, and green circles represent labeled training data learned during tasks $t - 1$, t , and $t + 1$, respectively. Dashed lines denote the decision boundaries inferred by the model during open-world continual learning. (a) The original embedding distributions for each task learned by the classical continual learning baseline EWC. (b) A current competitive open-world continual learning approach enforces tighter embedding cohesion for each task. (c) Our proposed method jointly addresses open detection and incrementally reduces known sample classification errors through knowledge transfer. [Best view in color]

Experimental results show that effective novelty detection in open-world continual learning must consider both inter-task and intra-task open samples. Furthermore, there exists a significant interaction between open detection and classification for known categories, challenging prior claims that the two are orthogonal.

As illustrated in Figure 6.2(a), using the classical continual learning baseline EWC [21] to learn the original embedding space under the KIRO scenario, we see that

with the incremental acquisition of new tasks and the recurring appearance of open samples, the embedding distributions of different tasks may overlap. This results in missed inter-task open samples and misclassified intra-task open samples into known categories due to the similarity between these classes.

In Figure 6.2 (b), the Pro-KT [30] baseline introduced in Chapter 4 improves task boundaries for detecting inter-task unknown samples but struggles with intra-task open detection. This is because intra-task open samples within the convex hull [128] of known categories (highlighted in pink) are likely to be misclassified as known categories.

Together, these findings show that both types (inter-task and intra-task) of open samples increase the risk of misclassification and degrade the incremental classification errors of knowns.

Our experimental results show a significant interaction between open detection and incremental prediction in open-world continual learning, whereas existing open-world continual learning works [19, 24, 25] consider them as orthogonal. As a consequence, existing open-world continual learning methods rely on a simple continual learning framework combined with an open detection module, which are only effective in the CINO scenario but fail in more complex open-world continual learning settings.

This motivates the need for models to enable effective knowledge transfer, not only to handle both known and unknown samples but also to update and expand the knowledge space for open-world continual learning incrementally. Accordingly, we establish a unified formulation to address the open-world continual learning problem, based on theoretical principles.

We introduce **nonlinear random projection (NRP)** for open-world continual learning to facilitate the learning of a more linearly separable embedding space, thereby reducing errors in knowledge transfer for known samples. At the same time, we propose a **distribution-aware prototypes (DAPs)** approach, using generative replay with novel pseudo-open samples to transfer knowledge related to novelties while updating the representation of open samples during the open-world continual learning process.

In the following, based on the experimental investigations conducted under our four open-world continual learning scenarios, we analyze the problems inherent to each scenario. Motivated by these insights, we further refine the problem formulation introduced in the previous chapters and conduct an in-depth exploration of the Knowns-Unknowns Knowledge Transfer problem, addressing both its algorithmic and theoretical aspects.

To address the challenges of open-world continual learning, this chapter introduces ***HoliTrans*** (holistic knowns-unknowns knowledge transfer), a unified framework de-

signed to tackle all four open-world continual learning scenarios by bridging theoretical analysis and practical performance. *HoliTrans* offers theoretical guarantees and is broadly applicable across diverse open-world continual learning settings. It integrates *nonlinear random projection* and *distribution-aware prototypes* to construct an adaptive knowledge space that enables effective knowledge transfer from both known and unknown samples, while dynamically updating representations throughout the continual acquisition of novel tasks.

The contributions of this chapter can be summarized as:

- A formal construction for open-world continual learning, introducing four distinct scenarios. Empirical experiments find a significant interaction between open detection and incremental prediction in open-world continual learning, challenging the current assumption of orthogonal decomposition.
- The introduction of a novel open-world continual learning framework, *HoliTrans*, which transfers knowledge from both known and unknown samples, and thus achieves open detection and incremental prediction quality in a unified manner.
- Experiments to validate our theoretical findings, with *HoliTrans* outperforming 22 baselines across different open-world continual learning scenarios.

Our code is open-source and available on <https://github.com/AIGNLAI/HoliTrans>. This chapter is based on the following paper [38]:

- **Li, Y.,** Lai, G., Yang, Y., Li, Y., Bonsangue, M., & Li, T.: Exploring Open-World Continual Learning with Knowns-Unknowns Knowledge Transfer. arXiv preprint arXiv:2502.20124.

6.2 Rethinking OWCL: A Scenario-Based Framework

In this section, we first refine the problem definition of open-world continual learning by integrating the optimization with open risk from unknowns and incremental prediction errors from knowns. Then, we provide an in-depth categorization scheme for different open-world continual learning scenarios. Subsequently, to explore potential issues in open-world continual learning, we conduct experiments with baselines using Softmax-based and NCM (Nearest Class Mean)-based classifiers on different baselines.

6.2.1 Problem Formulation

In open-world continual learning, samples not seen in the training set may appear during testing; thus, we use $\mathcal{D}_{tr}$ and $\mathcal{D}_{te}$ to distinguish between the training set and test set; all superscripts indicate the task order. The notations used throughout this chapter are listed in Table 6.2 for better clarification.

Table 6.2: Notations and explanations.

Notation	Explanation
t, i	Task identifiers
h^t	Neural network after task t
D_{tr}^i	Training set of task i
D_{te}^i	Testing set of task i
N_i	Number of training samples in D_{tr}^i
M_i	Number of seen classes in D_{tr}^i
$\mu(h(x), h(D_{tr}^i))$	Open risk of x in D_{te}^i
$\mathcal{H}$	Universal hypothesis space
$\epsilon_{\mathcal{D}_{tr}^i}(h)$	General prediction error on task i
λ	Balance for open risk and classification
$\mathcal{S}$	Universal space
$\mathcal{O}$	Open space for open-world continual learning
$\mathcal{K}$	Known space from each trained task
$\hat{\epsilon}_{\mathcal{D}_{tr}^i}(h^t)$	Empirical prediction error
$\mathcal{B}^t$	Experience buffers from $\mathcal{D}_{tr}^t$
s_x	Max similarity of x to prototypes
$g(\cdot)$	Nonlinear function
$\phi(\cdot)$	Pre-trained model
$\mathcal{P}$	Projection layer
G^i	Gram matrix of the NRP feature
$\mathbf{p}_k^i$	Prototype of class k in task i
$\mathcal{D}_{te}^i$	New test set with pseudo-samples

Recall from Definition 7 in Chapter 4 that we introduced a formal mathematical formulation of the objectives of OWCL. However, that definition captures only the basic objectives of OWCL and does not account for the four distinct OWCL scenarios developed in this chapter. Moreover, it does not formally characterize the open risk and incremental classification error. Therefore, this chapter presents a refined problem formulation that explicitly incorporates both the open risk and the incremental classification error, which are fundamental to all OWCL settings:

Definition 9 A OWCL Problem Formulation. *Each task i has N_i training samples and M_i classes $\mathcal{D}_{tr}^i = \{\mathbf{x}_j^i \in \mathcal{X}_{j=1}^{N_i}, y_j^i \in \mathcal{Y}_{j=1}^{M_i}\}$. In the training phase, only data pertinent to the current task is accessible, while the test samples $\mathcal{D}_{te}^i = \{\mathbf{x}_j^i \in \mathcal{X}_{j=1}^{N'_i}, y_j^i \in \mathcal{Y}_{j=1}^{M'_i}\}$ may contain unknowns/novelities, i.e., $\mathcal{Y}_{j=1}^{M'_i} - \mathcal{Y}_{j=1}^{M_i} \neq \emptyset$. Given the latent representation of the current task t 's training set $h(\mathcal{D}_{tr}^t)$, we denote $\mu(h(\mathbf{x}), h(\mathcal{D}_{tr}^i))$ as the open risk of a sample $\mathbf{x}$ from task i 's test set $\mathcal{D}_{te}^i$. The goal of open-world continual learning is to learn a uniform function h^* that minimizes the*

open risk and incremental prediction error jointly, across all seen tasks $[1, t]$:

$$\arg \min_{h^* \in \mathcal{H}} \left\{ \underbrace{(1 - \lambda) \sum_{i=1}^t \mu(h(\mathbf{x}), h(\mathcal{D}_{tr}^i))}_{(A)} + \lambda \underbrace{\sum_{i=1}^t \epsilon_{\mathcal{D}_{tr}^i}(h)}_{(B)} \right\}, \quad (6.1)$$

where $\mathcal{H}$ is a universal function space, $\epsilon_{\mathcal{D}_{tr}^i}(h)$ is the generally suggested prediction error term on task i [15], μ can be defined on any scoring function for out-of-distribution detection, λ is a hyper-parameter to balance the open risk and the incremental prediction error.

6.2.2 Four OWCL Scenarios

As discussed previously, traditional continual learning is usually categorized into three scenarios, i.e., *incremental learning*, *domain-incremental learning*, and *class-incremental learning*, based on whether task identifications are provided during tests [129]. However, in open-world continual learning, it is impossible for the model to know the task identifiers because unknown/open samples may randomly appear alongside known ones during testing. Furthermore, the shifting distributions of known classes and the repetitive presence of open samples hinder the current continual learning three scenarios [129] from being compatible with open-world continual learning problems. Consequently, four distinct scenarios of increasing difficulty (presented in Table 6.1).

First, in the class-incremental learning with non-repetitive open samples (**CINO**) scenario, the training set for each task is introduced incrementally without repeated classes. Once previously unseen samples are encountered during testing, they do not reappear in future tasks. This scenario is ideal for applications that require continual learning of new classes without retaining historical data, such as one-time classification tasks in fields like species identification in ecological monitoring or material recognition.

In many biodiversity surveys, for example, researchers deploy field sensors or drones to capture environmental data over time. Each survey (task) introduces images or signals corresponding to a distinct, non-overlapping set of species that were not observed previously. Moreover, due to seasonal or regional constraints, once certain rare or migratory species are encountered during a testing phase, they are unlikely to reappear in future surveys. In such settings, models must continually learn to classify newly discovered species without retaining raw data from earlier surveys, often due to storage or privacy limitations, while also correctly identifying when a sample belongs to an unknown, previously unseen species.

In the second scenario, class-incremental learning with repetitive open samples (**CIRO**), the training set also grows incrementally without repeated classes, but

open samples encountered in previous tasks may recur in future tests. This scenario is particularly suited to dynamic environments, such as autonomous driving or robotic systems, where the system must handle repetitive open samples over time, such as environmental objects like traffic signs or obstacles.

Expanding on the example of autonomous driving systems, the vehicle navigates through different regions or updates its software over time, and it incrementally learns to recognize new classes of road users, traffic signs, or environmental features (e.g., new types of vehicles or traffic control devices). However, during testing, it may repeatedly encounter previously unseen objects, such as rare construction signs, fallen debris, or unusual pedestrian behavior patterns, that were classified as unknowns in earlier tasks. These open samples do not belong to any known training class but may reappear across different geographic areas or at later times, requiring the system to consistently detect them as unknowns while continuing to learn new known classes. Since these repetitive open samples can influence safety-critical decisions (e.g., whether to stop, reroute, or initiate emergency protocols), the ability to maintain stable detection of recurring unknowns is vital.

The third scenario, knowledge-incremental learning with non-repetitive open samples (**KINO**), involves training sets that evolve incrementally, with repeated classes and shifting distributions, but open samples encountered during testing do not reappear. This approach applies to domains that need to adapt to changing class distributions without recalling past open samples, such as streaming data analysis or online content recommendation systems, where rapid adaptation to evolving user preferences is crucial.

As a concrete example, consider an online content recommendation system where user interactions, such as clicks, views, or likes, are streamed continuously. The system must incrementally update its understanding of known user classes as their preferences shift over time. For example, a user who was previously interested in science fiction content may gradually shift toward documentaries, reflecting a change in the distribution of known preferences. The system must adapt to these distributional shifts while retaining the ability to recommend content accurately across all previously encountered preference profiles. At the same time, new types of content (e.g., a completely new genre or format) may appear during testing without prior exposure. These open samples are one-time occurrences, such as a brief trend or a viral video format, and may not reoccur in future recommendation cycles. In such cases, the model must detect these novel user-content interactions as unknowns while continually adapting to the evolving knowledge of known user classes.

Finally, in knowledge-incremental learning with repetitive open samples (**KIRO**), the training set evolves incrementally with repeated classes and changing distributions, while previously unseen open samples encountered during testing may reoccur. This scenario is most relevant for complex, dynamic real-world applications, such as intelligent surveillance systems or financial risk analysis, where models need to

learn from historical patterns and manage repetitive open samples, such as shifts in customer behavior or market trends.

A concrete example can be found in intelligent surveillance systems deployed in smart cities. These systems incrementally learn to recognize known entities (e.g., registered vehicles or authorized personnel), where the appearance and behavior of these entities may change over time due to lighting, weather, or seasonal variations, reflecting distributional shifts across tasks. Meanwhile, the system must also detect and respond to previously unseen but potentially suspicious patterns, such as loitering behavior, unauthorized access attempts, or unusual object placements. Importantly, such open samples are not one-off occurrences; the same unknown individuals or suspicious behaviors may reappear in different contexts or at different times, making it essential for the system to consistently detect them as open while updating its knowledge base.

Remarks. Knowledge-incremental learning poses greater difficulty than task-, domain- and class-incremental learning settings. The primary difference is that knowledge-incremental learning emphasizes adapting to potential shifts in the data distributions of previously learned tasks while incrementally learning new classes without task identifiers. Due to the presence of open samples and complex task distribution shifts in open-world continual learning, along with the inability to know task identifiers during testing, knowledge-incremental learning requires the model to transfer knowledge not only about known samples but also about unknown samples.

6.2.3 Experiments on All OWCL Scenarios

In this subsection, we begin by thoroughly examining potential issues within open-world continual learning. Many of our empirical results challenge prevailing claims in current open-world continual learning research, questioning existing assumptions within the field. Our findings further inspire us to explore related theories and propose a novel open-world continual learning model accordingly.

Recently, with the increasing application of pre-trained models and growing research on nearest class mean classifiers, current continual learning studies [130] found that the NCM strategy has yet to reach its full potential for accuracy. Remarkably, it can achieve standout performance when combined with carefully tailored strategies that enhance feature representations extracted from pre-trained models. Hence, given the distinct characteristics and performance exhibited by different classifiers and loss calculation methods, we conduct experiments using various approaches with different types of classifiers (i.e., Softmax-based and NCM-based) across four open-world continual learning scenarios. Specifically, as typical methods utilizing a Softmax-based classifier, fine-tuning and L2P [32] are current state-of-the-art approaches for handling continual learning. SimpleCIL [131] and RanPAC [130] represent two of the

latest and most competitive continual learning baselines employing an NCM-based classifier.

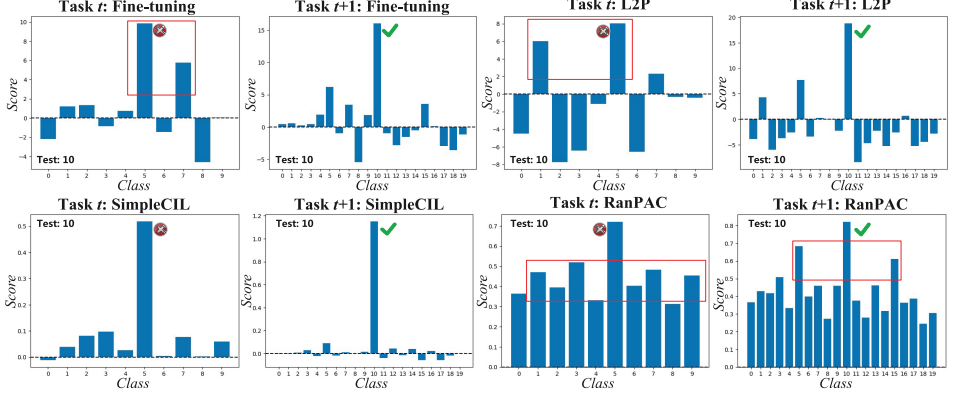


Figure 6.3: Unscaled classification scores on the CINO scenario.

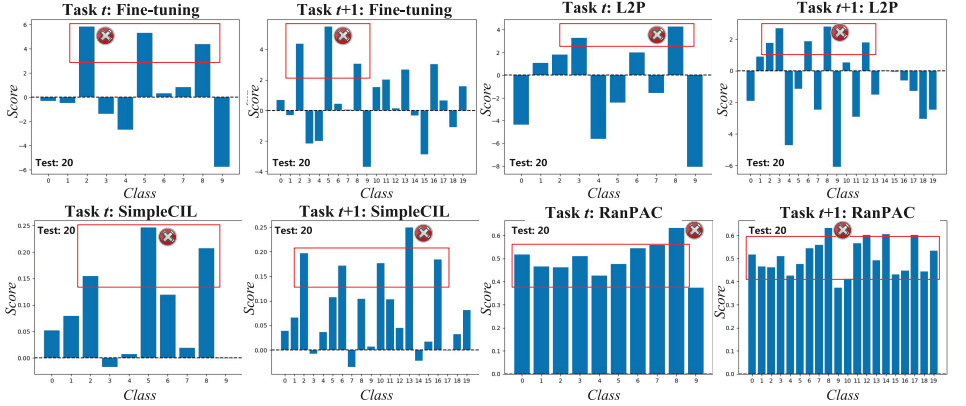


Figure 6.4: Unscaled classification scores on the CIRO scenario.

Figure 6.3 to Figure 6.6 respectively display the performance of different models across four distinct open-world continual learning scenarios, with the horizontal axis representing class labels and the vertical axis indicating unnormalized classification scores calculated by different types of classifiers.

From Figure 6.3, we observe that due to the CINO scenario, each task’s training set is presented incrementally without repeated classes, and previously unseen samples encountered during testing do not reappear subsequently. Therefore, open-world continual learning models fail to detect open samples appearing in task t , but after learning the corresponding training data for these open samples in task $t + 1$, the models are then able to classify them.

From Figure 6.4, in the CIRO scenario, because open samples encountered previously may reoccur in subsequent tests, we find that all the baseline models fail to detect open samples in both task t and task $t + 1$, erroneously classifying the inter-task and intra-task open samples into known classes, which results in significant open risk. Additionally, we notice that in the CIRO scenario, the unnormalized classification scores for each class are very close, reflecting the complexity of the data distributions among various tasks. By comparing the results in CINO and CIRO, we conclude that existing continual learning and incremental learning methods are unable to validate either inter-task or intra-task open samples and tend to classify open samples into seen categories erroneously.

Next, in the KINO and KIRO scenarios, each task’s training set evolves incrementally with repeated classes and changing distributions. Hence, from Figure 6.5 and Figure 6.6, we observe that the models fail to effectively distinguish open samples in task $(t + 1)$. Moreover, due to the presence of recurrent open samples in the KIRO scenario, the classification scores obtained by the models in this scenario are also closer. Consequently, the variation in data distribution between tasks poses greater challenges to open-world continual learning, and there exists a mutual influence between open risk and incremental prediction quality.

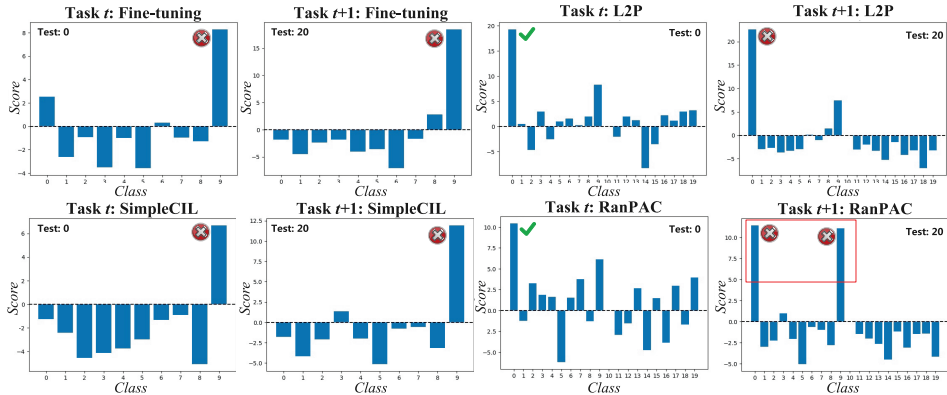


Figure 6.5: Unscaled classification scores on the KINO Scenario.

In conclusion, by comparing all the results from Figure 6.3 to Figure 6.6, as the difficulty of the four scenarios increases, the classification performance on known classes decreases because known classes repeatedly appear with changing data distributions.

6.2.3.1 Experimental Findings Summary

The empirical experiments show that in all open-world continual learning scenarios, the NCM-based classifier methods performed better than the Softmax-based classifier methods, especially when the data distributions changed across new tasks and

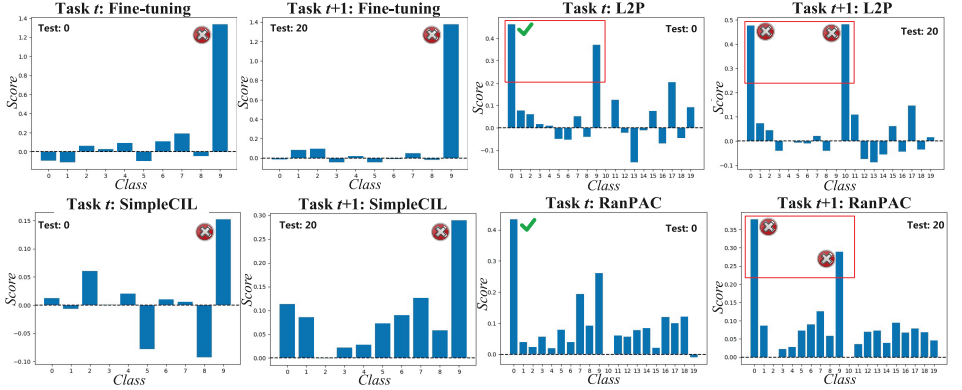


Figure 6.6: Unscaled classification scores on the KIRO Scenario.

open samples reappeared over time. In addition, we conclude with three key findings: (1) Vanilla continual learning and incremental learning methods are unable to validate open samples, either between tasks or within tasks, and tend to classify these open samples into previously seen categories mistakenly. (2) As the difficulty of the four scenarios increases, the classification performance on known samples decreases because known classes repeatedly appear with changing task distributions. (3) Open risk and incremental prediction error coexist in an interplay rather than independently, making them inseparable through orthogonal decomposition.

Our empirical observations demonstrate that the assumptions made in previous works [19, 24]—specifically, that the decomposition of open risk and incremental prediction error is orthogonal and unaffected by each other—are invalid. Our findings highlight the limitations of current open-world continual learning approaches and strongly motivate us to design a problem formulation with theoretical bases for open-world continual learning. Therefore, as open risk and known classification errors cannot be orthogonally decomposed, open-world continual learning models not only transfer knowledge from known samples but also obtain and transfer knowledge from unknown samples. Accordingly, we propose a novel framework, *HoliTrans*, which incorporates a nonlinear random projection method and distribution-aware prototypes to enable the knowledge transfer of both known and unknown samples (that we call **knowns-unknowns knowledge transfer**).

6.3 Theoretical Bases: Open Risk and Incremental Prediction Error

This section defines a new knowledge space and introduces a generic incremental prediction error term for open-world continual learning. We theoretically prove a tighter bound for incremental prediction error, providing several insights for design-

ing better open-world continual learning algorithms by considering knowledge from knowns and unknowns.

Following a general definition from [15], we extend the definition of **open space** $\mathcal{O}$ for open-world continual learning as follow:

Definition 10 *Open Space for OWCL.* *The open space for an open-world continual learning model is everything the model does not know yet at the current task t :*

$$\mathcal{O} = \mathcal{S} \setminus \bigcup_{\mathbf{x} \in \{\mathcal{D}_{tr}^i\}_{i=1}^t} \mathcal{K}(\mathbf{x}),$$

where $\mathcal{S}$ is the universal space, and $\mathcal{K}$ is the known space obtained by each trained task $i \in [1, t]$.

Note that it is not necessary to classify each open sample into exactly true unknown classes. For the sake of simplicity, we follow the general implementations [132], setting all unknown samples to one unified unknown as $[un]$. Hence, given the problem formulation in Definition 9, we assume that $\mathcal{Y}_{j=1}^{M'_i} - \mathcal{Y}_{j=1}^{M_i} = \{y_{un}\}$ where y_{un} represents the one unknown class and $M'_i = M_i + 1$. Therefore, from the probabilistic perspective, we can replace the left summand term (A) $\sum_{i=1}^t \mu(h(\mathbf{x}), h(\mathcal{D}_{tr}^i))$ in Equation 6.1 by the following term $R_{\mathbf{P},un}(h)$ redefining open risk in an open-world continual learning context:

Definition 11 *Open Risk of OWCL.* *The open risk $R_{\mathbf{P},un}(h)$ of the model h under the data distribution $\mathbf{P}$, specifically focusing on the unknown class y_{un} is defined by*

$$R_{\mathbf{P},un}(h) = \int_{\sum_{i=1}^t \mathcal{X}_{j=1}^{N'_i}} \ell(h(\mathbf{x}), y_{un}) d\mathbf{P}_{X|Y=y_{un}}(\mathbf{x}),$$

where $\mathbf{P}_{X|Y=y_{un}}$ is the conditional distribution of inputs $\mathbf{x}$ given that they belong to the unknown class y_{un} , ℓ is a loss satisfying $\ell(y, y') = 0$ if $y = y'$ and $\ell(y, y') = 1$ otherwise. It measures how far the model's prediction $h(\mathbf{x})$ is from correctly assigning the label y_{un} , and the integral computes the expected loss incurred by the model when it makes predictions on inputs from this unknown class.

This definition captures the expected loss a model incurs when it encounters unknown samples and is expected to classify them as belonging to the generic unknown class y_{un} . It quantifies the model's ability to distinguish between seen and novel inputs in a continual learning context without requiring precise labeling of all unknowns. Since all unknowns are collapsed into one generic class, the open risk $R_{\mathbf{P},un}(h)$ tells us how well the model identifies novel inputs as unknown instead of mistakenly classifying them as one of the known classes. It quantifies the model's ability to recognize and correctly handle inputs from unknown classes in a probabilistic and aggregated sense, across all tasks seen up to time t .

Previous works [19, 24] claimed that the decomposition in Equation 3.1 is orthogonal, meaning that the terms denoting the open-set risk and the incremental error do not

affect each other. However, from observations in Section 6.2.3, we find that the open risk affects the incremental prediction quality, and that the boundary across the learned tasks also affects the open detection performance. Therefore, we need to consider the interaction of open risk (term (A) in Definition 9) and incremental prediction error (term (B) in Definition 9) in a unified manner.

Next, we introduce a novel incremental prediction error term and theoretically derive a more compact bound for open-world continual learning. Basically, after the model finishes training on a new task t , we want to evaluate how well it performs not just on the current task, but also on any previous task i , with $i \leq t$. To do this, we compute the prediction error, which measures the proportion of mistakes with respect to the ground truth g_i that the model makes on the training data of task i using the model h^t trained up to task t .

Definition 12 Task-Specific Empirical Prediction Error. *After training a new task t , the current model can be denoted as h^t . Accordingly, the empirical prediction error $\hat{e}_{\mathcal{D}_{tr}^i}(h^t)$ of h^t in testing an arbitrary task $i(i \leq t)$ is defined as [133]:*

$$\hat{e}_{\mathcal{D}_{tr}^i}(h^t) = \begin{cases} \frac{1}{N_i} \sum_{\mathbf{x} \in \mathcal{X}_{j=1}^{N_i}} \mathbb{I}(h^t(\mathbf{x}) \neq g_i(\mathbf{x})) & \text{if } i = t, \\ \frac{1}{\tilde{N}_i} \sum_{\mathbf{x} \in \tilde{\mathcal{X}}_{j=1}^{\tilde{N}_i}} \mathbb{I}(h^t(\mathbf{x}) \neq g_i(\mathbf{x})) & \text{if } i < t. \end{cases}$$

where $\mathbb{I}(\cdot)$ is an indicator function that returns 1 if the input condition holds, and 0 otherwise. Only a small subset of data from previous tasks ($i < t$) is available at task t , i.e., $\tilde{N}_i \ll N_i$, and $\tilde{N}_i$ is the amount of replayed samples, leading the task i 's prediction error to deviate much from their overall real risks with severe forgetting issues.

The task-specific empirical prediction error above is defined differently depending on whether we are evaluating the current task ($i = t$) or a past task ($i < t$). In the case of the current task, we compute the error rate using the full training set, and we average over all samples whether the model's prediction matches the ground truth label given by g_i . Otherwise, in the case we are evaluating a previous task, we generally do not have access to the full training set anymore. Instead, we rely on a small subset of replayed samples from that task and compute the error rate in the same way as above, but only over this smaller dataset. Because the model is being evaluated on only a small sample of past data, and due to catastrophic forgetting, where models forget how to perform earlier tasks, the computed error rate can be quite different from the true error rate we would get if the full dataset were available. This often results in underestimating the model's real degradation on past tasks.

Hence, in open-world continual learning, when data from previous tasks is unavailable, using the historical model h^{t-1} from time $(t-1)$ rather than relying on replayed data offers a promising alternative for tightening the prediction error bound, thereby motivating the development of our improved method.

6. EXPLORING OPEN-WORLD CONTINUAL LEARNING WITH KNOWN-UNKNOWN KNOWLEDGE TRANSFER

Considering all the OWCL scenarios mentioned above, we first establish a general upper bound that accounts for the potential distributional shift between tasks:

Lemma 1 *Let h^t be the current model trained on task t , and h^{t-1} the model trained on the previous task $t-1$. Then, the task-specific prediction error of h^t on a previous task i (where $i \leq t$) can be bounded as follows:*

$$\epsilon_{\mathcal{D}_{tr}^i}(h^t) \leq \epsilon_{\mathcal{D}_{tr}^i}(h^t, h^{t-1}) + \frac{1}{2}d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_{tr}^i, \mathcal{D}_{tr}^t) + \epsilon_{\mathcal{D}_{tr}^i}(h^{t-1}).$$

Here $\mathcal{H}$ is a universal hypothesis space, i.e., a sufficiently expressive class of models under consideration, with a finite VC (Vapnik–Chervonenkis) dimension, which measures its capacity to shatter or fit arbitrary data; the term $\epsilon_{\mathcal{D}_{tr}^i}(h^t, h^{t-1})$ quantifies the disagreement between the current and previous models on task i ; and $\epsilon_{\mathcal{D}_{tr}^i}(h^{t-1})$ is the previous model’s prediction error on task i . Finally, $d_{\mathcal{H}\Delta\mathcal{H}}$ measures the divergence between the data distributions of task i and task t , based on the ability of classifiers in $\mathcal{H}$ to distinguish between the two domains.

From a Bayesian perspective, this divergence term can be approximated as:

$$2 \sup_{h^t \in \mathcal{H}} |\mathbf{P}_{\mathbf{x} \sim \mathcal{D}_{tr}^i}[h^t(\mathbf{x}) = 1] - \mathbf{P}_{\mathbf{x} \sim \mathcal{D}_{tr}^t}[h^t(\mathbf{x}) = 1]|.$$

This lemma provides an upper bound on the task-specific prediction error of the current model h^t after learning task t , when evaluated on any task $i \leq t$.

Specifically, it states that the prediction error of h^t on task i is no greater than the sum of two components: (1) $\epsilon_{\mathcal{D}_{tr}^i}(h^t, h^{t-1})$, the disagreement between the current model and the previous model on data from task i , and (2) $\epsilon_{\mathcal{D}_{tr}^i}(h^{t-1})$, the prediction error of the previous model on task i . Intuitively, this means that if the current model does not deviate much from the previous model on past tasks, and the previous model performed well on those tasks, then the new model is also likely to perform well. This bound highlights that the incremental prediction error will gradually accumulate as new tasks arrive.

More importantly, Lemma 1 also explicitly accounts for the potential distributional shift between tasks by considering $d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_{tr}^i)$. In practice, data distributions often differ significantly across tasks in open-world continual learning, making it essential to consider how such shifts affect generalization. Hence, this lemma introduces a distributional divergence term that refines the error bound by capturing the extent to which the current task distribution $\mathcal{D}_{tr}^t$ differs from a previous task distribution $\mathcal{D}_{tr}^i$. This divergence helps characterize when the historical model h^{t-1} can still provide reliable guidance for the current task, thus offering a tighter and more realistic prediction error bound in open-world settings.

In summary, Lemma 1 provides a theoretical basis for knowledge transfer in open-world continual learning. It shows that if the divergence between a previous task i and the current task t , as measured by $d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_{tr}^i, \mathcal{D}_{tr}^t)$, is sufficiently small, then

the predictions made by the earlier model h^{t-1} can be reliably used to guide learning on the current task. In particular, these predictions can act as a surrogate loss, a substitute for the true loss function that gives a practical way to enforce consistency with past knowledge that may no longer be accessible. This idea is used in many continual and open-world continual learning approaches, which introduce knowledge transfer mechanisms, such as distillation or regularization, that minimize distribution divergence. By doing so, they preserve task-specific knowledge while allowing new information to be integrated effectively, reducing catastrophic forgetting.

With the Lemma 1, we can derive the incremental prediction error bound as:

Theorem 1 Incremental Prediction Error Bound. *With a probability of at least $1 - \delta > 0$, the incremental prediction error bound can be refined as:*

$$\begin{aligned}
 \sum_{i=1}^t \epsilon_{\mathcal{D}_{tr}^i}(h^t) &\leq \left\{ \sum_{i=1}^{t-1} [\gamma_i \widehat{\epsilon}_{\mathcal{D}_{tr}^i}(h^t) + \alpha_i \widehat{\epsilon}_{\mathcal{D}_{tr}^i}(h^t, h^{t-1})] \right\} \\
 &\quad + \left\{ \widehat{\epsilon}_{\mathcal{D}_{tr}^t}(h^t) + \left(\sum_{i=1}^{t-1} \beta_i \right) \widehat{\epsilon}_{\mathcal{D}_{tr}^i}(h^t, h^{t-1}) \right\} \\
 &\quad + \frac{1}{2} \sum_{i=1}^{t-1} \beta_i d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_{tr}^i, \mathcal{D}_{tr}^t) \\
 &\quad + \sum_{i=1}^{t-1} (\alpha_i + \beta_i) \epsilon_{\mathcal{D}_{tr}^i}(h^{t-1}) + C_1 \\
 &= \underbrace{\left\{ \sum_{i=1}^{t-1} \beta_i \widehat{\epsilon}_{\mathcal{D}_{tr}^i}(h^t, h^{t-1}) \right\} + \left\{ \sum_{i=1}^{t-1} \alpha_i \widehat{\epsilon}_{\mathcal{D}_{tr}^i}(h^t, h^{t-1}) \right\}}_{\text{Model Discrepancy Terms}} \quad (6.2) \\
 &\quad + \underbrace{\left\{ \sum_{i=1}^{t-1} \gamma_i \widehat{\epsilon}_{\mathcal{D}_{tr}^i}(h^t) + \widehat{\epsilon}_{\mathcal{D}_{tr}^t}(h^t) \right\}}_{\text{ERM-like Terms}} \\
 &\quad + \underbrace{\left\{ \frac{1}{2} \sum_{i=1}^{t-1} \beta_i d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_{tr}^i, \mathcal{D}_{tr}^t) \right\}}_{\text{Data Discrepancy Term}} + C_2.
 \end{aligned}$$

where the constant C_1 can be calculated by ERM-Based generalization bound, $\gamma_i + \alpha_i + \beta_i = 1$, and the error term $\sum_{i=1}^{t-1} (\alpha_i + \beta_i) \epsilon_{\mathcal{D}_{tr}^i}(h^{t-1})$ is also a constant without any trainable parameters in the fixed (frozen) historical function h^{t-1} . The model discrepancy measures the weighted divergence between the current model h^t and the previous one h^{t-1} , and the data discrepancy measures the scaled separability between the data distributions of task i and task t .

Proof.

$$\begin{aligned}
 \epsilon_{\mathcal{D}_i}(h) &= (\gamma_i + \alpha_i + \beta_i) \epsilon_{\mathcal{D}_i}(h) \\
 &\leq \gamma_i \epsilon_{\mathcal{D}_i}(h) + \alpha_i [\epsilon_{\mathcal{D}_i}(h, H_{t-1}) + \epsilon_{\mathcal{D}_i}(H_{t-1})] \\
 &\quad + \beta_i \left[\epsilon_{\mathcal{D}_t}(h, H_{t-1}) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_i, \mathcal{D}_t) \right] \\
 &\quad + \beta_i \epsilon_{\mathcal{D}_i}(h^{t-1}), \\
 \Rightarrow \sum_{i=1}^t \epsilon_{\mathcal{D}_{tr}^i}(h^t) &\leq \left\{ \sum_{i=1}^{t-1} [\gamma_i \hat{\epsilon}_{\mathcal{D}_{tr}^i}(h^t) + \alpha_i \hat{\epsilon}_{\mathcal{D}_{tr}^i}(h^t, h^{t-1})] \right\} \\
 &\quad + \left\{ \hat{\epsilon}_{\mathcal{D}_t}(h) + \left(\sum_{i=1}^{t-1} \beta_i \right) \hat{\epsilon}_{\mathcal{D}_t}(h, H_{t-1}) \right\} \\
 &\quad + \frac{1}{2} \sum_{i=1}^{t-1} \beta_i d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_{tr}^i, \mathcal{D}_{tr}^t) \\
 &\quad + \sum_{i=1}^{t-1} (\alpha_i + \beta_i) \epsilon_{\mathcal{D}_i}(h^{t-1}) + C_1.
 \end{aligned}$$

Following [134], given each task i has a deterministic ground-truth classifier $f^i: \mathbb{R}^n \rightarrow \{0, 1\}$, the C_1 can be calculated by ERM-Based generalization bound:

$$\begin{aligned}
 \sum_{i=1}^t \epsilon_{\mathcal{D}_{tr}^i}(h^t) &\leq \sum_{i=1}^t \hat{\epsilon}_{\mathcal{D}_{tr}^i}(h^t) \\
 &\quad + \sqrt{\left(\frac{(1 + \sum_{i=1}^{t-1} \beta_i)^2}{N_t} + \sum_{i=1}^{t-1} \frac{(\gamma_i + \alpha_i)^2}{\tilde{N}_i} \right)} \\
 &\quad \times \sqrt{\left(8d \log \left(\frac{2eN}{d} \right) + 8 \log \left(\frac{2}{\delta} \right) \right)} \tag{6.3} \\
 \Rightarrow C_1 &= \sqrt{\left(\frac{(1 + \sum_{i=1}^{t-1} \beta_i)^2}{N_t} + \sum_{i=1}^{t-1} \frac{(\gamma_i + \alpha_i)^2}{\tilde{N}_i} \right)} \\
 &\quad \times \sqrt{\left(8d \log \left(\frac{2eN}{d} \right) + 8 \log \left(\frac{2}{\delta} \right) \right)}.
 \end{aligned}$$

Thus, C_1 is a constant with no trainable parameters, and the proof of Theorem 1 is concluded. $\square$

Notably, our theory is compatible with all the incremental scenarios (TIL, DIL, and CIL) and is capable of addressing all four open-world continual learning settings we discussed in the introduction. In addition to the data discrepancy term, another crucial component in Theorem 1 is the model discrepancy term, which effectively characterizes the optimization direction and extensions (or isolations) of the previously trained model from $t - 1$ to t . The model discrepancy term represents

a fundamental aspect of current CIL design, particularly in regularization-based methods and dynamic structural models. For the previous $t - 1$ tasks, the bound is given by:

$$\epsilon_{\mathcal{D}_{tr}^i}(h^t) \leq \mathbb{E}_{h \sim Q^t} [\widehat{\epsilon_{\mathcal{D}_{tr}^i}}(h)] + \sqrt{\frac{KL(Q^t \| P^{t-1}) + \log \frac{1}{\delta}}{2n_i}},$$

for $i = 1, \dots, t - 1$. Summing these bounds across all tasks, we obtain:

$$\sum_{i=1}^t \epsilon_{\mathcal{D}_{tr}^i}(h^t) \leq \sum_{i=1}^t \mathbb{E}_{h \sim Q^t} [\widehat{\epsilon_{\mathcal{D}_{tr}^i}}(h)] + \sum_{i=1}^t \sqrt{\frac{KL(Q^t \| P^{t-1}) + \log \frac{1}{\delta}}{2n_i}}.$$

By decomposing, we get:

$$\sum_{i=1}^t \epsilon_{\mathcal{D}_{tr}^i}(h^t) \leq \sum_{i=1}^t \mathbb{E}_{h \sim Q^t} [\widehat{\epsilon_{\mathcal{D}_{tr}^i}}(h)] + \sum_{i=1}^t \sqrt{\frac{KL(Q^t \| P^{t-1})}{2n_i}} + \sum_{i=1}^t \sqrt{\frac{\log \frac{1}{\delta}}{2n_i}}.$$

Here, the model discrepancy $\sum_{i=1}^t \sqrt{\frac{KL(Q^t \| P^{t-1})}{2n_i}}$ accounts for the divergence between the posterior distribution Q^t and the prior P^{t-1} , normalized by the number of training samples n_i .

6.4 Methodology

Motivated by empirical findings and theoretical analysis in the previous two sections, the interaction between open risk and incremental prediction error drives us to approach the open-world continual learning problem as an integrated whole. Hence, we design a model that effectively transfers knowledge for both known categories and unknown samples, i.e., **knowns-unknowns knowledge transfer**.

In open-world continual learning, where tasks may overlap in input space due to repeating classes or shared unknowns, we use a *projection* $P : \mathbb{R} \rightarrow \mathbb{R}'$ to map high-dimensional inputs into a lower-dimensional feature space where samples from different tasks are more easily separable. From previous theoretical analysis, we have:

Corollary 1 *Given a projection $\mathcal{P} : \mathbb{R} \rightarrow \mathbb{R}'$, an empirical estimate for the data discrepancy term is:*

$$\begin{aligned} & \frac{1}{2} \sum_{i=1}^{t-1} \beta_i d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_{tr}^i, \mathcal{D}_{tr}^t) \\ &= \frac{1}{2} \sum_{i=1}^{t-1} \beta_i \widehat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{P}(\mathcal{B}^i), \mathcal{P}(\mathcal{B}^t)) \\ &= \sum_{i=1}^{t-1} \beta_i \left(1 - \min_{d \in \mathcal{H}_d} \left[\frac{1}{\widehat{N}_i} \sum_{\mathbf{x} \in \mathcal{X}_i} [-\log([d(\mathcal{P}(\mathbf{x}))]_i)] + \frac{1}{N_t} \sum_{\mathbf{x} \in \mathcal{S}_t} [-\log([d(\mathcal{P}(\mathbf{x}))]_t)] \right] \right), \end{aligned}$$

where $\mathcal{B}^t$ and $\mathcal{B}^i$ are experience buffers drawn from training set $\mathcal{D}_{tr}^t$ and $\mathcal{D}_{tr}^i$ of tasks t and i , respectively.

The projected buffers $P(\mathcal{B}_i)$ and $P(\mathcal{B}_t)$ represent the embedded distributions of tasks i and t . The discrepancy $\hat{d}\mathcal{H}\Delta\mathcal{H}$ is estimated via a binary classifier $d \in \mathcal{H}_d$ trained to distinguish whether a sample originates from task i or task t . A smaller value of this discrepancy implies more similar task distributions in the projected space. Thus, learning a good projection P is crucial for achieving high inter-task discriminability and reliable estimation of transferability. Moreover, the choice of buffers $\mathcal{B}_i$ and $\mathcal{B}_t$ is important in representing the estimated distributions. This motivates the need for careful buffer construction and feature space design in open-world continual learning systems.

Given the most difficult scenario, KIRO, each task’s training set comes incrementally with repeated classes and changing distributions, which may result in overlapping across different tasks’ semantic spaces, thus increasing the inter-task open risk. Additionally, the repeated appearance of open samples may increase the intra-task open risk by identifying open/unknown samples as known. Therefore, it is not enough to pull together all seen/known samples belonging to the same class, but we need to (1) *find a proper projection with improved linear separability* and (2) *determine appropriate experience buffers to describe embedding distributions of different tasks*.

Next, we propose a novel open-world continual learning framework that can jointly learn knowledge about known classes and unknown samples, effectively generalizing under open-world assumptions while excelling in both open detection and incremental classification, called *HoliTrans*. It can infer general patterns from specific observed instances, which is important for adapting to new classes and detecting novel samples within all the open-world continual learning scenarios.

6.4.1 Initialization: Fine-tuning on the First Task

Given a sequence of tasks, the training data and labels from the first task may better represent the subsequent tasks than the data used to train the original pre-trained model. Hence, we first need to fine-tune the pre-trained model via a Parameter-Efficient Transfer Learning (PETL) strategy [33, 86]. Moreover, due to the advantages of the random projection layer being independent of the model, it can be applied to any feature extractor. Therefore, it can be applied orthogonally to the widely used PETL strategy, which does not alter any parameters of the original pre-trained model.

Here, we fine-tune the pre-trained model with the first task by learning PETL parameters and then freeze them for the subsequent tasks. Here, we use three PETL methods: AdaptFormer [94], SSF [135], and VPT [136]. After fine-tuning the pre-trained model with the first task, we use a novel two-stage training approach for each task, as shown in Algorithm 4.

6.4.2 The Proposed Method

Given a new task t , we extract features from its training set using the fine-tuned pre-trained model and obtain the corresponding embeddings. In the first stage training, we introduce the **nonlinear random projection** from the assumption that nonlinear random interactions between embedding distributions may be more linearly separable than the original in a higher-level embedding space, with corresponding mathematical properties that have been discussed extensively [130, 137, 138]. Thus, the embeddings can be improved as:

$$H^i = g(\phi(D_{tr}^i)W). \quad (6.4)$$

Then, we encode the labels into a one-hot format, obtaining the label matrix Y^i . Accordingly, we update the model’s Gram matrix G^i (i.e., the matrix of all pairwise inner products between projected feature vectors, capturing their similarity structure in the transformed higher-dimensional space) and class prototype matrix C^i as follows:

$$\begin{aligned} G^i &= G^{i-1} + H^{iT} H^i, \\ C^i &= C^{i-1} + H^{iT} Y. \end{aligned} \quad (6.5)$$

Given the training set of task t , each class’s prototype can be derived from Equation 6.5. Since the prototype for each known category is updated through nonlinear random projection, we refer to these prototypes as **distribution-aware prototypes**. Additionally, we compute the mean-variance of distances between samples of all classes and their prototypes:

$$\delta^i = \frac{\delta^{i-1} \sum_{j=1}^{i-1} N_j + \mathbb{E}[(H^i - H^{iT})^2]}{\sum_{j=1}^i N_j}. \quad (6.6)$$

To enable *HoliTrans* to acquire knowledge from known classes without forgetting prior tasks, we store the learned distribution-aware prototypes incrementally. In an open-world continual learning context, it is important not only to reduce incremental classification error but also to establish tighter decision boundaries for open detection in the testing phase. Therefore, we propose a novel pseudo-sample generation method that, unlike traditional replay-based continual learning approaches, which inject generated samples into the training set, enhances the test set instead. This allows us to make full use of the learned knowledge of both known and open classes, improving the model’s generalization and open-world detection capabilities.

Given a specific known class k with its distribution-aware prototype $\mathbf{p}_k$, we generate *positive pseudo-samples* by sampling from a multivariate Gaussian distribution $\mathcal{N}(\mathbf{p}_k, \delta^2 \mathbf{I})$, ensuring that the positive samples are centered around $\mathbf{p}_k$. Additionally, *negative pseudo-samples* are randomly generated from the distribution $\mathcal{N}(\mathbf{p}_{kl}, \delta^2 \mathbf{I})$,

where $\mathbf{p}_{kl} = \frac{\mathbf{p}_k + \zeta \mathbf{p}_l}{1 + \zeta}$ denotes an interpolated point between prototypes $\mathbf{p}_k$ and $\mathbf{p}_l$. These negative pseudo-samples constitute the pseudo-test set $\mathcal{D}_{te}^i$.

6.4.3 Overall Optimization and Complexity Analysis

Next, we propose a novel threshold learning approach with knowledge-adaptive capability, building on the NCM-based classifier. Our goal is to determine an appropriate threshold for each task, enabling the classifier to discern whether a test sample belongs to a known category. If it does, the NCM-based classifier aligns the sample with the corresponding label of the distribution-aware prototype and gets the output. If not, the test sample is identified as open.

Consequently, we design the following optimization problem to determine the threshold r :

$$\begin{aligned} & \arg \max_r \frac{1}{|\mathcal{D}_{te}^i|} \sum_{\mathbf{x} \in \mathcal{D}_{te}^i} \left[\mathbb{I} \left(\mathbf{s}_{\mathbf{x}} < r \frac{1}{|\mathcal{D}_{tr}^i|} \sum_{\mathbf{z} \in \mathcal{D}_{tr}^i} \mathbf{s}_{\mathbf{z}} | \text{known} \right) \right. \\ & \quad \left. + \mathbb{I} \left(\mathbf{s}_{\mathbf{x}} > r \frac{1}{|\mathcal{D}_{tr}^i|} \sum_{\mathbf{z} \in \mathcal{D}_{tr}^i} \mathbf{s}_{\mathbf{z}} | \text{open} \right) \right] \quad (6.7) \\ & \text{with } \mathbf{s}_{\mathbf{x}} = \max_{y \in \mathcal{Y}^{M_i}} g \left(\phi(\mathbf{x})^T \mathbf{W} \right) (G^i + \lambda^i I)^{-1} \mathbf{p}_y^i, \\ & \text{and } \mathbf{p}_y^i = \frac{\sum_{j=1}^{|\mathcal{X}^i|} \mathbb{I}(y_j^i = y) \phi(x_j^i)}{\sum_{j=1}^{|\mathcal{X}^i|} \mathbb{I}(y_j^i = y)}. \end{aligned}$$

Here $\mathbb{I}(\cdot)$ is an indicator function, $\mathbf{s}_{\mathbf{x}}$ is the maximum similarity between sample $\mathbf{x}$ with learned prototypes, $g(\cdot)$ is a nonlinear function, $\phi(\cdot)$ is the pre-trained model, G^i is the Gram matrix of the nonlinear random projection feature, and $\mathbf{p}_y^i$ is the prototype of class y in task i . Note the $\mathcal{D}_{te}^i$ is different from the original $\mathcal{D}_{te}$ in testing as we integrate a batch of generated pseudo-samples for each task to better represent data distribution via the embedding distribution over $\mathcal{D}_{tr}^i$, obtaining a distribution-aware prototype for each known class.

Naturally, in solving for r , we need further analysis as the function in r :

$$\frac{1}{|\mathcal{D}_{te}^i|} \sum_{\mathbf{x} \in \mathcal{D}_{te}^i} \left[\mathbb{I} \left(\mathbf{s}_{\mathbf{x}} < r \frac{1}{|\mathcal{D}_{tr}^i|} \sum_{\mathbf{z} \in \mathcal{D}_{tr}^i} \mathbf{s}_{\mathbf{z}} | \text{known} \right) + \mathbb{I} \left(\mathbf{s}_{\mathbf{x}} > r \frac{1}{|\mathcal{D}_{tr}^i|} \sum_{\mathbf{z} \in \mathcal{D}_{tr}^i} \mathbf{s}_{\mathbf{z}} | \text{open} \right) \right]$$

is *unimodal, long-tailed, and decreasing*. Consequently, in our *HoliTrans* framework, we employ the classic numerical technique of ternary search, an efficient optimization technique used to find the maximum of functions that have a single peak, particularly effective for long-tailed and monotone functions. In our case, the ternary search algorithm iteratively refines the search interval to adap-

tively identify the optimal value for the threshold r that maximizes accuracy for the NCM-based classifier.

Algorithm 4 summarizes the overall training process, which consists of two main phases. In the first phase, given a task i with its training set D_{tr}^i and test set D_{te}^i , the class prototypes G^i and $\{\mathbf{p}_k\}_{k=1}^{M^i}$ are updated using the training data, and the standard deviation δ between samples and their corresponding prototypes is calculated. In the second phase, the algorithm generates discriminative auxiliary pseudo-samples. Negative pseudo-samples are created by computing pseudo-prototypes from pairs of prototypes $(\mathbf{p}_k, \mathbf{p}_l)$ and sampling from a Gaussian distribution centered around the pseudo-prototype. Positive pseudo-samples are similarly generated from each prototype $\mathbf{p}_k$ using a Gaussian distribution. The pseudo-samples are then integrated with the test set D_{te}^i , and the decision threshold r is determined through a ternary search over the modified test set $\mathcal{D}_{te}^i$. This process enhances the classifier’s ability to distinguish between known and open classes by utilizing both positive and negative pseudo-samples.

Algorithm 4: The Training Process of *HoliTrans*

Input : Given a task i with its training set D_{tr}^i and test set D_{te}^i .

Phase 1: *NCM-based Classifier Training*

Update G^i and $\{\mathbf{p}_k\}_{k=1}^{M^i}$ by D_{tr}^i ;

Calculate standard deviation δ between samples and corresponding prototypes.

Phase 2: *Learning DAPs*

// Generate Negative Pseudo-samples

for an arbitrary pair of prototypes $(\mathbf{p}_k, \mathbf{p}_l)$ **do**

 Obtain pseudo-prototype $\mathbf{p}_{kl} = \frac{\mathbf{p}_k + \zeta \mathbf{p}_l}{1 + \zeta}$;

 Generate negative pseudo-samples via a Gaussian distribution $\mathcal{N}(\mathbf{p}_{kl}, \delta^2 \mathbf{I})$.

// Generate Positive Pseudo-samples

for each prototype $\mathbf{p}_k$ **do**

 Generate positive pseudo-samples via a Gaussian distribution $\mathcal{N}(\mathbf{p}_k, \delta^2 \mathbf{I})$.

Integrating all pseudo-samples with D_{te}^i ;

Determining r by ternary search over $\mathcal{D}_{te}^i$.

Complexity Analysis Following the design of our algorithm, we now provide an analysis of its computational complexity. The overall cost can be decomposed into three main components. First, the complexity of generating pseudo-samples via class-wise sampling and projection is $O(n^2 \times \text{dim})$, where n denotes the number of classes and dim is the dimension of the output features from the pre-trained backbone. Second, evaluating pseudo-samples within the nonlinear random projection space has a complexity $O(\text{dim} \times M \times n)$, where M represents the number of projections used in the nonlinear random projection module. This step is typically the most time-consuming in practice, as it involves computing the similarity between high-dimensional embeddings across all class pairs. Finally, the optimization of the DAP radius parameter r requires solving a ternary search over a bounded interval,

leading to a worst-case complexity of $O\left(\log_3\left(\frac{\max_{\mathbf{z} \in \mathcal{D}_{tr}^i} \mathbf{s}_{\mathbf{z}} - \min_{\mathbf{z} \in \mathcal{D}_{tr}^i} \mathbf{s}_{\mathbf{z}}}{\frac{1}{|\mathcal{D}_{tr}^i|} \sum_{\mathbf{z} \in \mathcal{D}_{tr}^i} \mathbf{s}_{\mathbf{z}}} / \epsilon\right)\right)$, where ϵ is a pre-defined convergence threshold, and $\mathbf{s}_{\mathbf{z}}$ denotes the similarity score of sample $\mathbf{z}$. This step ensures that the learned prototype radius adapts to the intrinsic distribution of training samples within each class.

6.4.4 Why NRP performs effectively in Knowledge Transfer for OWCL

Building on the algorithmic design and complexity analysis, we now present theoretical justifications for the effectiveness of nonlinear random projection in open-world continual learning, aiming to provide insights into its generalization and adaptability.

First, a nonlinear random projection approximates each class distribution as a Gaussian distribution. Given a vector f , the norm of its projected vector can be probabilistically bounded using the *Chernoff bound*, a concentration inequality that provides exponentially decreasing bounds on the tail probabilities of the sum of random variables, ensuring that deviations from the expected norm are highly unlikely.

$$\mathbb{P}(\|W^T f\| - \mathbb{E}_W[\|W^T f\|] > \epsilon \sigma^2) \leq 2e^{-\frac{\epsilon^2 \sigma^2}{2M + \epsilon}}.$$

This inequality bounds the probability that the norm of the projected vector $\|W^T f\|$ deviates from its expected value by more than $\epsilon \sigma^2$, where W is the random projection matrix, ϵ is a positive scalar that controls the deviation threshold, σ is the standard deviation of the original distribution, and M is the dimensionality of the projection space. The bound shows that, for fixed ϵ and σ , the deviation probability decays exponentially with increasing M . In other words, as the projection dimension M increases, the norm of the projected vectors becomes increasingly concentrated around its expected value, reinforcing the Gaussian approximation and making the projected representations more regular and predictable in high-dimensional space.

Secondly, nonlinear random projection can lead to better decision boundaries. Consider any two vectors f and f' , and analyze the Chernoff bound for their inner product before and after projection:

$$\mathbb{P}(|(W^T f)^T (W^T f') - \mathbb{E}_W[(W^T f)^T (W^T f')]| > \epsilon' M \sigma^2) \leq 2e^{-\frac{(\epsilon' M)^2 \sigma^2}{2M + \epsilon' M}},$$

where we made the tolerance level $\epsilon' M \sigma^2$ depending on both a given positive threshold ϵ' and the projection dimension M . Under the assumption that the entries of W are independent and identically distributed with zero mean and variance σ^2 , we have that

$$\mathbb{E}_W[(W^T f)^T (W^T f')] = \mathbb{E}_W[f^T W W^T f'] = f^T \mathbb{E}_W[W W^T] f' = f^T M \sigma^2 f' = M \sigma^2 f^T f'.$$

Our initial Chernoff bound can thus be rewritten as:

$$\mathbb{P}(|(W^T f)^T (W^T f') - M\sigma^2 f^T f'| > \epsilon' M\sigma^2) \leq 2e^{-\frac{(\epsilon' M)^2 \sigma^2}{2M + \epsilon' M}}.$$

Simplifying further, we get:

$$\mathbb{P}\left(\left|\frac{(W^T f)^T (W^T f')}{M\sigma^2} - f^T f'\right| > \epsilon'\right) \leq 2e^{-\frac{M\epsilon'^2 \sigma^2}{2 + \epsilon'}} \leq 2e^{-\frac{\epsilon'^2 \sigma^2}{2M + \epsilon'}}.$$

This bound shows that as the projection dimensionality M increases, the inner products of arbitrary vectors and their projections are less likely to be equal. In other words, higher dimensions reduce the probability that two vectors and their projections share the same inner product. This property can be used to establish better decision boundaries for open-world continual learning.

6.5 Experiments

To validate our theoretical findings and assess the practical effectiveness of *HoliTrans*, we execute some experiments across diverse open-world continual learning scenarios. These experiments are designed to evaluate both the incremental classification performance and the ability to detect unknown samples. In particular, we benchmark *HoliTrans* against state-of-the-art methods.

6.5.1 Experimental Settings

Baselines We compare *HoliTrans* with **22** baseline models, including open-world continual learning methods and various continual learning methods with several out-of-distribution (OOD) algorithms. We consider **5** state-of-the-art continual learning methods that use large-scale pre-training and prompt-tuning as well. Below is a summary of each baseline model:

- **L2P** [32]: This method uses prompt-based fine-tuning to adapt pre-trained models to continual learning tasks efficiently, leveraging large-scale pre-trained knowledge.
- **DualPrompt** [83]: This approach introduces dual prompt tuning, focusing on both task-specific and task-agnostic prompts to enhance generalization in continual learning scenarios.
- **CODA** [88]: CODA leverages contrastive learning and dynamic prompts to improve performance across continual learning tasks by preserving learned knowledge while adapting to new tasks.
- **RanPAC** [130]: A memory-efficient continual learning method that uses randomized partial attention to capture relevant task-specific knowledge while mitigating catastrophic forgetting.

- **ADAM** [131]: ADAM revisits the attention-based learning paradigm in continual learning, proposing an architecture that better preserves previous task knowledge through dynamic attention mechanisms.
- **OpenMax** [15]: This OOD detection method modifies the final layer of deep networks to identify unknown classes by adjusting softmax outputs to better handle open-set recognition.
- **MaxLogits** [101]: MaxLogits utilizes the maximum logit score to identify out-of-distribution samples, offering a simple and scalable OOD detection technique for continual learning.
- **Entropy** [74]: This method employs the entropy of the output distribution as a metric to distinguish between in-distribution and out-of-distribution data for OOD detection.
- **EnergyBased** [63]: Energy-based OOD detection uses the energy score derived from the output of the neural network to separate in-distribution samples from OOD instances, improving recognition of unseen classes.
- **MORE** [19]: MORE is an open-world continual learning method that introduces memory-efficient regularization techniques to better handle task transitions while preventing forgetting.
- **Pro-KT** [30]: Pro-KT is our prototype-based continual learning method we introduced in Chapter 4.4. It facilitates knowledge transfer across tasks, utilizing prototypes to manage open-world scenarios effectively.

Datasets and Scenario Setups As mentioned earlier, CIRO and KIRO are the two most challenging open-world continual learning scenarios. For the CIRO scenario, we divide the **Split-CIFAR100** dataset [98] into 10 tasks (10 classes for each task) and conduct testing using the entire testing set of all 10 tasks. For the KIRO scenario, we use the object recognition dataset, **Open-CORe50** [139]. In this dataset, new training samples of the same classes become available in a sequence of tasks with changing distributions (e.g., new conditions and backgrounds). Additionally, we specifically designate a certain class as open during testing.

As shown in Table 6.3, in the CIFAR100 dataset in the CIRO scenario, there are 90 known classes (M_{known}) and 10 unknown classes (M_{open}), with a total of 9 tasks and 55,000 samples. Each training epoch consists of 5000 samples (N_t), where $t = 1, \dots, T$. The validation set comprises 9000 samples from known classes ($N_{val,known}$) and 1000 samples from unknown classes ($N_{val,open}$).

In the CORe50 dataset in the KIRO scenario, there are 25 known classes and 25 unknown classes, with 4 tasks and a total of 59,936 samples. The number of samples in each training epoch is as follows: $N_1 = 14,989$, $N_2 = 14,986$, $N_3 = 14,995$, and

$N_4 = 14,966$. The validation set comprises 7528 samples from known classes and 7466 samples from unknown classes.

Table 6.3: Dataset description.

Dataset	M_{known}	M_{open}	T	N	$N_t, t = 1, \dots, T$	$N_{val,known}$	$N_{val,open}$
CIFAR100	90	10	9	55000	$N_t = 5000, t = 1, \dots, T$	9000	1000
CORe50	25	25	4	59936	$N_1 = 14989$ $N_2 = 14986$ $N_3 = 14995$ $N_4 = 14966$	7528	7466

Implementations *HoliTrans* implementation is based on PyTorch. All experiments are conducted on a single NVIDIA RTX 3090 GPU. The utilized pre-trained backbone is ViT-B/16, which is self-supervised and pre-trained on ImageNet-21K. In the testing phase of each task, we treat the test samples of classes that have not yet been learned as open/unknown classes when executing the CIRO and KIRO scenarios.

For the first stage of *HoliTrans*, we use SGD to train the parameters of the PETL method, namely AdaptFormer, SSF, and VPT. For each of these, we use a batch size of 48, a learning rate of 0.01, weight decay of 0.0005, momentum of 0.9, and cosine annealing with restarts ending at a learning rate of 0. Throughout, we typically train for 20 epochs, although in some experiments, this may be reduced if overfitting occurs. Softmax and cross-entropy loss are utilized when employing these methods. The number of classes equals the quantity from the first task, i.e., M_1 . Before commencing the second stage of *HoliTrans*, we discard the training weights and heads generated thereby.

During training, data augmentation for all datasets includes random resizing followed by cropping to 224x224 pixels and random horizontal flipping. For inference, images are resized to a short side of 256 pixels and then center-cropped to 224x224 for all datasets except CIFAR100, which is directly resized from its original 32x32 size to 224x224.

Metrics We employ three key metrics, namely ACC_t , AUC_t , and FPR_t , to provide an evaluation of model performance across T tasks. Specifically, ACC_t represents the average final accuracy concerning all previously encountered classes across the t tasks. It evaluates how well the model retains knowledge from earlier tasks without significant forgetting. AUC_t denotes the average area under the receiver operating characteristic (ROC) curve over all past tasks, reflecting the model’s ability to distinguish between known and unknown instances, thereby offering insights into its classification reliability. FPR_t , or the average false positive rate, quantifies the

error rate in open-set detection, indicating how often the model mistakenly classifies unknown instances as belonging to known categories.

6.5.2 Experimental Results and Analysis

Main Results Our evaluation approach distinguishes itself from prior works by addressing open-set detection and the classification of known samples simultaneously, rather than treating these tasks separately with distinct metrics. To ensure fairness and robust assessment, we use three random seeds with shuffled task orders, averaging the results and reporting the standard deviations. Table Table 6.4 presents the results on the CIRO and KIRO benchmark scenarios, where the proposed method is compared against 22 state-of-the-art approaches.

Table 6.4: Main Results. We evidence first, second, and third performances.

<i>Method</i>	Scenario: CIRO (Split-CIFAR100)			Scenario: KIRO (Open-CORE50)		
	$ACC_t(\uparrow)$	$AUC_t(\uparrow)$	$FPR_t(\downarrow)$	$ACC_t(\uparrow)$	$AUC_t(\uparrow)$	$FPR_t(\downarrow)$
<i>Continual learning methods equipped with out-of-distribution algorithms</i>						
L2P	OpenMax	0.436 $\pm$ 0.048	0.133 $\pm$ 0.051	0.937 $\pm$ 0.046	0.424 $\pm$ 0.063	0.244 $\pm$ 0.050
	MaxLogits	0.445 $\pm$ 0.041	0.103 $\pm$ 0.047	0.931 $\pm$ 0.045	0.412 $\pm$ 0.179	0.259 $\pm$ 0.053
	Entropy	0.445 $\pm$ 0.028	0.120 $\pm$ 0.038	0.932 $\pm$ 0.041	0.424 $\pm$ 0.063	0.257 $\pm$ 0.063
	EnergyBased	0.445 $\pm$ 0.038	0.102 $\pm$ 0.043	0.925 $\pm$ 0.040	0.425 $\pm$ 0.065	0.256 $\pm$ 0.057
DualPrompt	OpenMax	0.423 $\pm$ 0.031	0.139 $\pm$ 0.037	0.928 $\pm$ 0.045	0.402 $\pm$ 0.029	0.270 $\pm$ 0.024
	MaxLogits	0.425 $\pm$ 0.034	0.114 $\pm$ 0.039	0.933 $\pm$ 0.042	0.408 $\pm$ 0.113	0.261 $\pm$ 0.017
	Entropy	0.425 $\pm$ 0.045	0.204 $\pm$ 0.032	0.936 $\pm$ 0.048	0.408 $\pm$ 0.112	0.267 $\pm$ 0.026
	EnergyBased	0.425 $\pm$ 0.046	0.114 $\pm$ 0.049	0.942 $\pm$ 0.039	0.408 $\pm$ 0.112	0.260 $\pm$ 0.030
CODA	OpenMax	0.439 $\pm$ 0.036	0.125 $\pm$ 0.042	0.936 $\pm$ 0.046	0.413 $\pm$ 0.032	0.278 $\pm$ 0.045
	MaxLogits	0.439 $\pm$ 0.047	0.090 $\pm$ 0.039	0.942 $\pm$ 0.045	0.415 $\pm$ 0.033	0.183 $\pm$ 0.028
	Entropy	0.439 $\pm$ 0.036	0.106 $\pm$ 0.046	0.939 $\pm$ 0.047	0.415 $\pm$ 0.033	0.256 $\pm$ 0.056
	EnergyBased	0.439 $\pm$ 0.031	0.089 $\pm$ 0.053	0.946 $\pm$ 0.037	0.415 $\pm$ 0.033	0.169 $\pm$ 0.022
ADAM	OpenMax	0.440 $\pm$ 0.044	0.112 $\pm$ 0.040	0.876 $\pm$ 0.043	0.433 $\pm$ 0.025	0.283 $\pm$ 0.038
	MaxLogits	0.440 $\pm$ 0.042	0.099 $\pm$ 0.053	0.882 $\pm$ 0.033	0.433 $\pm$ 0.026	0.238 $\pm$ 0.029
	Entropy	0.440 $\pm$ 0.042	0.221 $\pm$ 0.043	0.876 $\pm$ 0.047	0.433 $\pm$ 0.025	0.197 $\pm$ 0.016
	EnergyBased	0.440 $\pm$ 0.035	0.386 $\pm$ 0.042	0.869 $\pm$ 0.041	0.433 $\pm$ 0.027	0.345 $\pm$ 0.023
RanPAC	OpenMax	0.467 $\pm$ 0.045	0.091 $\pm$ 0.044	0.879 $\pm$ 0.054	0.472 $\pm$ 0.022	0.201 $\pm$ 0.020
	MaxLogits	0.468 $\pm$ 0.042	0.092 $\pm$ 0.042	0.878 $\pm$ 0.039	0.472 $\pm$ 0.022	0.212 $\pm$ 0.017
	Entropy	0.468 $\pm$ 0.049	0.091 $\pm$ 0.041	0.871 $\pm$ 0.031	0.472 $\pm$ 0.022	0.219 $\pm$ 0.021
	EnergyBased	0.468 $\pm$ 0.046	0.227 $\pm$ 0.038	0.869 $\pm$ 0.041	0.472 $\pm$ 0.021	0.411 $\pm$ 0.025
<i>Open-world continual learning methods</i>						
MORE	0.716 $\pm$ 0.011	0.717 $\pm$ 0.127	0.492 $\pm$ 0.011	0.641 $\pm$ 0.077	0.641 $\pm$ 0.077	0.517 $\pm$ 0.011
Pro-KT	0.779 $\pm$ 0.010	0.745 $\pm$ 0.011	0.397 $\pm$ 0.026	0.635 $\pm$ 0.011	0.675 $\pm$ 0.125	0.451 $\pm$ 0.015
HoliTrans	0.851$\pm$0.010	0.858$\pm$0.012	0.193$\pm$0.024	0.748$\pm$0.010	0.748$\pm$0.010	0.097$\pm$0.031

Firstly, a comparison of the first 20 rows and the last 3 rows reveals that simply combining continual learning and out-of-distribution detection methods fails to achieve satisfactory performance under open-world continual learning settings. This underscores the necessity of addressing open-world continual learning as a unified problem, a central premise of our approach. Secondly, our *HoliTrans* method significantly outperforms all baseline models in both ACC_t and AUC_t , achieving a remarkable 7% improvement over the best existing methods. This establishes *HoliTrans*'s exceptional capability to enhance incremental classification accuracy and quality across diverse tasks. Thirdly, the proposed *HoliTrans* demonstrates outstanding performance in FPR_t , highlighting its superior ability to mitigate open-set

risks—a critical challenge in open-world continual learning problems. This performance advantage ensures that the model remains robust when encountering unseen or anomalous data.

Table 6.5 presents the performance of RanPAC equipped with the classic Local Outlier Factor (LOF) [140], a density-based unsupervised learning algorithm that detects outliers by comparing the local density of a data point with that of its neighbors. The different ratios indicate the proportion of replay samples utilized during training.

Table 6.5: Main Results. We format first, second, and third performances.

<i>Method</i>		Scenario: CIRO (Split-CIFAR100)			Scenario: KIRO (Open-CORe50)		
		$ACC_t(\uparrow)$	$AUC_t(\uparrow)$	$FPR_t(\downarrow)$	$ACC_t(\uparrow)$	$AUC_t(\uparrow)$	$FPR_t(\downarrow)$
RanPAC + LOF	100%	0.582 ± 0.009	0.575 ± 0.003	0.795 ± 0.007	0.505 ± 0.005	0.502 ± 0.005	0.459 ± 0.004
	75%	0.584 ± 0.004	0.565 ± 0.005	0.778 ± 0.002	0.501 ± 0.001	0.502 ± 0.004	0.466 ± 0.013
	50%	0.589 ± 0.009	0.564 ± 0.004	0.699 ± 0.045	0.504 ± 0.002	0.501 ± 0.002	0.478 ± 0.004
	25%	0.583 ± 0.015	0.566 ± 0.014	0.642 ± 0.049	0.509 ± 0.002	0.509 ± 0.003	0.467 ± 0.007
<i>HoliTrans</i>		0.851 ± 0.010	0.858 ± 0.012	0.193 ± 0.024	0.748 ± 0.010	0.748 ± 0.010	0.097 ± 0.031

Although RanPAC combined with LOF represents one of the current state-of-the-art methods in continual learning, its performance remains substantially inferior to the proposed *HoliTrans* under the open-world continual learning setting. Specifically, even with full memory replay (100%), RanPAC + LOF achieves only 0.582 in ACC_t and 0.575 in AUC_t on the **CIRO** scenario, and performs even worse in the **KIRO** scenario with $ACC_t = 0.505$ and $AUC_t = 0.502$. In stark contrast, *HoliTrans* significantly outperforms RanPAC across all metrics, achieving improvements of over 26.3% in ACC_t and 28.3% in AUC_t on **CIRO**, while simultaneously reducing FPR_t by more than 0.60. These consistent and substantial gains across different data scenarios and evaluation metrics demonstrate the superiority of *HoliTrans* in handling open-world class-incremental learning tasks. This further validates the effectiveness of our proposed architecture and its ability to generalize across diverse and challenging continual learning benchmarks.

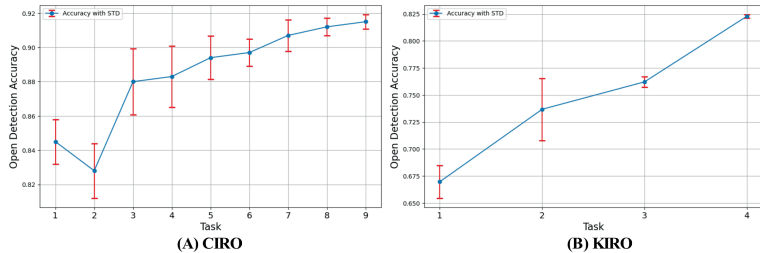


Figure 6.7: Visualizations of main results.

As visualized in Figure 6.7, the experimental results indicate that, for both CIRO and KIRO scenarios, the accuracy of open set detection shows an increasing trend

6. EXPLORING OPEN-WORLD CONTINUAL LEARNING WITH KNOWN-UNKNOWN KNOWLEDGE TRANSFER

as the incremental process progresses with new tasks appearing. The standard deviation of accuracy across different random seeds significantly decreases, suggesting that the proposed *HoliTrans* can incrementally learn the changing distributions of training data from new tasks, thereby accumulating knowledge, reducing open set risk, and thus demonstrating robustness.

In conclusion, across all evaluated metrics, *HoliTrans* consistently surpasses all baselines, setting a new benchmark for performance in all scenarios. From the practical point of view, our methodology bridges the gap between continual learning and out-of-distribution techniques, providing a unified framework to tackle real-world open-world continual learning challenges.

Ablation Study To underscore the significance of the key components in the *HoliTrans* framework, we conduct ablation studies, focusing on the impact of adaptively learning the optimal r through ternary search and the contribution of all distribution-aware prototypes. The results are detailed in Table 6.6.

Table 6.6: Ablation Study.

<i>Ablation Component</i>	Scenario: CIRO (Split-CIFAR100)			Scenario: KIRO (Open-CORe50)		
	$ACC_t(\uparrow)$	$AUC_t(\uparrow)$	$FPR_t(\downarrow)$	$ACC_t(\uparrow)$	$AUC_t(\uparrow)$	$FPR_t(\downarrow)$
w/o ternary search	$0.842_{\pm 0.010}$	$0.850_{\pm 0.012}$	$0.202_{\pm 0.030}$	$0.609_{\pm 0.020}$	$0.609_{\pm 0.020}$	$0.055_{\pm 0.029}$
w/o DAPs	$0.468_{\pm 0.042}$	$0.092_{\pm 0.010}$	$0.878_{\pm 0.039}$	$0.472_{\pm 0.022}$	$0.212_{\pm 0.017}$	$0.963_{\pm 0.015}$
<i>HoliTrans</i>	$0.851_{\pm 0.010}$	$0.859_{\pm 0.012}$	$0.193_{\pm 0.024}$	$0.748_{\pm 0.010}$	$0.748_{\pm 0.010}$	$0.054_{\pm 0.031}$

First, we examined the effect of replacing the ternary search strategy for adaptively determining r with the traditional equal interval search approach. As shown in Table 6.6, this substitution led to a noticeable decline in performance across all evaluation metrics, with the most pronounced degradation observed in the challenging KIRO scenario. This outcome highlights the importance of dynamic learning r , especially in scenarios characterized by complex and evolving distributional shifts across tasks.

Second, we assessed the contribution of distribution-aware prototypes by entirely removing them, thereby reverting the model to a baseline open-world continual learning framework employing a MaxLogits-like strategy. The absence of distribution-aware prototypes resulted in a marked reduction in performance, emphasizing their critical role in mitigating open risk and enhancing the quality of incremental classification. The inclusion of distribution-aware prototypes enables the model to address open-set challenges and maintain robustness in diverse and incrementally evolving environments, as evidenced by the substantial performance gains reported in Table 6.6.

These results demonstrate that both the adaptive learning of r and the integration of distribution-aware prototypes are indispensable for achieving superior performance in open-world continual learning tasks.

Generalization and Robustness We evaluate the impact of the size of nonlinear random projection (i.e., M) and report the final averaged performance across all tasks on different scenarios.

Table 6.7: Performance on Different NRP Size for CIRO (on Split-CIFAR100 dataset) and KIRO (on Open-CORe50 dataset).

Value of M	CIRO (Split-CIFAR100)			KIRO (Open-CORe50)		
	$ACC_t(\uparrow)$	$AUC_t(\uparrow)$	$FPR_t(\downarrow)$	$ACC_t(\uparrow)$	$AUC_t(\uparrow)$	$FPR_t(\downarrow)$
10000	0.851	0.825	0.142	0.748	0.748	0.097
5000	0.859	0.825	0.137	0.754	0.753	0.055
2500	0.863	0.877	0.124	0.744	0.744	0.063
1250	0.856	0.876	0.104	0.744	0.744	0.051
800	0.852	0.873	0.085	0.737	0.737	0.041
400	0.833	0.860	0.059	0.756	0.756	0.048

As shown in Table 6.7 and Figure 6.8, our *HoliTrans* remains robust across different values of M from 400 to 10000. Notably, a larger M for the random projection size does not necessarily lead to better performance: *HoliTrans* achieves better performance when M is between 1000 and 2000, which is a relatively small dimension compared to previous works and shows the adaptability of nonlinear random projection on the open-world continual learning problems.

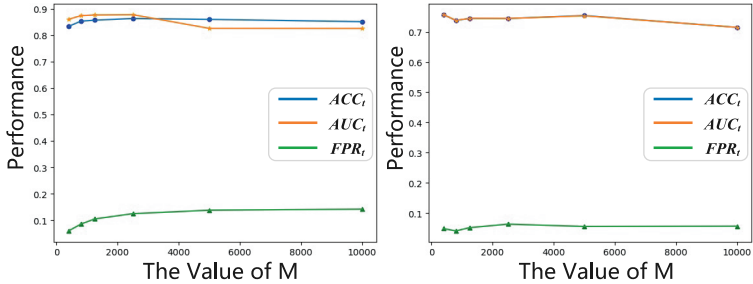


Figure 6.8: Performance with varying NRP size.

Additionally, we compare the actual embedding distribution with the embedding distribution learned by *HoliTrans*. As shown in Figure 6.9, *HoliTrans* increasingly aligns with the actual embedding distribution as new tasks come, with decreasing deviation. Such results highlight the potential of *HoliTrans* to adapt to new tasks and accumulate knowledge.

Furthermore, to verify the impact of the experience buffer $\mathcal{B}$, we replace our distribution-aware prototypes with a KNN-contrastive method [128] that replays embeddings in different sizes. The results illustrate that our *HoliTrans* with distribution-aware prototypes uses less memory and performs better in all open-world continual learning scenarios.

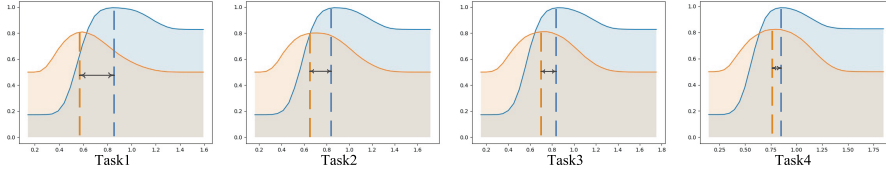


Figure 6.9: The divergence between the actual embedding distributions and the *HoliTrans*’s embedding distributions with distribution-aware prototypes (The horizontal axis is the values of r , the vertical axis is ACC_t , and the dashed line indicates the optimal value of r). Orange represents the actual ones, blue represents ours.

6.6 Summary

Despite the increasing interest in open-world continual learning research, existing approaches have merely focused on learning and transferring knowledge from known samples, while overlooking the knowledge derived from unknown samples. In this chapter, our empirical findings not only challenge the current assumption but also underscore the importance of addressing the open-world continual learning problem as a holistic paradigm. Following this, we introduce a novel open-world continual learning approach, *HoliTrans*, that supports the knowns-unknowns knowledge transfer and validates the proposed model through experiments.

The key insight of our work lies in discovering the interaction between open risk and incremental prediction error, viewing open-world continual learning as a holistic problem rather than a mere combination of continual learning and open-set recognition methods. Additionally, we cast open-world continual learning into four scenarios with increasing difficulty and conduct experiments to investigate the challenges of open-world continual learning. More importantly, our theoretical analysis gives guidance for our open-world continual learning framework. Specifically, we introduce the nonlinear random projection method, which provides nonlinear random feature interactions to enhance linear separability in a higher-dimensional space, and propose distribution-aware prototypes for seen classes to mitigate open risk and enhance incremental classification performance. Experimental results demonstrate the superiority of *HoliTrans* over 22 baseline models across various open-world continual learning scenarios and datasets, further underscoring its capability to effectively learn and transfer knowledge from both known and unknown samples.

While the theoretical results and the empirical validation underscore the superiority of *HoliTrans*, certain limitations need future consideration: 1) The practical implementation of theoretical results may encounter challenges due to the few-shot training and noise in real-world data, potentially affecting the algorithm’s performance. 2) The generalizability of the experimental findings might be constrained by task similarity, requiring further validation across diverse and complex situations that consider task similarity. 3) Experimental results show that using prototypes with

lower memory usage can achieve better performance than replaying a large number of samples in experience buffers. This phenomenon also needs further theoretical investigation.

Chapter 7

Cross-regional Fraud Detection via Continual Learning with Knowledge Transfer

While the preceding chapters have focused on advancing open-world continual learning through both theoretical insights and practical algorithms, the fundamental goal of **robust knowledge transfer**—from known to unknown classes, and from past to future data distributions—remains central to real-world applications. One such application domain is *fraud detection*, where the ability to adaptively transfer and retain knowledge across evolving environments is important for maintaining detection accuracy and operational efficiency.

In this chapter, we explore the challenge of **cross-regional fraud detection**, an applied setting that inherently demands effective knowledge transfer and continual adaptation. Traditional fraud detection systems are typically trained on static datasets from single regions, making them ill-suited for generalization to new cities or countries. When businesses expand geographically, retraining a model from scratch is not only computationally expensive but also risks *catastrophic forgetting* of previously learned fraud patterns.

To address these challenges, we formulate cross-regional fraud detection as an **incremental learning problem**, where the objective is to develop a unified model that incrementally learns from region-specific data while preserving past knowledge. We propose a novel paradigm called **Cross-Regional Continual Learning (CCL)**, which explicitly incorporates *knowledge transfer mechanisms* to support both forward adaptation and backward retention. Specifically, CCL employs *prototype-based knowledge replay* to distill transferable representations and utilizes a *parameter smoothing strategy* to mitigate forgetting during continual updates. Moreover, we construct **heterogeneous trade graphs (HTG)** to represent complex transaction structures and leverage graph-based backbones to enhance semantic understanding and facilitate transfer across regions with different fraud profiles.

Through extensive experiments, we demonstrate that CCL consistently outperforms baseline approaches in cross-regional fraud detection tasks. This application-driven chapter thus not only illustrates the practical utility of continual learning techniques, but also confirms the broader **transferability of knowledge-centric learning frameworks** developed throughout this thesis. Together, these insights lay a foundation for *adaptive, scalable fraud detection systems in dynamic, geographically diverse environments*.

This chapter is based on the following publications [39, 40]:

- **Li, Y.**, Yang, Y., Gao, Q., & Yang, X: Cross-Regional Fraud Detection via Continual Learning (Special Program). In: Proceedings of the AAAI Conference on Artificial Intelligence, 37(13): 16260-16261, 2023.
- **Li, Y.**, Yang, X., Gao, Q., Wang, H., Zhang, J., & Li, T.: Cross-Regional Fraud Detection via Continual Learning With Knowledge Transfer. In: IEEE Transactions on Knowledge and Data Engineering, 38(12): 7865-7877, 2024.

To contextualize and motivate our proposed framework, the following section outlines the core motivations, identifies the fundamental challenges in cross-regional fraud detection, and summarizes the key contributions of this chapter.

7.1 Continual Learning for Cross-Regional Fraud Detection

In recent years, the widespread application of Internet finance has become increasingly prevalent, concurrently leading to substantial financial losses due to fraudulent activities. Thus, it is important to effectively detect fraud for risk management, and there are substantial research efforts have been dedicated to this endeavor. Although existing fraud detection models can achieve promising performance, most solutions simply employ expert-driven rules or machine learning methods, using historical transaction data to discover potential anomalous behaviors [141]. Existing methods heavily depend on prior knowledge and fail to adapt dynamically to evolving data, particularly when financial operations expand into new regions. Furthermore, traditional methods suffer from the catastrophic forgetting issue, whereby previously acquired knowledge cannot be retained during the model’s evolutionary process.

In financial contexts, due to considerations of data privacy and security, data from different regions is not allowed to be shared. Therefore, we cannot combine data from different regions to train a unified fraud detection model. Moreover, training from scratch for each new region will result in forgetting previously acquired knowledge and increase computational costs rapidly. In addition, fraudsters usually adapt quickly to anti-fraud measures and detection methods [142], resulting in difficulty in capturing fraudulent behavioral patterns. As shown in Figure 7.1, when a financial

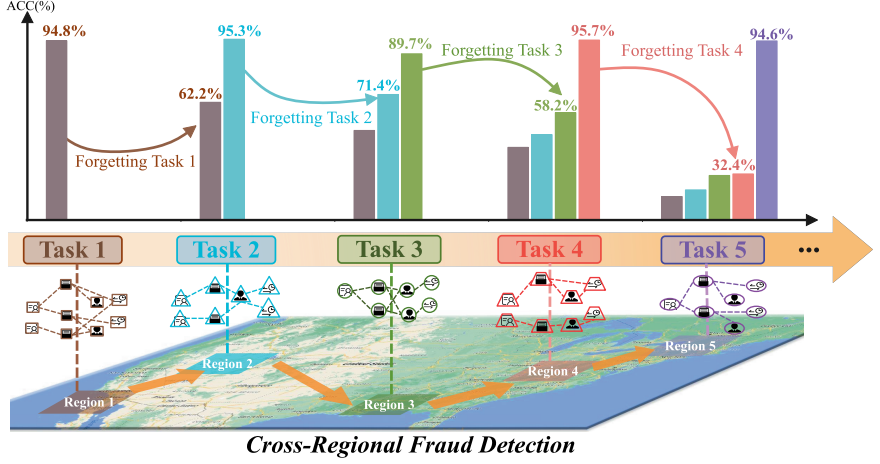


Figure 7.1: Bottom: An illustration of cross-regional fraud detection. The orange arrows indicate a financial business expansion from Region 1 to Region 5, where the data characteristics vary across different regions with different patterns. Our proposed method aims to incrementally learn a unified fraud detection model that can perform well in each task through knowledge transfer. Top: The histograms in different colors denote each corresponding region’s fraud detection performance (ACC) with training from scratch. The performance drops significantly after a model is trained using data from Region 2. Similarly, after training a model with data from Region 3, the performance of Region 1 and Region 2 fell further, respectively.

institution extends business to new regions, new data will appear incrementally, and the data of each region has its own characteristics. The fraud detection model observes cross-regional data sequentially as they arrive, thus any processing or learning must be done in a continual fashion. Hence, we formulate a new problem, **cross-regional fraud detection**, which aims to continually detect fraudulent activities in the process of business expansion to new regions with streaming data. The proposed problem stems from practical scenarios and sheds new light on integrating fraud detection with continual learning.

The cross-regional fraud detection problem is based on a real-world scenario that combines continual learning with knowledge transfer and poses two primary challenges. (1) *Regional Uniqueness*: The inherent geographical distances introduce differences between different regions. Due to the unique characteristics and region-specific knowledge of each region, the fraud detection model must adapt to region-specific knowledge in a continual learning framework. (2) *Regional Generality*: Fraud detection adheres to general knowledge applicable across all regions. Hence, the cross-regional fraud detection model should extract and transfer knowledge from a sequence of regions to preserve regional generality.

Our proposed cross-regional fraud detection problem requires both general knowledge with regional generality and specific knowledge with regional uniqueness. In response, we propose a novel **Cross-regional Continual Learning (CCL)** framework to address the previously mentioned challenges in a unified manner. CCL presents a novel plug-and-play prototype-based replay method for knowledge transfer and improves the regularization-based parameter smoothing strategy to alleviate forgetting. Specifically, to address challenge (1), we propose a prototype-based replay method to preserve the previously learned knowledge from formerly appeared regions with an incremental experience buffer. To deal with challenge (2), the proposed CCL alleviates the forgetting issue by utilizing an improved regularization-based parameter smoothing method to constrain the training process and parameter updating. Furthermore, we introduce a novel heterogeneous trade graph construction method to uncover hidden patterns of fraudulent activities and explore deeper semantic information to facilitate knowledge transfer.

Building upon the foundational insights and techniques developed in previous chapters, here we focus on extending continual learning principles to a real-world application: *cross-regional fraud detection*. In practical financial systems, models must operate across geographically diverse regions, each with unique behavioral patterns, data distributions, and entity relationships. This dynamic setting introduces a variety of challenges, including those we have seen before, such as distribution shift, catastrophic forgetting, and the need for efficient knowledge transfer across domains.

To address these issues, we propose a novel continual learning framework tailored for fraud detection tasks involving sequential regional adaptation. Our approach not only preserves previously learned knowledge but also facilitates forward transfer of useful patterns, enabling the model to maintain high detection accuracy in new regions without retraining from scratch. This is achieved through a unified integration of semantic graph modeling and continual learning techniques.

The remainder of this chapter presents the problem formulation, methodological design, and empirical validation of our approach. Below, we summarize the core contributions of this chapter:

- **Problem Formulation.** We introduce and formally define a novel task—*cross-regional fraud detection*—that integrates the dynamics of regional adaptation with the principles of continual learning. This formulation reflects real-world deployment scenarios where models must adapt to new geographic markets without retraining from scratch. To the best of our knowledge, this is the first work to investigate fraud detection in a continual learning setting, bridging a critical gap between static fraud detection systems and dynamic, evolving environments.

- **Graph-based Knowledge Representation.** To capture the diverse and intricate relationships among financial entities, we construct a *heterogeneous trade graph*. This graph structure encodes semantic information and models complex interdependencies between users, transactions, and institutions across regions, serving as a strong relational backbone for downstream learning and knowledge transfer.
- **Cross-Regional Knowledge Transfer Framework.** We propose a novel learning framework termed **Cross-Regional Continual Learning**, which enables adaptive and efficient transfer of knowledge across regions. CCL is built upon two complementary mechanisms: (1) a *prototype-based replay* module that preserves representative knowledge from previous regions to guide learning in new ones; and (2) an enhanced *parameter smoothing regularization* strategy that mitigates catastrophic forgetting by stabilizing model updates during sequential training.
- **Empirical Validation on Real-World Data.** We conduct experiments on a large-scale, real-world financial fraud dataset to validate the proposed method. Results show that CCL consistently outperforms state-of-the-art baselines across multiple evaluation tasks involving regional shifts. Additional ablation studies and qualitative analyses (e.g., representation visualization) further demonstrate the effectiveness and robustness of the proposed knowledge transfer mechanisms in mitigating catastrophic forgetting and enhancing generalization.

The remainder of this chapter is organized as follows. We review the recent studies of fraud detection and the advances of continual learning in section 7.2. In section 7.3, we first formalize the cross-regional fraud detection problem with fundamental definitions. Then, we introduce each component of the proposed CCL, involving graph construction, prototype-based knowledge replay, and regularization-based parameter smoothing. Subsequently, the experiments and results quantifying the proposed model are provided in section 7.4. Finally, in section 7.5, we conclude this work and remark on future works.

7.2 A Brief Recall on Fraud Detection

Fraud detection has always constituted an important task in risk management, demonstrating its efficacy in thwarting deleterious activities [143]. In previous works, rule-based methods predominantly relied on expert-crafted rules, presupposing that fraudulent activities have some apparent patterns. Across recent decades, the use of Deep Neural Networks (DNNs) has become ubiquitous in the field of fraud detection [144]. By refining informative representations from raw data, deep neural networks demonstrated their capabilities in the domain of fraud detection [145]. Several investigations harnessed temporal cues latent in sequential data through diverse iterations

of recurrent neural networks [146, 147]. Further explorations manifested in DNN-based methodologies [148], retrieving implicit features from data through feature engineering. [149] investigated fraud detection in the unsupervised framework and applied a long short-term memory mechanism. [142, 150] developed unsupervised schemes for sequential data in online anomaly detection. Auto-encoders, recognized for their ability to distill underlying patterns, have also found applications in fraud detection [145].

However, the aforementioned methods all have their shortcomings. Powered by rudimentary machine learning algorithms, the rule-based approaches exhibited a substantial reliance on expert experience [151], while addressing intricate and dynamic patterns. Furthermore, the current array of deep learning models still faces challenges in handling complex topological data with intricate relationships between various entities, posing an ongoing obstacle to fraud detection.

Recently, graph neural networks (GNNs) have emerged as a robust, reliable, and promising technique for financial fraud detection. Notable examples include GCN [152], GraphSAGE [153], and GAT [154], which utilize convolutional methods, the LSTM architecture, and self-attention mechanisms, respectively, to capture and aggregate complex information among nodes [155]. HAN [156] proposed a hierarchical attention aggregation that integrated node-level and semantic-level information from multiple meta-paths. In [157, 158], different types of objects were considered for building heterogeneous information networks. [159] proposed a graph construction method for GNN-based fraud detection on the non-attributed graph using a graph pre-training strategy. [160, 161] modeled dynamic connectivity patterns through bipartite graph embedding and then detected anomalies from the graph. [162] presented a hierarchical prototype network to represent the continuously expanded graphs. CARE-GNN [163] is a GNN-based fraud detector based on a multi-relation graph that enhances its performance against camouflaged fraudsters.

However, traditional GNN-based methods only deal with static tasks and cannot be directly applied to streaming data for fraud detection [164]. Recent works combined graph neural networks with continual learning to deal with streaming graphs and dynamic graphs [165, 166, 167, 168], while ignoring the knowledge extraction, memorization, and transferring during the process of learning new tasks. In contrast, our work here extends GNN-based approaches to the cross-regional fraud detection task with dynamic data, integrating with improved continual learning techniques. Given the power of providing neighborhood and fusing structure attributes in graph neural networks, we take a variant of GAT [154] as our backbone. Specifically, we generate multiple types of meta-paths and entities to capture intrinsic patterns, consider heterogeneous structures, and explore temporal-spatial semantics for cross-regional fraud detection. Different from previous works, the key insight of our approach is in extending GNN-based fraud detection within the CCL scenarios.

7.3 Methodology

In this section, we begin by presenting an overview of our proposed cross-regional fraud detection framework. Subsequently, we introduce the foundational concepts and notations that will be employed consistently throughout this paper. We also identify the research challenges in cross-regional fraud detection. Following this, we outline the procedure for constructing heterogeneous trade graphs for data engineering and provide a brief introduction to the Graph Neural Network backbone utilized in this study. Then, we investigate the technical particulars of the main continual learning components proposed. Finally, we conclude this section by detailing the algorithm with explanations.

7.3.1 Overview and Problem Formulation

The proposed framework for cross-regional fraud detection is designed to extract, retain, and transfer knowledge acquired from a sequence of region-specific learning tasks, thereby ensuring robust and consistent performance across all regions. As illustrated in Figure 7.3, the core of the framework—termed **Cross-Regional Continual Learning**—builds upon two foundational components: *heterogeneous graph construction* and *continual knowledge transfer*.

Specifically, the framework begins by transforming raw structured transaction data into **Heterogeneous Trade Graphs**, which encode multi-typed entities and their complex interactions, capturing rich relational and semantic information critical for fraud detection. These graphs serve as dynamic knowledge representations that facilitate both region-specific modeling and transferable insights across regions.

When faced with a sequence of cross-regional tasks, the model must incorporate knowledge from previously encountered regions while adapting to new ones. To this end, we introduce a novel **prototype-based knowledge replay** mechanism that maintains a buffer of representative samples from past tasks. This mechanism enables the model to selectively replay informative instances during the training of a new task l , thereby preserving prior knowledge and guiding the learning process with meaningful context. In addition, to address the issue of catastrophic forgetting, we integrate a parameter-level regularization strategy that constrains the update of crucial model parameters. This approach stabilizes learning across tasks and ensures that valuable knowledge is not overwritten during sequential adaptation. At the architectural level, CCL provides a powerful **graph-based representation learning backbone**, enhanced with attention mechanisms, to generate expressive and task-relevant embeddings from the constructed heterogeneous trade graphs. These embeddings are then passed through a binary classifier that predicts the likelihood of fraudulent activity at the transaction level.

For clarity and consistency, all notations used throughout the chapter are summarized in Table 7.1, along with detailed descriptions of their meanings and roles in the proposed framework.

Table 7.1: Notations and explanations.

Notation	Explanation
$\mathcal{D}_l$	The raw structured dataset of region l
$\mathcal{G}_l$	The heterogeneous trade graph of region l
$\mathcal{V}_l$	Node set of $\mathcal{G}_l$
$\mathcal{E}_l$	Edge set of $\mathcal{G}_l$
$\mathbf{X}_l$	Feature set of $\mathcal{G}_l$
$\mathcal{P}_m$	A meta-path m
$N_i^{\mathcal{P}_m}$	Neighbors of node i on Meta-path m
$\mathbf{W}$	A transformation matrix
$\mathbf{h}^i$	Initialized hidden representation of node i
$\mathbf{e}_{(i,j)}^{\mathcal{P}_m}$	The attention coefficient between node pair (i, j)
$\alpha_{(i,j)}^{\mathcal{P}_m}$	The normalized attention between node pair (i, j)
$\mathbf{z}_i^{\mathcal{P}_m}$	Meta-path based representation of target node i
$\mathbf{Z}_{\mathcal{P}_m}$	Semantic-specific node embedding of meta-path $\mathcal{P}$
$\beta_{\mathcal{P}_m}$	The importance of meta-path $\mathcal{P}_m$
$\mathbf{Z}$	The final representation
$\mathbb{B}_l$	The experience buffer collected from $\mathcal{D}_l$
$\mathbb{F}$	The Fisher Information matrix

The basic goal of cross-regional fraud detection is to continually detect fraudulent activities in the process of business expansion with a sequence of cross-regional data, which can be precisely formulated as follows:

Definition 13 Problem definition (Cross-Regional Fraud Detection). *Given a sequence of cross-regional datasets from regions $(\mathcal{R}_1, \dots, \mathcal{R}_l, \dots)$, the series of cross-regional datasets can be formulated as $(\mathcal{D}_1, \dots, \mathcal{D}_l, \dots)$. As the fraud detection task expands into the l -the region, $\mathcal{D}_l = (\mathcal{D}_l^r, \mathcal{D}_l^{te})$, we aim to update a unified neural function Φ using the dataset $\mathcal{D}_l^r$. The unified neural function $\Phi : \mathcal{D}_l^{te} \rightarrow \mathbb{R}^d$ maps input data to a low-dimensional space d . Subsequently, a dense layer is employed to identify whether a given sample is fraudulent.*

There are two important objectives: (1) upholding promising efficacy in fraud detection across all regions, and (2) improving the fraud detection performance within the present regional data through the incorporation of previously acquired knowledge.

7.3.2 Construct Heterogeneous Trade Graphs

Data acquisition. The business expansion strategies employed by financial institutions exemplify a geographically aware approach. Cross-regional tasks involve extending operations beyond the initial locale into diverse geographical areas. Financial institutions typically initiate a pilot phase within a specific city before expanding

their operations to other regions. Therefore, it is important to ensure the absence of data overlap among distinct regions during the data acquisition process.

In this work, based on a desensitized financial database that contains more than one million credit card transaction records in the United States Mainland from the duration of 1st Jan 2019 - 31st Dec 2020 ¹, we extract five regions' transaction data according to different geographical ranges: $\mathcal{R}_1$ ($30^\circ N - 40^\circ N, 95^\circ W - 100^\circ W$), $\mathcal{R}_2$ ($40^\circ N - 50^\circ N, 75^\circ W - 80^\circ W$), $\mathcal{R}_3$ ($30^\circ N - 40^\circ N, 75^\circ W - 80^\circ W$), $\mathcal{R}_4$ ($40^\circ N - 50^\circ N, 95^\circ W - 100^\circ W$), $\mathcal{R}_5$ ($30^\circ N - 40^\circ N, 90^\circ W - 95^\circ W$).

It is noteworthy that the raw data is structured and tabular, with each region being distinct and non-overlapping. The sequence of tasks is subjected to random shuffling to mitigate order effects. Simultaneously, to alleviate the issues posed by missing values in financial data and the sparsity of fraudulent data, we implement data imputation and data augmentation strategies to preprocess the raw financial datasets. Note that while the use of the above five regions serves an illustrative purpose, our proposed framework is designed to accommodate the integration of a greater number of regions in practical applications.

A solitary fraudulent incident can encompass numerous transactions, implicating multiple cardholders and/or merchants. This intricate interplay gives rise to a convoluted network of interactions that underlie each fraudulent occurrence [141]. Within this framework, the areas of feature engineering encompass the extraction of nodes and the exploration of relationships. To encapsulate both inherent and network-oriented attributes of credit card transactions, we introduce a graph-generation procedure that converts organized credit card transaction data into a topological graph.

Hence, each financial heterogeneous trade graph G_l stem from region $\mathcal{R}_l$ can be defined as below:

Definition 14 Let $\mathcal{G}_l = (\mathcal{V}_l, \mathcal{E}_l, \mathbf{X}_l)$ denote a financial heterogeneous trade graph extracted from l -th region's transaction data $\mathcal{D}_l$, where $\mathcal{V}_l$ is the set of nodes, $\mathcal{E}_l$ is the edge set and $\mathbf{X}_l$ is the initial feature set. $\mathcal{G}_l$ is associated with a node type mapping function $\phi : \mathcal{V}_l \rightarrow \mathcal{A}$ and an edge type mapping function $\psi : \mathcal{E}_l \rightarrow \mathcal{O}$, where the $\mathcal{A}$ and $\mathcal{O}$ respectively represent the set of node types and the set of edge types of $\mathcal{G}_l$. Notably, we have $|\mathcal{A}| + |\mathcal{O}| > 2$ since $\mathcal{G}_l$ is a heterogeneous graph, where nodes and edges are of different types.

Node Extraction. The main objective of our work is to detect whether a credit card transaction is fraudulent. Therefore, each transaction is represented as a node identified by its transaction ID. Additionally, cardholders and merchants play crucial roles in transactions, so they are also represented as nodes. In addition, [148] observed that fraudsters' activities are typically intensive and significantly different

¹<https://kaggle.com/kartik2112/fraud-detection>

from those of legitimate transactions. As shown in Figure 7.2, the fraudulent transactions and legitimate transactions exhibit markedly different behavioral patterns: the frequency of fraudulent transactions changes abruptly over time, whereas legitimate transactions remain significantly more stable, indicating the attackers operate within limited time frames for their campaigns [169]. Therefore, we need to consider the varying behavioral changes due to different transaction time intervals in a fraud detection system. To address this, we introduced time nodes into the heterogeneous trade graph. Our datasets include timestamped credit card transactions, logging the exact time of each transaction. To manage the scale of timestamps, we divided the time into 24 slices, with each time node representing one hour (e.g., [0:00, 1:00] and [1:00, 2:00]). Consequently, our heterogeneous graph G_l includes four types of nodes: cardholders C_l , merchants M_l , time slices S_l , and transactions T_l . Thus, $\mathcal{V}_l = \{T_l, C_l, M_l, S_l\}$, with $|\mathcal{A}_l| = 4$.

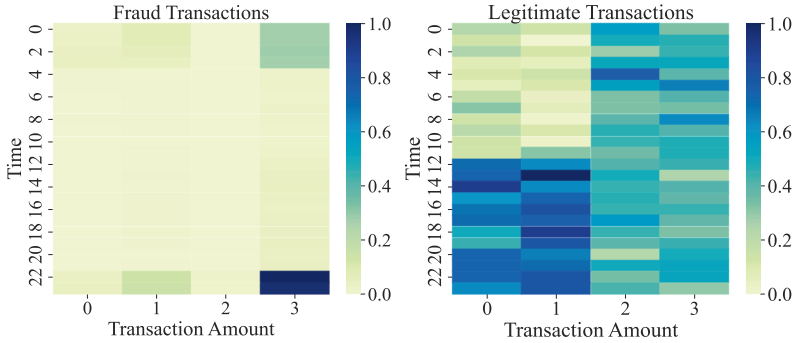


Figure 7.2: Heat maps of temporal slices. On the horizontal axis is the amount of fraudulent transactions, and on the vertical axis are time slices. The darker the color, the more fraudulent transactions occurred during the current period.

Relationship Exploration. To explore informative interactions among different types of nodes, meta-paths are widely used in graph generation. Meta-paths are basic logical routes that aim at connecting nodes with a composite relationship [170, 171]. Such a paradigm motivates us to generate meaningful meta-paths to explore the diverse semantic relationships among heterogeneous entities. Formally, the definition of a meta-path $\mathcal{P}$ is described as follows:

Definition 15 A meta-path [156] is a path $\mathcal{P}$ in form of $\mathcal{V}^1 \xrightarrow{R_1} \mathcal{V}^2 \xrightarrow{R_2} \dots \xrightarrow{R_k} \mathcal{V}^{k+1}$, where $\mathcal{V}^1$ to $\mathcal{V}^{k+1}$ are $(k+1)$ different types of nodes. The relations between node types $\mathcal{V}^1$ to $\mathcal{V}^{k+1}$ are denoted as a composite relation $R = R_1 \circ R_2 \circ \dots \circ R_k$, where the $\circ$ is the composition operator.

Given the four types of nodes $\mathcal{V}_l = \{T_l, C_l, M_l, S_l\}$, we generate three types of meta-paths: **TCT** (Transaction-Cardholder-Transaction), **TMT** (Transaction-Merchant-Transaction) and **TST** (Transaction-Time Slices-Transaction), respectively. To bet-

ter target transaction nodes, meta-paths are constructed through the underlying interactional logic between transaction nodes and the other three types of nodes. *TCT* depicts the connections between different transaction nodes that are initiated by the same cardholder. *TMT* describes the interactions of different transaction nodes that take place at the same merchant. *TST* implies that there are associations among transaction nodes that are launched in the same time interval. Finally, by defining the heterogeneous meta-paths, we complete the construction of the network structure for heterogeneous trade graphs.

Node features. Beyond nodes and the network structure, there are informative features with respect to transactions, such as transaction amount and transaction type. Additionally, the dataset contains numerous desensitization features that depict the inherent semantics of transactions, which could boost the downstream fraud detection task. Thus, we assign all these features to transaction nodes as corresponding feature vectors, i.e., each transaction node $v_l^i \in \mathcal{V}_l$ is associated with a feature vector $\mathbf{x}_l^i$. Hence, our generated financial HGTs are not only depicted as heterogeneous graphs with multiple types of nodes and relationships but also associated with different feature spaces.

7.3.3 The Proposed CCL Model

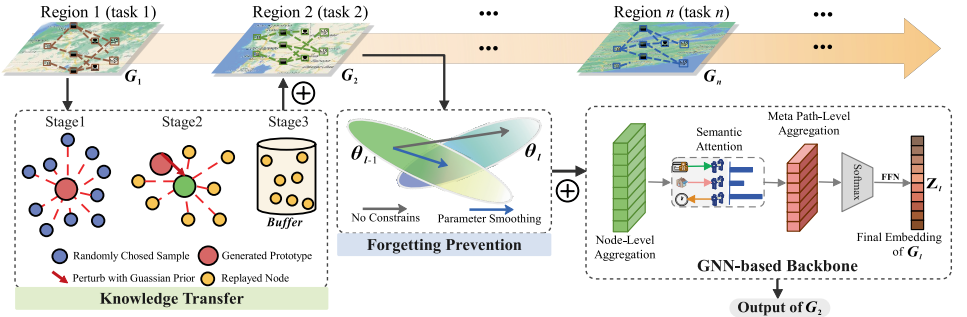


Figure 7.3: The overall framework of the proposed CCL. Given a sequence of cross-regional fraud detection tasks, the proposed prototype-based replay method for knowledge transfer and the forgetting prevention approach are illustrated. Take the learning process of G_2 as an example. In stage 1 of knowledge transfer, we randomly choose some target nodes from G_1 and generate a prototype $\mathbf{c}^1$. In stage 2, a Gaussian prior is added to perturb the prototype. In stage 3, nodes are sorted according to the distance from the perturbed prototype $\mathbf{c}^1$, and we select the replay nodes according to the replay ratio r . Then, all the selected nodes and their neighbors are stored in the experience buffer $\mathbb{B}_1$. Next, the replay buffer is appended to G_2 for the training, with parameter smoothing for forgetting prevention. Finally, the final embedding of Z_l is obtained by the GNN-based backbone accordingly.

Given a previously learned neural model Φ_{l-1} with parameters Θ_{l-1} , a naive application of Φ_{l-1} to data from a new region l often results in significant performance

degradation due to the *region-specific distribution shift*. Conversely, training a new model solely on region l 's data risks neglecting the valuable generalizable knowledge learned from earlier regions, and more critically, leads to catastrophic forgetting in the cross-regional continual learning process.

To address this trade-off between adaptation and retention, our objective is to construct a model that can not only acquire region-specific knowledge from the current region l , but also *retain and consolidate the experiences* accumulated across previously encountered regions $1, 2, \dots, l-1$. Importantly, this is achieved *without* retraining from scratch or deploying multiple region-specific models, thereby enabling a unified and memory-efficient continual learning process.

To this end, we introduce two complementary yet computationally efficient continual learning strategies:

- **Prototype-based Knowledge Replay.** Inspired by recent advances in continual learning [166], we maintain a compact experience buffer $\mathbb{B}_{l-1}$ that stores prototypical representations from previously encountered regions. During the learning of the new task, the model is co-trained on both the current heterogeneous trade graph $\mathcal{G}_l$ and the buffered samples in $\mathbb{B}_{l-1}$, enabling the model to revisit and reinforce previously acquired knowledge.
- **Parameter Regularization for Knowledge Retention.** In addition to data replay, we enforce consistency between the current model Φ_l and the preceding model Φ_{l-1} by applying a regularization-based constraint on critical parameters. This strategy preserves important weights associated with prior knowledge and guides the optimization process to prevent abrupt drift in the learned representations.

The technical implementation of these strategies, including buffer construction, replay sampling, and parameter regularization formulation, is detailed in the following subsections.

7.3.3.1 Prototype-Based Knowledge Replay

To mitigate catastrophic forgetting while maintaining a compact memory footprint, we propose a novel *prototype-based knowledge replay* mechanism for cross-regional continual learning. This strategy enables the current task model to selectively incorporate representative knowledge from previously learned regions without accessing the full historical data.

Given the heterogeneous trade graph $\mathcal{G}_{l-1}$ constructed from the $(l-1)$ -th region's data $\mathcal{D}_{l-1}$, we build a replay buffer $\mathbb{B}_{l-1}$ that stores a small subset of informative samples, with $|\mathbb{B}_{l-1}| \ll |\mathcal{D}_{l-1}|$. Unlike traditional methods that rely on random sampling or naive memory storage of original nodes, our approach generates **pseudo-samples**

guided by prototypes, which effectively summarize the semantic distribution of prior tasks while enhancing replay efficiency and generalization.

As illustrated in Stage 1–3 of Figure 7.3, the construction of $\mathbb{B}_{l-1}$ follows three sequential steps:

Stage 1: Prototype Construction. We randomly sample a set of labeled transaction nodes from $\mathcal{G}_{l-1}$ according to a predefined replay ratio r . Let each node be associated with an attribute vector $\mathbf{x}_i^{l-1}$. The prototype vector $\mathbf{c}^{l-1}$ is computed as the centroid of all selected feature vectors:

$$\mathbf{c}^{l-1} = \frac{1}{|\mathbb{B}_{l-1}|} \sum_{\mathbf{x}_i^{l-1} \in \mathbb{B}_{l-1}} \mathbf{x}_i^{l-1}. \quad (7.1)$$

This prototype acts as a compressed representation of the semantic center for region $(l-1)$ ’s fraud distribution.

Stage 2: Gaussian Perturbation. To improve the diversity and robustness of replayed samples, we perturb the prototype vector $\mathbf{c}^{l-1}$ with a Gaussian before obtaining a modified prototype $\mathbf{c}'^{l-1}$. This perturbation simulates realistic distributional variations and reduces overfitting to fixed replay centers:

$$\mathbf{c}'^{l-1} = \mathbf{c}^{l-1} + \mathcal{N}_0^{\sigma^2 \mathbf{I}}.$$

Stage 3: Distance-Based Sample Selection. Next, we compute the Euclidean distance between each transaction node in $\mathcal{G}_{l-1}$ and the perturbed prototype $\mathbf{c}'^{l-1}$. Nodes are ranked in ascending order of distance, and the top- r nodes are selected as replay candidates. To enhance contextual consistency, we also include their immediate neighbors in the final buffer:

$$\mathbb{B}_{l-1} = \text{Top-}r \left(\{ \mathbf{x}_i^{l-1} \in \mathcal{G}_{l-1} \mid \| \mathbf{x}_i^{l-1} - \mathbf{c}'^{l-1} \|_2 \} \right) \cup \mathcal{N}(\mathbf{x}_i^{l-1}),$$

where $\mathcal{N}(\cdot)$ denotes the set of neighboring nodes.

Security-Driven Replay Constraints. Considering the sensitive nature of financial fraud detection, we constrain the replay buffer to contain only labeled transaction nodes, avoiding the risk of propagating noise or unverifiable patterns from unlabeled data. This security-aware sampling improves interpretability and aligns with practical deployment requirements.

Integration into CCL. The resulting experience buffer $\mathbb{B}_{l-1}$ is appended to the current task graph $\mathcal{G}_l$ during training. This replayed knowledge is jointly optimized alongside region-specific data, enabling effective transfer and retention across tasks.

As shown in the left part of Figure 7.3, this mechanism plays a central role in the knowledge transfer pipeline of the proposed CCL framework.

7.3.3.2 Regularization-based Parameter Smoothing

Although the first strategy, to some extent, is capable of alleviating the forgetting issue from the sample consolidation aspect, it is unable to thoroughly familiarize itself with what knowledge is extracted from the previous regions. To further alleviate the knowledge forgetting, we are inspired by recent Elastic Weight Consolidation (EWC) [21], and use Fisher Information to evaluate the importance of each parameter in Θ_l w.r.t. Φ_l , which can be defined as:

$$\mathbb{F}_l = \frac{1}{|\mathbf{X}^{l-1}|} \sum_{x \in \mathbf{X}^{l-1}} \left[g(x; \Theta_l) \cdot g(x; \Theta_{l-1})^\top \right], \quad (7.2)$$

where g is the first-order derivative of the loss.

Hence, when the current task with the new heterogeneous trade graph $\mathcal{G}_l$ arrives, we add a smoothing term to constrain the parameter optimization based on $\mathbb{F}_l$, which can be formulated as:

$$\mathcal{L}_1 = \frac{\lambda}{2} \sum \mathbb{F}_l (\Theta_l(i) - \Theta_{l-1}(i))^2, \quad (7.3)$$

where λ is a hand-crafted weight.

Although financial transactions are usually region-related, fraud patterns have some common characteristics. For instance, we can observe that fraudulent transactions have high time correlations as shown in Figure 7.2. Thus, it is necessary to consider task similarity in the financial context. Inspired by multi-task learning [172], we add a regularization term to measure the parameter distance. Therefore, we can rewrite the Equation 7.3 as:

$$\mathcal{L}_2 = \frac{\lambda}{2} \sum \mathbb{F}_l (\Theta_l(i) - \Theta_{l-1}(i))^2 + \gamma (\|\Theta_l\|_2, \|\Theta_{l-1}\|_2)_1. \quad (7.4)$$

where the second term is the $l_{2,1}$ -norm and γ is a hand-craft weight.

In the end, we can summarize our objective for the region $l + 1$, that is:

$$\min_{\Theta_{l+1}} \{\mathcal{L}_1(\Theta_{l+1}) + \mathcal{L}_2(\Theta_{l+1})\}. \quad (7.5)$$

Herein, the loss function throughout this work is the cross-entropy loss defined as:

$$\mathcal{L}(\hat{y}, y) = -\frac{1}{N} \sum_{i=1}^N \left(y_i^l \log \hat{y}_i^l + (1 - y_i^l) \log \hat{y}_i^l \right), \quad (7.6)$$

where N is the amount of transaction nodes, y_i^l is the true label of node i , $\hat{y}_i^l$ denotes the prediction of node i .

7.3.4 Graph Representation Learning Backbone

Upon acquiring Heterogeneous Trade Graphs, the subsequent step involves procuring representations through a suitable GNN-based backbone. This preparation is vital for enabling cross-regional fraud detection through continual learning approaches.

As different types of nodes hold different semantic spaces, we first transfer each target node into a new unified semantic space. Given a node $\mathbf{x}_l^i$, it is the initial representation of a target node i in $\mathcal{G}_l$, and we define a learnable transformation matrix $\mathbf{W}_x$ to map the initial $\mathbf{x}_l^i$ to a new feature space:

$$\mathbf{h}_l^i = \mathbf{W}_x \cdot \mathbf{x}_l^i, \quad (7.7)$$

where $\mathbf{h}_l^i$ is the transformed representation of node i in a new semantic space. Notably, we omit l in the following parts for simplicity.

After projecting all the initial representations into a new latent space, we then leverage a node-oriented attention mechanism to automatically learn the weights among the target node and its neighbors. Following the recent attention mechanism in [154], the attention between a node pair (i, j) is the importance of the target node i for its meta-path-based neighbor node j . Given a pair of node (i, j) which connected via a meta-path $\mathcal{P}_m$, the attention coefficient $\mathbf{e}_{(i,j)}^{\mathcal{P}_m}$ can be formulated as follows:

$$\mathbf{e}_{(i,j)}^{\mathcal{P}_m} = \text{att}_\eta(\mathbf{h}^i, \mathbf{h}^j; \mathcal{P}_m), \quad (7.8)$$

where the $\mathbf{e}_{(i,j)}^{\mathcal{P}_m}$ is asymmetric and the att_η is an attention mechanism through the deep neural networks parameterized by η . Then, we perform normalization to get the weight coefficient $\alpha_{(i,j)}^{\mathcal{P}_m}$ via meta-path $\mathcal{P}_m$ using Softmax function:

$$\left[\alpha_{(i,j)}^{\mathcal{P}_m}\right] = \text{softmax}\left(\left[\mathbf{e}_{(i,j)}^{\mathcal{P}_m}\right]\right) = \frac{\exp\left(\left[\mathbf{e}_{(i,j)}^{\mathcal{P}_m}\right]\right)}{\sum_{k \in N_i^{\mathcal{P}_m}} \exp\left(\left[\mathbf{e}_{(i,k)}^{\mathcal{P}_m}\right]\right)}, \quad (7.9)$$

where $N_i^{\mathcal{P}_m}$ is the set of meta-path $\mathcal{P}_m$ based neighbors of target node i . Hence, the meta-path based representation of target node i , $\mathbf{z}_i^{\mathcal{P}_m}$, can be obtained from its neighbors with corresponding attention coefficients:

$$\mathbf{z}_i^{\mathcal{P}_m} = \text{Sigmoid}\left(\sum_{j \in N_i^{\mathcal{P}_m}} \alpha_{(i,j)}^{\mathcal{P}_m} \cdot \mathbf{h}^j\right). \quad (7.10)$$

Inspired by [156], we enhance the fusion of multiple semantics within a heteroge-

neous trade graph by incorporating various types of meta-paths. This approach allows us to assess the importance of different meta-paths and generate final representations. Once the meta-path-based representation $\mathbf{z}_i^{\mathcal{P}_m}$ of node i is obtained, the learned attention coefficient weights for each meta-path $\mathcal{P}_m$ can be formulated as follows:

$$\mathbf{w}^{\mathcal{P}_m} = \frac{1}{|\mathcal{V}|} \sum_{i \in \mathcal{V}} \mathbf{q}^T \cdot \tanh(\mathbf{W}' \cdot \mathbf{z}_i^{\mathcal{P}_m} + \mathbf{b}), \quad (7.11)$$

where $\mathbf{q}^T$ and $\mathbf{W}'$ are a learnable weight matrix, $\mathbf{b}$ is learnable bias, $\tanh$ refers to the hyperbolic tangent activation function. Notably, the average of all semantic-specific node embeddings can be interpreted as the importance of each meta-path.

After obtaining the importance of each meta-path, we normalize them via the Softmax function:

$$[\beta_{\mathcal{P}_m}] = \text{Softmax}(\mathbf{w}^{\mathcal{P}_m}) = \frac{\exp([\beta_{\mathcal{P}_m}])}{\sum_{k=1}^M \exp([\beta_{\mathcal{P}_k}])}, \quad (7.12)$$

where M is the total amount of all meta-paths, and the learned weight can be interpreted as the importance of each meta-path. Then, we finally obtain the representation $\mathbf{Z}$:

$$\mathbf{Z} = \sum_{i=1}^M \beta_{\mathcal{P}_m} \cdot \mathbf{Z}_{\mathcal{P}_m}. \quad (7.13)$$

By aggregating all semantic-specific embeddings, the final representation is formed and can be applied to cross-regional fraud detection. Subsequently, our attention shifts towards the exploration of the two continuous learning strategies that we have put forth for cross-regional fraud detection.

The time complexity of our algorithm accounts for the computational demands across multiple stages. For each of the n heterogeneous trade graphs, the process unfolds over each epoch. Within each epoch, operations are performed for N nodes and M meta-paths. Specifically, for each meta-path, the algorithm involves a projection of node features, extraction of meta-path-based neighbors, calculation of weight coefficients, and aggregation of node embeddings. These operations are conducted for an average of $N_{\mathcal{P}_m}$ neighbors per node, considering the complexity of d_h hidden representations. Additionally, the computational load includes back-propagation steps for parameter updates. Considering these factors, the overall time complexity is given by $O(N \cdot M \cdot (N_{\mathcal{P}_m} + d_h))$. This captures the intricate computations required for continual learning in the context of cross-regional fraud detection.

7.3.5 Overall Framework and Algorithm

We now present the complete procedural workflow of the proposed CCL framework, as outlined in Algorithm 5.

Given a sequence of heterogeneous trade graphs, the algorithm begins by randomly initializing the model parameters (Step 1) and preparing an empty experience buffer

(Step 2). For the first task—corresponding to the initial graph $\mathcal{G}_1$ —the model undergoes standard graph representation learning from Steps 3 to 22. At the end of this stage (Step 22), we obtain the optimal model parameters Θ_1^* .

Algorithm 5: The overall process of CCL.

Input: A sequence of heterogeneous trade graphs $\mathcal{G} = \{\mathcal{G}_1, \dots, \mathcal{G}_l, \dots, \mathcal{G}_n\}$.

Random initialize parameters Θ_1 of CCL;

Set $\mathbb{B} = \emptyset$;

if $l = 1$ **then**

for each epoch do

for a specific meta-path $\mathcal{P}_m$ do

 Projection $\mathbf{h} \leftarrow \mathbf{W}_x \cdot \mathbf{x}$;

for each node i do

 Extract meta-path based neighbors $N_i^{\mathcal{P}_m}$;

for $j \in N_i^{\mathcal{P}_m}$ do

 Calculate the weight coefficient $\alpha_{(i,j)}^{\mathcal{P}_m}$;

 Calculate the single type node embedding

$\mathbf{z}_i^{\mathcal{P}_m} \leftarrow \text{Sigmoid}(\sum_{j \in N_i^{\mathcal{P}_m}} \alpha_{(i,j)}^{\mathcal{P}_m} \cdot \mathbf{h}_j)$;

 Concatenate from all attention heads;

 Obtain the importance of each meta-path $\beta_{\mathcal{P}_m}$;

 Obtain the final representation by aggregating $\mathbf{Z} = \sum_{i=1}^M \beta_{\mathcal{P}_m} \cdot \mathbf{Z}_{\mathcal{P}_m}$;

 Infer node label via $\hat{y}^1 = f_1(\mathbf{Z}; \epsilon)$;

 Calculate Cross Entropy Loss via Equation 7.6;

 Back-propagation and update parameters;

 Obtain optimal Θ_1^* ;

for $l = 2, \dots, n - 1$ **do**

 Let $\Theta_l = \Theta_{l-1}^*$;

 Sample a experience buffer $\mathbb{B}_{l-1}$ from G_{l-1} for experience replay;

 Generate a twin experience buffer $\mathbb{B}'_{l-1}$ with Gaussian prior;

 Add the experience buffer $\mathbb{B}'_{l-1}$ to the current heterogeneous trade graphs $\mathcal{G}_l$;

for each epoch do

 Calculate graph representation and infer node label as steps 5 to 18;

 Calculate Loss via Equation 7.5;

 Back-propagation and update parameters;

 Obtain Optimal Θ_l^* ;

Output: $\{\Theta_n^*\}$.

Starting from the second graph $\mathcal{G}_2$, the continual learning procedure follows the pipeline defined in Steps 24 to 35. Specifically, Step 25 initializes the current model parameters Θ_l using the optimal parameters from the previous task, Θ_{l-1}^* . Steps 26 to 28 construct a prototype-based experience buffer by drawing samples from $\mathcal{G}_{l-1}$ and enhancing them via a Gaussian prior, thus facilitating cross-task knowledge transfer. The buffer is then integrated into the current task's data.

Subsequently, from Steps 29 to 33 (reusing the structure from earlier), the model performs representation learning on $\mathcal{G}_l$ while leveraging regularization-based smoothing to mitigate forgetting. The training concludes with the derivation of the updated model Θ_l^* for the current task. This process iterates across all graphs in the sequence, enabling CCL to accumulate and adapt knowledge over time.

7.4 Experiments

In this section, we present the experimental evaluation of the proposed CCL framework. The primary objective is to assess its effectiveness in addressing the challenges of continual learning in the context of cross-regional financial fraud detection. To this end, we conduct comparisons with a range of state-of-the-art continual learning baselines and perform ablation and sensitivity analyses.

Our experiments are designed to answer the following key questions specific to the CCL framework:

- **Q1:** Does the proposed CCL framework outperform state-of-the-art continual learning baselines on the cross-regional credit card fraud detection task?
- **Q2:** How do the core components of CCL—namely, prototype-based knowledge replay and parameter regularization—contribute to overall performance?
- **Q3:** How sensitive is the performance of CCL to different hyperparameter configurations (e.g., replay ratio, regularization weight)?
- **Q4:** Do the constructed Heterogeneous Trade Graphs provide interpretable insights relevant to financial fraud detection?

The remainder of this section details the experimental setup, evaluation metrics, and findings addressing the key questions posed above regarding the performance and behavior of the CCL framework.

7.4.1 Experimental Settings

7.4.1.1 Implementation Details and Datasets

We reproduce the baselines and implement our proposed model with the PyTorch library, accelerated by one NVIDIA RTX 3090 GPU. We use Adam as the optimizer with a learning rate of 0.05 and a weight decay of 0.001 as the default for comparison. We train 200 epochs for each region with early stopping to accelerate the training process, and randomly sample 15% of each region’s transactions for our experience replay buffer, λ is 1.5, and γ is 0.00025.

Table 7.2 summarizes the key statistics of the five regional datasets used in our experiments. Each region corresponds to a distinct financial ecosystem, with varying numbers of transactions, cardholders, and merchants. Notably, the number of

transactions ranges from 10,673 to 14,513, reflecting diverse market activity levels. The number of cardholders and merchants is relatively stable across regions, indicating similar population scales. In contrast, the total number of edges—representing all transactional interactions—varies more significantly, highlighting differences in graph connectivity and structural complexity. These variations emphasize the necessity of a continual learning framework capable of generalizing across heterogeneous regional distributions while preserving region-specific knowledge.

Table 7.2: The statistics of datasets for each region.

Region	$\mathcal{R}_1$	$\mathcal{R}_2$	$\mathcal{R}_3$	$\mathcal{R}_4$	$\mathcal{R}_5$
# Transactions	11,291	11,198	12,478	14,513	10,673
# Card Holders	103	124	126	115	118
# Merchants	700	698	698	699	700
# Total Edges	33,873	33,594	37,434	43,539	32,019

7.4.1.2 Baselines

We compare the proposed CCL with six representative continual learning baselines and adjust them for the proposed problem. The continual learning strategies used by these baselines include regularization-based methods, model structure-based methods, and replay-based methods.

- *EWC* [21]: A classical framework that protects the important weights of previous tasks to overcome catastrophic forgetting.
- *GEM* [173]: A model using gradient episodic memory to alleviate forgetting, while allowing beneficial transfer of knowledge to previous tasks.
- *MAS* [96]: A continual learning method called memory-aware synapses (MAS) computes the importance of parameters in an unsupervised online manner.
- *UCL* [46]: An uncertainty-regularized method that builds on traditional Bayesian online learning with variational inference.
- *DER++* [102]: An experience replay method that integrates knowledge distillation and regularization in rehearsal.
- *HAT* [43]: A task-based hard attention model that preserves previous tasks’ information without affecting the current task’s learning.

Following the learning paradigm of each continual learning baseline, we note that the first three baselines operate on graph representations from heterogeneous trade graphs as input, while the last three baselines take the pre-trained but learnable representations from the original transaction database as input.

7.4.1.3 Metrics

We note that fraud detection is a typical imbalanced classification problem, and the evaluation metrics should have no bias towards any class. Thus, we evaluate the proposed CCL and all the baselines with three widely used metrics: F1, Recall, and AUC. The higher scores of these metrics indicate better performance of the compared approaches.

7.4.2 Main Results

To assess the effectiveness of the proposed **Cross-Regional Continual Learning** framework, we conduct comparisons with six representative continual learning baselines, all implemented with a consistent graph representation learning backbone to ensure fairness. The results, summarized in Table 7.3, Table 7.4, and Table 7.5, reveal several noteworthy observations.

We adopt a standard continual learning protocol in our evaluation. For each region $\mathcal{R}_l$, the transaction data are randomly split into training, validation, and testing sets using a 6:1:3 ratio. During training, regional data are presented to the model sequentially in a randomly shuffled order, simulating the continual arrival of regional domains. Once a region’s training data are used, they become inaccessible to the model, and after all regional tasks have been trained, the final model is evaluated on the test sets of all regions.

Table 7.3: Recall scores (%) across regions $\mathcal{R}_1$ to $\mathcal{R}_5$.

Method	$\mathcal{R}_1$	$\mathcal{R}_2$	$\mathcal{R}_3$	$\mathcal{R}_4$	$\mathcal{R}_5$
EWC	88.58	88.97	89.03	91.37	90.17
GEM	81.18	78.73	79.49	76.73	74.60
MAS	88.66	88.89	89.10	90.69	89.34
UCL	72.20	71.96	71.69	73.14	69.19
DER++	69.84	70.16	68.76	70.57	66.87
HAT	75.37	74.65	73.01	77.24	71.40
CCL	95.51	91.43	90.67	96.69	92.59

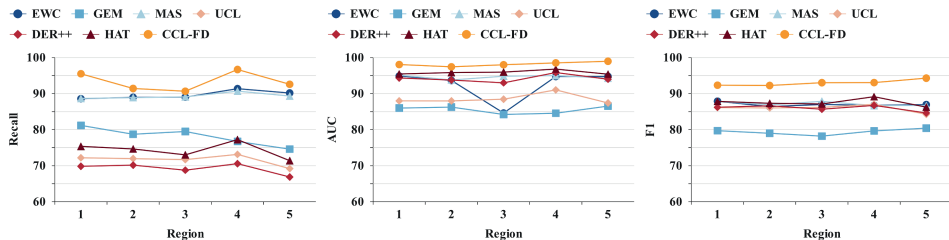


Figure 7.4: Left: Variation curves of recall. Middle: Variation curves of AUC. Right: Variation curves of F1.

Table 7.4: AUC scores (%) across regions $\mathcal{R}_1$ to $\mathcal{R}_5$.

Method	$\mathcal{R}_1$	$\mathcal{R}_2$	$\mathcal{R}_3$	$\mathcal{R}_4$	$\mathcal{R}_5$
EWC	94.89	93.66	84.62	94.75	94.77
GEM	85.99	86.26	84.22	84.56	86.54
MAS	<u>95.15</u>	93.78	94.83	95.00	94.29
UCL	88.00	87.98	88.45	91.04	87.41
DER++	94.34	93.77	93.00	95.85	94.01
HAT	95.45	<u>95.84</u>	<u>95.97</u>	<u>96.81</u>	<u>95.40</u>
CCL	98.08	97.44	98.01	98.53	98.98

Table 7.5: F1 scores (%) across regions $\mathcal{R}_1$ to $\mathcal{R}_5$.

Method	$\mathcal{R}_1$	$\mathcal{R}_2$	$\mathcal{R}_3$	$\mathcal{R}_4$	$\mathcal{R}_5$
EWC	<u>87.89</u>	86.40	86.99	86.75	<u>86.95</u>
GEM	79.72	79.02	78.21	79.66	80.42
MAS	87.87	86.90	<u>87.87</u>	86.68	86.70
UCL	86.13	86.03	86.18	87.02	84.29
DER++	86.22	86.62	85.68	86.73	84.64
HAT	87.85	<u>87.33</u>	87.13	<u>89.15</u>	86.25
CCL	92.34	92.26	93.03	93.06	94.30

Across all regions and evaluation metrics, CCL consistently achieves the best performance, outperforming the strongest baseline by a margin of 2%–8%. This clearly demonstrates that CCL not only excels in fraud detection accuracy but also effectively mitigates catastrophic forgetting. In particular, compared to regularization-based methods such as EWC and UCL, CCL yields substantially higher scores, highlighting the complementary benefits of integrating **knowledge replay** into a regularization-based framework. While methods like HAT are designed to prevent forgetting entirely, they fail to facilitate forward knowledge transfer; in contrast, CCL surpasses HAT across all regions, indicating that CCL not only preserves previously learned knowledge but also enhances the performance of earlier tasks through transferable representations.

Furthermore, although DER++ utilizes experience replay, its performance remains inferior to CCL, suggesting that directly applying off-the-shelf continual learning methods to financial tasks—without modeling the underlying structure and semantics of financial transactions—is insufficient. CCL’s design, which explicitly incorporates financial domain characteristics via graph modeling and task-aware replay strategies, proves more effective in knowledge consolidation and generalization.

In summary, these findings affirm that **effective knowledge transfer** is critical in preventing forgetting, and that the synergistic combination of replay and regularization leads to significant gains in continual learning performance.

7. CROSS-REGIONAL FRAUD DETECTION VIA CONTINUAL LEARNING WITH KNOWLEDGE TRANSFER

From a broader perspective, most continual learning methods alleviate catastrophic forgetting by constraining updates to important parameters, often resulting in a trade-off that sacrifices current task performance to preserve past knowledge. However, this rigid preservation strategy can hinder adaptability. A desirable continual learning framework should strike a balance between retaining historical knowledge and acquiring new information. Our experimental results suggest that the knowledge replay mechanism in CCL enables the model to achieve this balance more effectively.

As visualized in Figure 7.4, CCL maintains stable performance across all tasks as new regional domains are introduced. Unlike baseline methods that exhibit sharp performance drops for earlier tasks, CCL shows no abrupt degradation, indicating that knowledge from past regions is actively reinforced rather than forgotten. This confirms that CCL is inherently more robust and adaptive in dynamic, incremental settings and well-suited for real-world applications where new regional data continuously arrives.

7.4.3 Ablation Study

Table 7.6: Ablation study (Recall %) across regions $\mathcal{R}_1$ to $\mathcal{R}_5$.

Method	$\mathcal{R}_1$	$\mathcal{R}_2$	$\mathcal{R}_3$	$\mathcal{R}_4$	$\mathcal{R}_5$
EWC	88.58	88.97	<u>89.03</u>	91.37	90.17
DER++	69.84	70.16	68.76	70.57	66.87
CCL w/o RPS	85.47	82.83	<u>87.60</u>	<u>95.51</u>	94.58
CCL w/o PKR	<u>89.83</u>	<u>87.53</u>	83.97	93.80	92.58
CCL	95.51	91.43	90.67	96.69	<u>92.59</u>

Table 7.7: Ablation study (AUC %) across regions $\mathcal{R}_1$ to $\mathcal{R}_5$.

Method	$\mathcal{R}_1$	$\mathcal{R}_2$	$\mathcal{R}_3$	$\mathcal{R}_4$	$\mathcal{R}_5$
EWC	94.89	93.66	84.62	94.75	94.77
DER++	94.34	93.77	93.00	95.85	94.01
CCL w/o RPS	96.86	97.27	97.15	97.51	99.24
CCL w/o PKR	<u>97.77</u>	<u>98.68</u>	<u>98.01</u>	<u>97.76</u>	97.42
CCL	98.08	97.44	98.09	98.53	<u>98.98</u>

To address **Q2**, we conduct an ablation study to evaluate the individual contributions of the two core components in the proposed CCL framework: *Prototype-based Knowledge Replay (PKR)* and *Regularized Parameter Smoothing (RPS)*. Specifically, we define two CCL variants—**CCL w/o RPS** and **CCL w/o PKR**—by independently removing each module from the full model. For comparison, we also report the results of two representative continual learning baselines, EWC and DER++, all sharing the same graph-based representation learning backbone for fairness.

Table 7.8: Ablation study (F1 %) across regions $\mathcal{R}_1$ to $\mathcal{R}_5$.

Method	$\mathcal{R}_1$	$\mathcal{R}_2$	$\mathcal{R}_3$	$\mathcal{R}_4$	$\mathcal{R}_5$
EWC	87.89	86.40	86.99	86.75	86.95
DER++	86.22	86.62	85.68	86.73	84.64
CCL w/o RPS	90.81	90.84	91.47	86.91	95.66
CCL w/o PKR	<u>91.89</u>	<u>93.27</u>	<u>91.80</u>	<u>90.33</u>	93.60
CCL	92.34	92.26	93.03	93.06	<u>94.30</u>

The results are summarized in Table 7.6, Table 7.7, and Table 7.8, where several important observations can be made. First, **the full CCL model consistently outperforms all variants and baselines across most regions and metrics**, especially on the first four tasks, highlighting the effectiveness of jointly combining PKR and RPS for continual fraud detection. Notably, CCL yields the highest AUC and F1 scores across nearly all tasks, confirming its strong generalization and retention capabilities.

Second, when the **RPS module is removed** (CCL w/o RPS), performance on earlier tasks slightly degrades, while the last task— $\mathcal{R}_5$ —shows a **peak in F1 score**, even surpassing the full CCL. This phenomenon reflects the trade-off inherent in continual learning: while RPS helps prevent forgetting by constraining parameter drift, it may restrict full adaptation to the most recent task. In contrast, the absence of such constraints allows the model to overfit to the current region, boosting its short-term performance at the expense of prior knowledge.

Third, the variant **CCL w/o PKR** shows the largest drop in performance on the final task and lower stability across earlier tasks. Despite incorporating parameter regularization, the lack of informative sample replay limits the model’s ability to reconstruct and reinforce past knowledge. Interestingly, CCL w/o PKR still outperforms EWC and DER++, confirming that regularization alone is beneficial—but not sufficient—for stable knowledge consolidation. This further emphasizes the importance of leveraging task-specific data representations in continual learning, particularly under the non-stationary conditions typical in financial domains.

Overall, the ablation results clearly demonstrate that both PKR and RPS make complementary contributions to the effectiveness of CCL. PKR enables selective knowledge distillation via prototype-guided replay, while RPS ensures model stability through parameter preservation. Their joint integration not only alleviates catastrophic forgetting but also enhances cross-regional adaptability. These findings underscore the practical relevance of designing continual learning strategies tailored to the structural and semantic characteristics of financial fraud detection tasks.

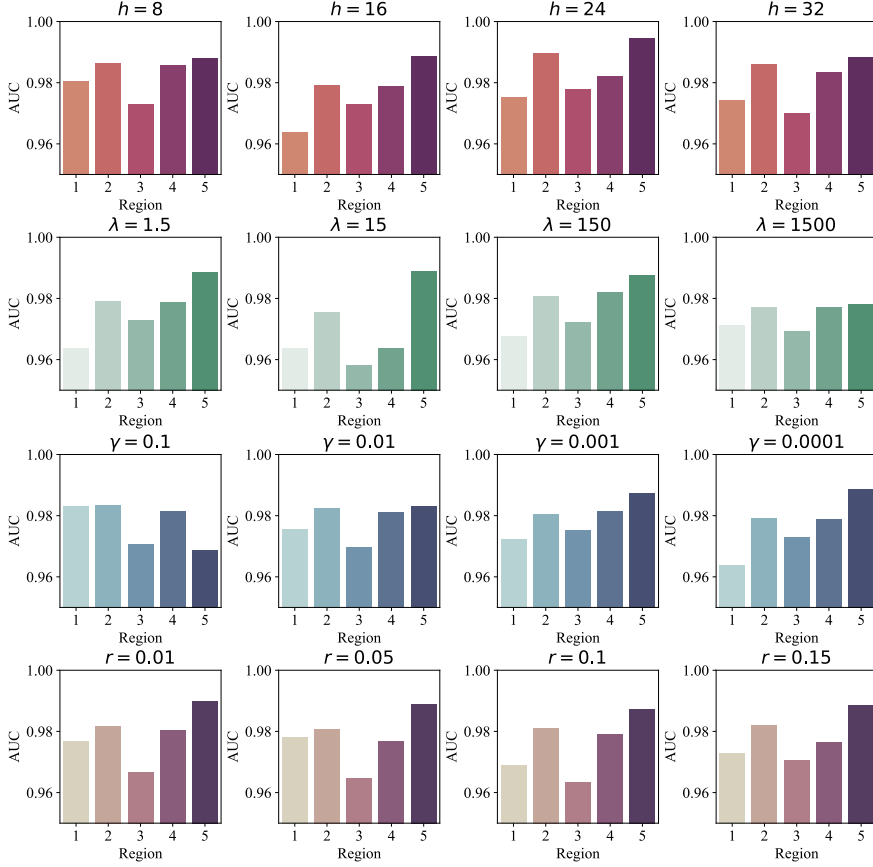


Figure 7.5: Sensitivity analysis of hyperparameters. We vary the hidden size $|\mathbf{h}|$ as $\{8, 16, 24, 32\}$, the smoothing term λ as $\{1.5, 15, 150, 1500\}$, the smoothing term γ as $\{0.1, 0.01, 0.001, 0.0001\}$, the replay ratio r as $\{0.01, 0.05, 0.1, 0.15\}$.

7.4.4 Sensitivity Analysis of Hyperparameters

To answer Q4, we analyze the impact of hyperparameters that could affect the proposed CCL: the hidden size $|\mathbf{h}|$, two hand-crafted weights of smoothing λ and γ , and the replay ratio r . In Figure 7.5, we find that the proposed CCL can keep satisfying performance (surpassing 0.96 of AUC) under different hyperparameter settings on all tasks. Thus, we can safely conclude that CCL is robust to a variety of hyperparameter settings for a sequence of tasks. However, the trade-off between the final task and previously learned tasks varies according to different hyperparameters, which shows that different parameters have a certain impact on the knowledge transfer function of the model, and can also reflect the degree of difference between different tasks.

Firstly, for the hidden size $|\mathbf{h}|$, the test performance of the last two tasks is always satisfying. However, there is a trade-off between the first and third tasks: when the

performance of the first task increases, the performance of the third task decreases, and vice versa. Generally, the hidden size $|\mathbf{h}|$ has an important impact on the capacity of the model. The results demonstrate that the hidden size $|\mathbf{h}|$ has an influence on the forgetting issue of a continual learning model, so it is important to choose an appropriate hidden size.

Secondly, the smoothing terms, λ and γ , are related to the Regularized Parameter Smoothing module. When $\lambda = 15$, the test result of the last task is the best, but the previous tasks perform badly, which shows that the forgetting issue has not been alleviated under this hyperparameter setting. As λ increases, the performance of the last task decreases slightly, but the performance of previously learned tasks gets better. Similarly, for the second smoothing term γ , there is a trade-off between the current task and previous tasks. When γ is 0.1, the test performance of the last task is the worst, but the performance of previous tasks achieves the best, which indicates that the model retains a lot of knowledge while sacrificing the test performance of the current task. As γ gets smaller, the performance of the last task gets better, while the performance of the first three learned tasks decreases significantly. To balance all tasks, these two smoothing terms must be chosen reasonably to minimize the trade-off between newly learned and previously learned tasks. Hence, the RECD-FD model takes $\lambda = 1.5$ and $\gamma = 0.00025$.

Finally, for the replay ratio r , we observe that a larger replay buffer will not necessarily bring better performance in the current task. But in the third region’s task, the greater replay ratio obviously led to better performance, which suggests the differences between the third region and the current region are relatively huge, so more replayed samples are needed to ensure the quality of knowledge transfer, and we choose the replay ratio safely as 15% for CCL.

7.4.5 Interpretability Analysis

To further investigate the interpretability of our proposed model, we analyze the contribution of each node type within the Heterogeneous Trade Graphs by quantifying their influence on the prediction outcomes. Specifically, we employ a grid search strategy in combination with the **SHAP** (Shapley Additive exPlanations) toolkit¹ to estimate and visualize the feature importance of different node types. Unlike traditional feature importance methods that merely identify influential features without explaining their directional effects, SHAP provides sample-specific explanations by quantifying both the positive and negative contributions of each feature to the model’s output.

In our experimental design, we apply node masking to simulate varying degrees of information loss for each node type in the heterogeneous trade graph. Specifically,

¹<https://pypi.org/project/shap>

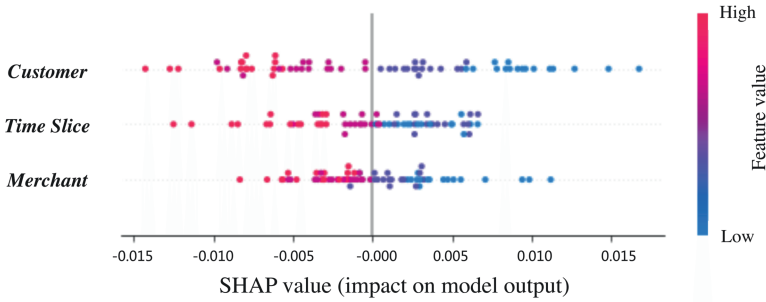


Figure 7.6: Node-type importance analysis via SHAP values. Each dot represents a sample’s SHAP value for a specific node type, indicating the feature’s impact on the model’s output. Higher SHAP values (toward the right) denote stronger positive contributions to fraud detection, while lower values (toward the left) denote negative influence. Node types with greater vertical dispersion and right-shifted distributions are more influential in the model’s decision-making. The color of each point corresponds to the masking level of that node type, where **red** indicates a higher feature value (i.e., lower masking), and **blue** indicates a lower feature value (i.e., more heavily masked). It can be observed that **Customer** nodes are the most impactful, followed by **Time Slice** and **Merchant**, underscoring the importance of temporal and behavioral semantics in cross-regional fraud detection.

we randomly mask 0%, 30%, 60%, and 90% of the nodes for each node type. The resulting SHAP values are illustrated in Figure 7.6, where points near the top indicate node types with a more substantial impact on model predictions.

Several key observations emerge from this analysis. First, **Customer nodes consistently exhibit the highest SHAP values**, indicating that they play a pivotal role in model decision-making. This aligns with practical expectations: fraudulent transactions are inherently tied to customer behavior, making their representation crucial for effective detection. Second, it is noteworthy that the **Time Slice** node type is ranked above **Merchant** in importance. This finding suggests that temporal patterns—such as the time of day or specific date clusters—carry strong signals in fraud detection, as malicious actors may exploit time-sensitive vulnerabilities to conduct fraudulent activities. By explicitly modeling these temporal semantics as nodes in heterogeneous trade graphs, our framework enhances both interpretability and predictive accuracy.

Furthermore, the SHAP visualization provides additional insights into how node masking affects model robustness. In Figure 7.6, **blue points** represent lower masking ratios, while **red points** correspond to higher masking levels. A clear trend can be observed: as more nodes of a given type are masked, the average SHAP value decreases and often turns negative, reflecting a degradation in model performance due to the loss of critical structural information. This pattern reinforces the intuition that the availability and integrity of key node types directly influence the model’s reasoning process.

In summary, this interpretability analysis reveals that CCL not only achieves high predictive performance but also captures meaningful semantic relationships in the data. By leveraging the structural diversity of heterogeneous trade graphs and preserving important node-type interactions, the model provides both actionable insights and strong robustness in real-world fraud detection scenarios.

7.5 Summary

In this chapter, we studied a novel perspective on the problem of **cross-regional fraud detection**, a practically important yet underexplored challenge originating from real-world financial systems. We proposed **CCL**, a continual learning framework designed to incrementally adapt across regional domains while effectively mitigating catastrophic forgetting. To the best of our knowledge, this is the first attempt to formulate cross-regional fraud detection as a continual learning problem. Our approach integrates **heterogeneous trade graphs**, constructed from structured financial transaction records, to model the complex semantic relationships among various entities such as users, merchants, and transactions.

CCL incorporates a prototype-based knowledge replay mechanism in conjunction with parameter smoothing techniques to facilitate knowledge retention and transfer. A GNN-based backbone is employed to support expressive and context-aware representation learning. Experimental evaluations on multi-region datasets demonstrate that CCL significantly outperforms established continual learning baselines in both single-region and cross-region scenarios, highlighting its robustness, adaptability, and practical applicability.

Despite its promising performance, CCL has several limitations as follows. First, the construction and training of heterogeneous trade graphs require additional data preprocessing procedures, and the generation of graph-structured data also imposes extra memory demands for storage, especially when dealing with large-scale structured datasets. Second, the model’s robustness may be impacted by the inherent class imbalance in financial fraud data, which poses challenges for stable generalization. As part of our future work, we plan to augment the semantic richness of heterogeneous trade graphs by incorporating external contextual information, such as **social relationships or user behavioral traces**, to uncover deeper inter-entity dependencies and enable more nuanced fraud detection. In addition, we aim to explore the development of **expandable or modular neural architectures** that can dynamically grow with new regional domains, further enhancing adaptability without retraining from scratch.

While this study focuses on a specific applied domain, it also reveals broader implications for the general field of open-world continual learning. Our earlier chapters explored open-world continual learning in abstract formulations and theoretical settings; however, the integration of open-world continual learning into real-world,

high-stakes domains remains underdeveloped. The complex, evolving, and partially observable nature of financial fraud aligns closely with the open-world assumptions of open-world continual learning, where unknown classes or domains emerge over time. Nonetheless, most existing open-world continual learning methods lack tailored mechanisms for task-specific constraints or domain-specific semantic modeling.

As a promising future direction, we plan to extend open-world continual learning frameworks toward a broader set of realistic applications, including but not limited to fraud detection, risk prediction, and cross-domain recommendation systems, by incorporating domain-specific inductive biases and designing more interpretable, modular, and resource-aware learning strategies. We envision a unified learning paradigm that jointly addresses open-world generalization, task continuity, and real-time deployment needs, bridging the gap between open-world continual learning theory and impactful real-world solutions.

Chapter 8

Conclusions and Future Work

This thesis is an investigation of **Open-world Continual Learning**, an emerging paradigm at the intersection of continual learning and open-set recognition, including out-of-distribution detection. By formalizing open-world continual learning with a mathematical definition and identifying four real-world scenarios, we have laid a foundational framework in this domain. Our empirical analyses reveal the unique challenges of open-world continual learning, particularly its inherent open risks and the dynamic interplay between sequential tasks, distinguishing it from conventional deep learning approaches.

From a technical perspective, we have pioneered a knowledge transfer-centric approach to open-world continual learning, addressing aspects such as knowledge representation, known-unknown boundary discovery, and forward/backward transfer mechanisms. Our proposed models *Pro-KT* and *PEAK* demonstrate the effectiveness of this perspective in both large-scale and few-shot learning scenarios. Furthermore, we introduced *HoliTrans* to provide a unified and theoretically based framework that integrates open-set detection with continual learning, offering a holistic solution to open-world continual learning challenges.

Beyond theoretical contributions, we have validated the practical relevance of open-world continual learning through a real-world application in cross-regional fraud detection. By introducing *CCL*, a continual graph learning model, we address the critical issue of distributional shifts in fraud patterns across regions, demonstrating the adaptability of open-world continual learning principles in complex, evolving environments.

Collectively, the four research questions explored in this thesis form an evolving narrative of open-world continual learning. *Pro-KT* (RQ1) initiates the study by establishing prompt-enhanced knowledge transfer for maintaining stability and adaptability. *PEAK* (RQ2) advances this direction under few-shot constraints, enhancing robustness with limited labeled data. *HoliTrans* (RQ3) unifies open-set recognition

and continual learning into a single theoretical and architectural framework, addressing knowledge transfer between known and unknown samples. Finally, *CCL* (RQ4) grounds these principles in a real-world scenario, i.e., cross-regional fraud detection, showcasing the framework’s practical impact. Together, these contributions chart a consistent progression from theoretical foundations to holistic methodology and applied validation.

This chapter summarizes the main contributions of this thesis and reflects on their implications for the field of open-world continual learning. We then discuss open challenges and propose future research directions to guide further progress in this evolving paradigm.

8.1 Summary of Contributions

In Chapter 1, we formulated four key research questions. We now revisit them to summarize this thesis’s contributions to each.

[RQ1] *Can we develop an effective open-world continual learning model to accumulate the knowledge gained in the past and to use the knowledge to help divide the test data into knowns and unknowns/opens, especially for enlarging the open-set boundary between knowns and unknowns?*

This thesis addressed the stability-plasticity trade-off in open-world continual learning through a novel perspective on knowledge representation and transfer. We proposed *Pro-KT* in Chapter 4, a prompt-learning approach that encodes both task-specific and task-agnostic knowledge into a unified repository, enabling dynamic knowledge transfer across tasks while maintaining clear boundaries between known and unknown classes. Our experiments demonstrated that *Pro-KT* achieves superior performance in both knowledge retention (stability) and open-set detection (plasticity), while remaining practically deployable as a plug-and-play framework for pre-trained models. These results not only establish new benchmarks for open-world continual learning systems but also provide foundational insights for handling increasingly complex open-world scenarios where knowledge must evolve continuously yet robustly.

[RQ2] *How can we enhance the performance of existing open-world continual learning models under the constraint of limited labeled data by employing an improved knowledge transfer approach?*

Addressing the critical challenge of data scarcity in open-world continual learning, in Chapter 5 we introduced *PEAK*, the first dedicated framework for few-shot open-world continual learning scenarios, where data is often sparse and tasks are dynamic. Building on the foundation of Chapter 4, we identified three fundamental obstacles

that emerge when learning from limited data: representation collapse under sparse data distributions, ambiguous decision boundaries for open-set detection, and degradation of knowledge transfer across tasks. *PEAK*'s architecture combines three novel components specifically designed to overcome these challenges simultaneously. Experimental results demonstrated consistent superiority over both state-of-the-art few-shot learning and open-world continual learning baselines, validating *PEAK*'s dual capability for robust open-set recognition and continual adaptation.

More importantly, *PEAK* establishes a methodological bridge between theoretical open-world continual learning research and practical applications where data is inherently scarce and learning requirements evolve continuously. In medical diagnosis, for instance, the ability to learn from sparse labeled examples while detecting novel pathologies would address the critical need for adaptive systems in rare disease identification. Similarly, in fraud detection, where new attack patterns emerge with minimal labeled samples, robust knowledge transfer would prevent catastrophic forgetting while flagging unseen threats. Autonomous systems operating in dynamic environments, such as disaster-response robots, would benefit from few-shot adaptation to novel objects without losing prior knowledge, a capability equally important for cybersecurity (zero-day attack detection) and biodiversity monitoring when tracking rare or even unknown species. Even in personalized AI assistants, where user preferences shift incrementally with limited feedback, it is important to balance stability and plasticity. By tackling the real-world challenges of data scarcity, open-set recognition, and non-stationary task distributions, *PEAK* opens new possibilities for deployment in resource-constrained environments.

[RQ3] *How can we leverage knowledge from both known and unknown samples to enable knowledge transfer in open-world continual learning, effectively improving open-set detection with incremental classification at both theoretical and architectural levels?*

In Chapter 6, we presented a transformative reconceptualization of open-world continual learning, improving on the limitations of existing approaches that merely combine continual learning and out-of-distribution detection techniques. Building on the foundations established in Chapter 4 and Chapter 5, we first exposed a critical shortcoming in current methods: their reliance on the flawed orthogonal decomposition assumption, which artificially separates open-set recognition from incremental classification. Through an analysis of four open-world continual learning scenarios, we demonstrated how this oversimplification leads to failures in capturing the intricate interactions and structural challenges inherent to open-world continual learning.

Our theoretical contribution in Chapter 6 started with a novel problem formulation that captures the intrinsic coupling between open-set dynamics and continual task arrival. We proposed a unified decision space for open-world continual learning and

derived, through theoretical analysis, the key optimization factors that govern effective learning in open-world environments. These theoretical insights were used for the development of *HoliTrans*, an architectural solution that integrates nearest class mean principles into open-world continual learning. The elegance of *HoliTrans* lies in its dual capacity: it simultaneously achieves state-of-the-art empirical performance across all four open-world continual learning scenarios while providing provable guarantees on decision boundary stability and knowledge preservation.

By bridging theoretical insights with practical implementation, we established a new foundation for developing robust learning systems capable of operating in truly open-ended environments. We demonstrated the success of *HoliTrans* across diverse experiments, showing its potential to address adaptive yet stable learning, with applicability from evolving cybersecurity threats to dynamic medical diagnostic systems.

[RQ4] *How can we improve the model performance by formulating practical problems with an open-world continual learning framework and developing effective models for challenging real-world applications?*

In Chapter 8, we applied our theoretical and algorithmic results on open-world continual learning to the financial domain of cross-regional fraud detection. By formulating this task as an incremental learning problem, we introduced *CCL*, a framework that addresses two challenges: adapting to region-specific fraud patterns while preserving knowledge across geographically shifting distributions. The integration of prototype-based knowledge replay, parameter smoothing, and heterogeneous graph modeling enabled *CCL* to achieve simultaneous forward adaptability and backward stability, a balance that is missing in conventional fraud detection systems.

Experimental results demonstrated that *CCL* reduced catastrophic forgetting when compared to baseline methods while improving cross-regional knowledge transfer accuracy. These findings not only validate the practical utility of open-world continual learning principles but also reveal how knowledge-centric designs can overcome distributional shifts in real-world scenarios. More broadly, this part of our work shows that the theoretical frameworks developed earlier in the thesis translate effectively to applications where data evolves continuously.

Note. All methods proposed in this thesis were released as open-source software. We believe that open release promotes reproducibility, enables practical deployment, and provides an inspiration for further research and practical applications.

8.2 Future Work

Building upon the contributions of this thesis, we identify several promising directions to advance open-world continual learning under modern deep learning paradigms.

a) Modular OWCL with Sparse Expert LLMs

The rapid evolution of sparse Mixture-of-Experts (MoE) language models, such as Mixtral, GLaM, and DBMoE, presents a transformative opportunity for open-world continual learning. While these models have demonstrated scalability in static NLP tasks, their potential for dynamic, open-ended learning is still largely unexplored.

Current open-world continual learning methods face two critical limitations when scaled to modern LLMs: monolithic architectures hinder efficient knowledge reuse, and full fine-tuning becomes computationally prohibitive. Although MoE-based LLMs provide a modular structure that can be exploited for dynamic task specialization, it is unclear how expert activation, routing strategies, and parameter sharing behave under sequential and open-world task distributions.

To address these challenges, future research could explore the design of a paradigm where MoE-based LLMs serve as modular knowledge repositories, with task-conditioned gating mechanisms dynamically routing known inputs and emerging tasks to specialized experts. By combining prompt routing with task-adaptive gating networks, it should be possible to selectively activate subsets of experts, reducing computation time while promoting transferability. Additionally, incorporating temporal-aware gating mechanisms, such as dynamic routing (as in DRMoE) or hierarchical gating (as in HyperMoE), may enhance the system’s adaptability to evolving task sequences and distribution shifts.

Another promising direction is in the integration of MoE architectures with parameter-efficient adaptation techniques. By combining dynamic expert routing with adapter tuning or expert masking strategies, we could achieve both flexible knowledge organization and computationally sustainable continual learning. This hybrid approach would enable on-demand expert reconfiguration for novel tasks, preservation of core competencies without catastrophic forgetting, and elimination of the need for expensive data replay. The development of such architectures should be accompanied by novel evaluation frameworks, including quantitative metrics for expert utilization diversity (measuring how effectively specialized knowledge is deployed) and transfer stability (assessing consistency of performance across task sequences). These metrics would provide quantitative ways to evaluate the efficiency and robustness of modular knowledge routing under open-world continual learning regimes.

This direction, bridging sparse LLM architectures with open-world continual learning, could yield systems capable of incremental specialization, where new experts are added for novel tasks while maintaining access to foundational knowledge. Such

architectures would offer scalable, adaptable, and interpretable solutions to many real-world applications.

b) OWCL under Heterogeneous Multimodal Environments

Another underexplored direction involves extending open-world continual learning to handle heterogeneous task streams with multimodal data distributions. Current open-world continual learning frameworks mainly operate under the limiting assumption of homogeneous, unimodal task sequences, a condition rarely satisfied in practical deployment scenarios. While some recent works in continual learning have started to explore multimodal task streams, to the best of our knowledge, there has been no dedicated research within the open-world continual learning setting addressing heterogeneity and multimodality (e.g., visual, textual, graph-structured) with divergent statistical properties and annotation protocols.

The main challenge is in developing architectures capable of maintaining stable knowledge transfer across modality boundaries while preserving sensitivity to open-set novelty within each data domain. Potential solutions may involve integrating cross-modal prompt embeddings with multimodal transformers to facilitate shared representations across tasks involving text, vision, and graph-based data. Graph-structured memory modules or modality-aware task encoders can further assist in modeling the structural relationships between tasks and promoting generalization.

Another challenge is the mitigation of modality interference and ensuring stable transfer across heterogeneous tasks. Adaptive task selection mechanisms, curriculum-based training strategies, or graph-guided expert activation may help address this issue. Equally important is the development of evaluation methodologies for multimodal open-world continual learning systems. Current benchmarks fail to capture essential dimensions such as cross-modal interference effects, modality-specific catastrophic forgetting rates, and open-set detection performance across heterogeneous data types. New metrics and testing protocols should account for these composite challenges introduced by simultaneous modality shifts and task boundary ambiguity.

This research direction carries significant implications for domains where multimodal data evolves continuously, including clinical decision support systems integrating imaging and electronic health records, or autonomous platforms processing sensor data in non-stationary environments. The insights gained from studying heterogeneous open-world continual learning would be of interest to the broader machine learning community because of the recent interest in multimodal representation learning.

c) Scalable and Privacy-Preserving OWCL

The increasing deployment of large-scale machine learning models necessitates the development of parameter-efficient adaptation strategies that support continual learning while maintaining computational efficiency. Current approaches based on full model fine-tuning are impractical for open-world continual learning scenarios due to their substantial computational and storage requirements. Recent advances in parameter-efficient fine-tuning techniques, including low-rank adaptation, Adapter-Fusion, and Prefix Tuning, may be promising alternatives to investigate within the open-world continual learning paradigm.

Future research on integrating parameter-efficient fine-tuning mechanisms with prompt-based learning architectures could achieve both efficient task adaptation and stable knowledge retention. Potential solutions may include dynamic growing adapter modules, selectively routing adapters based on task identity, or compressing knowledge into modular parameter blocks for knowledge consolidation. The theoretical properties of such hybrid architectures, as well as their capacity bounds and forgetting dynamics, are important open questions.

The combination of parameter-efficient fine-tuning methods with federated open-world continual learning frameworks presents another promising research direction where models are updated across distributed clients under data privacy constraints. By implementing parameter-efficient fine-tuning’s lightweight update mechanisms, it should be possible to enable continual knowledge sharing across edge nodes without transmitting sensitive data. In particular, designing parameter-efficient fine-tuning-compatible synchronization protocols and communication-efficient update strategies will be crucial for scalable and secure open-world continual learning under federated settings. Furthermore, exploring how expert modularity or adapter routing can be personalized across users or client populations offers a compelling direction for federated personalization in open-world contexts.

By uniting the principles of open-world continual learning, parameter-efficient fine-tuning, and federated learning, this research direction has particular relevance for application domains requiring both continuous adaptation and strict data privacy in a decentralized environment, such as healthcare frameworks, industrial IoT systems, and financial services.

d) Explainability and Trustworthiness in OWCL

The development of explainable and trustworthy learning systems represents an important challenge in open-world continual learning, particularly for deployment in safety-critical domains. Current open-world continual learning methodologies predominantly focus on improving task performance metrics while not considering model interpretability and decision transparency. This is increasingly problematic as open-world continual learning systems encounter novel inputs and evolving task distributions, but understanding model behavior is often essential for practical adoption.

The interaction between continual adaptation mechanisms and open-set recognition components creates complex dynamics that existing interpretation methods are not yet able to analyze. Also, the dynamic evolution of decision boundaries and feature representations in continual learning scenarios lacks any formal frameworks for temporal analysis.

Future research should develop theoretical approaches to explainability that address these OWCL-specific challenges. Promising directions include the incorporation of explainability tools such as attention entropy visualization, gating path trace analysis, or counterfactual generation to present a decision-making process. By identifying which experts or prompts are activated under different tasks, one could gain insights into the model’s internal reasoning and evaluate the consistency of knowledge reuse. Causal inference techniques may also play a role in uncovering the underlying factors driving prediction shifts or classification errors. Finally, visual explanations that highlight influential features or task-specific representations can enhance interpretability and inform model debugging.

As for many deep learning systems, ultimately, explainability-enhanced open-world continual learning systems will bridge the gap between black-box model behavior and human understanding, allowing humans in the loop, interactive learning, and regulatory transparency needed, for example, in healthcare diagnostics, autonomous systems, and financial risk assessment.

e) Meta-Memory and Adaptive Sampling for Long-Horizon OWCL

In real-world dynamic and evolving environments, open-world continual learning systems must go beyond passive adaptation and develop mechanisms for strategic memory management and sample prioritization. One promising direction is to integrate meta-memory architectures and adaptive sampling strategies to guide long-term knowledge retention and continual adaptation under constrained resources.

Meta-memory architectures offer a promising solution by maintaining abstracted, task-generalizable knowledge representations that support efficient retrieval during subsequent learning phases. These architectures differ from conventional memory buffers by enabling selective access to relevant prior knowledge without requiring extensive storage of raw data or brute force replay. When combined with adaptive sampling strategies (for example, guided by task uncertainty, representation novelty, or learning progress), the model can dynamically select the most informative or high-impact samples for fine-tuning, thus improving both data efficiency and generalization.

The unique requirements of open-world continual learning introduce several important considerations for these approaches. First, unlike traditional active learning that assumes access to an external oracle, adaptive sampling in open-world continual learning can be self-supervised or uncertainty-aware, making it particularly

valuable in environments where annotations are costly or delayed. Second, the interplay between meta-memory and adaptive sampling requires an architectural design of open-world continual learning as autonomous systems that learn efficiently, prioritize effectively, and operate reliably over extended time horizons. Third, the integration of curriculum learning principles with adaptive sampling could provide systematic control over task difficulty progression while mitigating catastrophic forgetting.

The development of such architectures requires addressing fundamental questions about memory compression, retrieval efficiency, and sample selection criteria in open-world scenarios. Theoretical analysis of these systems' long-term behavior and empirical validation under realistic non-stationary conditions are important open challenges in the field that could find applications in domains characterized by gradual distribution shifts and sparse supervision signals. Examples include personalized education systems where meta-memory could enable long-term student modeling while adaptive sampling would focus on knowledge gaps, and autonomous agents operating in physical environments that could benefit from prioritized experience retention and efficient sample selection for continuous adaptation.

Bibliography

- [1] Parisi, G.I., Kemker, R., Part, J.L., Kanan, C., Wermter, S.: Continual lifelong learning with neural networks: A review. *Neural Networks* **113** (2019) 54–71
- [2] Ke, Z., Liu, B., Huang, X.: Continual learning of a mixed sequence of similar and dissimilar tasks. *Advances in Neural Information Processing Systems* **33** (2020) 18493–18504
- [3] Han, R., Ren, X., Peng, N.: Econet: Effective continual pretraining of language models for event temporal reasoning. In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. (2021) 5367–5380
- [4] Kudithipudi, D., Aguilar-Simon, M., Babb, J., Bazhenov, M., Blackiston, D., Bongard, J., Brna, A.P., Chakravarthi Raja, S., Cheney, N., Clune, J., et al.: Biological underpinnings for lifelong learning machines. *Nature Machine Intelligence* **4** (2022) 196–210
- [5] Wang, L., Zhang, X., Su, H., Zhu, J.: A comprehensive survey of continual learning: theory, method and application. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2024)
- [6] Van de Ven, G.M., Tolias, A.S.: Three scenarios for continual learning. *arXiv:1904.07734* (2019)
- [7] Van de Ven, G.M., Siegelmann, H.T., Tolias, A.S.: Brain-inspired replay for continual learning with artificial neural networks (2020)
- [8] Liu, B., Mazumder, S., Robertson, E., Grigsby, S.: AI autonomy: Self-initiated open-world continual learning and adaptation. *AI Magazine* **44** (2023) 185–199
- [9] Chen, Z., Liu, B.: *Lifelong machine learning*. Morgan & Claypool Publishers (2018)
- [10] Yang, G., Fini, E., Xu, D., Rota, P., Ding, M., Nabi, M., Alameda-Pineda, X., Ricci, E.: Uncertainty-aware contrastive distillation for incremental semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **45** (2022) 2567–2581
- [11] McClelland, J.L., McNaughton, B.L., O’Reilly, R.C.: Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychological Review* **102** (1995) 419
- [12] McCloskey, M., Cohen, N.J.: Catastrophic interference in connectionist networks: The sequential learning problem. In: *Psychology of learning and motivation*. Volume 24. Elsevier (1989) 109–165
- [13] Wang, H., Liu, B., Wang, S., Ma, N., Yang, Y.: Forward and backward knowledge transfer for sentiment classification. In: *Proceedings of the Asian Conference on Machine Learning*, PMLR (2019) 457–472
- [14] Zhou, D.W., Wang, Q.W., Qi, Z.H., Ye, H.J., Zhan, D.C., Liu, Z.: Class-incremental learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2024)
- [15] Bendale, A., Boulton, T.: Towards open world recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. (2015) 1893–1902
- [16] Fei, G., Wang, S., Liu, B.: Learning cumulatively to become more knowledgeable. In: *Proceedings of the 22nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. (2016) 1565–1574

- [17] Tulving, E., Kroll, N.: Novelty assessment in the brain and long-term memory encoding. *Psychonomic bulletin & review* **2** (1995) 387–390
- [18] Murty, V.P., Ballard, I.C., Macduffie, K.E., Krebs, R.M., Adcock, R.A.: Hippocampal networks habituate as novelty accumulates. *Learning & Memory* **20** (2013) 229–235
- [19] Kim, G., Xiao, C., Konishi, T., Ke, Z., Liu, B.: A theoretical study on solving continual learning. *Advances in Neural Information Processing Systems* **35** (2022) 5065–5079
- [20] Scheirer, W.J., de Rezende Rocha, A., Sapkota, A., Boulton, T.E.: Toward open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **35** (2012) 1757–1772
- [21] Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A.A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., et al.: Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences* **114** (2017) 3521–3526
- [22] Zenke, F., Poole, B., Ganguli, S.: Continual learning through synaptic intelligence. In Precup, D., Teh, Y.W., eds.: *Proceedings of the 34th International Conference on Machine Learning*. Volume 70 of PMLR., PMLR (2017) 3987–3995
- [23] Kishida, I., Chen, H., Baba, M., Jin, J., Amma, A., Nakayama, H.: Object recognition with continual open set domain adaptation for home robot. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. (2021) 1517–1526
- [24] Truong, T.D., Nguyen, H.Q., Raj, B., Luu, K.: Fairness continual learning approach to semantic scene understanding in open-world environments. *Advances in Neural Information Processing Systems* **36** (2023) 65456–65467
- [25] Kim, G., Xiao, C., Konishi, T., Ke, Z., Liu, B.: Open-world continual learning: Unifying novelty detection and continual learning. *Artificial Intelligence* **338** (2025) 104237
- [26] Dhar, P., Singh, R.V., Peng, K.C., Wu, Z., Chellappa, R.: Learning without memorizing. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE (2019) 5138–5146
- [27] Vijayan, P., Bhat, P., Zonooz, B., Arani, E.: Trire: A multi-mechanism learning paradigm for continual knowledge retention and promotion. In: *Advances in Neural Information Processing Systems*. (2023)
- [28] Masana, M., Liu, X., Twardowski, B., Menta, M., Bagdanov, A.D., Van De Weijer, J.: Class-incremental learning: survey and performance evaluation on image classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **45** (2022) 5513–5533
- [29] Lo, S.Y., Oza, P., Patel, V.M.: Adversarially robust one-class novelty detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **45** (2022) 4167–4179
- [30] Li, Y., Yang, X., Wang, H., Wang, X., Li, T.: Learning to prompt knowledge transfer for open-world continual learning. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Volume 38(12). (2024) 13700–13708
- [31] Mundt, M., Pliushch, I., Majumder, S., Hong, Y., Ramesh, V.: Unified probabilistic deep continual learning through generative replay and open set recognition. *Journal of Imaging* **8** (2022) 93
- [32] Wang, Z., Zhang, Z., Lee, C.Y., Zhang, H., Sun, R., Ren, X., Su, G., Perot, V., Dy, J., Pfister, T.: Learning to prompt for continual learning. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. (2022) 139–149
- [33] Housby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., Gelly, S.: Parameter-efficient transfer learning for nlp. In: *International conference on machine learning*, PMLR (2019) 2790–2799
- [34] Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., Neubig, G.: Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys* **55** (2023) 1–35
- [35] Li, Y., Wang, X., Yang, X., Bonsangue, M., Zhang, J., Li, T.: Improving open-world continual learning under the constraints of scarce labeled data. In: *Proceedings of the 31st*

- ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. (2025)
- [36] Fei, N., Gao, Y., Lu, Z., Xiang, T.: Z-score normalization, hubness, and few-shot learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. (2021) 142–151
 - [37] Trosten, D.J., Chakraborty, R., Løkse, S., Wickstrøm, K.K., Jenssen, R., Kampffmeyer, M.C.: Hubs and hyperspheres: Reducing hubness and improving transductive few-shot learning with hyperspherical embeddings. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2023) 7527–7536
 - [38] Li, Y., Lai, G., Yang, X., Li, Y., Bonsangue, M., Li, T.: Exploring open-world continual learning with knowns-unknowns knowledge transfer. arXiv:2502.20124 (2025)
 - [39] Li, Y., Yang, Y., Gao, Q., Yang, X.: Cross-regional fraud detection via continual learning (special program). In: Proceedings of the AAAI Conference on Artificial Intelligence. Volume 37. (2023) 16260–16261
 - [40] Li, Y., Yang, X., Gao, Q., Wang, H., Zhang, J., Li, T.: Cross-regional fraud detection via continual learning with knowledge transfer. IEEE Transactions on Knowledge and Data Engineering (2024)
 - [41] Shin, H., Lee, J.K., Kim, J., Kim, J.: Continual learning with deep generative replay. Advances in Neural Information Processing Systems **30** (2017)
 - [42] Rebuffi, S.A., Kolesnikov, A., Sperl, G., Lampert, C.H.: icarl: Incremental classifier and representation learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2017) 2001–2010
 - [43] Serra, J., Suris, D., Miron, M., Karatzoglou, A.: Overcoming catastrophic forgetting with hard attention to the task. In: International conference on machine learning, PMLR (2018) 4548–4557
 - [44] Lopez-Paz, D., Ranzato, M.: Gradient episodic memory for continual learning. Advances in Neural Information Processing Systems **30** (2017)
 - [45] Mirzadeh, S.I., Farajtabar, M., Pascanu, R., Ghasemzadeh, H.: Understanding the role of training regimes in continual learning. Advances in Neural Information Processing Systems **33** (2020) 7308–7320
 - [46] Ahn, H., Cha, S., Lee, D., Moon, T.: Uncertainty-based continual learning with adaptive regularization. Advances in Neural Information Processing Systems **32** (2019)
 - [47] Chaudhry, A., Ranzato, M., Rohrbach, M., Elhoseiny, M.: Efficient lifelong learning with a-gem. In: International Conference on Learning Representations. (2018)
 - [48] Yang, R., Wang, S., Zhang, H., Xu, S., Guo, Y., Ye, X., Hou, B., Jiao, L.: Knowledge decomposition and replay: A novel cross-modal image-text retrieval continual learning method. In: Proceedings of the 31st ACM International Conference on Multimedia, New York, NY, USA, Association for Computing Machinery (2023) 6510–6519
 - [49] Li, Z., Hoiem, D.: Learning without forgetting. IEEE Transactions on Pattern Analysis and Machine Intelligence **40** (2017) 2935–2947
 - [50] Wu, Y., Chen, Y., Wang, L., Ye, Y., Liu, Z., Guo, Y., Fu, Y.: Large scale incremental learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2019) 374–382
 - [51] Lai, G., Li, Y., Wang, X., Zhang, J., Li, T., Yang, X.: Order-robust class incremental learning: Graph-driven dynamic similarity grouping. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. (2025)
 - [52] Xu, J., Zhu, Z.: Reinforced continual learning. Advances in Neural Information Processing Systems **31** (2018)
 - [53] Aljundi, R., Chakravarty, P., Tuytelaars, T.: Expert gate: Lifelong learning with a network of experts. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2017) 3366–3375
 - [54] Mallya, A., Lazebnik, S.: Packnet: Adding multiple tasks to a single network by iterative

- pruning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2018) 7765–7773
- [55] Joseph, K.J., Khan, S., Khan, F.S., Balasubramanian, V.N.: Towards open world object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2021) 5830–5840
- [56] Dhamija, A.R., Günther, M., Boulton, T.: Reducing network agnostophobia. *Advances in Neural Information Processing Systems* **31** (2018)
- [57] Bendale, A., Boulton, T.E.: Towards open set deep networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2016) 1563–1572
- [58] Shu, L., Xu, H., Liu, B.: Doc: Deep open classification of text documents. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. (2017) 2911–2916
- [59] Dang, S., Cao, Z., Cui, Z., Pi, Y., Liu, N.: Open set incremental learning for automatic target recognition. *IEEE Transactions on Geoscience and Remote Sensing* **57** (2019) 4445–4456
- [60] Zhu, X., Zhang, R., He, B., Guo, Z., Zeng, Z., Qin, Z., Zhang, S., Gao, P.: Pointclip v2: Prompting clip and gpt for powerful 3d open-world learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. (2023) 2639–2650
- [61] Zhao, X., Ma, Y., Wang, D., Shen, Y., Qiao, Y., Liu, X.: Revisiting open world object detection. *IEEE Transactions on Circuits and Systems for Video Technology* (2023)
- [62] Liang, S., Li, Y., Srikant, R.: Enhancing the reliability of out-of-distribution image detection in neural networks. In: International Conference on Learning Representations. (2018)
- [63] Liu, W., Wang, X., Owens, J., Li, Y.: Energy-based out-of-distribution detection. *Advances in Neural Information Processing Systems* **33** (2020) 21464–21475
- [64] Lin, Z., Roy, S.D., Li, Y.: Mood: Multi-level out-of-distribution detection. In: Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition. (2021) 15313–15323
- [65] Djuricic, A., Bozanic, N., Ashok, A., Liu, R.: Extremely simple activation shaping for out-of-distribution detection. In: The Eleventh International Conference on Learning Representations. (2023)
- [66] Morteza, P., Li, Y.: Provable guarantees for understanding out-of-distribution detection. In: Proceedings of the AAAI Conference on Artificial Intelligence. Volume 36. (2022) 7831–7840
- [67] Hsu, Y.C., Shen, Y., Jin, H., Kira, Z.: Generalized odin: Detecting out-of-distribution image without learning from out-of-distribution data. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. (2020) 10951–10960
- [68] Lee, K., Lee, K., Min, K., Zhang, Y., Shin, J., Lee, H.: Hierarchical novelty detection for visual object recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2018) 1034–1042
- [69] Huang, R., Li, Y.: Mos: Towards scaling out-of-distribution detection for large semantic space. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2021) 8710–8719
- [70] Linderman, R., Zhang, J., Inkawich, N., Li, H., Chen, Y.: Fine-grain inference on out-of-distribution data with hierarchical classification. In: Conference on Lifelong Learning Agents, PMLR (2023) 162–183
- [71] Liu, X., Zhai, J.T., Bagdanov, A.D., Li, K., Cheng, M.M.: Task-adaptive saliency guidance for exemplar-free class incremental learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2024) 23954–23963
- [72] Wang, Y., Ding, Y., He, X., Fan, X., Lin, C., Li, F., Wang, T., Luo, Z., Luo, J.: Novelty detection and online learning for chunk data streams. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **43** (2020) 2400–2412
- [73] Yu, L., Liu, X., Van de Weijer, J.: Self-training for class-incremental semantic segmentation. *IEEE Transactions on Neural Networks and Learning Systems* **34** (2022) 9116–9127

- [74] Chan, R., Rottmann, M., Gottschalk, H.: Entropy maximization and meta classification for out-of-distribution detection in semantic segmentation. In: *Proceedings of the IEEE/CVF international conference on computer vision*. (2021) 5128–5137
- [75] Mazumder, S., Liu, B.: *Lifelong and Continual Learning Dialogue Systems*. Springer Nature (2024)
- [76] Schlachter, P., Liao, Y., Yang, B.: Deep open set recognition using dynamic intra-class splitting. *SN Computer Science* **1** (2020) 1–12
- [77] Aljundi, R., Reino, D.O., Chumerin, N., Turner, R.E.: Continual novelty detection. In: *Conference on Lifelong Learning Agents*, PMLR (2022) 1004–1025
- [78] Oza, P., Patel, V.M.: C2ae: Class conditioned auto-encoder for open-set recognition. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. (2019) 2307–2316
- [79] Lee, K., Lee, K., Lee, H., Shin, J.: A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In: *Advances in Neural Information Processing Systems*. Volume 31. (2018) 7167–7177
- [80] Hendrycks, D., Gimpel, K.: A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv:1610.02136* (2016)
- [81] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv:2010.11929* (2020)
- [82] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: *Advances in Neural Information Processing Systems*. (2017) 5998–6008
- [83] Wang, Z., Zhang, Z., Ebrahimi, S., Sun, R., Zhang, H., Lee, C.Y., Ren, X., Su, G., Perot, V., Dy, J., et al.: Dualprompt: Complementary prompting for rehearsal-free continual learning. In: *European conference on computer vision*, Springer (2022) 631–648
- [84] Wang, L., Xie, J., Zhang, X., Huang, M., Su, H., Zhu, J.: Hierarchical decomposition of prompt-based continual learning: Rethinking obscured sub-optimality. *Advances in Neural Information Processing Systems* **36** (2023) 69054–69076
- [85] Lester, B., Al-Rfou, R., Constant, N.: The power of scale for parameter-efficient prompt tuning. *arXiv:2104.08691* (2021)
- [86] Li, X.L., Liang, P.: Prefix-tuning: Optimizing continuous prompts for generation. In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. (2021) 4582–4597
- [87] Zhou, D.W., Sun, H.L., Ning, J., Ye, H.J., Zhan, D.C.: Continual learning with pre-trained models: a survey. In: *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*. (2024) 8363–8371
- [88] Smith, J.S., Karlinsky, L., Gutta, V., Cascante-Bonilla, P., Kim, D., Arbelle, A., Panda, R., Feris, R., Kira, Z.: Coda-prompt: Continual decomposed attention-based prompting for rehearsal-free continual learning. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. (2023) 11909–11919
- [89] Jung, D., Han, D., Bang, J., Song, H.: Generating instance-level prompts for rehearsal-free continual learning. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. (2023) 11847–11857
- [90] Wang, Y., Huang, Z., Hong, X.: S-prompts learning with pre-trained transformers: An occam’s razor for domain incremental learning. *Advances in Neural Information Processing Systems* **35** (2022) 5682–5695
- [91] Zhou, D.W., Cai, Z.W., Ye, H.J., Zhan, D.C., Liu, Z.: Revisiting class-incremental learning with pre-trained models: Generalizability and adaptivity are all you need. *International Journal of Computer Vision* **133** (2025) 1012–1032

- [92] Zhou, D.W., Zhang, Y., Wang, Y., Ning, J., Ye, H.J., Zhan, D.C., Liu, Z.: Learning without forgetting for vision-language models. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2025)
- [93] Zhou, D.W., Sun, H.L., Ye, H.J., Zhan, D.C.: Expandable subspace ensemble for pre-trained model-based class-incremental learning. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. (2024) 23554–23564
- [94] Chen, S., Ge, C., Tong, Z., Wang, J., Song, Y., Wang, J., Luo, P.: Adaptformer: Adapting vision transformers for scalable visual recognition. *Advances in Neural Information Processing Systems* **35** (2022) 16664–16678
- [95] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: *Proceedings of the 38th International Conference on Machine Learning*. Volume 139 of *Proceedings of Machine Learning Research*. (2021) 8748–8763
- [96] Aljundi, R., Babiloni, F., Elhoseiny, M., Rohrbach, M., Tuytelaars, T.: Memory aware synapses: Learning what (not) to forget. In: *Proceedings of the European conference on computer vision (ECCV)*. (2018) 139–154
- [97] Quanz, B., Huan, J., Mishra, M.: Knowledge transfer with low-quality data: A feature extraction issue. *IEEE Transactions on Knowledge and Data Engineering* **24** (2012) 1789–1802
- [98] Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images. Master’s thesis, University of Toronto (2009)
- [99] Ebrahimi, S., Meier, F., Calandra, R., Darrell, T., Rohrbach, M.: Adversarial continual learning. In: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, Springer (2020) 386–402
- [100] Yu, Q., Aizawa, K.: Unsupervised out-of-distribution detection by maximum classifier discrepancy. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. (2019) 9518–9526
- [101] Basart, S., Mantas, M., Mohammadreza, M., Jacob, S., Dawn, S.: Scaling out-of-distribution detection for real-world settings. In: *International Conference on Machine Learning*. (2022)
- [102] Buzzega, P., Boschini, M., Porrello, A., Abati, D., Calderara, S.: Dark experience for general continual learning: a strong, simple baseline. *Advances in Neural Information Processing Systems* **33** (2020) 15920–15930
- [103] Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. *Journal of Machine Learning Research* **9** (2008)
- [104] McInnes, L., Healy, J., Saul, N., Großberger, L.: Umap: Uniform manifold approximation and projection. *Journal of Open Source Software* **3** (2018)
- [105] Ahmad, T., Dhamija, A.R., Cruz, S., Rabinowitz, R., Li, C., Jafarzadeh, M., Boulton, T.E.: Few-shot class incremental learning leveraging self-supervised features. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. (2022) 3900–3910
- [106] Yoon, J., Madjid, S., Hwang, S.J., Yoo, C.D., et al.: On the soft-subnetwork for few-shot class incremental learning. In: *International Conference on Learning Representations, International Conference on Learning Representations* (2023)
- [107] Qin, C., Joty, S.: Lfpt5: A unified framework for lifelong few-shot language learning based on prompt tuning of t5. In: *International Conference on Learning Representations*. (2021)
- [108] Tao, X., Hong, X., Chang, X., Dong, S., Wei, X., Gong, Y.: Few-shot class-incremental learning. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. (2020) 12183–12192
- [109] Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The caltech-ucsd birds-200-2011 dataset (2011)
- [110] Vinyals, O., Blundell, C., Lillicrap, T., Wierstra, D., et al.: Matching networks for one shot

- learning. *Advances in Neural Information Processing Systems* **29** (2016)
- [111] Hendrycks, D., Basart, S., Mazeika, M., Zou, A., Kwon, J., Mostajabi, M., Steinhardt, J., Song, D.: Scaling out-of-distribution detection for real-world settings. In: *International Conference on Machine Learning*, PMLR (2022) 8759–8773
 - [112] Sehwag, V., Chiang, M., Mittal, P.: Ssd: A unified framework for self-supervised. In: *International Conference on Machine Learning*. (2021)
 - [113] Wang, H., Li, Z., Feng, L., Zhang, W.: Vim: Out-of-distribution with virtual-logit matching. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. (2022) 4921–4930
 - [114] Sun, Y., Ming, Y., Zhu, X., Li, Y.: Out-of-distribution detection with deep nearest neighbors. In: *International Conference on Machine Learning*, PMLR (2022) 20827–20840
 - [115] Park, J., Jung, Y.G., Teoh, A.B.J.: Nearest neighbor guidance for out-of-distribution detection. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. (2023) 1686–1695
 - [116] Zhang, C., Song, N., Lin, G., Zheng, Y., Pan, P., Xu, Y.: Few-shot incremental learning with continually evolved classifiers. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. (2021) 12455–12464
 - [117] Chi, Z., Gu, L., Liu, H., Wang, Y., Yu, Y., Tang, J.: Metafscl: A meta-learning approach for few-shot class incremental learning. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. (2022) 14166–14175
 - [118] Peng, C., Zhao, K., Wang, T., Li, M., Lovell, B.C.: Few-shot class-incremental learning from an open-set perspective. In: *European Conference on Computer Vision*, Springer (2022) 382–397
 - [119] Zhou, D.W., Ye, H.J., Ma, L., Xie, D., Pu, S., Zhan, D.C.: Few-shot class-incremental learning by sampling multi-phase tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2022)
 - [120] Yang, Y., Yuan, H., Li, X., Lin, Z., Torr, P., Tao, D.: Neural collapse inspired feature-classifier alignment for few-shot class-incremental learning. In: *The Eleventh International Conference on Learning Representations*. (2022)
 - [121] Ji, Z., Hou, Z., Liu, X., Pang, Y., Li, X.: Memorizing complementation network for few-shot class-incremental learning. *IEEE Transactions on Image Processing* **32** (2023) 937–948
 - [122] Wang, X., Liu, Y., Fan, J., Wen, W., Xue, H., Zhu, W.: Continual few-shot learning with transformer adaptation and knowledge regularization. In: *Proceedings of the ACM Web Conference*. Volume 2023. (2023)
 - [123] Li, J., Bai, Y., Lou, Y., Linghu, X., He, J., Xu, S., Bai, T.: Memory-based label-text tuning for few-shot class-incremental learning. *arXiv:2207.01036* (2022)
 - [124] Agarwal, A., Banerjee, B., Cuzzolin, F., Chaudhuri, S.: Semantics-driven generative replay for few-shot class incremental learning. In: *Proceedings of the 30th ACM International Conference on Multimedia*. (2022) 5246–5254
 - [125] D’Alessandro, M., Alonso, A., Calabrés, E., Galar, M.: Multimodal parameter-efficient few-shot class incremental learning. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. (2023) 3393–3403
 - [126] Zhu, P., Zhang, W., Wang, Y., Hu, Q.: Multi-granularity inter-class correlation based contrastive learning for open set recognition. *Int. J. Softw. Informatics* **12** (2022) 157–175
 - [127] Huang, H., Wang, Y., Hu, Q., Cheng, M.M.: Class-specific semantic reconstruction for open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **45** (2022) 4214–4228
 - [128] Zhou, Y., Liu, P., Qiu, X.: Knn-contrastive learning for out-of-domain intent classification. In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. (2022) 5129–5141
 - [129] van de Ven, G.M., Tolias, A.S.: Three continual learning scenarios. In: *Advances in Neural*

- Information Processing Systems. Volume 1(9). (2018)
- [130] McDonnell, M.D., Gong, D., Parvaneh, A., Abbasnejad, E., van den Hengel, A.: Ranpac: Random projections and pre-trained models for continual learning. *Advances in Neural Information Processing Systems* **36** (2024)
 - [131] Zhou, D.W., Ye, H.J., Zhan, D.C., Liu, Z.: Revisiting class-incremental learning with pre-trained models: Generalizability and adaptivity are all you need. *arXiv:2303.07338* (2023)
 - [132] Fang, Z., Lu, J., Liu, A., Liu, F., Zhang, G.: Learning bounds for open-set learning. In: *International conference on machine learning*, PMLR (2021) 3122–3132
 - [133] Shi, H., Wang, H.: A unified approach to domain incremental learning with memory: Theory and algorithm. *Advances in Neural Information Processing Systems* **36** (2024)
 - [134] Anthony, M., Bartlett, P.L., Bartlett, P.L., et al.: *Neural network learning: Theoretical foundations*. Volume 9. cambridge university press Cambridge (1999)
 - [135] Lian, D., Zhou, D., Feng, J., Wang, X.: Scaling & shifting your features: A new baseline for efficient model tuning. *Advances in Neural Information Processing Systems* **35** (2022) 109–123
 - [136] Jia, M., Tang, L., Chen, B.C., Cardie, C., Belongie, S., Hariharan, B., Lim, S.N.: Visual prompt tuning. In: *European Conference on Computer Vision*, Springer (2022) 709–727
 - [137] Chen, C.P.: A rapid supervised learning neural network for function interpolation and approximation. *IEEE Transactions on Neural Networks* **7** (1996) 1220–1230
 - [138] Huang, G.B., Zhu, Q.Y., Siew, C.K.: Extreme learning machine: theory and applications. *Neurocomputing* **70** (2006) 489–501
 - [139] Lomonaco, V., Maltoni, D.: Core50: a new dataset and benchmark for continuous object recognition. In: *Conference on robot learning*, PMLR (2017) 17–26
 - [140] Breunig, M.M., Kriegel, H.P., Ng, R.T., Sander, J.: Lof: identifying density-based local outliers. In: *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*. (2000) 93–104
 - [141] Shi, H., Tayebi, M.A., Pei, J., Cao, J.: Cost-sensitive learning for medical insurance fraud detection with temporal information. *IEEE Transactions on Knowledge and Data Engineering* (2023) 1–14
 - [142] Kerpici, M., Ozkan, H., Kozat, S.S.: Online anomaly detection with bandwidth optimized hierarchical kernel density estimators. *IEEE Transactions on Neural Networks and Learning Systems* **32** (2021) 4253–4266
 - [143] Ma, X., Wu, J., Xue, S., Yang, J., Zhou, C., Sheng, Q., Xiong, H., Akoglu, L.: A comprehensive survey on graph anomaly detection with deep learning. *IEEE Transactions on Knowledge and Data Engineering* (2021)
 - [144] Li, Z., Liu, G., Jiang, C.: Deep representation learning with full center loss for credit card fraud detection. *IEEE Transactions on Computational Social Systems* **7** (2020) 569–579
 - [145] Fanai, H., Abbasimehr, H.: A novel combined approach based on deep autoencoder and deep classifiers for credit card fraud detection. *Expert Systems with Applications* (2023) 119562
 - [146] Chung, J., Gulcehre, C., Cho, K., Bengio, Y.: Empirical evaluation of gated recurrent neural networks on sequence modeling. In: *NIPS 2014 Workshop on Deep Learning*, December 2014. (2014)
 - [147] Benchaji, I., Douzi, S., El Ouahidi, B., Jaafari, J.: Enhanced credit card fraud detection based on attention mechanism and lstm deep model. *Journal of Big Data* **8** (2021) 1–21
 - [148] Xie, Y., Liu, G., Yan, C., Jiang, C., Zhou, M.: Time-aware attention-based gated network for credit card fraud detection by extracting transactional behaviors. *IEEE Transactions on Computational Social Systems* (2022)
 - [149] Ergen, T., Kozat, S.S.: Unsupervised anomaly detection with lstm neural networks. *IEEE Transactions on Neural Networks and Learning Systems* **31** (2019) 3127–3141
 - [150] Wu, W., He, L., Lin, W., Su, Y., Cui, Y., Maple, C., Jarvis, S.: Developing an unsupervised

- real-time anomaly detection scheme for time series with multi-seasonality. *IEEE Transactions on Knowledge and Data Engineering* **34** (2020) 4147–4160
- [151] Wang, D., Lin, J., Cui, P., Jia, Q., Wang, Z., Fang, Y., Yu, Q., Zhou, J., Yang, S., Qi, Y.: A semi-supervised graph attentive network for financial fraud detection. In: *ICDM, IEEE* (2019) 598–607
 - [152] Kipf, T., Welling, M.: Semi-supervised classification with graph convolutional networks. In: *International Conference on Learning Representations*. (2016)
 - [153] Hamilton, W.L., Ying, R., Leskovec, J.: Inductive representation learning on large graphs. In: *NIPS*. (2017) 1025–1035
 - [154] Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., Bengio, Y.: Graph attention networks. In: *International Conference on Learning Representations*. (2018)
 - [155] Peng, L., Mo, Y., Xu, J., Shen, J., Shi, X., Li, X., Shen, H.T., Zhu, X.: Grlc: Graph representation learning with constraints. *IEEE Transactions on Neural Networks and Learning Systems* (2023)
 - [156] Wang, X., Ji, H., Shi, C., Wang, B., Ye, Y., Cui, P., Yu, P.S.: Heterogeneous graph attention network. In: *The world wide web conference*. (2019) 2022–2032
 - [157] Cheng, D., Wang, X., Zhang, Y., Zhang, L.: Graph neural network for fraud detection via spatial-temporal attention. *IEEE Transactions on Knowledge and Data Engineering* (2020)
 - [158] Hei, Y., Yang, R., Peng, H., Wang, L., Xu, X., Liu, J., Liu, H., Xu, J., Sun, L.: Hawk: Rapid android malware detection through heterogeneous graph attention networks. *IEEE Transactions on Neural Networks and Learning Systems* (2021)
 - [159] Wang, C., Dou, Y., Chen, M., Chen, J., Liu, Z., Philip, S.Y.: Deep fraud detection on non-attributed graph. In: *Big Data, IEEE* (2021) 5470–5473
 - [160] Chen, Z., Sun, A.: Anomaly detection on dynamic bipartite graph with burstiness. In: *2020 IEEE International Conference on Data Mining (ICDM), IEEE* (2020) 966–971
 - [161] Gan, J., Hu, R., Mo, Y., Kang, Z., Peng, L., Zhu, Y., Zhu, X.: Multigraph fusion for dynamic graph convolutional network. *IEEE Transactions on Neural Networks and Learning Systems* **35** (2022) 196–207
 - [162] Zhang, X., Song, D., Tao, D.: Hierarchical prototype networks for continual graph representation learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **45** (2023) 4622–4636
 - [163] Dou, Y., Liu, Z., Sun, L., Deng, Y., Peng, H., Yu, P.S.: Enhancing graph neural network-based fraud detectors against camouflaged fraudsters. In: *CIKM*. (2020) 315–324
 - [164] Lin, H., Liu, G., Wu, J., Zuo, Y., Wan, X., Li, H.: Fraud detection in dynamic interaction network. *IEEE Transactions on Knowledge and Data Engineering* **32** (2020) 1936–1950
 - [165] Wang, J., Song, G., Wu, Y., Wang, L.: Streaming graph neural networks via continual learning. In: *Proceedings of the 29th ACM international conference on information & knowledge management*. (2020) 1515–1524
 - [166] Zhou, F., Cao, C.: Overcoming catastrophic forgetting in graph neural networks with experience replay. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Number 5 (2021) 4714–4722
 - [167] Zhang, P., Yan, Y., Li, C., Wang, S., Xie, X., Song, G., Kim, S.: Continual learning on dynamic graphs via parameter isolation. In: *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. (2023) 601–611
 - [168] Zhu, Y., Ma, J., Yuan, C., Zhu, X.: Interpretable learning based dynamic graph convolutional networks for alzheimer’s disease analysis. *Information Fusion* **77** (2022) 53–61
 - [169] Liu, Z., Chen, C., Yang, X., Zhou, J., Li, X., Song, L.: Heterogeneous graph neural networks for malicious account detection. In: *Proceedings of the 27th ACM international conference on information and knowledge management*. (2018) 2077–2085
 - [170] Mo, Y., Chen, Y., Lei, Y., Peng, L., Shi, X., Yuan, C., Zhu, X.: Multiplex graph represen-

- tation learning via dual correlation reduction. *IEEE Transactions on Knowledge and Data Engineering* **35** (2023) 12814–12827
- [171] Perozzi, B., Al-Rfou, R., Skiena, S.: Deepwalk: online learning of social representations. In: *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM (2014) 701–710
- [172] Zhang, J., Zhang, J., Ghosh, S., Li, D., Zhu, J., Zhang, H., Wang, Y.: Regularize, expand and compress: Nonexpansive continual learning. In: *Proceedings of the IEEE/CVF winter conference on applications of computer vision*. (2020) 854–862
- [173] Tang, B., Matteson, D.S.: Graph-based continual learning. In: *International Conference on Learning Representations*. (2020)

List of Figures

1.1	A conceptual framework of continual learning. a. Continual learning requires adapting to incremental tasks with dynamic data distributions. b. A desirable solution should ensure an appropriate trade-off between stability (orange arrow) and plasticity (green arrow), as well as a generalizability to intra-task (blue arrow) and inter-task (red arrow) distribution differences.	2
1.2	An illustration of the open-world continual learning problem and comparison between continual learning. In open-world continual learning, the test set of each task may include samples from classes unseen during training, deviating from the closed-world assumption of traditional continual learning. As shown in the figure, this setting imposes two key challenges: (1) preserving knowledge from previously learned tasks to mitigate catastrophic forgetting, and (2) performing robust open-set recognition to correctly identify and reject instances from unknown classes. While continual learning models tend to misclassify open samples as one of the known classes (e.g., the green pentagon and orange square), open-world continual learning models are explicitly designed to distinguish between known and unknown classes, thereby enabling more reliable and adaptive decision-making in dynamic environments.	4
3.1	In contrast to rehearsal-based continual learning approaches that sequentially adapt full or partial model parameters to new tasks and rely on a rehearsal buffer to alleviate catastrophic forgetting, prompt-based methods adopt a unified backbone model augmented by learned prompts [32, 83]. Task-specific information is encapsulated within the prompts, thereby eliminating the necessity of a rehearsal buffer. Furthermore, different methods have different ways of selecting and updating prompts dynamically, usually enabling task-agnostic inference at test time without explicit task identity.	32
3.2	Different kinds of prompt selection, including key-value matching, shared and task-specific retrieval, attention-based combination, and instance-specific prompt generation.	33

4.1	A simple illustration of OWCL.	37
4.2	Distributions of the outputted logits scores of test samples, using MAS (a typical continual learning baseline) and the proposed <i>Pro-KT</i> on Task 1 and Task 2, respectively. The x -axis is the unscaled class-maximum logits of samples, and the y -axis shows the number of samples.	38
4.3	Overview of the proposed <i>Pro-KT</i> framework. Left: Training phase. Right: Testing phase. During training, input samples are first projected using a query function. <i>Pro-KT</i> learns a new set of prompts and stores them in a prompt bank. Each sample is then enriched with selected prompts, and the resulting prompt-enhanced embeddings are processed by a frozen pre-trained backbone. The final representations are passed to a trainable classifier. Additionally, an adaptive threshold for a task-aware open-set boundary is learned based on the softmax entropy of these enhanced representations. In the <i>testing phase</i> , test samples are similarly projected and matched with prompts from the prompt bank using a sample-wise selection strategy, enabling transfer of both task-specific and generic knowledge. The enriched embeddings are fed into the unified classifier to obtain unscaled logit scores. Finally, <i>Pro-KT</i> classifies each sample as either <i>[known]</i> or <i>[unknown]</i> , using the learned adaptive threshold.	42
4.4	(a) and (b): Variation curve of A_N on Split CIFAR-100 and 5-Datasets, the x -axis indicates the task IDs. (c) and (d): Variation curve of F_N on Split CIFAR-100 and 5-Datasets, the x -axis indicates the task IDs.	53
4.5	Visualization of the prompt bank via T-SNE and UMAP after training on all tasks (Split CIFAR-100 dataset).	55
4.6	Left and middle are performance heatmaps (A_N) showing interaction effects between prompt length (L_P , y -axis) and selection size (K , x -axis) on both datasets, with fixed $M = 25$. Right is A_N variation across prompt counts (M) with optimal $L_P = 5$ and $K = 3$	56
5.1	Open-world few-shot continual learning framework overview. The model incrementally learns new tasks (task 1, ..., task n , etc.) with few-shot class examples (class 3, ..., class i , etc.) while simultaneously detecting unknown samples during testing and overcoming severe forgetting and overfitting challenges inherent in data-scarce continual learning scenarios.	60

5.2	The overall <i>PEAK</i> model consists of three key components: (1) ITA (Yellow): enhancing the samples by matching them with appropriate additional tokens; (2) MOB (Purple): constructing compact decision boundaries of knowns by margin-based loss function for open detection; (3) AKS (Red): incorporating knowledge learned from both knowns and unknowns, and facilitating the transition from unknowns to knowns.	64
5.3	Visualizations of model performance on known classification.	76
5.4	(a): ACC_N w.r.t token length L_P and additional token size K , given $l = 25$. (b): ACC_N w.r.t. l (i.e., the total number of additional tokens of each task) with $L_P = 5$ and $K = 5$ (take CUB200 for illustration).	78
5.5	(a) and (b): AUC_N and FPR_N w.r.t constraint governing deviations σ and margin m (CUB200 dataset).	78
5.6	Visual comparison of test sample representations: (left) baseline Transformer backbone versus (right) our proposed <i>PEAK</i> framework.	79
6.1	A toy example of the KIRO scenario in OWCL. During training Task 1 and Task t , Robots A and B learn <i>Golden Retrievers</i> and <i>Labradors</i> , both labeled as $[dog]$. In testing Task 1, both robots correctly classified $[dog]$ and $[bus]$ samples while identifying the unknown wolf without errors. However, in testing Task t , due to the emergence of <i>Labrador</i> , Robot A experienced distributional shifts when recognizing <i>Golden Retrievers</i> . This also led to the misclassification of <i>Wolves</i> as <i>Labradors</i> , likely due to visual similarity. In contrast, Robot B accurately classifies all known samples and consistently detects the wolf as an open sample via knowledge transfer.	82
6.2	From task t to task $t + 1$ in the KIRO scenario, we illustrate the embedding space using different baselines. Red circles and red triangles denote unknown samples encountered in tasks t and $t - 1$, respectively. Orange, blue, and green circles represent labeled training data learned during tasks $t - 1$, t , and $t + 1$, respectively. Dashed lines denote the decision boundaries inferred by the model during open-world continual learning. (a) The original embedding distributions for each task learned by the classical continual learning baseline EWC. (b) A current competitive open-world continual learning approach enforces tighter embedding cohesion for each task. (c) Our proposed method jointly addresses open detection and incrementally reduces known sample classification errors through knowledge transfer. [Best view in color]	84
6.3	Unscaled classification scores on the CINO scenario.	91
6.4	Unscaled classification scores on the CIRO scenario.	91
6.5	Unscaled classification scores on the KINO Scenario.	92
6.6	Unscaled classification scores on the KIRO Scenario.	93

6.7	Visualizations of main results.	109
6.8	Performance with varying NRP size.	111
6.9	The divergence between the actual embedding distributions and the <i>HoliTrans</i> 's embedding distributions with distribution-aware prototypes (The horizontal axis is the values of r , the vertical axis is ACC_t , and the dashed line indicates the optimal value of r). Orange represents the actual ones, blue represents ours.	112
7.1	Bottom: An illustration of cross-regional fraud detection. The orange arrows indicate a financial business expansion from Region 1 to Region 5, where the data characteristics vary across different regions with different patterns. Our proposed method aims to incrementally learn a unified fraud detection model that can perform well in each task through knowledge transfer. Top: The histograms in different colors denote each corresponding region's fraud detection performance (ACC) with training from scratch. The performance drops significantly after a model is trained using data from Region 2. Similarly, after training a model with data from Region 3, the performance of Region 1 and Region 2 fell further, respectively. . . .	117
7.2	Heat maps of temporal slices. On the horizontal axis is the amount of fraudulent transactions, and on the vertical axis are time slices. The darker the color, the more fraudulent transactions occurred during the current period.	124
7.3	The overall framework of the proposed CCL. Given a sequence of cross-regional fraud detection tasks, the proposed prototype-based replay method for knowledge transfer and the forgetting prevention approach are illustrated. Take the learning process of $\mathcal{G}_2$ as an example. In stage 1 of knowledge transfer, we randomly choose some target nodes from $\mathcal{G}_1$ and generate a prototype $\mathbf{c}^1$. In stage 2, a Gaussian prior is added to perturb the prototype. In stage 3, nodes are sorted according to the distance from the perturbed prototype $\mathbf{c}^1$, and we select the replay nodes according to the replay ratio r . Then, all the selected nodes and their neighbors are stored in the experience buffer $\mathbb{B}_1$. Next, the replay buffer is appended to $\mathcal{G}_2$ for the training, with parameter smoothing for forgetting prevention. Finally, the final embedding of $\mathbf{Z}_t$ is obtained by the GNN-based backbone accordingly.	125
7.4	Left: Variation curves of recall. Middle: Variation curves of AUC. Right: Variation curves of F1.	134
7.5	Sensitivity analysis of hyperparameters. We vary the hidden size $ \mathbf{h} $ as $\{8, 16, 24, 32\}$, the smoothing term λ as $\{1.5, 15, 150, 1500\}$, the smoothing term γ as $\{0.1, 0.01, 0.001, 0.0001\}$, the replay ratio r as $\{0.01, 0.05, 0.1, 0.15\}$	138

- 7.6 Node-type importance analysis via SHAP values. Each dot represents a sample's SHAP value for a specific node type, indicating the feature's impact on the model's output. Higher SHAP values (toward the right) denote stronger positive contributions to fraud detection, while lower values (toward the left) denote negative influence. Node types with greater vertical dispersion and right-shifted distributions are more influential in the model's decision-making. The color of each point corresponds to the masking level of that node type, where **red** indicates a higher feature value (i.e., lower masking), and **blue** indicates a lower feature value (i.e., more heavily masked). It can be observed that **Customer** nodes are the most impactful, followed by **Time Slice** and **Merchant**, underscoring the importance of temporal and behavioral semantics in cross-regional fraud detection. . . . 140

List of Tables

1.1	Overview of different OWCL scenarios.	5
2.1	Comparison of continual learning, open-set learning, and open-world continual learning.	21
3.1	Overview of the three CL scenarios.	25
3.2	Comparison between OSL and OOD Detection.	27
3.3	Notations used in the OWCL problem formulation.	29
4.1	Notation used in this chapter.	41
4.2	Evaluation results of unknown detection performance (%) across 10 sequential tasks (10 classes each) on Split CIFAR-100 and 5 tasks on 5-Datasets.	51
4.3	Classification accuracy (%) on known classes for Split CIFAR-100 and 5-Datasets benchmarks. Experiments evaluate continual learning performance across 10 sequential tasks (10 classes each) on Split CIFAR-100, and 5 sequential tasks for 5 datasets. We used the ResNet32 and ViT backbones in the experiments.	52
4.4	Ablation study on the variants of <i>Pro-KT</i>	54
5.1	Notations and explanations.	63
5.2	Results (%) regarding unknown detection. We report the results over 10 tasks for CUB200 (10-way 5-shot) and 8 tasks for MiniImageNet (5-way 5-shot).	74
5.3	Averaged classification accuracy (%) on CUB200 (10-way 5-shot).	75
5.4	Averaged classification accuracy (%) on MiniImageNet (5-way 5-shot).	76
5.5	Ablation studies.	77
5.6	Performance under different γ configurations on the CUB200 dataset.	79
6.1	Overview of the four OWCL scenarios.	83
6.2	Notations and explanations.	87
6.3	Dataset description.	107
6.4	Main Results. We evidence first , <u>second</u> , and <u>third</u> performances.	108
6.5	Main Results. We format first , <u>second</u> , and <u>third</u> performances.	109
6.6	Ablation Study.	110

6.7	Performance on Different NRP Size for CIRO (on Split-CIFAR100 dataset) and KIRO (on Open-CORE50 dataset).	111
7.1	Notations and explanations.	122
7.2	The statistics of datasets for each region.	133
7.3	Recall scores (%) across regions $\mathcal{R}_1$ to $\mathcal{R}_5$	134
7.4	AUC scores (%) across regions $\mathcal{R}_1$ to $\mathcal{R}_5$	135
7.5	F1 scores (%) across regions $\mathcal{R}_1$ to $\mathcal{R}_5$	135
7.6	Ablation study (Recall %) across regions $\mathcal{R}_1$ to $\mathcal{R}_5$	136
7.7	Ablation study (AUC %) across regions $\mathcal{R}_1$ to $\mathcal{R}_5$	136
7.8	Ablation study (F1 %) across regions $\mathcal{R}_1$ to $\mathcal{R}_5$	137

English Summary

This thesis explores and advances Open-World Continual Learning (OWCL), a new learning paradigm that addresses the core limitations of traditional deep learning in dynamic, uncertain, and non-stationary environments. Unlike conventional models that assume closed-world training with fixed task boundaries and complete label space, OWCL seeks to endow learning systems with the ability to continuously adapt to new tasks, detect and accommodate unknown classes, and retain prior knowledge without retraining from scratch. This paradigm fundamentally blends elements of continual learning, open-set recognition, and out-of-distribution generalization, and introduces new conceptual and algorithmic challenges at their intersection.

The theoretical importance of OWCL lies in its disruption of the classical i.i.d. learning assumption, motivating new mathematical formulations of decision spaces, transferability, and uncertainty. From a research perspective, OWCL redefines the structure of lifelong learning and opens new questions on how knowledge should be represented, stored, and adapted over time. From a practical perspective, OWCL is directly relevant to numerous real-world applications such as fraud detection, autonomous systems, clinical diagnostics, and recommendation systems—scenarios where the data evolves, classes appear or disappear, and full supervision is rarely guaranteed.

To systematically address the multifaceted challenges of OWCL, this thesis presents a coherent line of contributions centered around four proposed models, each designed to target a core bottleneck in the field:

- **Pro-KT (Prompt-enhanced Knowledge Transfer)** tackles the fundamental issue of balancing knowledge retention (stability) and plasticity in OWCL. It introduces a principled use of prompt learning to modularize and organize past knowledge into a reusable prompt space, enabling dynamic task adaptation. By leveraging a prompt repository that encodes both task-specific and task-agnostic information, Pro-KT enables flexible transfer across evolving tasks and achieves robust performance without catastrophic forgetting.
- **OFCL (Open-World Few-Shot Continual Learning)** expands OWCL into the low-resource regime, where labeled data is sparse or expensive. OFCL

identifies and mitigates three core challenges: representation collapse, ambiguity in open-set boundaries, and degraded transferability under supervision scarcity. Through a modular design incorporating contrastive learning, prototype replay, and boundary-aware training, our proposed model in this chapter, PEAK, demonstrates the feasibility and importance of scalable OWCL even under extreme label constraints, opening the door for real-world deployments where the annotation is costly or delayed.

- **HoliTrans (Holistic Knowns-Unknowns Knowledge Transfer)** formalizes OWCL at both a theoretical and algorithmic level. It begins by identifying four practically motivated OWCL scenarios that have not been rigorously captured in prior literature, and proposes a unified decision space formulation. Algorithmically, HoliTrans bridges continual learning with open-set detection using Nearest Class Mean (NCM) dynamics and transformer-based architectures. The model offers strong empirical performance while aligning with a theoretically grounded optimization formulation, serving as a comprehensive OWCL solution with provable generalization and boundary consistency.
- **CCL (Cross-regional Continual Learning for Fraud Detection)** demonstrates the practical significance of OWCL in real-world decision-making, specifically through the lens of financial fraud detection. By modeling distribution shifts across geographic regions and temporal phases, CCL introduces a graph-based continual learner that adapts to evolving fraud patterns without discarding prior experience. Its ability to generalize across regional boundaries highlights the robustness and deployability of OWCL principles in dynamic, mission-critical applications.

Collectively, these contributions not only advance the algorithmic frontier of OWCL but also serve to redefine the learning protocol under open-world conditions. The thesis advocates for a shift from monolithic, static models toward modular, adaptive, and transparent learning systems, capable of learning continuously, transferring knowledge purposefully, and responding intelligently to novel conditions.

The work concludes with several forward-looking research directions, including MoE-based modular architectures, PEFT-integrated federated OWCL systems, cross-modal learning under task heterogeneity, and explainable continual reasoning. These proposed directions reflect a broader aspiration: to build intelligent systems that support lifelong learning, that understand their limitations, and that serve real-world goals with robustness, transparency, and adaptability.

Nederlandse Samenvatting

Deze scriptie verkent en ontwikkelt het concept van **Open-World Continual Learning (OWCL)** — een veelbelovend nieuw leerparadigma dat de kernbependingen van traditionele deep learning aanpakt in dynamische, onzekere en niet-stationaire omgevingen. In tegenstelling tot conventionele modellen die uitgaan van een gesloten wereld met vaste taakgrenzen en een volledige labelruimte, streeft OWCL ernaar leersystemen in staat te stellen zich continu aan te passen aan nieuwe taken, onbekende klassen te detecteren en zich hierop aan te passen, en eerder verworven kennis te behouden zonder volledige hertraining.

Dit paradigma combineert fundamenteel elementen van *continual learning*, *open-set recognition* en *out-of-distribution generalisatie*, en introduceert nieuwe conceptuele en algoritmische uitdagingen op het snijvlak van deze gebieden.

De theoretische relevantie van OWCL ligt in het doorbreken van de klassieke *i.i.d.*-aanname (independent and identically distributed), wat aanleiding geeft tot nieuwe wiskundige formuleringen van beslissingsruimten, overdraagbaarheid en onzekerheid. Vanuit een onderzoeksperspectief herdefinieert OWCL de structuur van *lifelong learning* en opent het nieuwe vragen over hoe kennis in de tijd moet worden gerepresenteerd, opgeslagen en aangepast. Vanuit een praktisch perspectief is OWCL direct toepasbaar op talloze realistische scenario's zoals fraudedetectie, autonome systemen, klinische diagnostiek en aanbevelingssystemen — situaties waarin data evolueert, klassen verschijnen of verdwijnen, en volledige supervisie zelden is gegarandeerd.

Om de veelzijdige uitdagingen van OWCL systematisch aan te pakken, presenteert deze scriptie een consistente reeks bijdragen, gecentreerd rond vier voorgestelde modellen, die elk gericht zijn op een cruciale bottleneck binnen het veld:

- **Pro-KT (Prompt-enhanced Knowledge Transfer)** richt zich op de fundamentele uitdaging van het balanceren van kennisbehoud (stabiliteit) en aanpasbaarheid (plasticiteit) in OWCL. Het introduceert een principiële toepassing van *prompt learning* om eerder geleerde kennis modulair te organiseren in een herbruikbare promptruimte, waarmee dynamische taakaanpassing mogelijk wordt. Door een prompt-repository te benutten die zowel taakspecifieke als

taak-onafhankelijke informatie encodeert, maakt Pro-KT flexibele kennisoverdracht mogelijk over veranderende taken en voorkomt het *catastrophic forgetting*.

- **OFCL (Open-World Few-Shot Continual Learning)** breidt OWCL uit naar het *low-resource* domein, waarin gelabelde data schaars of duur is. OFCL identificeert en pakt drie kernproblemen aan: *representation collapse*, onduidelijkheid in *open-set* grenzen en verminderde overdraagbaarheid bij beperkte supervisie. Door een modulaire ontwerpbenadering die *contrastive learning*, *prototype replay* en *boundary-aware training* combineert, toont ons voorgestelde model, PEAK, de haalbaarheid en het belang aan van schaalbare OWCL zelfs onder extreme labelbeperkingen, wat de weg opent naar implementaties in de praktijk waarin annotatie kostbaar of vertraagd is.
- **HoliTrans (Holistic Knowns-Unknowns Knowledge Transfer)** formaliseert OWCL op zowel theoretisch als algoritmisch niveau. Het begint met het identificeren van vier praktisch gemotiveerde OWCL-scenario's die in eerdere literatuur niet adequaat zijn geformaliseerd, en stelt een uniforme formulering van de beslissingsruimte voor. Op algoritmisch niveau verbindt HoliTrans *continual learning* met *open-set detection* door middel van *Nearest Class Mean (NCM)*-dynamiek en transformer-gebaseerde architecturen. Het model levert sterke empirische prestaties, terwijl het in lijn is met een theoretisch onderbouwde optimalisatieformulering, en fungeert zo als een allesomvattende OWCL-oplossing met aantoonbare generalisatie en grensconsistentie.
- **CCL (Cross-regional Continual Learning for Fraud Detection)** illustreert de praktische waarde van OWCL binnen besluitvorming in de echte wereld, met name in de context van financiële fraudedetectie. Door distributieveverschuivingen tussen geografische regio's en temporele fasen te modelleren, introduceert CCL een graaf-gebaseerde *continual learner* die zich aanpast aan veranderende fraudepatronen zonder eerder opgedane kennis te verliezen. De mogelijkheid om over regionale grenzen heen te generaliseren, benadrukt de robuustheid en toepasbaarheid van OWCL-principes in dynamische en kritieke toepassingen.

Deze bijdragen verleggen niet alleen de algoritmische grenzen van OWCL, maar herdefiniëren ook de leerprotocollen onder *open-world* condities. De scriptie pleit voor een verschuiving weg van monolithische, statische modellen naar modulaire, adaptieve en transparante leersystemen die continu leren, kennis doelgericht overdragen en intelligent reageren op onbekende situaties.

Tot slot worden er verschillende toekomstgerichte onderzoekslijnen voorgesteld, waaronder modulaire *Mixture-of-Experts* architecturen, *PEFT*-geïntegreerde federatieve OWCL-systemen, *crossmodaal* leren onder taakheterogeniteit, en verklaarbaar *continual reasoning*. Deze richtingen weerspiegelen een bredere ambitie: het bouwen

van intelligente systemen die *lifelong learning* ondersteunen, hun eigen beperkingen begrijpen, en realistische doelen dienen met robuustheid, transparantie en aanpasbaarheid.

Acknowledgements

This dissertation marks the end of a long journey that I could never have completed without the guidance, kindness, and support of many people to whom I owe my deepest gratitude.

First and foremost, I would like to express my heartfelt gratitude to Prof. Marcello Bonsangue, my supervisor, for his support, encouragement and patience. His kindness and wisdom have not only guided my research but also deeply influenced my attitude toward science and life. I am truly thankful for his trust and continuous support, which have helped me a lot. I would also like to extend my sincere gratitude to Prof. Xin Yang from the Southwestern University of Finance and Economics, whose guidance and inspiration laid the foundation for my research journey. His insightful advice and rigorous attitude toward his academic pursuit have continued to influence me.

Beyond academia, I want to thank my friends near and far for their warmth, humor, and understanding, which carried me through difficult times and reminded me to stay human amidst the pursuit of knowledge. To my dear friends in China, thank you for always being there despite the distance, for your messages, care, and quiet encouragement that always support me. To my friends at LIACS, thank you for the laughter, drinks, and countless small moments that made research life lighter and more meaningful. I am also grateful to the friends I met through climbing, whose energy, trust, and love for challenge have given me strength, balance, and peace beyond the world of papers and deadlines.

Above all, I owe my deepest gratitude to my family. To my parents, for their endless love, sacrifices, and faith in me when I doubted myself. To my grandparents, whose love and memory continue to inspire me every day. Their warmth, strength, and faith have shaped who I am and will forever guide my path. To my loved ones who stood by me with patience and unwavering belief: this achievement is as much yours as it is mine.

Finally, I thank life itself for its uncertainties—the setbacks, detours, and moments of doubt that taught me humility, perseverance, and gratitude. Completing this thesis is not an end but a new beginning, and I face the future with the same curiosity and hope that first set me on this path.

Curriculum Vitae

Yujie Li was born in Ganzhou, Jiangxi Province, China, in the winter of 1996. After receiving her Bachelor’s degree in Management Science and Engineering in 2019, she was admitted to a direct Ph.D. program (2+3 scheme) at the Southwestern University of Finance and Economics, where she pursued her Master’s and subsequently her Ph.D. studies. In 2023, she joined Leiden University, the Netherlands, as a Ph.D. candidate in Computer Science.

Her research interests lie at the intersection of artificial intelligence, machine learning, and data-driven decision-making, with a particular focus on continual learning, open-world continual learning, graph neural networks, federated learning and multi-granularity learning. In terms of applied research, she is especially interested in fintech, anonymous and fraud detection, and human-centered machine learning.

During her Ph.D. studies, she took courses in *Machine Learning*, *Deep Learning*, *Statistical Learning Theory*, *Optimization Methods*, *Data Mining* and scientific conduct, which provided solid theoretical foundations and strengthened her analytical and research capabilities in data-driven AI.