# Continual Few-shot Learning with Transformer Adaptation and Knowledge Regularization

Xin Wang
Tsinghua University
xin_wang@tsinghua.edu.cn

Yue Liu
Tsinghua University
liuyue17@mails.tsinghua.edu.cn

Jiapei Fan
Alibaba Group
jiapei.fjp@alibaba-inc.com

Weigao Wen
Alibaba Group
weigao.wwg@alibaba-inc.com

Hui Xue
Alibaba Group
hui.xueh@alibaba-inc.com

Wenwu Zhu*
Tsinghua University
wwzhu@tsinghua.edu.cn

## ABSTRACT

Continual few-shot learning, as a paradigm that simultaneously solves continual learning and few-shot learning, has become a challenging problem in machine learning. An eligible continual few-shot learning model is expected to distinguish all seen classes upon new categories arriving, where each category only includes very few labeled data. However, existing continual few-shot learning methods only consider the visual modality, where the distributions of new categories often indistinguishably overlap with old categories, thus resulting in the severe catastrophic forgetting problem. To tackle this problem, in this paper we study continual few-shot learning with the assistance of semantic knowledge by simultaneously taking both visual modality and semantic concepts of categories into account. We propose a Continual few-shot learning algorithm with Semantic knowledge Regularization (**CoSR**) for adapting to the distribution changes of visual prototypes through a Transformer-based prototype adaptation mechanism. Specifically, the original visual prototypes from the backbone are fed into the well-designed Transformer with corresponding semantic concepts, where the semantic concepts are extracted from all categories. The semantic-level regularization forces the categories with similar semantics to be closely distributed, while the opposite ones are constrained to be far away from each other. The semantic regularization improves the model's ability to distinguish between new and old categories, thus significantly mitigating the catastrophic forgetting problem in continual few-shot learning. Extensive experiments on CIFAR100, miniImageNet, CUB200 and an industrial dataset with long-tail distribution demonstrate the advantages of our **CoSR** model compared with state-of-the-art methods.

## CCS CONCEPTS

• **Computing methodologies** → **Computer vision**; **Learning paradigms**; **Machine learning approaches**; **Knowledge representation and reasoning**.

*Corresponding author



*WWW '23, April 30–May 04, 2023, Austin, TX, USA*

ACM ISBN 978-1-4503-9416-1/23/04.
https://doi.org/10.1145/3543507.3583262

## KEYWORDS

Continual Few-shot Learning, Semantic Knowledge Regularization

**ACM Reference Format:**
Xin Wang, Yue Liu, Jiapei Fan, Weigao Wen, Hui Xue, and Wenwu Zhu. 2023. Continual Few-shot Learning with Transformer Adaptation and Knowledge Regularization. In *Proceedings of the ACM Web Conference 2023 (WWW '23), April 30–May 04, 2023, Austin, TX, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3543507.3583262

## 1 INTRODUCTION

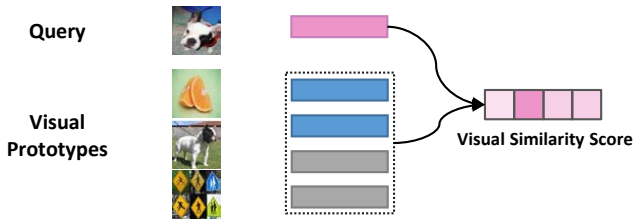
Deep neural networks (DNNs) have achieved great success when a large amount of labeled data are available. For example, DNNs are able to accurately conduct image classifications upon well training over enough labeled data. In practice, we always have new data and tasks arriving in sequence, craving for an ideal machine learning model which is able to recognize newly-arrived data associated with new classes and maintain the knowledge of previous classes simultaneously. Continual learning is such a learning paradigm, aiming to alleviate the catastrophic forgetting of old classes upon sequential arrival of new classes [16, 23]. Nevertheless, the amount of newly-arrived data is usually limited, thus requiring that the model can quickly adapt to new classes with few-shot data. Continual few-shot learning, as a paradigm that simultaneously solves continual learning and few-shot learning [1, 36], has attracted an increasing number of attentions in the research community recently.

Compared with the traditional learning paradigm, continual few-shot learning is more analogous to human learning since humans can learn new concepts from a limited amount of data and maintain most of the previously old knowledge simultaneously. There are two learning phases in continual few-shot learning, the base learning phase and the continual learning phase. In the base learning phase, the model is trained with base classes with full labeled data for each class. Then in the continual learning phase, the model is expected to learn new classes with a small amount of labeled data while maintaining the knowledge of base classes. Given that the labels from new classes have not appeared in the base classes, continual few-shot learning requires that the model should quickly adapt to the distribution changes from new classes as well as maintain the ability to distinguish all old classes. Due to the unavailable old classes during the continual learning phase, the distributions fitted based on new classes tend to overlap with those fitted based on old classes, which results in the failure of traditional machine learning approaches in distinguishing the new classes from old classes. In sum, there exist two key challenges in continual few-shot learning.

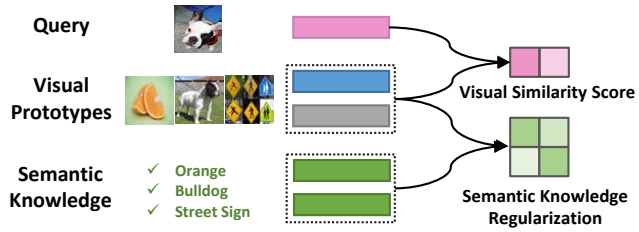
(1) The catastrophic forgetting issue that traditional deep learning models tend to over-fit the new classes while forgetting the knowledge of old classes.
(2) The requirement of fast learning ability with only a limited amount of labeled data from new classes.

On the one hand, there are existing works in continual learning, focusing on the problem of catastrophic forgetting. Some works [4, 16, 19] tackle the forgetting problem through constraining the parameter shifts in the deep learning model. Several works [2, 22, 23, 30] also propose to extend the model or learn parameter masks for new classes. Other works [8, 21, 27, 32] suggest using rehearsal memory by storing samples from previous classes or generating samples to alleviate forgetting. On the other hand, the above models on continual learning suffer from relatively large estimation errors when only a limited number of samples are available, motivating the advent of continual few-shot learning which learns prototypes based on the given support images, and classifies the input image according to distance criteria such as Euclidean distance and cosine distance etc. Several existing approaches [5, 24, 28, 44, 45, 47] generate adaptive prototypes through well-designed adaptive mechanisms which only utilize visual signals of images without considering the semantic knowledge hidden in text as well as the semantic association between the base and new classes. Other work [7] makes use of word embeddings to align visual prototypes in textual feature space, where the sparsely distributed textual space may not be suitable for visual prototype learning. However, the existing works severely ignore the importance of semantic knowledge in helping to distinguish both new and old classes upon continually arriving tasks, thus failing to solve the catastrophic forgetting and the fast learning problems simultaneously.

To solve the problem, we propose to extract the semantic knowledge from categorical information as a regularization for learning the consistent semantic visual prototypes. This design enables us to conduct continual classification through nearest neighbor instead of fine-tuning the model with a limited amount of data. Figure 1 demonstrates the comparison between existing works with only visual prototypes matching and our proposed model with semantic knowledge regularization. Given that the learnable visual prototypes can significantly reduce the estimation error with a limited amount of data samples, a Transformer-based adaptation mechanism is designed to align the visual and textual prototypes. Specifically, we propose a **co**ntinual few-shot learning model with **s**emantic **r**egularization, **CoSR**, which is a Transformer-based prototype adaptation mechanism adapting the distributions of visual prototypes with semantic knowledge regularization. As the sequential data from new classes arrive, our proposed CoSR model is able to achieve fast adaption through re-calculating the previous visual prototypes with few-shot labeled samples. With the single visual modality, the estimation of prototypes usually suffers from large errors. Thus, we employ semantic knowledge to assist the learning of visual prototypes, where the semantic knowledge is extracted from the textual prototype of each class through a linear affine layer. The previous visual prototypes and the textual prototypes are fed into a Transformer layer to obtain the adaptive prototypes, where the Transformer module aims at mining complex relationships between visual and textual prototypes to generate enhanced




**(a) Baselines with Only Visual Prototypes Matching**




**(b) Our CoSR Model with Semantic Knowledge Regularization**

**Figure 1: The concept of our proposed CoSR model. (a) Existing works usually learn prototypes using a single modality to classify the query image. (b) Our CoSR model learns the visual prototypes with semantic knowledge regularization. This Transformer-based prototype adaptation mechanism enhances visual prototypes with the semantic association among classes and thus alleviates the forgetting issue.**

prototypes for continual classification learning. The textual prototypes in the common latent space can provide valid anchors for the visual prototypes during the continual learning process, thus alleviating catastrophic forgetting significantly. Experiments on both public and industrial datasets demonstrate the advantages of the proposed CoSR model against baseline approaches.

The contributions of this work can be summarized as follows,

- We propose a continual few-shot learning algorithm with semantic knowledge regularization, CoSR, to tackle the catastrophic forgetting of previously old classes when quickly learning the new classes.
- We utilize semantic knowledge from each class to mine the natural semantic association among classes, significantly enhancing the learning process for visual features.
- We design a Transformer-based prototype adaptation mechanism to adapt visual prototypes with semantic knowledge regularization, improving the ability to distinguish old and new classes in continual few-shot learning.
- We conduct extensive experiments on both public and industrial datasets to demonstrate the superiority of our CoSR model over the state-of-the-art approaches.

## 2 RELATED WORK

In this section, we review related works on traditional continual learning, few-shot learning and the most recent continual few-shot learning, which are most relevant to our work.

**Continual Learning**. Continual learning aims to continually learn new categories and classify all seen categories. There are three main solutions from the perspective of reducing catastrophic forgetting or interference. Regularization-based methods [4, 16, 19] introduce prior distribution of parameters to penalize the shifts of important parameters or use knowledge distillation as data regularization. Model expansion-based methods [2, 22, 23, 30] propose to dedicate different parameters to each task by freezing the old task parameters or adding parameter masks. These works usually suffer from large model storage, with new classes continually arriving. Replay-based methods [8, 21, 27, 32, 42] store or generate old samples in memory to balance the old and new tasks and alleviate forgetting. Our work is related to continual learning, but more challenging with fewer samples in the new classes. With a few training samples, fine-tuning the feature space will bring a relatively large estimation error. Therefore, we solve the problem through prototype adaptation with semantic knowledge.

**Few-shot Learning**. Traditional few-shot learning aims to distinguish unseen classes with few given data, but ignores the distinction between the new and old categories [38]. There are three mainstreams of few-shot learning. Data generation based methods [12, 13, 29, 39] usually apply a pre-defined data generation function to expand the training data of unseen categories. Metric based methods [6, 25, 33, 34, 37] focus on embedding learning or refining for the few-shot tasks. Meta-learning based methods [11, 20, 40, 46] use prior knowledge to search an optimal initialized parameter. Our work is more related to metric-based methods since we use semantic knowledge as regularization to refine the prototypes in the feature space. Different from traditional few-shot learning, continual few-shot learning requires that the model should not only recognize new categories but also maintain the ability to distinguish between all seen categories.

**Continual Few-shot Learning**. Continual few-shot learning has been studied recently. F2M [31] suggests that flat local minima in the training of base classes can facilitate the learning of new classes. TOPIC [36] and TPCIL [35] are both topology preserved methods that learn and preserve the topology of the feature manifold to mitigate forgetting of the old classes. ERDIL [10] also selects representative samples in old categories to construct a relation graph for knowledge distillation, which eases the forgetting problem in continual few-shot learning. Some works [18, 43] solve the problem through calibrating features or classifiers of new categories to alleviate forgetting. Other works [5, 24, 28, 44, 45, 47] propose to generate adaptive prototypes (reference vectors or classifiers) through well-designed adaptive mechanism. However, they mostly focus on the single modality of images for feature learning, ignoring the semantic association among classes. Semantic knowledge is used in works [1, 7] as external information for continual few-shot learning. Cheraghia et al. [7] make use of word embeddings as semantic information to align visual and semantic vectors with an attention mechanism. The visual features are directly projected to the sparse textual space, which may suppress the effect of visual information. Akyürek et al. [1] propose to learn the new classifiers with the regularization of semantic similarity. The semantic association among different categories is linearly affine to the space of classifiers.

Our work is different from the above methods in semantic regularization terms. We design two schemes for the use of semantic knowledge. One is aligning the visual and textual vectors of each category with a Transformer. Another is extracting semantic prototypes in the visual space using textual information and then enhancing the visual prototypes with semantic ones.

## 3 THE PROPOSED METHOD

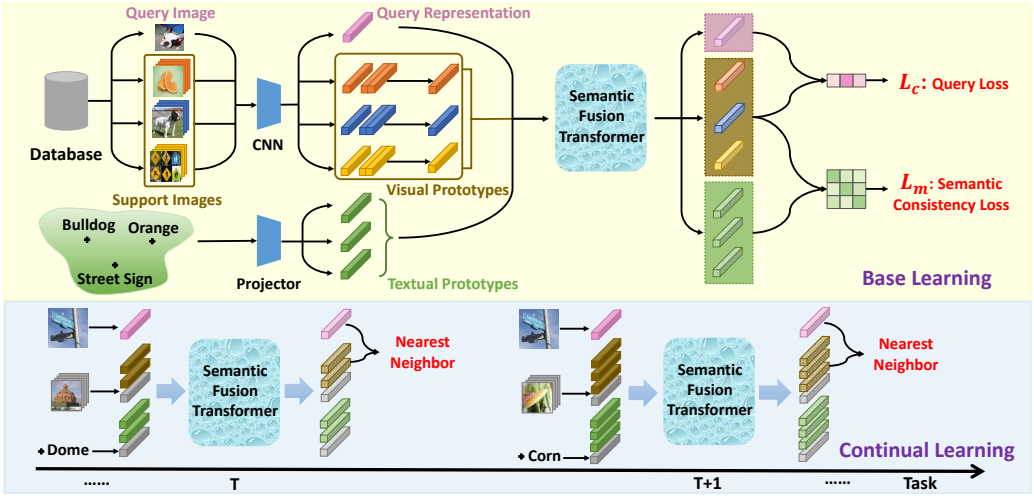
Our proposed CoSR model consists of two stages, the base learning phase and the continual learning phase. The CoSR model is first trained with base data and then continually learns with newly-arrived few-shot data.

### 3.1 Problem Formulation

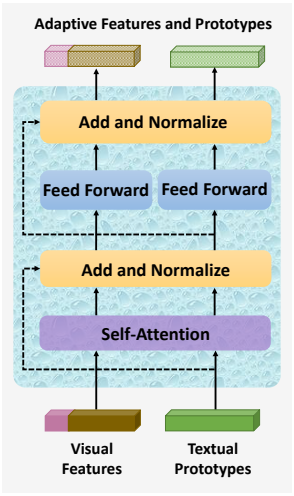
We define the continual few-shot learning setting as follows. Given a sequence of tasks $\{T_0, T_1, ..., T_I\}$, where the number of tasks is $I$. Each task $T_i$ ($0 \le i \le I$) contains a test set $D_{test}^{(i)}$ and a training set $D_{train}^{(i)}$, where $D_{train}^{(i)} = \{X_k^{(i)}, y_k^{(i)}\}_{k=1}^{K_i}$, $y_k^{(i)} \in C^{(i)}$. Here, $K_i$ is the number of samples in $D_{train}^{(i)}$ and $C^{(i)}$ is the label set of task $T_i$. Considering that there is no category overlap between different tasks, we have $\forall i, j, C^{(i)} \cap C^{(j)} = \emptyset$. In continual few-shot learning, task $T_0$ is named as the base task, including a full training set of base classes. Each task $T_{i,i>0}$ is a novel task or new task, which only includes a few samples for training. Usually, if there are $K$ data samples within each of the $C$ new classes in each novel task, then the setting is called $C$-way $K$-shot problem. To follow the common practice in continual few-shot learning, the training samples in the base task and novel tasks are severely unbalanced with $K$ normally being smaller than 5, while the size of the test set in each task is balanced. Therefore, we should first train a continual few-shot learning model on the base task and then continue to learn novel tasks in sequence. After training on $D_{train}^{(i)}$ for the $i$-th task, a continual few-shot learning model is evaluated on the test set for both the current task and those previous tasks, i.e., $\{D_{test}^{(0)}, D_{test}^{(1)}, ..., D_{test}^{(i)}\}$. This challenging setting requires a good continual few-shot learning model to quickly learn with a very small number of data samples for the novel task while maintaining the capability of distinguishing between previous classes and new classes simultaneously.

### 3.2 Continual Few-shot Learning with Semantic Knowledge Regularization

The challenge of continual few-shot learning comes from two aspects: i) the fast adaptive learning of new classes and ii) the catastrophic forgetting of old classes. In the base learning phase, the model is trained with base classes to learn the base task distribution in the latent feature space. In the continual learning phase, the distributions of new classes should be quickly learned given only a few samples, where the distributions of new classes usually tend to overlap with previously old classes in the latent feature space, thus leading to the catastrophic forgetting problem. To tackle the challenging issue, we extract semantic information from the textual modality to help the model to discover better feature space in which the new and previous classes do not overlap.




(a) Continual Few-shot Learning with Semantic Knowledge Regularization (CoSR)




(b) Semantic Fusion Transformer

**Figure 2: (a) The framework of our proposed CoSR model. In the base learning phase, the backbone and the semantic fusion Transformer are trained according to the query classification loss and semantic consistency loss. In the continual learning phase, the model obtains the semantically consistent visual prototypes through the Transformer. The nearest neighbor principle is used for classification. (b) The structure illustration of our semantic fusion Transformer. Best view in color.**

As shown in Figure 2 (a), the base learning phase includes multiple training episodes. In each episode, a support set of $C$-way $K$-shot is sampled from the base database. The query set is also sampled from the same class in the base database as the support set. Both the query and support images are fed into the learnable *CNN* backbone to obtain the corresponding visual features. According to the latent representations of support images transformed via the *CNN* backbone, the visual prototype, denoted as $e_v$, can be computed as the center of each category. Intuitively, we can calculate the Euclidean distance or cosine distance between the query representation $e_q$ and each visual prototype $e_v$ to decide which category the query belongs to. However, the representation space may change in the course of time because the distributions of new classes can drift away from the base classes. Thus, the visual prototypes obtained from previous categories may shift and even overlap with new prototypes in the latent space, resulting in the performance drop of the continual learning model.

Naturally, the semantic knowledge of each class can provide useful information for learning visual prototypes. For example, given that "bulldog" and "cat" both belong to "animal", the prototype of cats should be quickly learned and adapted with the help of semantic similarity upon learning the semantic concept of dogs. On the other hand, the semantic similarity can provide anchors to the visual prototypes, capable of reducing the distribution drift from old prototypes. Therefore, we introduce semantic knowledge regularization to prevent the prototype from drifting with the help of Transformer adaptation. The word embedding of each category can be employed as semantic knowledge, since the distribution of word embedding may reflect the semantic association among different categories. Specifically, the word embedding of each category is calculated via a pre-trained model before being projected to the latent space via *Projector*, the learnable backbone with a linear affine layer, where the projected vector $e_t$ is denoted as the textual prototype of each category.

Upon obtaining the visual prototypes and textual prototypes of all support classes, we design a semantic fusion Transformer to model the complex semantic relationships among different classes. As shown in Figure 2 (a), the visual features which concatenate the query representation and the visual prototypes, as well as the textual prototypes, are fed into the semantic fusion Transformer to obtain the adaptive query representation and prototypes. The detailed structure of the semantic fusion Transformer is illustrated in Figure 2 (b). Through the semantic fusion Transformer, the multimodal information containing visual features as well as textual prototypes can be fused and enhanced with each other via the self-attention mechanism. The feed-forward layer is designed to map the multimodal information into a common latent space. The output of the semantic fusion Transformer is the adaptive query representation $e'_q$, visual prototype $e'_v$ and textual prototype $e'_t$, resulting in the following expression illustrated by Eq.(1):

$$e'_q, e'_v, e'_t = \mathcal{T}(e_q, e_v, e_t), \tag{1}$$

where $\mathcal{T}$ represents the semantic fusion Transformer. We employ the adaptive textual prototype $e'_t$ as the anchor in the common latent space. The learnable visual prototype $e'_v$ is expected to align with anchor $e'_t$. Therefore, we propose the semantic knowledge regularization term in Eq.(2) to align $e'_v$ and $e'_t$, indicating whether the visual and textual prototypes are semantically consistent.

$$\begin{aligned} m_v &= \underset{t}{\text{argmax}}\,(e'_v \cdot e'_t), \\ L_m &= \sum_v CrossEntropy(m_v, g_v), \end{aligned} \tag{2}$$

where $m_v$ is the maximum calculated matching probability between the visual prototype $e'_v$ and textual prototype $e'_t$. $g_v$ is the ground

truth label which indicates the true matching between the visual prototype $e'_v$ and its corresponding textual prototype. The cross entropy loss is used for the semantic consistency loss $L_m$.

Besides, the adaptive query representation $e'_q$ is categorized utilizing the nearest neighbor principle. The distance between the adaptive query representation $e'_q$ and each visual prototype $e'_v$ is calculated so that $e'_q$ is assigned to the class with the minimum distance in the latent space. Without loss of generality, we use cosine distance as the metric and cross-entropy loss as the query loss term $L_q$, as shown in Eq.(3),

$$\begin{aligned} c_q &= \underset{v}{\operatorname{argmax}} \left(e'_q \cdot e'_v\right), \\ L_c &= \sum_q CrossEntropy(c_q, y_q), \end{aligned} \tag{3}$$

where $c_q$ is the predicted class label of the query representation $e'_q$ and $y_q$ is the ground truth class label of the query representation, indicating which class the query belongs to. The overall training objective is illustrated in Eq.(4) as follows,

$$L = L_c + \lambda \cdot L_m, \tag{4}$$

where $\lambda$ is the controlling factor to balance the query loss $L_c$ and the semantic consistency loss $L_m$.

## 3.3 Discussions

The continual few-shot learning procedure consists of two phases: i) base learning on the base task, where the backbones and the semantic fusion Transformer are trained on the full dataset of the base task; ii) continual learning on sequentially coming new tasks, where the model first learns visual and textual prototypes of new classes and then obtains the semantically consistent visual prototypes through Transformer adaptation. Upon learning every observable task, the proposed CoSR model is tested with all available classes.

**Base Learning on Base Task.** As shown in Figure 2 (a), the base learning phase includes multiple episodes. In each episode, the $C$-way $K$-shot support images as well as the query image are sampled from the database. The word embedding of all categories are precalculated to be employed as semantic information. The support images and query image are fed into the *CNN* backbone to obtain the visual features. We calculate the average latent representations of images in each class as the visual prototype for the corresponding category. The word embedding of each category is projected into the shared latent space with the visual prototypes through a linear affine layer. Then we use the semantic fusion Transformer depicted in Figure 2 (b) to generate adaptive visual and textual prototypes as well as the adaptive query representation. The whole model is trained in an end-to-end manner with the total objective $L$, where the query loss $L_c$ aims to distinguish from different categories in the latent space. The proposed semantic consistency loss $L_m$ will encourage the alignment of the visual and textual prototypes, as well as enhance the visual prototypes with semantic information indicated in the textual prototypes. Compared to existing methods that ignore the semantic association among classes, our proposed CoSR model utilizes the semantic association among categories to facilitate the learning of visual features. The textual prototypes learned in this phase provide anchors in the latent space, which simultaneously alleviates the forgetting issue in continual learning and accelerates the learning of new knowledge.

**Continual Learning on New Tasks.** After the base learning phase, the new tasks are expected to arrive in sequence. We assume that the backbones have been well-trained during the base learning phase, since the visual features usually reflect low-level visual information. The subsequent classification module then tends to play a more important role in achieving the continual few-shot learning. For the $t$-th task, we first obtain the visual prototypes through the newly arriving $C$-way $K$-shot samples and the textual prototypes of corresponding categories. The old and new textual prototypes may contain semantic information and reflect relationships between different classes. We use the semantic fusion Transformer to produce adaptive visual prototypes, which are enhanced by the semantic information carried in the textual prototypes. The adaptive visual prototypes benefit from the semantic information regularization and thus are capable of reducing the estimation errors with only a very small number of samples. Finally, the query image is assigned to its best matching category whose adaptive prototype has the minimum cosine distance from the adaptive query representation via the nearest neighbor principle. The proposed semantic fusion Transformer is able to learn adaptive prototypes for novel and unseen tasks, alleviating the catastrophic forgetting issue.

# 4 EXPERIMENTS

We conduct extensive experiments to compare the proposed CoSR model with several baselines on three public datasets. To verify the effectiveness of CoSR in the real-world scenario, we further test CoSR on an industrial dataset as well.

## 4.1 Experimental Settings

**Datasets.** Following existing literature [36, 44], we conduct continual few-shot learning experiments on three popular datasets, i.e., CIFAR100 [17], MiniImageNet [37] and CUB200 [41]. For CIFAR100 and MiniImageNet, we sample 60 classes as the base learning task. The novel task includes 5-way 5-shot samples each, and there are 8 tasks in the continual learning stage. As for CUB200, the base learning task contains 100 classes and each novel task has 10-way 5-shot samples, with the total number of novel tasks being set to 10. In addition, we test our proposed CoSR model on Goofish[1], an industrial dataset for online commodity purchase services, where the goal of the service provider is to recognize prohibited commodities through continual classification of the new items. The dataset includes images, titles, and descriptions for online items, serving as an appropriate scenario for multi-modal continual classification. The label of each data sample indicates the category of the item with multi-modal information, and each category has a human understandable name as semantic knowledge in the dataset. The size of Goofish dataset is 1.8 million items with 161 classes. We split the dataset into the training and test set according to the timestamp, resulting in the training set of 1.5 million items, and the test set of $300K$ items. Note that the training set and test set are split without overlap. There are several classes with only 30 samples in the test set which never appear in the training set, serving as a typical continual few-shot learning setting.

[1] https://goofish.com/

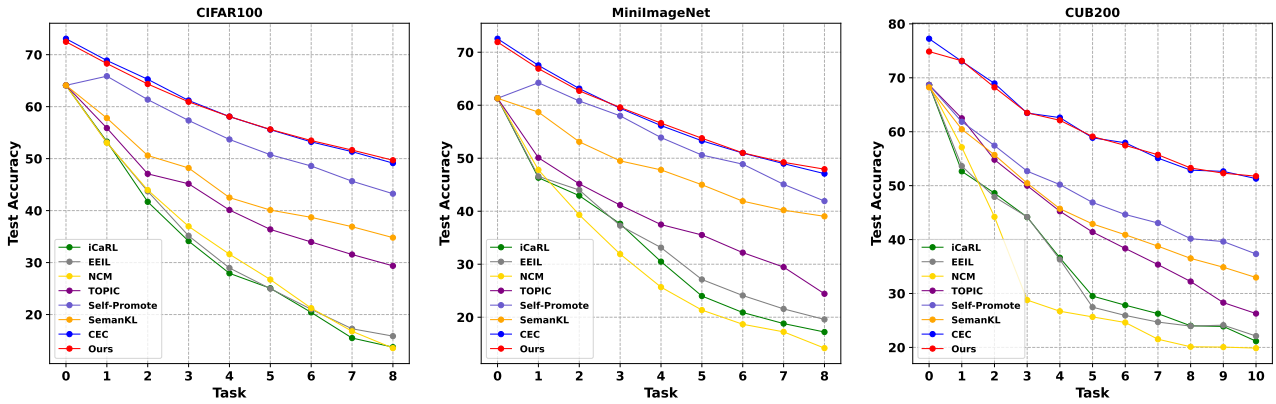



**Figure 3: The visualization of experimental results on CIFAR100, MiniImageNet and CUB200. We compare the proposed CoSR model with state-of-the-art baselines in continual few-shot learning. Best view in color.**

**Semantic Knowledge.** We use pre-trained vectorized word embedding as the semantic knowledge in the experiments. For CIFAR100 and MiniImageNet, we employ 300 dimensional GloVe [26] vectors as the extracted semantic knowledge. For CUB200, the 768 dimensional Bert [9] vectors are utilized to extract semantic knowledge. For the industrial dataset, we use the pre-trained 512 dimensional embedding of class names as semantic knowledge.

**Evaluations.** After training on the base learning task $T_0$, the test accuracy is calculated on the test set of the base learning task. Then the model sequentially gets trained upon the arrival of new tasks. After learning the $i$-th new task, the mean test accuracy is calculated on all observable tasks $T_0, T_1, ..., T_i$. The evaluation metric is the final test accuracy over all categories when the learning of the last task finishes. Besides, the performance drop rate (PD rate), i.e., the percentage of average accuracy drops in the last task w.r.t. the accuracy after the base task learning, is also used to measure the ability to learn new tasks while alleviating the catastrophic forgetting issue.

**Baselines.** We compare our CoSR with several state-of-the-art baselines. We take the "fine-tune" approach as the lower bound of the model performances, which just fine-tunes the model for each task. Several methods [3, 15, 27] for continual learning are compared as baselines. Existing state-of-the-art approaches [7, 36, 44, 47] for continual few-shot learning are also tested in the experiments.

**Implementations.** We conduct experiments using PyTorch library. Following the common practice [44], ResNet20 [14] is employed as *CNN* backbone for CIFAR100 and ResNet18 [14] is utilized as *CNN* backbone on MiniImageNet as well as CUB200. The *Projector* consists of one linear affine layer, the output of which is the same as the *CNN* backbone. We use SGD with momentum for optimization and the learning rate is set to 0.0001. The learning rate is decayed by 0.5 every 20 epochs. The total number of training epochs in the base learning phase is 100. For all experiments, we take the average performance of 5 runs and report the final results.

## 4.2 Experimental Results on Public Datasets

**CIFAR100.** We conduct continual few-shot learning experiments on CIFAR100 and visualize the results in Figure 3. The figure demonstrates the test accuracy over all observable categories after each task. We compare our method CoSR with other baselines. As shown in the figure, the initial test accuracies after learning the base task of CEC and our CoSR are higher than other models whose test accuracy is about 64%. The reason is that CEC and our CoSR both train the model through sampling dataset in multiple episodes, while other methods train the base model with the full dataset. With the increasing of novel tasks, the test accuracy of each model drops due to the forgetting of old classes and inefficient learning of new classes with only a few samples. The traditional continual learning approaches such as iCaRL, EEIL and NCM perform worse than other methods since they usually need a large amount of data to learn new classes. The continual few-shot learning approaches can learn the novel tasks efficiently, thus adapting quickly to the few-shot scenarios with a well-designed learning mechanism. Among all the methods, our CoSR has the best performances after learning all tasks. The performance drop of CoSR is also less than other methods, which indicates that our method has better ability to alleviate forgetting with the semantic knowledge regularization.

More specifically, we report the detailed experimental results in Table 1. In addition to the test accuracy after each task, we also report the performance drop rate to demonstrate the effect of different approaches on the forgetting issue. As shown in the table, our CoSR performs best in the continual learning phase with the final test accuracy 49.69%. Among the baselines, the fine-tuning method only has 2.65% accuracy as the lower bound since it fine-tunes the model with only a few samples and ignores the learning of old classes. Similar to our CoSR, SemanKL also uses word embeddings as semantic knowledge. But SemanKL performs worse than our CoSR since it projects the visual features to the word embedding space to regularize the learning of visual features. Due to the sparsity of the semantic space, this method cannot effectively utilize semantic information to learn good visual representations. Differently, we design a semantic fusion Transformer to fuse and align the visual and textual features, which is more effective than SemanKL and achieves higher accuracy. CEC, as a strong baseline, performs very close to our method. But our CoSR has a lower performance drop rate 31.45% compared to CEC with 32.75%, and outperforms CEC beginning at the arriving of the fourth task. The experimental

**Table 1: Experimental results on CIFAR100.**

| Model | Test accuracy after each task | | | | | | | | | PD rate (%)↓ |
|---|---|---|---|---|---|---|---|---|---|---|
| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | |
| fine-tune | 64.10 | 36.91 | 15.37 | 9.80 | 6.67 | 3.80 | 3.70 | 3.14 | 2.65 | 95.87 |
| iCaRL [27] | 64.10 | 53.28 | 41.69 | 34.13 | 27.93 | 25.06 | 20.41 | 15.48 | 13.73 | 78.58 |
| EEIL [3] | 64.10 | 53.11 | 43.71 | 35.15 | 28.96 | 24.98 | 21.01 | 17.26 | 15.85 | 75.27 |
| NCM [15] | 64.10 | 53.05 | 43.96 | 36.97 | 31.61 | 26.73 | 21.23 | 16.78 | 13.54 | 78.88 |
| TOPIC [36] | 64.10 | 55.88 | 47.07 | 45.16 | 40.11 | 36.38 | 33.96 | 31.55 | 29.37 | 54.18 |
| Self-promote [47] | 64.10 | 65.86 | 61.36 | 57.34 | 53.69 | 50.75 | 48.58 | 45.66 | 43.25 | 32.53 |
| SemanKL [7] | 64.10 | 57.80 | 50.60 | 48.20 | 42.50 | 40.10 | 38.70 | 36.90 | 34.80 | 45.71 |
| CEC [44] | **73.07** | **68.88** | **65.26** | **61.19** | **58.09** | 55.57 | 53.22 | 51.34 | 49.14 | 32.75 |
| **ours (CoSR)** | 72.48 | 68.29 | 64.36 | 60.93 | **58.09** | **55.63** | **53.48** | **51.64** | **49.69** | **31.44** |

**Table 2: Experimental results on MiniImageNet.**

| Model | Test accuracy after each task | | | | | | | | | PD rate (%)↓ |
|---|---|---|---|---|---|---|---|---|---|---|
| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | |
| fine-tune | 61.31 | 27.22 | 16.37 | 6.08 | 2.54 | 1.56 | 1.93 | 2.60 | 1.40 | 97.72 |
| iCaRL [27] | 61.31 | 46.32 | 42.94 | 37.63 | 30.49 | 24 | 20.89 | 18.80 | 17.21 | 71.93 |
| EEIL [3] | 61.31 | 46.58 | 44 | 37.29 | 33.14 | 27.12 | 24.1 | 21.57 | 19.58 | 68.06 |
| NCM [15] | 61.31 | 47.80 | 39.31 | 31.91 | 25.68 | 21.35 | 18.67 | 17.24 | 14.17 | 76.89 |
| TOPIC [36] | 61.31 | 50.09 | 45.17 | 41.16 | 37.48 | 35.52 | 32.19 | 29.46 | 24.42 | 60.17 |
| Self-promote [47] | 61.31 | 64.24 | 60.8 | 58.0 | 53.9 | 50.6 | 48.88 | 45.07 | 41.92 | **31.63** |
| SemanKL [7] | 61.31 | 58.70 | 53.10 | 49.50 | 47.80 | 45.00 | 41.90 | 40.20 | 39.04 | 36.32 |
| CEC [44] | **72.55** | **67.52** | **63.14** | 59.42 | 56.16 | 53.29 | 50.97 | 48.97 | 47.09 | 35.09 |
| **ours (CoSR)** | 71.92 | 66.91 | 62.71 | **59.59** | **56.63** | **53.78** | **51.01** | **49.23** | **47.93** | 33.36 |

results on CIFAR100 indicate the effectiveness of our CoSR with semantic knowledge regularization.

**MiniImageNet.** To evaluate the performance of CoSR, we conduct experiments on MiniImageNet. There are eight novel tasks in the experiment, and the visualization results of the average test accuracy after each task are shown in Figure 3. The phenomenon of performance drop is similar to the result on CIFAR100. These methods, which are specially designed for continual few-shot learning, such as CEC [44], SemanKL [7], Self-promote [47] and TOPIC [36] have better performance than others for continual learning, indicating that the well-designed algorithms for few-shot learning have the ability to learn fast from a few new samples. After learning the base task, the initial test accuracies of the CEC algorithm and our CoSR algorithm are higher than other models. Our CoSR begins to achieve the best test accuracy after learning the third task, which verifies the effectiveness of our Transformer-based prototypes adaptation mechanism in few shot continual learning.

Furthermore, we report detailed experimental results on MiniImageNet in Table 2. In addition to the test accuracy after each task, we also report the performance drop rate to demonstrate the severity of the forgetting problem for different methods during the continual learning phase. As shown in the table, our CoSR algorithm has a final test accuracy of 47.93% after learning all tasks, outperforming other benchmark models. In the baseline model, the fine-tuning method has an accuracy of only 1.40%, and the accuracies of other methods designed only for continual learning scenarios are lower than 20%. The algorithms designed for continual few-shot learning scenarios perform relatively well. The final accuracy of the CEC model reached 47.09%, which is the best method among the existing baselines. In terms of performance drop rate, our algorithm has a performance drop rate of 31.35%, which is better among all algorithms. It is worth noting that the performance drop rate of the Self-promote [47] is 31.63%, which is also very close to CoSR. However, in the absolute value of the final accuracy, CoSR clearly outperforms the Self-promote algorithm as new tasks coming. Overall, the experimental results on MiniImageNet show the effectiveness of our CoSR algorithm in continual few-shot learning scenarios. Using semantic knowledge regularization to enhance the visual feature learning can significantly improve the generalization performance of the model.

**CUB200.** We conduct experiments on CUB200 to verify the effectiveness of the CoSR algorithm. In the CUB200 experiment, the model needs to learn ten new tasks, and the average test accuracy after learning each task is shown in Figure 3. In the continual learning stage, the average accuracy of all models decreased to varying degrees. Among them, the performance of the iCaRL [27], EEIL [3] and NCM [15] algorithms decreased faster, indicating that they cannot quickly learn knowledge of new categories and distinguish new categories from old ones in the continual few-shot learning scenario. The reason is that the training of these algorithms on new tasks relies on a large amount of labeled data and thus has poor performance in scenarios with only a few new samples. In contrast, methods for continual few-shot learning perform much better, among which our CoSR algorithm achieves the highest accuracy among all algorithms. This demonstrates that using semantic knowledge regularization to constrain the learning of visual features is effective. The generalization performance of the model on new categories with few samples is improved in CoSR. Meanwhile, our CoSR algorithm also outperforms other benchmark algorithms in terms of performance drop rate.

We show more detailed experimental results in Table 3, including the test accuracy after each task and the final performance drop rate. Through the performance drop rate, we can evaluate how well the model alleviates the catastrophic forgetting problem in the continual few-shot learning scenario. From the experimental results of CUB200, we can conclude that our CoSR algorithm performs

**Table 3: Experimental results on CUB200.**

| Model | Test accuracy after each task | | | | | | | | | | | PD rate (%)↓ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | |
| fine-tune | 68.68 | 43.70 | 25.05 | 17.72 | 18.08 | 16.95 | 15.10 | 10.60 | 8.93 | 8.93 | 8.47 | 87.67 |
| iCaRL [27] | 68.68 | 52.65 | 48.61 | 44.16 | 36.62 | 29.52 | 27.83 | 26.26 | 24.01 | 23.89 | 21.16 | 69.19 |
| EEIL [3] | 68.68 | 53.63 | 47.91 | 44.20 | 36.30 | 27.46 | 25.93 | 24.70 | 23.95 | 24.13 | 22.11 | 67.81 |
| NCM [15] | 68.68 | 57.12 | 44.21 | 28.78 | 26.71 | 25.66 | 24.62 | 21.52 | 20.12 | 20.06 | 19.87 | 71.07 |
| TOPIC [36] | 68.68 | 62.49 | 54.81 | 49.99 | 45.25 | 41.4 | 38.35 | 35.36 | 32.22 | 28.31 | 26.28 | 61.74 |
| Self-promote [47] | 68.68 | 61.85 | 57.43 | 52.68 | 50.19 | 46.88 | 44.65 | 43.07 | 40.17 | 39.63 | 37.33 | 45.65 |
| SemanKL [7] | 68.23 | 60.45 | 55.70 | 50.45 | 45.72 | 42.90 | 40.89 | 38.77 | 36.51 | 34.87 | 32.96 | 51.69 |
| CEC [44] | **77.24** | **73.04** | **68.95** | 63.45 | 62.62 | 58.87 | **57.95** | 55.11 | 52.87 | **52.65** | 51.30 | 33.58 |
| **ours (CoSR)** | 74.87 | 73.15 | 68.23 | **63.50** | **62.72** | **59.10** | 57.46 | **55.73** | **53.28** | 52.31 | **51.75** | **30.87** |

**Table 4: Experimental results on industrial dataset Goofish. We use the precision and recall on the test set as metrics.**

| Model | Full categories | | Few-shot categories | |
|---|---|---|---|---|
| | Precision | Recall | Precision | Recall |
| Binary-class | 0.37 | 0.45 | - | - |
| Multi-class | 0.4026 | 0.4569 | 0.0470 | 0.3277 |
| Multi-class++ | 0.3958 | 0.4607 | 0.0453 | 0.3109 |
| **ours (CoSR)** | **0.4349** | **0.4647** | **0.0538** | **0.3277** |

the best in the continual learning stage, with a final test accuracy of 51.75%. Among the benchmark models, the final accuracy of the fine-tuning method is only 8.47%, which is the lower bound of the benchmark models. The Self-promote, SemanKL and CEC algorithms perform relatively well, and the final accuracy rates are all above 30%. The final accuracy of the CEC algorithm is 51.30%, which is lower than the final accuracy of our algorithm CoSR, i.e., 51.75%. From the perspective of performance drop rate, our CoSR has the lowest performance drop rate of 30.87%, indicating that under the constraint of semantic knowledge regularization, CoSR can learn a better visual feature space and alleviate the catastrophic forgetting problem in few shot continual learning. In contrast, the performance drop rate of the CEC algorithm is 33.58%, which is worse than our CoSR algorithm.

## 4.3 Experimental results on industrial datasets

**Goofish.** To further evaluate the learning ability of CoSR on the industrial dataset, we conduct experiments on a multi-modal classification dataset with a long-tailed distribution. This dataset is a multi-modal commodity dataset in the real world, which contains multi-modal information, such as images, titles, and descriptions of online commodities. The labels of online commodities include whether one particular product is illegal and which category it belongs to. Each category has a name as the semantic knowledge corresponding to the illegal product. The base task here is to determine whether a target product belongs to the illegal categories, which is a binary classification task. The illegal product category has a corresponding name as the semantic knowledge for base learning and continual learning. Since the dataset has long-tail distribution, we split it into the base classes and the few-shot classes. We take the category with a relatively small amount of data as few-shot classes and learn the few-shot classes in the continual learning phase. The other categories in the full dataset are used as the base classes to train the model in the base learning phase. Note that the categories in the few-shot classes do not overlap with those in the base classes.

In order to conduct comparative experiments, we design three baseline models on the industrial dataset, namely the binary classification model, multi-classification model, and semantic-assisted multi-classification model (multi-classification++). Among them, the binary classification model directly classifies the given product to determine whether it is in the illegal category. The multi-classification model aims to classify whether the given product is illegal and if so, which illegal category it belongs to. The semantic-assisted multi-classification model (multi-classification++) utilizes semantic knowledge from categorical information to assist the classification. Our proposed CoSR model concatenates the semantic vectors with the original visual features in order to better handle the real-world industrial scenario. We use the classification precision and recall on the full test set as evaluation metrics.

For the full categories, we can observe that the performance of our proposed CoSR on Goofish dataset is improved, indicating that our CoSR model has better classification ability on the long-tail distributed dataset. For the few-shot categories, our CoSR model can also achieve improvement over baselines in terms of both precision and recall. Due to the small number of labeled samples in the few-shot categories, traditional classification models usually fail to learn effective information from these few labeled samples and thus ignore them. CoSR extracts semantic knowledge from the long-tailed distributed categories to enhance the learning of multi-modal representations. By adding constraints to multi-modal representation learning, our model can quickly estimate the optimal category prototypes in the continual few-shot learning scenario. The performance of CoSR on the categories with long-tail distribution shows that our proposed model can significantly enhance the continual few shot classification ability through semantic knowledge regularization.

## 5 CONCLUSION

In this paper, we propose a novel approach, CoSR, to tackle the problem of continual few-shot learning. The well-designed Transformer adaptation in CoSR mines complex relationships between visual signals and semantic knowledge to generate suitable visual prototypes for continual few-shot classification. The semantic knowledge actually provides valid anchors for the visual prototypes in continual learning, thus alleviating catastrophic forgetting significantly with only a limited amount of data. Extensive experiments on both public and industrial data demonstrate the superiority of our proposed CoSR model over the state-of-the-art models. For future work, more types of semantic knowledge such as knowledge graph and commonsense knowledge can be explored for further investigations.

## ACKNOWLEDGMENTS

This work was supported by the National Key Research and Development Program of China No. 2020AAA0106300 and National Natural Science Foundation of China (No. 62222209, 62102222, 62250008).

## REFERENCES

[1] Afra Feyza Akyürek, Ekin Akyürek, Derry Wijaya, and Jacob Andreas. [n. d.]. Subspace Regularizers for Few-Shot Class Incremental Learning. In *International Conference on Learning Representations*.

[2] Rahaf Aljundi, Punarjay Chakravarty, and Tinne Tuytelaars. 2017. Expert gate: Lifelong learning with a network of experts. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3366–3375.

[3] 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.

[4] Arslan Chaudhry, Puneet K Dokania, Thalaiyasingam Ajanthan, and Philip HS Torr. 2018. Riemannian walk for incremental learning: Understanding forgetting and intransigence. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 532–547.

[5] Kuilin Chen and Chi-Guhn Lee. 2020. Incremental few-shot learning via vector quantization in deep embedded space. In *International Conference on Learning Representations*.

[6] Yudong Chen, Xin Wang, Miao Fan, Jizhou Huang, Shengwen Yang, and Wenwu Zhu. 2021. Curriculum meta-learning for next POI recommendation. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 2692–2702.

[7] Ali Cheraghian, Shafin Rahman, Pengfei Fang, Soumava Kumar Roy, Lars Petersson, and Mehrtash Harandi. 2021. Semantic-aware Knowledge Distillation for Few-Shot Class-Incremental Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2534–2543.

[8] Matthias De Lange and Tinne Tuytelaars. 2021. Continual prototype evolution: Learning online from non-stationary data streams. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 8250–8259.

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*. 4171–4186.

[10] Songlin Dong, Xiaopeng Hong, Xiaoyu Tao, Xinyuan Chang, Xing Wei, and Yihong Gong. 2021. Few-Shot Class-Incremental Learning via Relation Knowledge Distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 1255–1263.

[11] 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.

[12] Hang Gao, Zheng Shou, Alireza Zareian, Hanwang Zhang, and Shih-Fu Chang. 2018. Low-shot learning via covariance-preserving adversarial augmentation networks. *Advances in Neural Information Processing Systems* 31 (2018).

[13] Bharath Hariharan and Ross Girshick. 2017. Low-shot visual recognition by shrinking and hallucinating features. In *Proceedings of the IEEE International Conference on Computer Vision*. 3018–3027.

[14] 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.

[15] 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.

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

[18] Anna Kukleva, Hilde Kuehne, and Bernt Schiele. 2021. Generalized and incremental few-shot learning by explicit learning and calibration without forgetting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 9020–9029.

[19] Sang-Woo Lee, Jin-Hwa Kim, Jaehyun Jun, Jung-Woo Ha, and Byoung-Tak Zhang. 2017. Overcoming catastrophic forgetting by incremental moment matching. *Advances in neural information processing systems* 30 (2017).

[20] Yoonho Lee and Seungjin Choi. 2018. Gradient-based meta-learning with learned layerwise metric and subspace. In *International Conference on Machine Learning*. PMLR, 2927–2936.

[21] David Lopez-Paz and Marc'Aurelio Ranzato. 2017. Gradient episodic memory for continual learning. *Advances in neural information processing systems* 30 (2017), 6467–6476.

[22] 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.

[23] 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.

[24] Pratik Mazumder, Pravendra Singh, and Piyush Rai. [n. d.]. Few-Shot Lifelong Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

[25] Nikhil Mishra, Mostafa Rohaninejad, Xi Chen, and Pieter Abbeel. 2018. A Simple Neural Attentive Meta-Learner. In *International Conference on Learning Representations*.

[26] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.

[27] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. 2017. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 2001–2010.

[28] Mengye Ren, Renjie Liao, Ethan Fetaya, and Richard Zemel. 2019. Incremental few-shot learning with attention attractor networks. *Advances in Neural Information Processing Systems* 32 (2019).

[29] Eli Schwartz, Leonid Karlinsky, Joseph Shtok, Sivan Harary, Mattias Marder, Abhishek Kumar, Rogerio Feris, Raja Giryes, and Alex Bronstein. 2018. Delta-encoder: an effective sample synthesis method for few-shot object recognition. *Advances in neural information processing systems* 31 (2018).

[30] Joan Serra, Didac Suris, Marius Miron, and Alexandros Karatzoglou. 2018. Overcoming catastrophic forgetting with hard attention to the task. In *International Conference on Machine Learning*. PMLR, 4548–4557.

[31] 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. *Advances in Neural Information Processing Systems* 34 (2021).

[32] Hanul Shin, Jung Kwon Lee, Jaehong Kim, and Jiwon Kim. 2017. Continual learning with deep generative replay. *Advances in neural information processing systems* 30 (2017).

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

[34] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. 2018. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1199–1208.

[35] Xiaoyu Tao, Xinyuan Chang, Xiaopeng Hong, Xing Wei, and Yihong Gong. 2020. Topology-preserving class-incremental learning. In *European Conference on Computer Vision*. Springer, 254–270.

[36] Xiaoyu Tao, Xiaopeng Hong, Xinyuan Chang, Songlin Dong, Xing Wei, and Yihong Gong. 2020. Few-shot class-incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12183–12192.

[37] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. 2016. Matching networks for one shot learning. *Advances in neural information processing systems* 29 (2016), 3630–3638.

[38] Yaqing Wang and Quanming Yao. 2019. Few-shot learning: A survey. (2019).

[39] Yu-Xiong Wang, Ross Girshick, Martial Hebert, and Bharath Hariharan. 2018. Low-shot learning from imaginary data. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7278–7286.

[40] Yu-Xiong Wang and Martial Hebert. 2016. Learning to learn: Model regression networks for easy small sample learning. In *European Conference on Computer Vision*. Springer, 616–634.

[41] Peter Welinder, Steve Branson, Takeshi Mita, Catherine Wah, Florian Schroff, Serge Belongie, and Pietro Perona. 2010. Caltech-UCSD birds 200. (2010).

[42] Xuanrong Yao, Xin Wang, Yue Liu, and Wenwu Zhu. 2022. Continual Recognition with Adaptive Memory Update. *ACM Transactions on Multimedia Computing, Communications and Applications* (2022).

[43] Sung Whan Yoon, Do-Yeon Kim, Jun Seo, and Jaekyun Moon. 2020. Xtarnet: Learning to extract task-adaptive representation for incremental few-shot learning. In *International Conference on Machine Learning*. PMLR, 10852–10860.

[44] Chi Zhang, Nan Song, Guosheng Lin, Yun Zheng, Pan Pan, and Yinghui Xu. 2021. Few-shot incremental learning with continually evolved classifiers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12455–12464.

[45] Hanbin Zhao, Yongjian Fu, Mintong Kang, Qi Tian, Fei Wu, and Xi Li. 2021. Mgsvf: Multi-grained slow vs. fast framework for few-shot class-incremental learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021).

[46] Linjun Zhou, Peng Cui, Shiqiang Yang, Wenwu Zhu, and Qi Tian. 2019. Learning to learn image classifiers with visual analogy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11497–11506.

[47] Kai Zhu, Yang Cao, Wei Zhai, Jie Cheng, and Zheng-Jun Zha. 2021. Self-Promoted Prototype Refinement for Few-Shot Class-Incremental Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6801–6810.