Causal inference-based few-shot class-incremental learning

Zhou, W.; Xiao, G.; Lew, M.S.K.; Wu, S.

Causal Inference-based Few-Shot Class-Incremental Learning

Weiwei Zhou
Southwest University
Chongqing, China
vivichou1@email.swu.edu.cn

Guoqiang Xiao
Southwest University
Chongqing, China
gqxiao@swu.edu.cn

Michael S. Lew
Leiden University
Leiden, Netherlands
m.s.lew@liacs.leidenuniv.nl

Song Wu*
Southwest University
Chongqing, China
songwuswu@swu.edu.cn

ABSTRACT

Few-Shot Class-Incremental Learning (FSCIL) aims to keep recognizing novel classes from a limited number of samples after training on abundant data from base classes while maintaining the performance of the old classes. The challenge, however, is that limited data from new classes not only leads to the issue of overfitting but also catastrophic forgetting. To address these two issues, we propose a causal inference strategy in the mainstream FSCIL framework, which encourages the model to learn significant knowledge in the base training session and enhance the model's ability to extract features to cope with the emergence of unseen classes in the incremental session, by improving the learning of causal relationships between features and predictions for perturbed samples. In addition, to improve the effectiveness of learning new tasks in the incremental sessions while preventing the model from overfitting to the novel class data, we freeze the feature extractor while adding a Fourier transform after the feature extractor in the incremental session. It can denoise the features, strengthen the features of the novel classes, and suppress the error in extracting the features of the limited number of samples directly from the feature extractor. Extensive experiments on CIFAR100, Caltech-USCD Birds-200-2011, and miniImageNet datasets show that our proposed framework achieves state-of-the-art performance on FSCIL. The source code of our designed framework is at https://github.com/SWU-CS-MediaLab/CIFSCIL.

CCS CONCEPTS

• **Computing methodologies** → **Computer vision**.

KEYWORDS

Casual Inference, Few-Shot Class-Incremental Learning, Image Classification

*Corresponding author

ICMR '24, June 10–14, 2024, Phuket, Thailand.

ACM ISBN 979-8-4007-0619-6/24/06
https://doi.org/10.1145/3652583.3658098

ACM Reference Format:
Weiwei Zhou, Guoqiang Xiao, Michael S. Lew, and Song Wu. 2024. Causal Inference-based Few-Shot Class-Incremental Learning. In *Proceedings of the 2024 International Conference on Multimedia Retrieval (ICMR '24), June 10–14, 2024, Phuket, Thailand.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3652583.3658098

1 INTRODUCTION

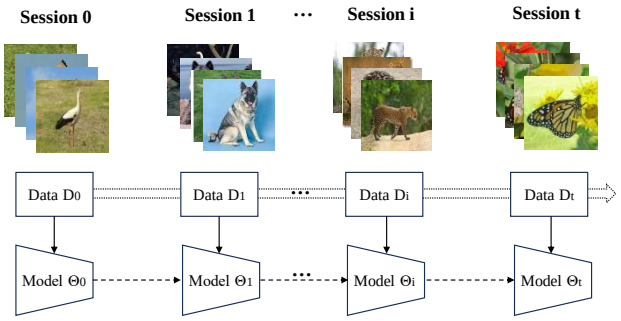


Figure 1: Class-Incremental Learning. At each phase, the model receives sufficient training data for the new task and uses them for training.

With the rapid development of deep learning [20], current learning systems have performed very well in many tasks [24, 43, 45]. However, real-world applications are often confronted with streaming data [10] and novel classes that keep arriving [56], in which case the ideal neural network model should not only recognize the novel classes, but also remain distinguishable from the old ones [1, 52]. This learning paradigm which continuously learns in the face of successive tasks and overcomes the forgetting of old knowledge, known as Class-Incremental Learning (CIL). And the situation of forgetting old tasks after learning new tasks, known as catastrophic forgetting [11], is a primary challenge in CIL. As shown in Figure 1, CIL [28] usually makes only the novel classes of the current task visible in each round of incremental training, and cannot reread data from the old task. In CIL, there is often a trade-off between learning novel classes and keeping the old classes, which is the stability-plasticity dilemma [29]. Few-Shot Class-Incremental

Learning (FSCIL) [41] further imposes a constraint on data availability. In FSCIL, only a few data samples, i.e., few-shot, are allowed for each novel class to be trained in incremental sessions, leading to an even more challenging incremental learning problem. In FSCIL, after the model learns a new task, the performance of the old (base) task decreases significantly. This is due to the lack of training data for the old task, which causes the model to focus only on the new task and become oblivious to the old task. While learning new tasks directly is somewhat beneficial for predicting newly learned tasks, model training is prone to overfitting as training samples for new tasks become scarce, and the issue of catastrophic forgetting in FSCIL thus becomes apparent faster than in traditional CIL settings.

To address the catastrophic forgetting, many traditional and outstanding incremental learning methods [21, 21, 33, 33, 52, 52] have been proposed. However, due to the limitation of sparse training samples in FSCIL, the traditional incremental learning methods can no longer be fully adapted to the context in FSCIL. Some recent studies [37, 39, 41] have proposed freezing the base class training model and only fine-tuning or learning only the classifiers of the novel classes when learning the new task, and these approaches can undoubtedly maximize the retention of the memory of the old task and mitigate the issue of catastrophic forgetting. However, due to the stability-plasticity dilemma in incremental learning, the model's learning of novel classes will be more limited.

Starting from the above problem, to improve the model's ability to recognize novel classes during subsequent incremental sessions, we should emphasize the generalization of the model training for base classes. This enables the model can pay attention to some "extracurricular knowledge" while learning the base class task, preventing confusion when encountering the novel classes in incremental sessions. However, we observe that many existing deep learning methods essentially seek correlations, which harms model training and generalization. Because correlation does not mean causality, it is crucial to prioritize the exploration of causal relationships over mere correlations between features and outcomes.

In this paper, we propose to use the causal inference strategy [48] of counterfactual intervention [19] on the base classes data after adding perturbations to learn diversified features and explore the causal relationship between features and predictions, aiming to enhance the model's ability to learn features under challenging situations and improve the model's generalization ability to few-shot novel classes. In addition, in order to obtain more significant feature information from novel classes in incremental sessions, the Fourier transform is introduced in incremental sessions to improve the model's feature representation ability for novel classes.

The main contributions are summarized as follows:

- We propose to improve the model's ability to learn diverse discriminative features and facilitate the model's identification of few-shot novel classes in the incremental session by using causal inference to explore the causal relationship between perturbed sample features and predictions.
- In order to facilitate the classifier to better capture the features of the few-shot data, the extracted feature representations of few-shot data are further optimized by the Fourier transform to enhance the discriminative information in the features.
- We validate the effectiveness of our method on three benchmark datasets, and the experimental results show that our method significantly outperforms the baseline and yields better performance compared with several state-of-the-art FSCIL algorithms.

2 RELATED WORK

2.1 Class-Incremental Learning

Class-Incremental Learning (CIL) [4, 13, 21, 35] aims to recognize novel classes without forgetting the knowledge of old classes. There are three main approaches to solving the catastrophic forgetting problem in CIL: regularization-based approaches [6], such as EWC [16] uses the Fisher information matrix to estimate parameter importance, expecting that the important parameters change slightly with the regularization term. parameter-isolation-based approaches [26, 27], for example, LwF [21] proposes that each network is responsible for different tasks, and reduces forgetting by sharing a portion of the parameters. Replay-based approaches [33, 52], such as iCaRL [35] replays and performs knowledge distillation to maintain old knowledge. However, CIL typically requires a large number of novel class training samples, which makes it unsuitable for many practical applications such as incremental anomaly detection [2].

2.2 Few-Shot Learning

Few-Shot Learning (FSL) [25, 38] is a machine learning task in which only scarce training are available for learning and training the model during the training process. Current FSL methods can be broadly classified into two categories: optimization-based methods [8, 14, 22] and metric-based methods [9, 49, 50]. Optimization-based approaches aim to quickly adapt to new few-shot tasks by learning an optimization algorithm. Metric-based approaches consider learning a suitable distance metric between support instances and query instances. However, FSL aims to adapt to novel classes with limited samples, ignoring the ability to handle previously learned classes.

2.3 Causal Inference

Causal Inference (CI) [32] differs from traditional correlation analysis, which studies causal relationships between variables to explain the causal relationships behind the data and improve feature learning about attention by analyzing the causal relationships. It shows promising results in a variety of computer vision tasks, including few-shot image classification [7], long-tailed distributions [23], incremental learning, augmented learning [17], and natural language processing [30]. The counterfactual learning strategy we use is causal inference-based attentional learning that encourages the network to learn more effective attention and improves network generalization.

2.4 Few-Shot Class-Incremental Learning

Few-Shot Class-Incremental Learning (FSCIL) [41] considers both the FSL and CIL challenges described above. Specifically, FSCIL aims to learn incrementally from limited novel class samples while retaining what has already been learned [54]. In this case, it is difficult to improve the performance of FSCIL using traditional

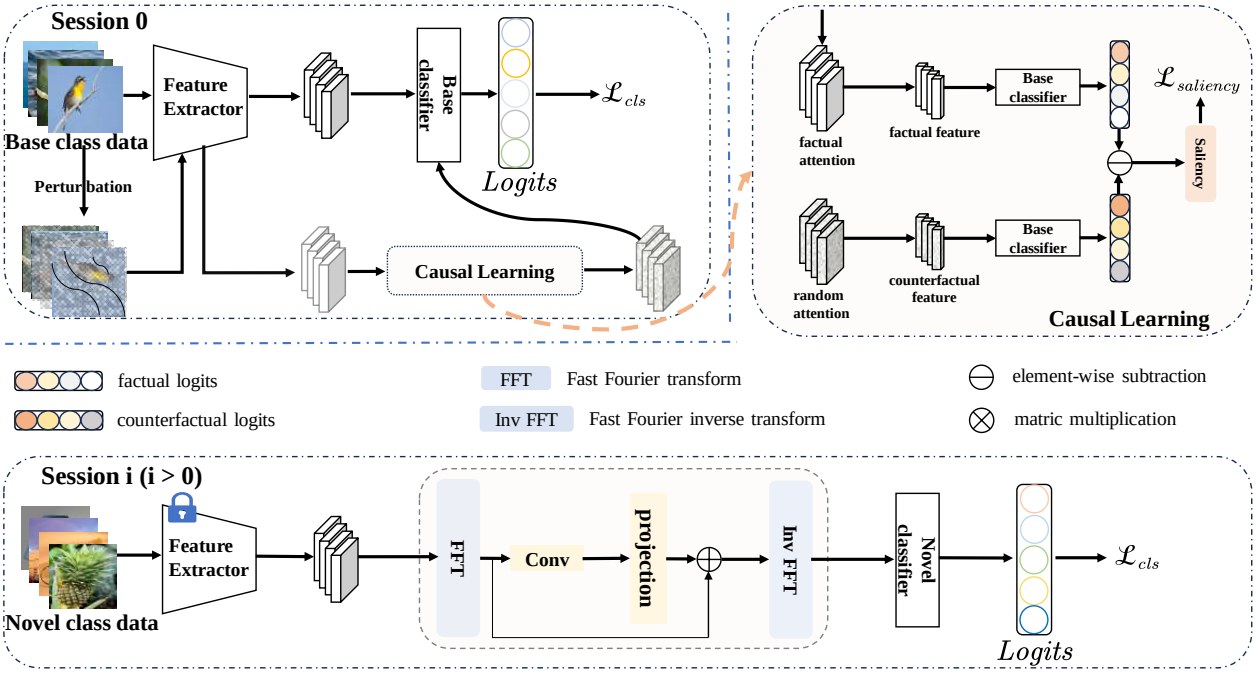


Figure 2: An overview of our proposed CIFSCIL. The lock symbol indicates that the model parameters are frozen. The feature extractor is first trained with base class data (original and perturbed data) and then the perturbed features are learned by counterfactual learning. For incremental training, few-shot data are passed through the feature extractor to get the base features, and then the features are enhanced by the Fourier transform.

incremental learning frameworks because none of these methods consider incremental few-shot overfitting, which can easily exacerbate catastrophic forgetting. The TOPIC [41] framework mitigates the forgetting problem by stabilizing the neural gas network topology. CEC [51] separates each class with independent classifiers and employs a graph model to propagate contextual information among the classifiers. F2M [37] overcomes catastrophic forgetting by finding a flat minimum. It does this by injecting noise during base training and argues that the focus of FSCIL should be on the basic training session. Our Causal Inference-based Few-shot Class-Incremental Learning (CIFSCIL) considers incorporating causal inference in the learning of the base classes to improve the generalization ability of the feature extractor, and to compensate for the poorer performance of the few-shot novel class in the incremental session of learning, we propose to perform further Fourier transform learning on the few-shot features.

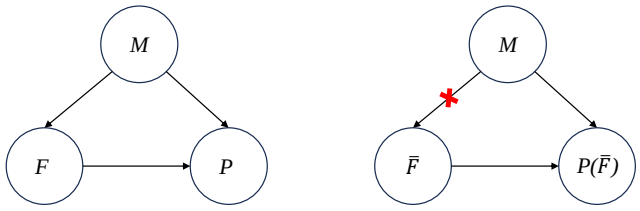
3 PRELIMINARY

FSCIL [38, 41] aims to design a machine learning algorithm that learns a series of tasks continuously to obtain a model that does not forget the knowledge of the old classes; meanwhile, it also performs well for learning novel classes [54]. Specifically, the model is usually initially trained by first providing enough base class data in a basic session, which is denoted by $\mathcal{D}_{train} = \{\mathcal{D}^t_{train}\}^T_{t=0}$. Subsequent incremental sessions provide only a few samples, i.e., $\mathcal{D}^t_{train} = \{(\mathbf{x}_i,\ y_i)\}^{N_t}_{i=0}$, which represents the training samples from session t, with $\mathbf{x}_i$ and y_i being the i-th image and the corresponding labels, respectively, assuming that the labeling configuration for the t-th task is C^t and that there is no overlap in labeling space from different tasks. Once the model training moves to the next session, the training dataset from the previous learning session is no longer available. The evaluation of the FSCIL task in each session involves all classes from previous sessions and the current session. That is, the model trained on $\mathcal{D}^t_{train}$ should be evaluated on $\mathcal{D}^t_{test}$, which contains all encountered classes $C^0 \cup C^1 \cdots \cup C^t$ in the t-th session. In particular, the incremental data is always involved in the training in the form of an N-way K-shot, i.e., there are N classes, and each class contains K training data. For example, in the commonly used benchmark dataset CIFAR100, there are 60 classes in the base session with 500 training images for each class. In contrast, in the incremental session, there are only 5 classes available for training and only 5 images for each class. FSCIL defines a harsh problem setting, where severe data scarcity and data imbalance issues further exacerbate knowledge forgetting in incremental learning.

4 METHODOLOGY

This section provides the algorithmic details of our proposed Causal Inference-based Few-Shot Class-Incremental Learning (CIFSCIL). Firstly, CIFSCIL uses causal learning during the base class task training, which allows the model to be better generalized by base classes learning. Subsequently, the Fourier transform is used on the novel class data in the incremental sessions to enhance and filter

the features to improve the model's ability to effectively classify the novel classes. The overview of our CIFSCIL framework is shown in Figure 2.



(a) Confounded (b) Deconfounded

Figure 3: The counterfactual intervention $P(do(F))$. The backdoor path is $F \leftarrow M \rightarrow P$. And $F \rightarrow P$ is the frontdoor path. The causal relationship between F and P is obtained by cutting off $M \rightarrow F$ in the backdoor path.

4.1 Causal learning

In general FSCIL algorithms, it is common to use the strategy of incremental-frozen framework, and many methods [39, 41] have demonstrated the superiority of this operation over other incremental training operations based on fine-tuning, etc., so this strategy is also utilized in our framework. The use of incremental-frozen framework strategy prompts us to focus more on the subsequent generalization of the novel classes: i.e., how can we better achieve future generalization for Few-Shot Learning in this case?

In recent years, many studies [34, 53] proposed to apply causal inference to deep-learning methods based on causal inference. These methods actually solve a series of problems by exploring causal relationships to overcome the limitations of the current deep-learning methods that only consider correlations. Causal inference usually illustrates the effect of features on the predictions by inferring the difference between counterfactual logic and factual logic, i.e., the backdoor adjustment, as illustrated by the causal graph in Figure 3, which is structured as a directed acyclic graph $G = \{N, E\}$, where each variable in the model has a corresponding node in N, and the causal relation E describes how these variables interact with each other, and the causal relationship between a variable and an outcome is obtained by cutting the connection between two variables (e.g., M and F). This analysis of the causal relationship between features and outcomes is beneficial for the robustness of the learned feature and the optimization of the model training. Inspired by these successful studies [34, 40], we propose to embed causal inference in the training of the base classes to facilitate the model's learning capability of the base classes and to enhance the generalization of the feature extractor to novel classes.

Our CIFSCIL framework is first trained on the base tasks; here, we generally use the common CNN structure as a feature extractor, followed by a fully connected layer as a classifier. $\boldsymbol{W}_{NET}$ and $\boldsymbol{W}_{FC}$ are used as weights for the feature extractor and the fully connected layer, respectively. Given an input $\boldsymbol{X} \in \mathcal{D}^0_{train} \sim \{\mathcal{X}^0_{train}, \mathcal{Y}^0_{train}\}$, $\mathcal{D}^0_{train} = \{(x_i, y_i)\}_{i=1}^n$, where $x_i \in \mathcal{X}^0_{train}$ is the instance and $y_i \in \mathcal{Y}^0_{train}$ is the corresponding label, and the output of the network is:

$$\mathcal{P} = f_{FC}(f_{flat}(f_{NET}(\boldsymbol{X}, \boldsymbol{W}_{NET})), \boldsymbol{W}_{FC}), \tag{1}$$

Where $f_{NET}(\boldsymbol{X}, \boldsymbol{W}_{NET})$ indicates the output of the feature extractor after feeding the input $\boldsymbol{X}$; $f_{FC}(\cdot, \boldsymbol{W}_{FC})$ is the output of the classifier; f_{flat} is the flattening operator, which flattens an m–D tensor into a 1-D vector. Then, the model training by optimizing the per-sample loss is:

$$\mathcal{L}_{ce}(y, \mathcal{P}) = -\frac{1}{|\mathcal{X}_{train}|} \sum_{m=1}^{|\mathcal{X}_{train}|} \log \frac{\exp\left(\eta \mathcal{P}^{(m)}\right)}{\sum_{i \neq m} \exp\left(\eta \mathcal{P}^{(i)}\right)}, \tag{2}$$

Where $\mathcal{L}_{ce}$ is the cross entropy loss function, P indicates the predictions obtained from the model. Aiming to enable causal inference to increase the generalization of the model as we train the base classes, we further learn causality from the features extracted by backbone (feature extractor) from the base classes data after adding perturbations (e.g., rotations, translations, adding noise, and other operations). Thus, we denote the perturbed training sample as $\widetilde{\boldsymbol{X}} \in \widetilde{\mathcal{D}}^0_{train} \sim \{\widetilde{\mathcal{X}}_{train}, \widetilde{\mathcal{Y}}_{train}\}$, and similarly, the perturbed sample is characterized by the feature extractor to obtain the feature $\mathcal{F}(\widetilde{\boldsymbol{X}})$:

$$\mathcal{F}(\widetilde{\boldsymbol{X}}) = f_{NET}(\widetilde{\boldsymbol{X}}, \boldsymbol{W}_{NET}), \tag{3}$$

Subsequently, we perform a counterfactual intervention on $\mathcal{F}(\widetilde{\boldsymbol{X}})$. As shown in Figure 3, we establish a causal relationship for the variables of the visual feature $\mathcal{F}(\widetilde{\boldsymbol{X}})$, the visual confounder M of the image, and the prediction P, where the direct edges denote the causality between the two variables. By cutting the $M \rightarrow F$ connection in the backdoor path, a counterfactual feature $\mathcal{F}(\widetilde{\boldsymbol{X}})^{coun}$ can be obtained (in practice, we implement this using randomized attention). In other words, we intervene counterfactually by imagining nonexistent features instead of learned ones.

As shown in Figure 2, in order to enhance the learning for attention, the features of attention factual attention $\mathcal{W}(\widetilde{\boldsymbol{X}})$ and random attention $\mathcal{W}(\widetilde{\boldsymbol{X}})^{coun}$ are enhanced to participate in the computation of the corresponding features, respectively.

$$\mathcal{F}_a(\widetilde{\boldsymbol{X}}) = \mathcal{W}(\widetilde{\boldsymbol{X}}) \otimes \mathcal{F}(\widetilde{\boldsymbol{X}}), \tag{4}$$

$$\mathcal{F}_a(\widetilde{\boldsymbol{X}})^{coun} = \mathcal{W}(\widetilde{\boldsymbol{X}})^{coun} \otimes \mathcal{F}(\widetilde{\boldsymbol{X}})^{coun}, \tag{5}$$

Where $\mathcal{W}(\widetilde{\boldsymbol{X}})$ and $\mathcal{W}(\widetilde{\boldsymbol{X}})^{coun}$ are obtained by putting $\mathcal{F}(\widetilde{\boldsymbol{X}})$ and $\mathcal{F}(\widetilde{\boldsymbol{X}})^{coun}$ through a convolution operation, respectively. $\otimes$ denotes matrix multiplication. In order to effectively analyze the causal relationship between features and predictions, we also need to obtain the corresponding factual predictions $\widetilde{\mathcal{P}}$ and counterfactual predictions $\widetilde{\mathcal{P}}_{coun}$.

$$\widetilde{\mathcal{P}} = f_{FC}(\mathcal{F}_a(\widetilde{\boldsymbol{X}}), \boldsymbol{W}_{FC}), \tag{6}$$

$$\widetilde{\mathcal{P}}_{coun} = f_{FC}(\mathcal{F}_a(\widetilde{\boldsymbol{X}})^{coun}, \boldsymbol{W}_{FC}), \tag{7}$$

To further eliminate "pseudo-correlation" and find the true causal relationship between factual features and factual predictions, we calculate the difference between factual predictions and counterfactual predictions to account for the effect of the features on the predictions, which is a common tool in causal inference.

$$\widetilde{\mathcal{P}}_{saliency} = \widetilde{\mathcal{P}} - \widetilde{\mathcal{P}}_{coun}, \tag{8}$$

The obtained $\widetilde{\mathcal{P}}_{saliency}$ is fed into the cross-entropy loss and the process is supervised using corresponding labels.

$$\begin{aligned}&\mathcal{L}_{ce}\left(y, \widetilde{\mathcal{P}}_{saliency}\right)\\ &= -\frac{1}{\left|\widetilde{\mathcal{X}}_{train}\right|}\sum_{m=1}^{\left|\widetilde{\mathcal{X}}_{train}\right|}\log\frac{\exp\left(\xi\widetilde{\mathcal{P}}^{(m)}_{saliency}\right)}{\sum\limits_{i\neq m}\exp\left(\xi\widetilde{\mathcal{P}}^{(i)}_{saliency}\right)},\end{aligned} \tag{9}$$

It is worth noting that the predictions for the perturbed training samples $\widetilde{\boldsymbol{X}}$ are also optimized using the cross-entropy loss.

$$\mathcal{L}_{ce}\left(y, \widetilde{\mathcal{P}}\right) = -\frac{1}{\left|\widetilde{\mathcal{X}}_{train}\right|}\sum_{m=1}^{\left|\widetilde{\mathcal{X}}_{train}\right|}\log\frac{\exp\left(\delta\widetilde{\mathcal{P}}^{(m)}\right)}{\sum\limits_{i\neq m}\exp\left(\delta\widetilde{\mathcal{P}}^{(i)}\right)}, \tag{10}$$

Thus, the total loss during base training can be expressed as:

$$\mathcal{L}^{base}_{total} = \alpha\mathcal{L}_{ce}\left(y, \mathcal{P}\right) + \beta\mathcal{L}_{ce}\left(y, \widetilde{\mathcal{P}}\right) + \gamma\mathcal{L}_{ce}\left(y, \widetilde{\mathcal{P}}_{saliency}\right), \tag{11}$$

Where α, β, and γ are hyperparameters. Our causal learning strategy can determine the impact of learned features on classification by subtracting counterfactual predictions from factual predictions, thus encouraging the model to learn more influential features. Additionally, we continuously interfere with the model's learning of discriminative features by causally learning from perturbed sample features, prompting the model to learn robust features from more difficult environments as a way to improve the model's generalization ability and the model's capability to recognize novel classes in incremental sessions.

4.2 Fourier transform

To further overcome catastrophic forgetting in FSCIL, similar to many methods, we choose to freeze the obtained feature extractor $\boldsymbol{W}_{NET}$ after training from the base training data, and do not update it again in the subsequent incremental sessions. However, since the feature extractor has not learned the novel classes, a feature extractor that performs well in base tasks may be overwhelmed when facing data from the novel classes. In other words, the features extracted by the feature extractor directly from the novel classes will inevitably be biased and deformed. When faced with limited data from the novel classes, the frozen feature extractor cannot update the parameters to optimize the learning of the features from the novel classes. Therefore, it becomes more challenging to learn the new tasks. Thus, we believe that additional re-learning of the features of the limited novel class data is particularly important for recognizing novel classes.

In recent years, the Fourier transform [46, 47] has been widely used in image processing. At its core, the Fourier transform decomposes a time function (signal) into its constituent frequencies. This is critical in many practical scenarios; analyzing the frequency components of a signal provides more insight than examining the signal in the original time domain. If a convolution calculation is performed on a signal, it is equivalent to performing a multiplication calculation in the spectrum. A time domain convolution calculation is a frequency domain multiplication calculation, which is why convolution is referred to as filtering in many neural networks. When processing images, the presence of noise is inevitable. The Fourier transform can help to separate the noise component from the image (signal), thus making it easier to reduce or eliminate noise and improve the quality of features. Therefore, it is beneficial in image processing for edge detection and image filtering. Therefore, to further improve the model's ability to recognize novel classes, we introduce the Fast Fourier Transform [3, 31] into our CIFSCIL framework to enhance the features ignored by the feature extractor in the previous part of the module.

First, we get the features $\mathcal{F}(\boldsymbol{X}_n)$ extracted by the feature extractor $\boldsymbol{W}_{NET}$:

$$\mathcal{F}(\boldsymbol{X}_n) = f_{NET}(\boldsymbol{X}_n, \boldsymbol{W}_{NET}), \tag{12}$$

Where $\boldsymbol{X}_n \in \mathcal{D}^t_{train} \sim \{\mathcal{X}^t_{train}, \mathcal{Y}^t_{train}\}$, $\mathcal{D}^t_{train} = \{(x_i, y_i)\}^n_{i=1}$, and $x_i \in \mathcal{X}^t_{train}$ is the instance and $y_i \in \mathcal{Y}^t_{train}$ is the corresponding label, and $t > 0$.

Then, the Fourier transform is utilized to obtain the frequency domain features and further go through the residual block.

$$\mathcal{F}_{fri}(\boldsymbol{X}_n) = F_{ou}(\mathcal{F}(\boldsymbol{X}_n)), \tag{13}$$

$$\mathcal{F}_{res}(\boldsymbol{X}_n) = project(conv(\mathcal{F}(\boldsymbol{X}_n))), \tag{14}$$

Where F_{ou} is the Fourier transform, $conv$ is a convolution operation, and $project$ is a feature mapping layer for further extraction of detailed information in the features. In the residual block, the learned residuals can effectively harmonize the frequency domain features, e.g., to recover lost or damaged textures in the foreground region in the frequency domain, and to compensate for feature information ignored by the feature extractor.

$$\mathcal{F}_{aug}(\boldsymbol{X}_n) = \mathcal{F}_{res}(\boldsymbol{X}_n) + \mathcal{F}_{fri}(\boldsymbol{X}_n), \tag{15}$$

After passing through the residual block, the output $\mathcal{F}_{aug}(\boldsymbol{X}_n)$ is then transformed back to the spatial domain by Fourier inverse transform to obtain the enhanced features.

$$\mathcal{F}_{cls}(\boldsymbol{X}_n) = inv(\mathcal{F}_{aug}(\boldsymbol{X}_n)), \tag{16}$$

Where inv is the Fourier inverse transform operation. After the final feature $\mathcal{F}_{cls}(\boldsymbol{X}_n)$ is obtained, it is fed directly into the novel classifier, which is optimized by using the cross-entropy loss.

$$\mathcal{P}^n = f^n_{FC}(\mathcal{F}_{cls}(\boldsymbol{X}_n), \boldsymbol{W}^n_{FC}), \tag{17}$$

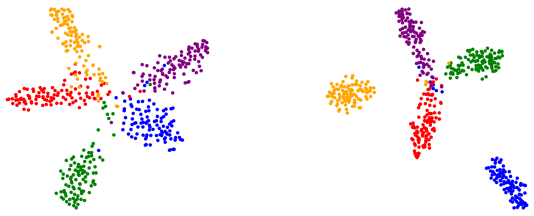
Where $\boldsymbol{W}^n_{FC}$ indicates the weights of the novel classifier.

$$\mathcal{L}_{ce}\left(y^n, \mathcal{P}^n\right) = -\frac{1}{\left|\mathcal{X}^t_{train}\right|}\sum_{m=1}^{\left|\mathcal{X}^t_{train}\right|}\log\frac{\exp\left(\lambda\mathcal{P}^{(m)}\right)}{\sum\limits_{i\neq m}\exp\left(\lambda\mathcal{P}^{(i)}\right)}, \tag{18}$$

After the Fourier transformation process, the significant features in the model can be found more efficiently, and the edge part can be strengthened to some extent, which is more conducive for our classifier to recognize the features of the novel classes. In addition, since the feature extractor has not actually been trained with the

Table 1: Comparison with SOTA methods on CUB200 dataset for FSCIL. The $*$ denotes the result report in the corresponding paper.

Method	Venue/Year	Acc. in each session (%) ↑											PD ↓
		0	1	2	3	4	5	6	7	8	9	10	
Finetune*	CVPR2020	68.68	43.70	25.05	17.72	18.08	16.95	15.10	10.06	8.93	8.93	8.47	60.21
CEC* [51]	CVPR2021	75.85	71.94	68.50	63.50	62.43	58.27	57.73	55.81	54.83	53.52	52.28	23.57
F2M* [37]	NIPS2021	81.07	78.16	75.57	72.89	70.86	68.17	67.01	65.26	63.36	61.76	60.26	20.81
CLOM* [58]	CVPR2022	79.57	76.07	72.94	69.82	67.80	65.56	63.94	62.59	60.62	60.34	59.58	19.99
MetaFSCIL [5]	CVPR2022	75.90	72.41	68.78	64.78	62.96	59.99	58.30	56.85	54.78	53.82	52.64	23.26
FACT [55]	CVPR2022	75.90	73.23	70.84	66.13	65.56	62.15	61.74	59.83	58.41	57.89	56.94	18.96
LIMIT [56]	TPAMI2023	75.89	73.55	71.99	68.14	67.42	63.61	62.40	61.35	59.91	58.66	57.41	18.48
GKEAL[57]	CVPR2023	78.88	75.62	72.32	68.62	67.23	64.26	62.98	61.89	60.20	59.21	58.67	20.21
MCNet[15]	TIP2023	77.57	73.96	70.47	65.81	66.16	63.81	62.09	61.82	60.41	60.09	59.08	18.49
SAVC[39]	CVPR2023	81.85	77.92	74.95	70.21	69.96	67.02	66.16	65.30	63.84	63.15	62.50	19.35
CIFSCIL (Ours)	–	**82.23**	**79.70**	**76.92**	**71.89**	**71.90**	**68.99**	**68.89**	**67.54**	**65.78**	**65.31**	**64.41**	**17.82**



(a) SAVC (b) Ours

Figure 4: The t-SNE visualization on the CIFAR100 dataset of the embeddings learned by various methods. Dots with different colors represent data points from different classes.

data of the new class, the features of the new class data extracted directly by it will tend to be biased toward the features of the base class data, i.e., there will be biases and distortions to a certain extent. After passing the Fourier transform, we can alleviate this deformation to some extent, and also help to separate the noise component in features, remove the features belonging to the base classes information in the feature extractor, and improve the quality of the features.

5 EXPERIMENTS

In this section, extensive experiments are conducted on three commonly used CIL datasets, and we compared our proposed CIFSCIL with state-of-the-art FSCIL approaches. In addition, ablation studies were also performed to demonstrate the impact of the causal inference learning strategy and the Fourier transform process in our CIFSCIL on the final performance of FSCIL.

5.1 Datasets

Following the benchmark setting [41], the CIFAR100 [18], miniImageNet [36], and CUB200 [44] are used.

- CIFAR100: It contains a total of 100 different classes, and each class contains 600 32×32 RGB images, of which 500 are used as training images and 100 as test images. We follow the division in [41] where 60 classes and 40 classes are used as base and novel classes, respectively. The 40 novel classes are further divided into 8 incremental sessions, and each new session is a 5-way 5-shot classification task.
- miniImageNet: This is a subset of the ImageNet dataset. It contains 100 classes, each with 600 color images of 84×84 size. We divide the 100 classes into 60 base classes and 40 incremental classes according to [41]. The 40 novel classes are further divided into 8 sessions of 5 classes each, and each class has 5 training images in the incremental session.
- Caltech-UCSD Birds-200-2011 (CUB200): CUB200 contains 200 bird categories with about 60 images per category, totaling 11,788 images, each with a size of 224×224. We divide the 200 categories into 100 base categories and 100 new categories according to the division in [41]. The 100 new categories are further ambiented into 10 incremental sessions, each of which is a 10-way 5-shot task.

5.2 Experiments details and evaluation metric

Following the setting in [41], we adopt ResNet18 [12] backbone for miniImageNet and CUB200, and ResNet20 [16] for experiments on CIFAR100. We use SGD with 0.9 momentum to optimize the model. The initial learning rate is 0.1 for CIFAR100 and miniImageNet, and 0.002 for CUB200 in the base session. We evaluate the proposed method in terms of performance dropping rate (PD) and Top 1 accuracy (Acc) [41] obtained in each session, and PD is used to measure the absolute accuracy in the last session, i.e., PD = A_0-A_N, where A_0 is the classification accuracy in the base session and A_N is the accuracy in the last session. Reporting the performance drop is meaningful as some methods may give good results mainly due to a well-trained network on the base dataset.

5.3 Comparison with state-of-the-art methods

We compared the performance of our CIFSCIL with the state-of-the-art methods at each session on the miniImageNet, CIFAR100, and CUB200 datasets. The accuracy of all the methods decreases with incremental sessions for two reasons. Firstly, the model takes in new class data in each phase, giving it more choices to distinguish from, naturally leading to decreased accuracy. The second reason is caused by the main dilemma of FSCIL, which is catastrophic forgetting, where the constant forgetting of knowledge about old tasks leads to a constant decrease in accuracy as well [57]. The

Table 2: Comparison with SOTA methods on CIFAR100 dataset for FSCIL. The $*$ denotes the result report in the corresponding paper.

Method	Venue/Year	Acc. in each session (%) ↑									PD ↓
		0	1	2	3	4	5	6	7	8	
Finetune*	CVPR2020	64.10	39.61	15.37	9.80	6.67	3.80	3.70	3.14	2.65	61.45
CEC* [51]	CVPR2021	73.07	68.88	65.26	61.19	58.09	55.57	53.22	51.34	49.14	23.93
F2M* [37]	NIPS2021	64.71	62.05	59.01	55.58	52.55	49.96	48.08	46.28	44.67	20.04
CLOM* [58]	CVPR2022	74.20	69.83	66.17	62.39	59.26	56.48	54.36	52.16	50.25	23.95
MetaFSCIL [5]	CVPR2022	74.50	70.10	66.84	62.77	59.48	56.52	54.36	52.56	49.97	24.53
FACT [55]	CVPR2022	74.60	72.09	67.56	63.52	61.38	58.36	58.26	54.24	52.10	23.5
LIMIT [56]	TPAMI2023	73.81	72.09	67.87	63.89	60.70	57.77	55.67	53.52	51.23	22.58
GKEAL[57]	CVPR2023	74.01	70.45	67.01	63.08	60.01	57.30	55.50	53.39	51.40	22.61
MCNet[15]	TIP2023	73.30	69.34	65.72	61.70	58.75	56.44	54.59	53.01	50.72	22.58
SAVC [39]	CVPR2023	**78.77**	73.31	69.31	64.93	61.70	59.25	57.13	55.19	53.12	25.65
CIFSCIL (Ours)	–	78.47	**77.05**	**74.49**	**70.88**	**67.94**	**65.80**	**64.30**	**62.08**	**61.02**	**17.45**

Table 3: Comparison with SOTA methods on miniImageNet dataset for FSCIL. The $*$ denotes the result report in the corresponding paper.

Method	Venue/Year	Acc. in each session (%) ↑									PD ↓
		0	1	2	3	4	5	6	7	8	
Finetune*	CVPR2020	61.31	27.22	16.37	6.08	2.54	1.56	1.93	2.60	1.40	59.91
CEC* [51]	CVPR2021	72.00	66.83	62.97	59.43	56.70	53.73	51.19	49.24	47.63	24.37
F2M* [37]	NIPS2021	67.28	63.80	60.38	57.06	54.08	51.39	48.82	46.58	44.65	22.63
CLOM* [58]	CVPR2022	73.08	68.09	64.16	60.41	57.41	54.29	51.54	49.37	48.00	25.08
MetaFSCIL [5]	CVPR2022	72.04	67.94	63.77	60.29	57.58	55.16	52.9	50.79	49.19	22.85
FACT [55]	CVPR2022	75.56	69.63	66.38	62.77	60.60	57.33	54.34	52.16	50.49	25.07
LIMIT [56]	TPAMI2023	72.32	68.47	64.30	60.78	57.95	55.07	52.70	50.72	49.19	23.13
GKEAL[57]	CVPR2023	73.59	68.90	65.33	62.29	59.39	56.70	54.20	52.59	51.31	22.28
MCNet[15]	TIP2023	72.33	67.70	63.50	60.34	57.59	54.70	52.13	50.41	49.08	23.25
SAVC [39]	CVPR2023	**81.12**	76.14	72.43	68.92	66.48	62.95	59.92	58.39	57.11	24.01
CIFSCIL (Ours)	–	80.9	**77.85**	**75.34**	**72.93**	**70.94**	**67.78**	**65.07**	**63.58**	**62.21**	**18.69**

methods we compare include: classical incremental-frozen FSCIL methods, i.e., F2M [37], CEC [51], and FACT [55], and recent state-of-the-art methods, i.e., SAVC [39], GKEAL [57], MCNet [15]. We also show a naive baseline that directly finetunes the model with limited data as "finetune". As observed in the Figure 5, we report the performance curve of each method. The most recent FSCIL teniques like SAVC [39], MCNet [15] and GKEAL [57], they significantly outperforms CEC [51] and F2M [37]. Our CIFSCIL method shows the best results in overcoming forgetting compared to the state-of-the-art methods, and it also can be seen in the line graph that the trend of our method is much flatter in subsequent incremental sessions, which indicates a good trade-off between mitigating the catastrophic forgetting and the capability of learning new tasks.

We report our experimental results on the CUB200, CIFAR100, and miniImageNet datasets in detail in Table 1, Table 2, and Table 3, respectively. Our approach effectively improves the model's ability to recognize novel classes through generalized learning of the base classes and additional learning of novel classes, effectively mitigating catastrophic forgetting, as demonstrated by the consistently highest accuracy of incremental sessions in the experimental results. For example, our method achieves final accuracy of 61.02%, 64.41%, and 62.21% for CIFAR100, CUB200, and miniImageNet, respectively, which are 7.90%, 1.91%, and 5.10% better than the current SOTA method SAVC, respectively. However, we can note that, except for the CUB200 dataset, the other two datasets do not have the highest accuracy in the base session in either case, which we speculate is due to the fact that the module discards some information about the base classes samples to learn features of a more generalized nature, and so has a very subtle effect on the recognition of the base classes. The CUB200 dataset, on the other hand, exhibits slightly higher accuracy in the base session, which is due to the good learning of the base class data caused by the higher image resolution of the CUB200 dataset. Additionally, for the evaluation metric of PD, the PD value of our CIFSCIL compared with the classical and the latest methods is the lowest on all three datasets, proving that our method achieves a more appropriate balance between alleviating forgetting and learning novel knowledge.

Moreover, to further demonstrate that our causal inference learning strategy (CI) and the Fourier transform process (FT) can better overcome catastrophic forgetting and learn incremental classes, we introduce the widely used t-SNE [42] tool to visualize feature distribution maps in 2D space. We visualize the embedding space on the CIFAR100 dataset in Figure 4. As it can be observed from the figure, our approach allows for more precise separation of different classes and tighter clustering of the same class.

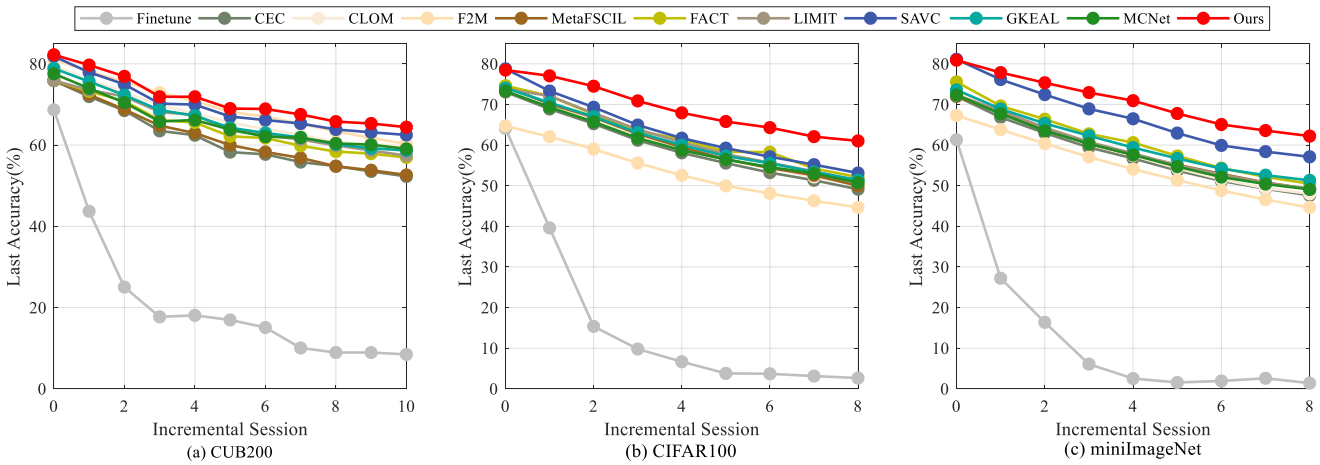


Figure 5: Comparison with SOTA methods on CUB200, CIFAR100, and miniImageNet benchmarks.

Table 4: Ablation studies on miniImageNet benchmark. CI, FT denote causal inference learning, Fourier transform, respectively.

CI	FT	Acc. in each session (%) ↑									PD ↓
		0	1	2	3	4	5	6	7	8	
		81.12	76.14	72.43	68.92	66.48	62.95	59.92	58.39	57.11	24.01
✓		80.72	77.68	74.67	71.95	69.80	66.37	63.69	62.24	60.76	19.96
	✓	80.72	77.40	74.14	71.28	69.19	65.75	63.02	61.39	59.89	20.83
✓	✓	80.90	**77.85**	**75.34**	**72.93**	**70.94**	**67.78**	**65.07**	**63.58**	**62.21**	**18.69**

5.4 Ablation study

We conduct ablation studies to prove the importance of our proposed components. Our ablation experiments include adding only the CI strategy, adding only the FT process, and both the CI and FT acting together. Both modules are important to CIFSCIL. In particular, the CL makes a critical contribution.

As shown in Table 4, we validate the role played by our different modules in the miniImageNet dataset. In the case of adding only the CI strategy, the CI improves the model's generalization ability when the model is trained for the base classes by learning the causal relationship of the interference samples and has a lasting good impact on the subsequent recognition of the novel classes. Our experimental results clearly show that our CIFSCIL exhibits the highest accuracy at each incremental session, and this gap becomes more pronounced the further the incremental sessions go. This proves the significant role played by our CI strategy in allowing the model to recognize novel classes and overcome forgetfulness. In the case of adding only the FT process, the accuracy in the incremental sessions is somewhat improved. The two modules have different focuses; the CI mainly allows the model to improve its ability to cope with unfamiliar information and avoid catastrophic forgetting by learning sufficient base class data, while the FT mainly focuses on the features of the novel class data, allowing the model to ultimately represent the features more accurately. The CI and FT work together to make the model not only mitigate catastrophic forgetting, but also show excellent performance for learning novel classes. As demonstrated in the last line in Table 4, we show the performance of the CI and FT acting together, and the results indicate that a new state-of-the-art performance can be achieved on the FSCIL.

6 CONCLUSION

In this paper, we designed a CIFSCIL framework to effectively address the issues of catastrophic forgetting and overfitting in the task of FSCIL. Through experimental performance evaluation on three CIL datasets commonly used in FSCIL, our CIFSCIL achieves the best performance and effectively mitigates the problems of catastrophic forgetting and overfitting compared to both classical baseline and state-of-the-art methods. The experiment results demonstrated that incorporating a causal inference strategy in our CIFSCIL to learn the causal relationship between features and outcomes during the base phase training can effectively generalize the feature-capturing ability to incremental sessions and improve the recognition of novel classes while alleviating the model's forgetfulness of old tasks. Additionally, the Fourier transform process in our CIFSCIL can effectively strengthen feature learning for novel classes and facilitate the classifier to better capture the features of the few-shot data in CIL.

ACKNOWLEDGMENTS

This work was supported by the Fundamental Research Funds for the Central Universities, China (SWU-KT22032).

REFERENCES

[1] Eden Belouadah, Adrian Popescu, and Ioannis Kanellos. 2021. A comprehensive study of class incremental learning algorithms for visual tasks. *Neural Networks* 135 (2021), 38–54.

[2] Monowar H Bhuyan, Dhruba K Bhattacharyya, and Jugal K Kalita. 2012. Survey on incremental approaches for network anomaly detection. *arXiv preprint arXiv:1211.4493* (2012).

[3] Junyan Cao, Yan Hong, and Li Niu. 2023. Painterly image harmonization in dual domains. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 268–276.

[4] Francisco M Castro, Manuel J Marín-Jiménez, Nicolás Guil, Cordelia Schmid, and Karteek Alahari. 2018. End-to-end incremental learning. In *Proceedings of the European conference on computer vision (ECCV)*. 233–248.

[5] Zhixiang Chi, Li Gu, Huan Liu, Yang Wang, Yuanhao Yu, and Jin Tang. 2022. Metafscil: A meta-learning approach for few-shot class incremental learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 14166–14175.

[6] Prithviraj Dhar, Rajat Vikram Singh, Kuan-Chuan Peng, Ziyan Wu, and Rama Chellappa. 2019. Learning without memorizing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 5138–5146.

[7] Guneet S Dhillon, Pratik Chaudhari, Avinash Ravichandran, and Stefano Soatto. 2019. A baseline for few-shot image classification. *arXiv preprint arXiv:1909.02729* (2019).

[8] Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*. PMLR, 1126–1135.

[9] Spyros Gidaris and Nikos Komodakis. 2018. Dynamic Few-Shot Visual Learning without Forgetting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4367–4375.

[10] Heitor Murilo Gomes, Jean Paul Barddal, Fabrício Enembreck, and Albert Bifet. 2017. A survey on ensemble learning for data stream classification. *ACM Computing Surveys (CSUR)* 50, 2 (2017), 1–36.

[11] Ian J Goodfellow, Mehdi Mirza, Da Xiao, Aaron Courville, and Yoshua Bengio. 2013. An empirical investigation of catastrophic forgetting in gradient-based neural networks. *arXiv preprint arXiv:1312.6211* (2013).

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[13] Saihui Hou, Xinyu Pan, Chen Change Loy, Zilei Wang, and Dahua Lin. 2019. Learning a unified classifier incrementally via rebalancing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 831–839.

[14] Muhammad Abdullah Jamal and Guo-Jun Qi. 2019. Task agnostic meta-learning for few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11719–11727.

[15] Zhong Ji, Zhishen Hou, Xiyao Liu, Yanwei Pang, and Xuelong Li. 2023. Memorizing complementation network for few-shot class-incremental learning. *IEEE Transactions on Image Processing* 32 (2023), 937–948.

[16] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences* 114, 13 (2017), 3521–3526.

[17] Eric Klopfer. 2008. *Augmented learning: Research and design of mobile educational games*. MIT press.

[18] Alex Krizhevsky, Geoffrey Hinton, et al. 2009. Learning multiple layers of features from tiny images. (2009).

[19] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. *Advances in neural information processing systems* 30 (2017).

[20] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *nature* 521, 7553 (2015), 436–444.

[21] Zhizhong Li and Derek Hoiem. 2017. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence* 40, 12 (2017), 2935–2947.

[22] Yaoyao Liu, Bernt Schiele, and Qianru Sun. 2020. An Ensemble of Epoch-wise Empirical Bayes for Few-shot Learning. In *European Conference on Computer Vision (ECCV)*.

[23] Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X Yu. 2019. Large-scale long-tailed recognition in an open world. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2537–2546.

[24] Dengsheng Lu and Qihao Weng. 2007. A survey of image classification methods and techniques for improving classification performance. *International journal of Remote sensing* 28, 5 (2007), 823–870.

[25] Xu Luo, Jing Xu, and Zenglin Xu. 2022. Channel Importance Matters in Few-Shot Image Classification. In *International Conference on Machine Learning*.

[26] Arun Mallya, Dillon Davis, and Svetlana Lazebnik. 2018. Piggyback: Adapting a single network to multiple tasks by learning to mask weights. In *Proceedings of the European conference on computer vision (ECCV)*. 67–82.

[27] Arun Mallya and Svetlana Lazebnik. 2018. Packnet: Adding multiple tasks to a single network by iterative pruning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 7765–7773.

[28] Marc Masana, Xialei Liu, Bartłomiej Twardowski, Mikel Menta, Andrew D Bagdanov, and Joost Van De Weijer. 2022. Class-incremental learning: survey and performance evaluation on image classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 5 (2022), 5513–5533.

[29] Martial Mermillod, Aurélia Bugaiska, and Patrick Bonin. 2013. The stability-plasticity dilemma: Investigating the continuum from catastrophic forgetting to age-limited learning effects. , 504 pages.

[30] Prakash M Nadkarni, Lucila Ohno-Machado, and Wendy W Chapman. 2011. Natural language processing: an introduction. *Journal of the American Medical Informatics Association* 18, 5 (2011), 544–551.

[31] Henri J Nussbaumer and Henri J Nussbaumer. 1982. *The fast Fourier transform*. Springer.

[32] Judea Pearl. 2019. The seven tools of causal inference, with reflections on machine learning. *Commun. ACM* 62, 3 (2019), 54–60.

[33] Amal Rannen, Rahaf Aljundi, Matthew B Blaschko, and Tinne Tuytelaars. 2017. Encoder based lifelong learning. In *Proceedings of the IEEE international conference on computer vision*. 1320–1328.

[34] Yongming Rao, Guangyi Chen, Jiwen Lu, and Jie Zhou. 2021. Counterfactual Attention Learning for Fine-Grained Visual Categorization and Re-identification. In *ICCV*.

[35] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H. Lampert. 2017. iCaRL: Incremental Classifier and Representation Learning. In *Proceedings of the IEEE international conference on computer vision*.

[36] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision* 115 (2015), 211–252.

[37] Guangyuan Shi, Jiaxin Chen, Wenlong Zhang, Li-Ming Zhan, and Xiao-Ming Wu. 2021. Overcoming catastrophic forgetting in incremental few-shot learning by finding flat minima. In *NeurIPS*. 6747–6761.

[38] Jake Snell, Kevin Swersky, and Richard Zemel. 2017. Prototypical networks for few-shot learning. *Advances in neural information processing systems* 30 (2017).

[39] Zeyin Song, Yifan Zhao, Yujun Shi, Peixi Peng, Li Yuan, and Yonghong Tian. 2023. Learning with Fantasy: Semantic-Aware Virtual Contrastive Constraint for Few-Shot Class-Incremental Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 24183–24192.

[40] Adith Swaminathan and Thorsten Joachims. 2015. The self-normalized estimator for counterfactual learning. *advances in neural information processing systems* 28 (2015).

[41] Xiaoyu Tao, Xiaopeng Hong, Xinyuan Chang, Songlin Dong, Xing Wei, and Yihong Gong. 2020. Few-shot class-incremental learning. In *CVPR*. 12183–12192.

[42] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. journal of Machine Learning Research 9. *Nov (2008)* (2008).

[43] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[44] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. 2011. The caltech-ucsd birds-200-2011 dataset. (2011).

[45] Panqu Wang, Pengfei Chen, Ye Yuan, Ding Liu, Zehua Huang, Xiaodi Hou, and Garrison Cottrell. 2018. Understanding convolution for semantic segmentation. In *2018 IEEE winter conference on applications of computer vision (WACV)*. Ieee, 1451–1460.

[46] Zhi-Qin John Xu, Yaoyu Zhang, Tao Luo, Yanyang Xiao, and Zheng Ma. 2019. Frequency principle: Fourier analysis sheds light on deep neural networks. *arXiv preprint arXiv:1901.06523* (2019).

[47] Zhi-Qin John Xu, Yaoyu Zhang, and Yanyang Xiao. 2019. Training behavior of deep neural network in frequency domain. In *Neural Information Processing: 26th International Conference, ICONIP 2019, Sydney, NSW, Australia, December 12–15, 2019, Proceedings, Part I 26*. Springer, 264–274.

[48] Liuyi Yao, Zhixuan Chu, Sheng Li, Yaliang Li, Jing Gao, and Aidong Zhang. 2021. A survey on causal inference. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 15, 5 (2021), 1–46.

[49] Han-Jia Ye, Hexiang Hu, and De-Chuan Zhan. 2021. Learning Adaptive Classifiers Synthesis for Generalized Few-Shot Learning. *International Journal of Computer Vision* 129, 6 (2021), 1930–1953.

[50] Han-Jia Ye, Hexiang Hu, De-Chuan Zhan, and Fei Sha. 2020. Few-Shot Learning via Embedding Adaptation with Set-to-Set Functions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 8808–8817.

[51] Chi Zhang, Nan Song, Guosheng Lin, Yun Zheng, Pan Pan, and Yinghui Xu. 2021. Few-shot incremental learning with continually evolved classifiers. In *CVPR*. 12455–12464.

[52] Junting Zhang, Jie Zhang, Shalini Ghosh, Dawei Li, Serafettin Tasci, Larry Heck, Heming Zhang, and C-C Jay Kuo. 2020. Class-incremental learning via deep model consolidation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 1131–1140.

[53] Xiheng Zhang, Yongkang Wong, Xiaofei Wu, Juwei Lu, Mohan Kankanhalli, Xiangdong Li, and Weidong Geng. 2021. Learning causal representation for training cross-domain pose estimator via generative interventions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 11270–11280.

[54] Linglan Zhao, Jing Lu, Yunlu Xu, Zhanzhan Cheng, Dashan Guo, Yi Niu, and Xiangzhong Fang. 2023. Few-Shot Class-Incremental Learning via Class-Aware Bilateral Distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11838–11847.

[55] Da-Wei Zhou, Fu-Yun Wang, Han-Jia Ye, Liang Ma, Shiliang Pu, and De-Chuan Zhan. 2022. Forward compatible few-shot class-incremental learning. In *CVPR*. 9046–9056.

[56] Da-Wei Zhou, Han-Jia Ye, Liang Ma, Di Xie, Shiliang Pu, and De-Chuan Zhan. 2023. Few-Shot Class-Incremental Learning by Sampling Multi-Phase Tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 11 (2023), 12816–12831. https://doi.org/10.1109/TPAMI.2022.3200865

[57] Huiping Zhuang, Zhenyu Weng, Run He, Zhiping Lin, and Ziqian Zeng. 2023. GKEAL: Gaussian Kernel Embedded Analytic Learning for Few-Shot Class Incremental Task. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7746–7755.

[58] Yixiong Zou, Shanghang Zhang, Yuhua Li, and Ruixuan Li. 2022. Margin-based few-shot class-incremental learning with class-level overfitting mitigation. *Advances in neural information processing systems* 35 (2022), 27267–27279.