

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).

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 codebase 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.

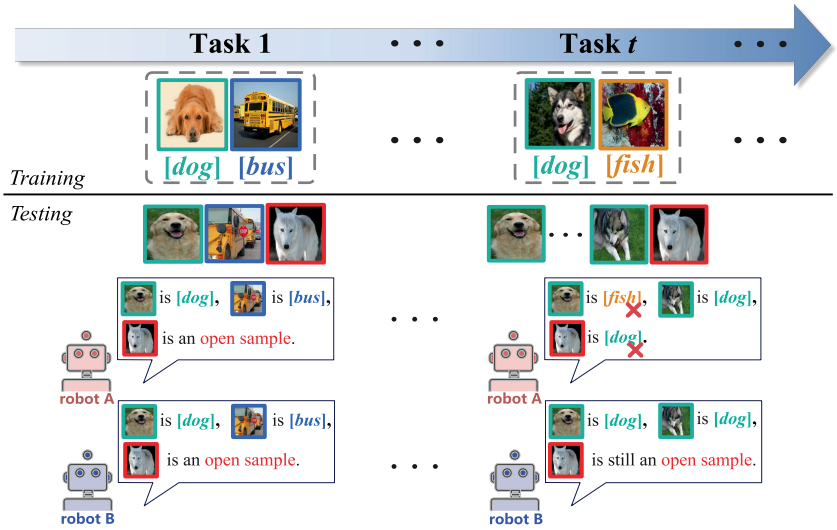


Figure 6.1: A toy example of the KIRO scenario in OWCL. During training Task 1 and Task t, Robots A and B learn *Golden Retrievers* and *Labradors*, both labeled as [*dog*]. In testing Task 1, both robots correctly classified [*dog*] and [*bus*] samples while identifying the unknown wolf without errors. 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.

Scenario	**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.

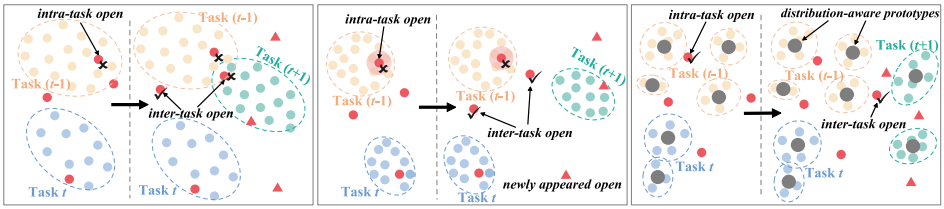
To overcome this, we introduce a novel scenario definition: **knowledge-incremental learning (KIL)**. In knowledge-incremental learning, models must solve each previously encountered task without knowing the task identity, and training sets across

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.



(a) CL Baseline Method (b) Pro-KT (c) HoliTrans

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^i_{tr}	Training set of task i
D^i_{te}	Testing set of task i
N_i	Number of training samples in D^i_{tr}
M_i	Number of seen classes in D^i_{tr}
$\mu(h(x), h(D^i_{tr}))$	Open risk of x in D^i_{te}
$\mathcal{H}$	Universal hypothesis space
$\epsilon_{\mathcal{D}^i_{tr}}(h)$	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
$\widehat{\epsilon}_{\mathcal{D}^i_{tr}}(h^t)$	Empirical prediction error
$\mathcal{B}^t$	Experience buffers from $\mathcal{D}^t_{tr}$
$\boldsymbol{s_x}$	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
$\boldsymbol{p}^i_k$	Prototype of class k in task i
$\mathcal{D}'^i_{te}$	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 $\mathcal{D}^i_{tr} = \{\boldsymbol{x}^i_j \in \mathcal{X}^{N_i}_{j=1}, y^i_j \in \mathcal{Y}^{M_i}_{j=1}\}$. In the training phase, only data pertinent to the current task is accessible, while the test samples $\mathcal{D}^i_{te} = \{\boldsymbol{x}^i_j \in \mathcal{X}^{N'_i}_{j=1}, y^i_j \in \mathcal{Y}^{M'_i}_{j=1}\}$ may contain unknowns/novelties, i.e., $\mathcal{Y}^{M'_i}_{j=1} - \mathcal{Y}^{M_i}_{j=1} \neq \emptyset$. Given the latent representation of the current task t's training set $h(\mathcal{D}^t_{tr})$, we denote $\mu(h(\boldsymbol{x}), h(\mathcal{D}^i_{tr}))$ as the open risk of a sample $\boldsymbol{x}$ from task i's test set $\mathcal{D}^i_{te}$. 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* $[1, t]$*:*

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

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.

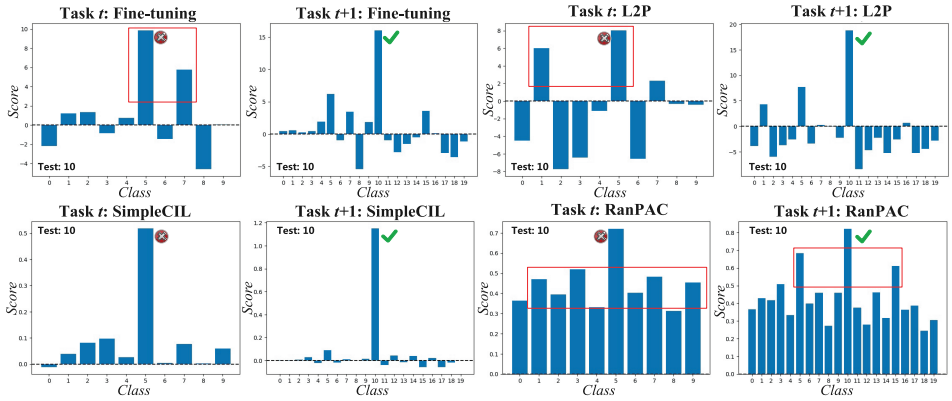


Figure 6.3: Unscaled classification scores on the CINO scenario.

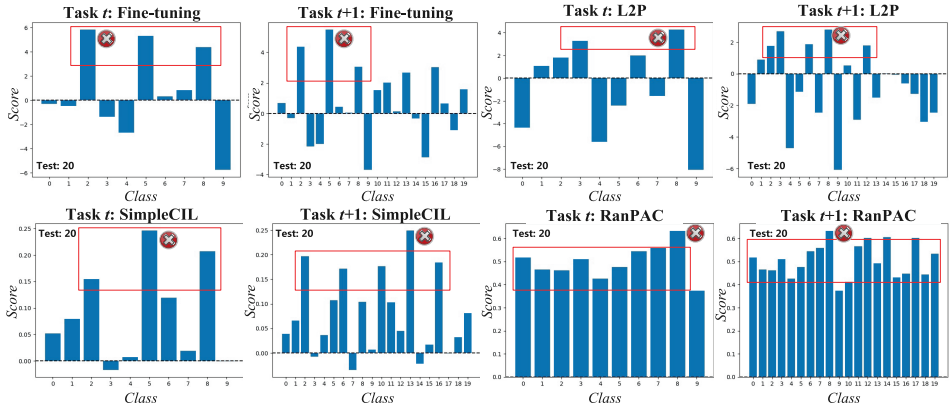


Figure 6.4: Unscaled classification scores on the CIRO scenario.

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.

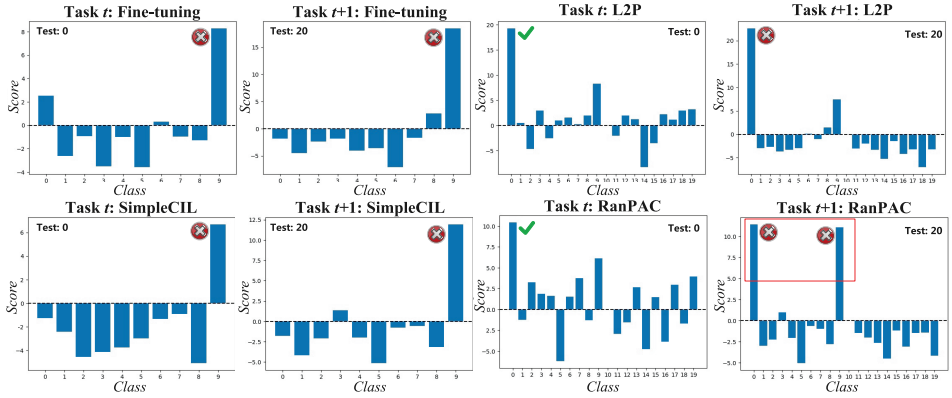


Figure 6.5: Unscaled classification scores on the KINO Scenario.

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

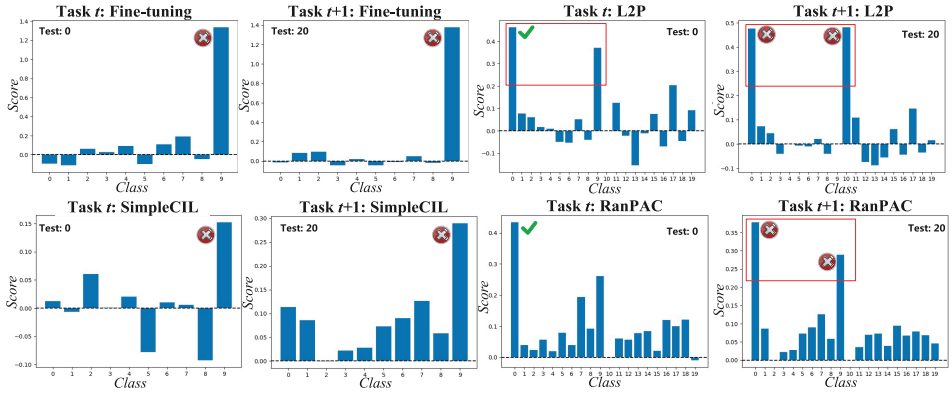


Figure 6.6: Unscaled classification scores on the KIRO Scenario.

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:*

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

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) $\sum_{i=1}^{t} \mu(h(\boldsymbol{x}), h(\mathcal{D}_{tr}^{i}))$ 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*

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

where $\mathbf{P}_{X|Y=y_{un}}$ is the conditional distribution of inputs $\boldsymbol{x}$ 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 $h(\boldsymbol{x})$ 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 $\widehat{\epsilon}_{\mathcal{D}_{tr}^i}(h^t)$ of h^t in testing an arbitrary task $i (i \leq t)$ is defined as [133]:*

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

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., $\widetilde{N}_i \ll N_i$, and $\widetilde{N}_i$ 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 he 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.

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:*

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

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:

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

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:*

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

where the constant C_1 can be calculated by ERM-Based generalization bound, $\gamma_i + \alpha_i + \beta_i = 1$, and the error term $\sum_{i=1}^{t-1} \left(\alpha_i + \beta_i\right) \epsilon_{\mathcal{D}_{tr}^{i}}\left(h^{t-1}\right)$ 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.

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

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

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

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

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:

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

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

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

By decomposing, we get:

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

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* $P : \mathbb{R} \to \mathbb{R}'$ 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 $\mathcal{P} : \mathbb{R} \to \mathbb{R}'$, an empirical estimate for the data discrepancy term is:*

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

where $\mathcal{B}^t$ and $\mathcal{B}^i$ are experience buffers drawn from training set $\mathcal{D}^t_{tr}$ and $\mathcal{D}^i_{tr}$ 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 $\widehat{d\mathcal{H}\Delta\mathcal{H}}$ 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 non-linear 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:

$$H^i = g(\phi(D^i_{tr})W). \tag{6.4}$$

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^ias follows:

$$\begin{aligned} G^i &= G^{i-1} + H^{i^T} H^i, \\ C^i &= C^{i-1} + H^{i^T} Y. \end{aligned} \tag{6.5}$$

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:

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

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

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

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:

$$\begin{aligned} \arg\max_r \frac{1}{|\mathcal{D}'^i_{te}|} \sum_{\boldsymbol{x} \in \mathcal{D}'^i_{te}} \Bigg[& \mathbb{I}\left(\boldsymbol{s}_{\boldsymbol{x}} < r \frac{1}{|\mathcal{D}'^i_{tr}|} \sum_{\boldsymbol{z} \in \mathcal{D}'^i_{tr}} \boldsymbol{s}_{\boldsymbol{z}} \,|\, \text{known} \right) \\ & + \mathbb{I}\left(\boldsymbol{s}_{\boldsymbol{x}} > r \frac{1}{|\mathcal{D}'^i_{tr}|} \sum_{\boldsymbol{z} \in \mathcal{D}'^i_{tr}} \boldsymbol{s}_{\boldsymbol{z}} \,|\, \text{open} \right) \Bigg] \\ \text{with} \quad & \boldsymbol{s}_{\boldsymbol{x}} = \max_{y \in \mathcal{Y}^{M_i}} g\left(\phi(x)^T W\right) (G^i + \lambda^i I)^{-1} \boldsymbol{p}^i_y, \\ \text{and} \quad & \boldsymbol{p}^i_y = \frac{\sum_{j=1}^{|\mathcal{X}^i|} \mathbb{I}(y^i_j = y) \phi(x^i_j)}{\sum_{j=1}^{|\mathcal{X}^t|} \mathbb{I}(y^i_j = y)}. \end{aligned} \tag{6.7}$$

Here $\mathbb{I}(\cdot)$ is an indicator function, $\boldsymbol{s}_{\boldsymbol{x}}$ is the maximum similarity between sample 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 $\boldsymbol{p}^i_y$ is the prototype of class y in task i. Note the $\mathcal{D}'^i_{te}$ is different from the original $\mathcal{D}^i_{te}$ in testing as we integrate a batch of generated pseudo-samples for each task to better represent data distribution via the embedding distribution over $\mathcal{D}^i_{tr}$, obtaining a distribution-aware prototype for each known class.

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

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

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^i_{tr} and test set D^i_{te}, the class prototypes G^i and $\{\boldsymbol{p_k}\}_{k=1}^{M^i}$ 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 $(\boldsymbol{p}_k, \boldsymbol{p}_l)$ and sampling from a Gaussian distribution centered around the pseudo-prototype. Positive pseudo-samples are similarly generated from each prototype $\boldsymbol{p_k}$ using a Gaussian distribution. The pseudo-samples are then integrated with the test set $\mathcal{D}^i_{te}$, and the decision threshold r is determined through a ternary search over the modified test set $\mathcal{D'}^i_{te}$. 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^i_{tr} and test set D^i_{te}.
Phase 1: *NCM-based Classifier Training*
Update G^i and $\{\boldsymbol{p_k}\}_{k=1}^{k=M^i}$ by D^i_{tr};
Calculate standard deviation δ between samples and corresponding prototypes.
Phase 2: *Learning DAPs*
`// Generate Negative Pseudo-samples`
for *an arbitrary pair of prototypes* $(\boldsymbol{p}_k, \boldsymbol{p}_l)$ **do**
  Obtain seudo-prototype $\boldsymbol{p}_{kl} = \frac{\boldsymbol{p}_k + \zeta \boldsymbol{p}_l}{1+\zeta}$;
  Generate negative pseudo-samples via a Gaussian distribution $\mathcal{N}(\boldsymbol{p}_k, \delta^2\mathbf{I})$.
`// Generate Positive Pseudo-samples`
for *each prototype* $\boldsymbol{p_k}$ **do**
  Generate positive pseudo-samples via a Gaussian distribution $\mathcal{N}(\boldsymbol{p}_k, \delta^2\mathbf{I})$.
Integrating all pseudo-samples with $\mathcal{D}^i_{te}$;
Determining r by ternary search over $\mathcal{D'}^i_{te}$.

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 $O\left(\log_3\left(\frac{\max_{\boldsymbol{z}\in\mathcal{D}tr^i}\boldsymbol{s_z}-\min\boldsymbol{z}\in\mathcal{D}tr^i\boldsymbol{s_z}}{\frac{1}{|\mathcal{D}tr^i|}\sum_{\boldsymbol{z}\in\mathcal{D}^i_{tr}}\boldsymbol{s_z}}\Big/\epsilon\right)\right)$, where ϵ is a pre-defined convergence threshold, and $\boldsymbol{s_z}$ denotes the similarity score of sample $\boldsymbol{z}$. 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.

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

This inequality bounds the probability that the norm of the projected vector $\|W^Tf\|$ 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:

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

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

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

Our initial Chernoff bound can thus be rewritten as:

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

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 $t = 1, \ldots, T$. 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	$N_t, t = 1, ..., T$	$N_{val,known}$	$N_{val,open}$
CIFAR100	90	10	9	55000	$N_t = 5000, t = 1, ..., T$	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.

Method		Scenario: CIRO (Split-CIFAR100)			Scenario: KIRO (Open-CORe50)		
		$ACC_t(\uparrow)$	$AUC_t(\uparrow)$	$FPR_t(\downarrow)$	$ACC_t(\uparrow)$	$AUC_t(\uparrow)$	$FPR_t(\downarrow)$
Continual learning methods equipped with out-of-distribution algorithms							
L2P	OpenMax	$0.436_{\pm 0.048}$	$0.133_{\pm 0.051}$	$0.937_{\pm 0.046}$	$0.424_{\pm 0.063}$	$0.244_{\pm 0.050}$	$0.968_{\pm 0.039}$
	MaxLogits	$0.445_{\pm 0.041}$	$0.103_{\pm 0.047}$	$0.931_{\pm 0.045}$	$0.412_{\pm 0.179}$	$0.259_{\pm 0.053}$	$0.964_{\pm 0.037}$
	Entropy	$0.445_{\pm 0.028}$	$0.120_{\pm 0.038}$	$0.932_{\pm 0.041}$	$0.424_{\pm 0.063}$	$0.257_{\pm 0.063}$	$0.958_{\pm 0.046}$
	EnergyBased	$0.445_{\pm 0.038}$	$0.102_{\pm 0.043}$	$0.925_{\pm 0.040}$	$0.425_{\pm 0.065}$	$0.256_{\pm 0.057}$	$0.970_{\pm 0.057}$
DualPrompt	OpenMax	$0.423_{\pm 0.031}$	$0.139_{\pm 0.037}$	$0.928_{\pm 0.045}$	$0.402_{\pm 0.029}$	$0.270_{\pm 0.024}$	$0.960_{\pm 0.036}$
	MaxLogits	$0.425_{\pm 0.034}$	$0.114_{\pm 0.039}$	$0.933_{\pm 0.042}$	$0.408_{\pm 0.113}$	$0.261_{\pm 0.017}$	$0.975_{\pm 0.022}$
	Entropy	$0.425_{\pm 0.045}$	$0.204_{\pm 0.032}$	$0.936_{\pm 0.048}$	$0.408_{\pm 0.112}$	$0.267_{\pm 0.026}$	$0.961_{\pm 0.019}$
	EnergyBased	$0.425_{\pm 0.046}$	$0.114_{\pm 0.049}$	$0.942_{\pm 0.039}$	$0.408_{\pm 0.112}$	$0.260_{\pm 0.030}$	$0.973_{\pm 0.020}$
CODA	OpenMax	$0.439_{\pm 0.036}$	$0.125_{\pm 0.042}$	$0.936_{\pm 0.046}$	$0.413_{\pm 0.032}$	$0.278_{\pm 0.045}$	$0.953_{\pm 0.020}$
	MaxLogits	$0.439_{\pm 0.047}$	$0.090_{\pm 0.039}$	$0.942_{\pm 0.045}$	$0.415_{\pm 0.033}$	$0.183_{\pm 0.028}$	$0.964_{\pm 0.018}$
	Entropy	$0.439_{\pm 0.036}$	$0.106_{\pm 0.046}$	$0.939_{\pm 0.047}$	$0.415_{\pm 0.033}$	$0.256_{\pm 0.056}$	$0.953_{\pm 0.021}$
	EnergyBased	$0.439_{\pm 0.031}$	$0.089_{\pm 0.053}$	$0.946_{\pm 0.037}$	$0.415_{\pm 0.033}$	$0.169_{\pm 0.022}$	$0.962_{\pm 0.018}$
ADAM	OpenMax	$0.440_{\pm 0.044}$	$0.112_{\pm 0.040}$	$0.876_{\pm 0.043}$	$0.433_{\pm 0.025}$	$0.283_{\pm 0.038}$	$0.960_{\pm 0.029}$
	MaxLogits	$0.440_{\pm 0.042}$	$0.099_{\pm 0.053}$	$0.882_{\pm 0.033}$	$0.433_{\pm 0.026}$	$0.238_{\pm 0.029}$	$0.965_{\pm 0.025}$
	Entropy	$0.440_{\pm 0.042}$	$0.221_{\pm 0.043}$	$0.876_{\pm 0.047}$	$0.433_{\pm 0.025}$	$0.197_{\pm 0.016}$	$0.967_{\pm 0.021}$
	EnergyBased	$0.440_{\pm 0.035}$	$0.386_{\pm 0.042}$	$0.869_{\pm 0.041}$	$0.433_{\pm 0.027}$	$0.345_{\pm 0.023}$	$0.906_{\pm 0.035}$
RanPAC	OpenMax	$0.467_{\pm 0.045}$	$0.091_{\pm 0.044}$	$0.879_{\pm 0.054}$	$0.472_{\pm 0.022}$	$0.201_{\pm 0.020}$	$0.952_{\pm 0.013}$
	MaxLogits	$0.468_{\pm 0.042}$	$0.092_{\pm 0.042}$	$0.878_{\pm 0.039}$	$0.472_{\pm 0.022}$	$0.212_{\pm 0.017}$	$0.963_{\pm 0.015}$
	Entropy	$0.468_{\pm 0.049}$	$0.091_{\pm 0.041}$	$0.871_{\pm 0.031}$	$0.472_{\pm 0.022}$	$0.219_{\pm 0.021}$	$0.962_{\pm 0.010}$
	EnergyBased	$0.468_{\pm 0.046}$	$0.227_{\pm 0.038}$	$0.869_{\pm 0.041}$	$0.475_{\pm 0.021}$	$0.411_{\pm 0.025}$	$0.939_{\pm 0.021}$
Open-world continual learning methods							
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		$\mathbf{0.851_{\pm 0.010}}$	$\mathbf{0.858_{\pm 0.012}}$	$\mathbf{0.193_{\pm 0.024}}$	$\mathbf{0.748_{\pm 0.010}}$	$\mathbf{0.748_{\pm 0.010}}$	$\mathbf{0.097_{\pm 0.031}}$

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.

Method		**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%	$0.582_{\pm 0.009}$	$0.575_{\pm 0.003}$	$0.795_{\pm 0.007}$	$0.505_{\pm 0.005}$	$0.502_{\pm 0.005}$	$0.459_{\pm 0.004}$
	75%	$0.584_{\pm 0.004}$	$0.565_{\pm 0.005}$	$0.778_{\pm 0.002}$	$0.501_{\pm 0.001}$	$0.502_{\pm 0.004}$	$0.466_{\pm 0.013}$
	50%	$0.589_{\pm 0.009}$	$0.564_{\pm 0.004}$	$0.699_{\pm 0.045}$	$0.504_{\pm 0.002}$	$0.501_{\pm 0.002}$	$0.478_{\pm 0.004}$
	25%	$0.583_{\pm 0.015}$	$0.566_{\pm 0.014}$	$0.642_{\pm 0.049}$	$0.509_{\pm 0.002}$	$0.509_{\pm 0.003}$	$0.467_{\pm 0.007}$
HoliTrans		$\mathbf{0.851}_{\pm 0.010}$	$\mathbf{0.858}_{\pm 0.012}$	$\mathbf{0.193}_{\pm 0.024}$	$\mathbf{0.748}_{\pm 0.010}$	$\mathbf{0.748}_{\pm 0.010}$	$\mathbf{0.097}_{\pm 0.031}$

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 $\text{ACC}_t = 0.505$ and $\text{AUC}_t = 0.502$. 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.

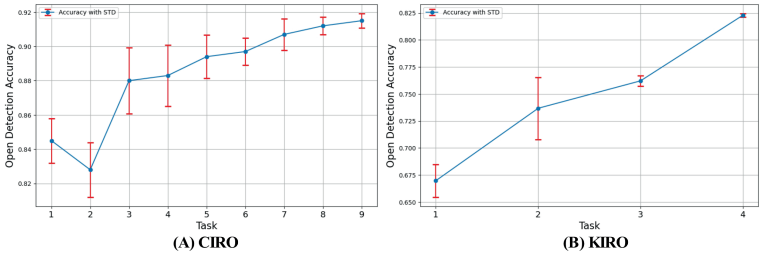


Figure 6.7: Visualizations of main results.

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

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.

Ablation Component	**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	$0.468_{\pm 0.042}$	$0.092_{\pm 0.010}$	$0.878_{\pm 0.039}$	$0.472_{\pm 0.022}$	$0.212_{\pm 0.017}$	$0.963_{\pm 0.015}$
HoliTrans	$\mathbf{0.851_{\pm 0.010}}$	$\mathbf{0.859_{\pm 0.012}}$	$\mathbf{0.193_{\pm 0.024}}$	$\mathbf{0.748_{\pm 0.010}}$	$\mathbf{0.748_{\pm 0.010}}$	$\mathbf{0.054_{\pm 0.031}}$

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.

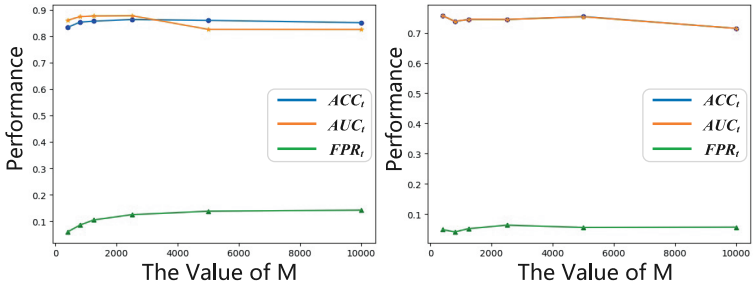


Figure 6.8: Performance with varying NRP size.

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.

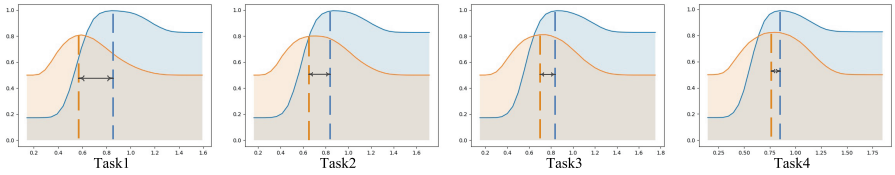


Figure 6.9: The divergence between the actual embedding distributions and the *HoliTrans*'s embedding distributions with distribution-aware prototypes (The horizontal axis is the values of r, the vertical axis is ACC_t, and the dashed line indicates the optimal value of r). Orange represents the actual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