With a Little Help from Language: Semantic Enhanced Visual Prototype Framework for Few-Shot Learning

Hecheng Cai[1,2], **Yang Liu**[1,2], **Shudong Huang**[1,2*] and **Jiancheng Lv**[1,2]
[1]College of Computer Science, Sichuan University, Chengdu 610065, China
[2]Engineering Research Center of Machine Learning and Industry Intelligence, Ministry of Education, Chengdu, 610065 China
caihecheng@stu.scu.edu.cn, liuyyy111@gmail.com, {huangsd, lvjiancheng}@scu.edu.cn

Abstract

Few-shot learning (FSL) aims to recognize new categories given limited training samples. The core challenge is to avoid overfitting to the minimal data while ensuring good generalization to novel classes. One mainstream method employs prototypes from visual feature extractors as classifier weight and the performance depends on the quality of the prototype. Since different categories may have similar visual features, the visual prototype has limitations. This is because existing methods only learn a simple visual feature extractor during the pre-training stage but neglect the importance of a well-developed feature space for the prototype. We introduce the Semantic Enhanced Visual Prototype framework (SEVpro) to address this issue. SEVpro refines prototype learning from the pre-training stage and serves as a versatile plug-and-play framework for all prototype-based FSL methods. Specifically, we enhance prototype discriminability by transforming semantic embeddings into the visual space, aiding in separating categories with similar visual features. For novel class learning, we leverage knowledge from base classes and incorporate semantic information to elevate prototype quality further. Meanwhile, extensive experiments on FSL benchmarks and ablation studies demonstrate the superiority of our proposed SEVpro for FSL.

1 Introduction

Modern artificial intelligence (AI) has achieved impressive results in many areas, such as speech recognition, natural language processing, and computer vision [Dong *et al.*, 2021]. However, there is still a gap between AI and human learning behaviour [Fei-Fei *et al.*, 2006; Miller *et al.*, 2000]. Humans can grasp new concepts with minimal instructions effortlessly, whereas AI models often demand thousands of labeled training samples. Since data collection and labeling are essential steps in developing AI models, it's no secret that data labeling can be very difficult, expensive, time-consuming, or all of the above. A naive solution is to train directly on limited data, but it always leads to overfitting and generalization issues. To give AI human-like capabilities, few-shot learning (FSL) is proposed and has received much attention from researchers. Few-shot learning is a subfield of machine learning and deep learning that aims to teach AI models how to learn from only a small number of labeled training data. The goal of few-shot learning is to enable models to generalize new, unseen categories based on a small number of samples we give them during the training process. One typical framework of FSL consists of two stages: the first stage, also called the pre-training stage, involves constructing a feature extractor on the base class with abundant data. In the second stage, a classifier is fine-tuned to recognize novel classes with only a few data.

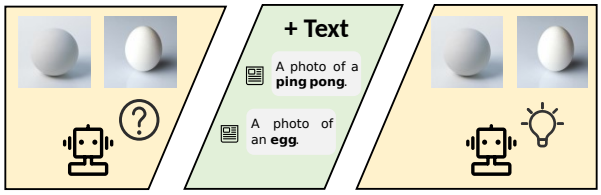


Figure 1: Egg and ping pong may have similar visual features. However, they have dissimilar semantic features.

one of the most promising approaches for FSL is the metric-based method [Zhang *et al.*, 2021]. The goal of metric learning is to learn a representation function that maps objects into an embedded space. It has shown promising results that investigate class neighborhood relationships with latent feature representations, where features from the same class should keep high similarity. As one of the mainstream technologies, Prototypical Networks [Snell *et al.*, 2017] and its follow-up works have achieved notable advancements in the few-shot image classification task. The goal of these models is to learn a prototype as classifier weight for each class. The prototype is a representative or central point in the feature space for instances belonging to a particular class. The original method obtained the prototype by averaging the features extracted from support examples. Recent progress introduces many techniques to obtain representative prototypes, for instance, self-attention mechanism [Gidaris and Komodakis, 2018], meta-learning paradigm [Chen *et al.*, 2021], or Graph Neural Networks [Garcia and Bruna, 2017].

*Corresponding author.

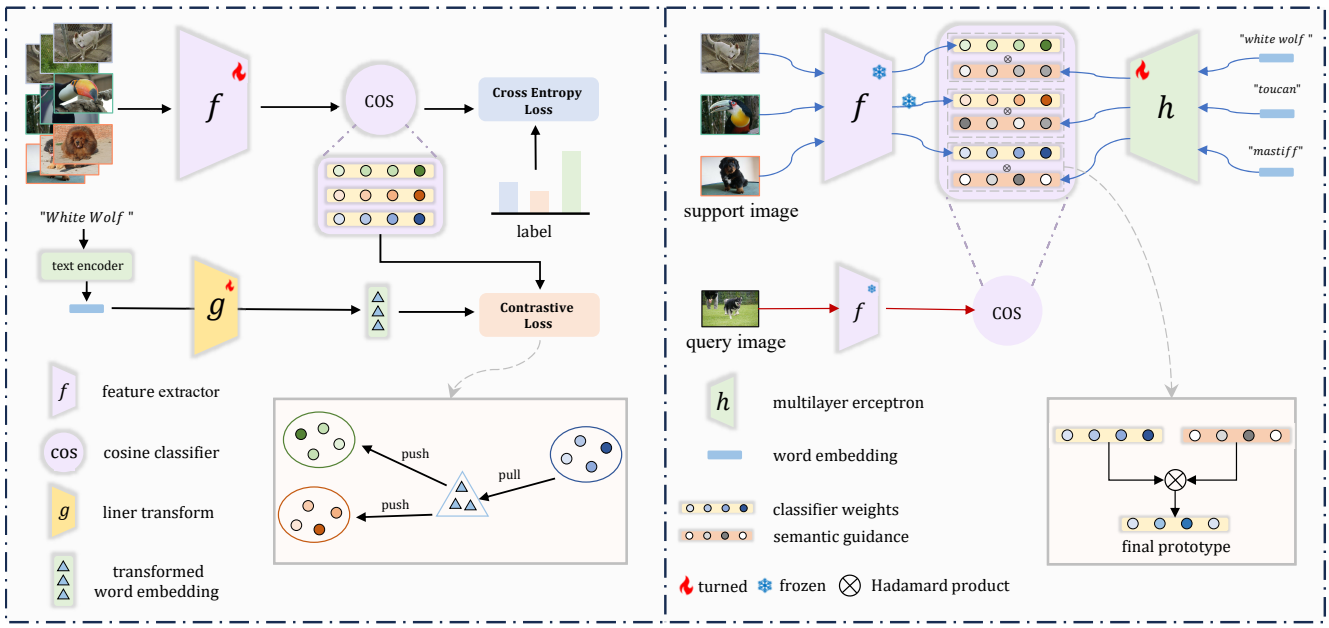


Figure 2: **Framework of our method. Left:** The semantically enhanced feature extractor learning procedure. **Right:** The learning procedure of a few-shot classifier. We use semantics to guide the learning of the final prototype.

Existing prototype-based methods use the two-stage framework mentioned above. Convolutional Neural Networks (CNN), such as ResNet or Convnet, are widely used to extract the visual features of images within a novel class and then the prototypes are calculated for image classification. While they have achieved remarkable performance, one major challenge is the unimodal visual representation can not always guarantee high-quality prototypes. The two classes with approximate visual features may lead to close prototypes and wrong classification. In other words, existing approaches neglect the effective utilization of the pre-training stage. This sample visual feature extractor hinders the optimization of prototypes, which could substantially contribute to the overall success of few-shot learning. In the real world, humans typically construct new concepts by integrating visual and semantic signals. For instance, a child can learn the concept of ”apple” by looking at its picture and listening to the description (”An apple is typically a round or oval red fruit”) from his parents. Motivate by this, as shown in Figure 1. The model may initially struggle to distinguish between an egg and a ping-pong ball, but with the addition of text, it can effectively differentiate between them. In recent years, researchers have integrated semantic knowledge into the second stage to fine-tune a robust and enhanced prototype [Xing *et al.*, 2019]. Similar to the previous method, a feature extractor like CNN is learned during the pre-training stage. After that, the semantic feature derived from textual descriptions is used to strengthen the discriminability of prototype [Huang *et al.*, 2022]. The quality of the feature extractor significantly influences classification performance as it defines the embedding space and serves as the cornerstone for computing prototypes.

Instead of solely relying on semantic knowledge for fine-tuning prototypes in the second stage, we propose a plug-and-play framework to incorporate semantic knowledge during the pre-training stage. This approach aids in constructing a more dispersed embedding space. A contrastive loss enables the feature extractor to learn from both visual and semantic aspects. As shown in Figure 3, a ping-pong ball and an egg may share similar visual features, making their prototypes close in the embedding space. In this case, semantic knowledge may offer robust prior knowledge support to facilitate learning and help to facilitate the distinctiveness between their prototypes. With a high-quality embedded space, better knowledge can be transferred from the base class to the new class, and the existing semantic fine-tuning method can also work better. To sum up, our contributions are:

- We propose a plug-and-play framework that can be integrated into any existing prototype-based FSL architecture. It enhances the feature extractor by incorporating semantic knowledge during the pre-training stage. In the few-shot learning phase, the high-quality representation generates robust prototypes and facilitates learning novel classes.
- A simple and efficient semantic and visual contrastive loss is designed. It can harness semantic representations to achieve more outstanding distinctiveness among visual features within the embedding space.
- Extensive experiments and ablation studies on three few-shot benchmarks, demonstrating the effectiveness and superiority of our method.

2 Related Work

2.1 Few-shot Learning

Few-shot learning focuses on learning novel classes with a few labeled examples. Existing methodologies in this domain can be broadly categorized into four types. Metric-

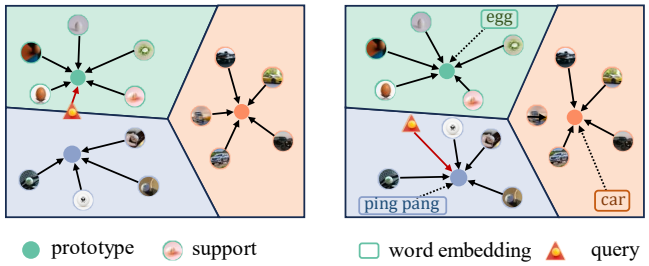


Figure 3: **Left:** Visual information may lead to imperfect prototypes. **Right:** our proposed enhanced prototype with semantics in embedding space, which uses semantic knowledge to help learn prototypes, making the prototype more discriminative and separated (Best viewed in color, colors represent different categories).

based methods aim to learn a metric space, which query example can be classified via a nearest neighbor [Koch *et al.*, 2015]. A framework for variational metric scaling is developed through metric-based meta-learning from a Bayesian perspective [Chen *et al.*, 2020a], which can measure the metric scaling parameter automatically. Optimization-based methods use meta-learning to obtain better initializing parameters and then adapt to novel classes with a few optimization steps. In [Finn *et al.*, 2017], an optimization-based approach is crafted for effectively tackling novel learning tasks with minimal training samples, providing an easily fine-tunable solution. Graph-based methods are devoted to constructing a graph to propagate knowledge from the base classes. Integrating conventional message-passing inference algorithms with the neural network counterparts, a graph neural network architecture is formulated, offering a generalized framework for several recently proposed few-shot learning models [Satorras and Estrach, 2018]. Semantic-based methods employ semantic knowledge to enhance performance in FSL. In [Garcia and Bruna, 2017], the Adaptive Modality Mixture Mechanism (AM3) is designed to intelligently integrate information from visual and semantic modalities based on the new image categories to be learned.

2.2 Semantics in Few-Shot Learning

Semantics is firstly widely used in the zero-shot learning field [Han *et al.*, 2021; Guan *et al.*, 2020], which aims to classify the non-observed classes through some form of auxiliary information. Inspired by this, researchers have explored the potential of semantics in the context of FSL. For example, [Chen *et al.*, 2019]employs semantic embeddings of labels or attributes to guide the latent representation of an autoencoder, functioning as a regularizing mechanism. [Xing *et al.*, 2019] dynamically combines visual information and semantic representations to effectively adjust the focus on two modalities. [Schwartz *et al.*, 2022] extended the previous one in a joint framework to exploit different semantics such as class labels, text descriptions, and manual attributes in a joint framework. Thus, these methods build upon visual features extracted in CNNs, leveraging semantic knowledge as auxiliary information to acquire improved classification weights. However, the existing methods seldom consider adding semantic knowledge in the training of visual feature extractors.

2.3 Contrastive Learning

Contrastive learning is a type of unsupervised learning method used in machine learning [Oord *et al.*, 2018; Khosla *et al.*, 2020]. The goal of contrastive learning is to bring similar instances closer in a learned feature space while pushing dissimilar instances apart. The process involves comparing and contrasting representations of different data samples. [He *et al.*, 2020] construct a dynamic dictionary to facilitate the contrastive learning process. [Chen *et al.*, 2020b] propose a simple but effective framework with larger batch sizes and data augmentation, which does not require specialized architectures or a memory bank. [Park *et al.*, 2020] propose employing contrastive learning between the patches of the target image and source images. In the paper, contrastive learning is employed to train an improved feature extractor. This process facilitates the separation of confusable visual features with the aid of semantic knowledge.

3 Preliminaries

3.1 Problem Setting

In the standard few-shot classification setting, the dataset is divided into a base set $\mathcal{D}_{base}$ with M classes and a novel set $\mathcal{D}_{novel}$ with N classes, where $\mathcal{C}_{\text{base}} \cap \mathcal{C}_{\text{novel}} = \emptyset$. The base set includes a sufficient number of labeled examples for each class for training, while the novel set is divided into two parts. The support set $\mathcal{S} = \{(x_i, y_i)\}_{i=0}^{N \times K}$ contains a few of labeled samples for training and a query set $\mathcal{Q} = \{(x_i, y_i)\}_{i=0}^{Q}$ consisting of unlabeled samples for testing. N denotes the number of novel classes in S, and K indicates the number of images for each class, thus we call it the N-way K-shot problems. x_i represents the input image and y_i is its corresponding label.

3.2 Visual Prototype

Prototypical Network [Snell *et al.*, 2017] is a simple and effective metric-based method, which has garnered significant attention for its ability to learn discriminative and compact embeddings (prototype) for few-shot classification tasks. A prototypical network first trains a visual feature extractor, such as CNNs [LeCun *et al.*, 1998] $f : \mathbb{R}^{d_x} \rightarrow \mathbb{R}^{d_v}$ and then obtain the classifier. The classification weight of a class is called prototype, which is the mean vector of the embedded support samples belonging to its class. The feature extractor f with parameter θ constructs an embedding space of dimension d_v in which samples of the same class are brought closer together, and conversely, samples from different classes are separated.

Following the classical work, the visual prototype $\mathbf{p}^c$ for class c can be computed by averaging the feature representations of all support samples that belong to class c.

$$\mathbf{p}^c = \frac{1}{|\mathcal{D}_c|} \sum_{(\boldsymbol{x_i}, y_i) \in \mathcal{D}_c} f_\theta \left(\boldsymbol{x_i} \right), \tag{1}$$

where $\mathcal{D}_c$ is the support images of class c.

3.3 Cosine Classifier

After extract the feature vector $f_\theta(\boldsymbol{x})$, the standard classifier have to step to estimate the classification probablility vector p. Firstly compute the row classification score and then applying the softmax oprator across all classification scores. However, the A standard classifier is not suited for FSL because limited samples can easily lead to overfitting. Therefore, the Cosine Classifier is more widely adopted in FSL setting [Gidaris and Komodakis, 2018; Qi *et al.*, 2018]. The classification score of image $\boldsymbol{x}$ for class c is calculated based on the cosine similarity score:

$$s_c = \cos\left\langle f_\theta(\boldsymbol{x}), \mathbf{p}^c \right\rangle = \left\langle \frac{f_\theta(\boldsymbol{x})}{\|f_\theta(\boldsymbol{x})\|}, \frac{\mathbf{p}^c}{\|\mathbf{p}^c\|} \right\rangle. \quad (2)$$

Unlike the dot-product, with the additional l_2 normalization, the classification score relies solely on the angle, thus enhancing robustness in data sparsity.

4 Method

4.1 Learning of Feature Extractor

As shown in Figure 2 left, in the pre-training stage, all samples from the base dataset $\mathcal{D}_{\text{base}}$ and their corresponding labels are fed into the feature extractor. The extractor, which is a convolutional neural network, is already capable of capturing latent visual representations. Our framework augments it with semantic knowledge to obtain a better and more discrete feature space.

Source of semantic knowledge: To attain semantic enhancement, it is imperative to address the source of semantic knowledge initially. According to previous experience, it can be category labels, text descriptions, and manual attributes. For simplicity, we utilize class label as a semantic knowledge source, i.e. $\mathbf{L} = \{l_c\}_{c=1}^{M}$, where l_c is the word label of class c and M is the total number of classes in $\mathcal{D}_{base}$. Specifically, we first prompt each class label with a sentence template, such as *"A photo of a (an) label"*, then it is fed into a text embedding model $e : \mathbb{R}^{d_l} \rightarrow \mathbb{R}^{d_s}$. We use the text encoder model of CLIP [Radford *et al.*, 2021] to get the embedding of every prompted sentence. In this way, we can obtain the semantic representation of each class by the following operation:

$$\mathbf{S} = e(\mathbf{L}), \quad (3)$$

Following Eq. (1) and Eq. (3), for class c, we have obtain its visual representation $\mathbf{p}^c \in \mathbb{R}^{d_v}$ and the semantic representation $\mathbf{s}^c \in \mathbb{R}^{d_s}$ for class c, respectively. Due to their inconsistent dimensions, we need to introduce an additional linear transform g to map the semantic representation $\mathbf{s}^c$ into the visual embedding space. After that, the transformed semantic feature $\mathbf{z}^c = g(\mathbf{s}^c) \in \mathbb{R}^{d_v}$ is obtained. Finally, we have addressed the issue of the source of semantic knowledge.

Comparative loss: We design a simple and effective hybrid loss to leverage semantic knowledge to guide the training of the feature extractor and effectively distinguish some categories that are easily confused at the visual level. Specifically, an infoNCE loss [Oord *et al.*, 2018] is added during the training procedure. Given a transformed semantic representation in embedding space, only the corresponding visual feature will be pulled closer together, while others will be pushed away. For instance, the visual feature of the egg and ping-pong ball may stay close in the embedding space. However, their semantic feature may keep enough distance. The two types of confusing features can be separated with the traction of semantic knowledge. Finally, the classification centroids of all classes will be separate. The infoNCE loss is formulated as follows:

$$L_{nce} = -\log \frac{\exp(\mathbf{p}^c \cdot \mathbf{s}^+/\tau)}{\sum_{c=0}^{k} \exp(\mathbf{p}^c \cdot \mathbf{s}^c/\tau)}, \quad (4)$$

where τ is a temperature hyper-parameter [Wu *et al.*, 2018], and k is the number of classes in a batch.

Classification loss: To solve the classification task, we use the common cross-entropy loss:

$$L_{ce} = \log \frac{\exp(\cos\langle f_\theta(\boldsymbol{x}), \mathbf{p}^c\rangle)}{\sum_{c\in\mathcal{C}_{\text{base}}} \exp(\cos\langle f_\theta(\boldsymbol{x}), \mathbf{p}^c\rangle)}. \quad (5)$$

Based on the above two losses, the total loss can be defined as follows:

$$L = L_{ce} + \lambda L_q, \quad (6)$$

where λ is a balance coefficient between classification loss and constructive loss. Our proposed loss contributes to constructing the embedding space by incorporating additional semantics instruction. The enhanced feature extractor helps prevent the network from overfitting due to limited novel class data. An advantageous feature space is conducive to learning from small samples and can mitigate overfitting. This effect can be heightened when introducing semantic knowledge during the few-shot learning phase.

4.2 Learning of Classifier

After learning the feature extractor, we fix it and fine-tune the classifier (shown in Figure 2 right). We employ a cosine classifier in line with conventional practices, where the classification weight $\mathcal{W}$ is equivalent to the class prototype. In the training of the classifier, following Eq. 1, we firstly compute the primary prototype $\mathbf{p}^c_{avg}$ by averaging the feature values of images belonging to category c. We follow our baseline method [Yang *et al.*, 2022] to complete knowledge transfer and semantic guidance, which can make full use of the semantic enhanced visual prototype learned from pre-training.

Transfer knowledge from base class: It is worth noting that some features of the new class may already be present in the base class. For instance, if the classifier needs to identify a new category, zebras, the knowledge acquired from horses from the base class can be beneficial. To promote the transfer of prior knowledge from base class $\mathcal{W}^{base}$ to novel class, we get the valuable knowledge by attention mechanism:

$$\mathbf{p}^c_{att} = \frac{1}{|\mathcal{D}^n_c|} \sum_{(\boldsymbol{x_i}, y_i)\in\mathcal{D}^n_c} \sum_{j\in\mathcal{Y}^b} \text{Att}\left(\boldsymbol{\phi_q} f_\theta(\boldsymbol{x_i}), \boldsymbol{k_j}\right) \cdot \boldsymbol{w}_j, \quad (7)$$

where $\boldsymbol{\phi_q} \in \mathbb{R}^{n_v \times n_v}$ is a learnable weight matrix and $\boldsymbol{k_j} \in \mathbb{R}^{n_v}$ is a set of learnable keys for base prototypes. Specifically, the feature $f_\theta(\boldsymbol{x_i})$ is transformed to a query vector by $\boldsymbol{\phi_q}$, then perform attention with $\boldsymbol{k_j}$ by a cosine based attention kernel $\text{Att}(\cdot,\cdot)$. Take the learnable coefficients between

the original term and transfer term as λ_1 and λ_2, we have the combination prototype:

$$\mathbf{p}^c = \lambda_1 \times \mathbf{p}^c_{\text{avg}} + \lambda_2 \times \mathbf{p}^c_{\text{att}}\,. \tag{8}$$

Semantic guided enhancement: Given that the embedding space is constructed with the assistance of semantics during pre-training, we also incorporate semantic knowledge to guide the learning of novel class weights. This approach allows for the optimal utilization of the space. Here we also use the word embedding as our semantic source: $\mathcal{Z} = \left\{\boldsymbol{z}^c \in \mathbb{R}^{d_s}\right\}_{c=1}^{N}$, where $\boldsymbol{z}^c$ represent the sentence embedding of class c and d_s is the dimension of primordial semantic space. To apply $\mathcal{Z}$ in classification weight space, we use an MLP to transform it: $h : \mathbb{R}^{d_s} \rightarrow \mathbb{R}^{d_v}$. We have semantic attention $\boldsymbol{a}^c = g\left(\boldsymbol{h}^c\right)$. The last layer of g is a sigmoid function, so $\boldsymbol{a}^c$ is bounded between $[0, 1]$. The elements of the vector $\boldsymbol{a}^c$ with values close to 1 represent the components that should be given more attention. Finally, we obtain the classification weight for class c by:

$$\boldsymbol{w}^c = \boldsymbol{a}^c \otimes \mathbf{p}^c, \tag{9}$$

where $\otimes$ is the Hadamard product. For a query image, we must apply the cosine classifier with weight $\boldsymbol{w}_c$ to get the probability of belonging to class c.

4.3 Combine with Exiting FSCIL Architectures

Our proposed framework is plug-and-play, which can be seamlessly integrated with current FSL architectures in the context of image classification tasks (i.e., for FSL models with semantic knowledge). Our module can be effectively embedded in feature extraction training, leading to a better classifier.

5 Experiment

In this section, we present experimental results to demonstrate the efficacy of our proposed methodology on three benchmark datasets. This is accomplished through a comprehensive comparison with alternative state-of-the-art FSL architectures. Additionally, we conduct ablation studies to examine the effects of different contrastive loss in our method. Finally, we provide t-SNE visualizations to substantiate our findings.

5.1 Datasets

Three common FSL benchmarks are used in our experiment, i.e., miniImageNet [Vinyals *et al.*, 2016], tieredImageNet [Ren *et al.*, 2018] and CIFAR-FS [Bertinetto *et al.*, 2018]

miniImageNet is a popular dataset for FSL classification task. It consists of 100 classes extracted from the original ImageNet dataset proposed by Matching Networks [Deng *et al.*, 2009]. Each class has 600 images, each size 84×84. We utilize 64 classes for training, allocate 16 classes for validation, and reserve the remaining 20 classes for testing.

tieredImageNet dataset represents an expanded subset of the original ImageNet dataset, encompassing a total of 608 distinct image categories, each accompanied by an average of 1281 images. The image dimensions are set at 84×84 pixels, with 351 classes allocated for training, 97 for validation, and the remaining 160 for testing purposes.

CIFAR-FS dataset, a subset of CIFAR-100 [Krizhevsky *et al.*, 2010], has explicitly been introduced to address the Few-Shot Learning (FSL) challenge. It consists of 100 unique classes, each accompanied by 600 high-quality, 32×32 natural color images. These classes are randomly partitioned, with 64 allocated for training, 16 for validation, and 20 for testing.

5.2 Setting

Our model is an independent framework that can be combined with existing FSL architecture. In this paper, we use the recent method SEGA [Yang *et al.*, 2022] as our baseline model. Following previous works [Li *et al.*, 2020; Liu *et al.*, 2020; Xing *et al.*, 2019], we use ResNet-12 as the backbone in the pre-training stage. In alignment with the prior settings, adjustments have been made to the filter dimensions, evolving from (64, 128, 256, 512) to (64, 160, 320, 640). We have also implemented enhancements to address overfitting concerns, including random cropping, color jittering, erasing, and Dropblock regularization. For our semantic knowledge source, we select the text encoder derived from CLIP [Radford *et al.*, 2021], a multi-modal model adept in vision and language, trained on a wide array of images with rich linguistic annotations. We leverage this encoder to obtain 512-dimensional word embeddings for each class label. All experiments are conducted within the PyTorch framework. We employ an AdamW optimizer with a 5×10^{-2} weight decay rate. In the training stage, we train the Feature Extractor for 60 epochs (90 for tieredImageNet), with each epoch comprising 1000 episodes. Subsequently, in the second stage, we refer to the operation of SEGA, utilizing attention-based knowledge transfer and semantic guidance to get the final prototypes. For the setting of hyperparameters, The λ in Eq. 6 is a hyperparameter, we set it to 0.5 to achieve the best performance. The λ_1 and λ_2 are learnable coefficients. After setting the initial values, they will automatically update during iteration.

5.3 Baseline

State-of-the-art Few-Shot Learning (FSL) methods are applied to the few-shot classification task to facilitate a thorough comparison. The methods included in the comparison are: Matching Networks [Vinyals *et al.*, 2016], MAML [Finn *et al.*, 2017], ProtoNet [Snell *et al.*, 2017], D-FSL [Gidaris and Komodakis, 2018], Relation NetWorks [Sung *et al.*, 2018], wDAE-CNN [Gidaris and Komodakis, 2019], MetaOptNet [Lee *et al.*, 2019], TEWAM [Qiao *et al.*, 2019], Shot-Free [Ravichandran *et al.*, 2019], KTN [Peng *et al.*, 2019], TriNet [Chen *et al.*, 2019], AM3 [Xing *et al.*, 2019], DeepEMD [Zhang *et al.*, 2020], RFS [Tian *et al.*, 2020], NegCosine [Liu *et al.*, 2020], SEGA [Yang *et al.*, 2022].

5.4 Experimental Results and Analysis

Comparison with State-of-the-art Methods

In this subsection, we evaluate the performance of our method by compared with several recent state-of-the-art approaches.

Models	Backbone	Sem	miniImageNet		tieredImageNet	
			5Way-1Shot	5Way-5Shot	5Way-1Shot	5Way-5Shot
Matching Networks	4 Conv	No	43.56 ± 0.84	55.31 ± 0.73	-	-
MAML	4 Conv	No	48.70 ± 1.84	63.11 ± 0.92	51.67 ± 1.81	70.30 ± 1.75
ProtoNet	4 Conv	No	49.42 ± 0.78	68.20 ± 0.66	53.31 ± 0.89	72.69 ± 0.74
D-FSL	4 Conv	No	56.20 ± 0.86	72.81 ± 0.62	-	-
wDAE-CNN	WRN-28-10	No	61.07 ± 0.15	76.75 ± 0.11	68.18 ± 0.16	83.09 ± 0.12
MetaOptNet	ResNet-12	No	62.24 ± 0.61	78.63 ± 0.46	65.99 ± 0.72	81.56 ± 0.53
DeepEMD	ResNet-12	No	65.91 ± 0.82	$\mathbf{82.41 \pm 0.43}$	71.16 ± 0.87	$\mathbf{86.03 \pm 0.58}$
RFS	ResNet-12	No	64.82 ± 0.60	82.14 ± 0.56	71.52 ± 0.69	$\mathbf{86.03 \pm 0.49}$
Neg-Cosine	ResNet-12	No	63.85 ± 0.81	81.57 ± 0.56	-	-
KTN	4 Conv	Yes	64.42 ± 0.72	74.16 ± 0.56	-	-
TriNet	ResNet-18	Yes	58.12 ± 1.37	76.92 ± 0.69	-	-
AM3	ResNet-12	Yes	65.30 ± 0.49	78.10 ± 0.36	69.08 ± 0.47	82.58 ± 0.31
SEGA **(our baseline)**	ResNet-12	Yes	69.04 ± 0.26	79.03 ± 0.18	72.18 ± 0.30	84.28 ± 0.21
SEVPro	ResNet-12	Yes	$\mathbf{71.81 \pm 0.22}$	78.88 ± 0.18	$\mathbf{72.77 \pm 0.30}$	84.04 ± 0.21

Table 1: **Comparison to prior work on miniImageNet and tieredImageNet.** Average classification accuracies (%) with 5000 test episodes of novel categories (with 95% confidence intervals). "Sem" indicates whether semantic knowledge was utilized.

Models	CIFAR-FS	
	5Way 1Shot	5Way 5Shot
MAML	58.9 ± 1.9	71.5 ± 1.0
ProtoNet	55.5 ± 0.7	72.0 ± 0.6
Relation Networks	55.0 ± 1.0	69.3 ± 0.8
Shot-Free	$69.2 \pm$ n/a	$84.7 \pm$ n/a
TEWAM	$70.4 \pm$ n/a	$81.3 \pm$ n/a
MetaOptNet	72.0 ± 0.7	84.2 ± 0.5
RFS	73.9 ± 0.8	$\mathbf{86.9 \pm 0.5}$
SEGA **(our baseline)**	78.5 ± 0.2	86.0 ± 0.2
SEVPro	$\mathbf{80.36 \pm 0.24}$	86.12 ± 0.20

Table 2: **Comparison on CIFAR-FS.** The setting is same as above.

In Table 1, we report the results for the 5-Way 1-Shot and 5-Way 5-Shot classification tasks conducted on miniImageNet and tieredImageNet. Additionally, Table 2 showcases the results for CIFAR-FS. Our method demonstrates comparable performance in most scenarios. Our approach stands out with the highest performance in the 5-Way 1-Shot setting across all three datasets, outperforming our baseline method SEGA. Note that previous works typically exclude the incorporation of semantic knowledge during the pre-training of feature extractors. For instance, AM3 and SEGA independently utilize a pre-trained CNN to extract visual features, which are combined with semantic knowledge to enhance prototype quality. In contrast, our approach focuses on training the CNN using semantic knowledge to encourage separating features with similar visual characteristics within the embedding space. This strategy affords us the flexibility to integrate with existing prototype-based architectures seamlessly.

Ablation Studies

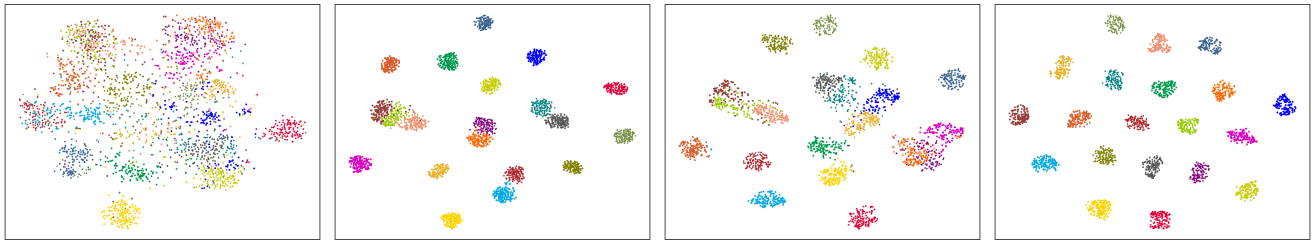
Effective of plug-and-play ability: We verify the plug-and-play ability of our approach on two classic prototype-based approaches (ProtNet and SEGA). The ProtNet computes the prototype by averaging the visual embedding of support images, SEGA uses semantics to emphasize which parts of the prototype are essential. Based on these two networks, we added our proposed SEVPro framework, so we have four settings as follows: (1) ProtNet; (2) ProtNet + SEVPro; (3) SEGA; (4) SEGA + SEVPro.

Table 3 shows the detailed results on miniImageNet, tieredImageNet, and CIFAR-FS. With the incorporation of the additional infoNCE loss, one can observe that our method transcends the ProtoNet model, generally delivering improvements across the three datasets in both 1-shot and 5-shot tasks. This can be attributed to our approach's unique capacity to infuse semantic understanding into the feature space learning process, facilitating the disambiguation of categories that may be visually challenging to distinguish. Besides, when introducing semantic knowledge in the few-shot learning phase, our SEVPro can work better. Note that our feature extractor combined with SEGA achieves, on average 2.77%, 0.59%, and 1.91 % on the 1-Shot setting, respectively. Because visual information is limited in this setting, the decentralized prototypes learned from the base class can transfer clearer knowledge to the learning of new categories, in particular under the guidance of semantics.

Effective of different loss: We study the effect of different contrastive loss functions on the CIFAR-FS dataset: (a) **infoNCE loss:** It formulates a binary classification task where the model distinguishes between positive pairs and negative pairs; (b) **cosine similarity loss:** It measures the cosine of the angle between two vectors. If the vectors point in the same direction (high cosine similarity), they are considered similar. (c) **triplet loss:** It comprises three elements: an anchor, a positive example (similar to the anchor), and a negative example (dissimilar to the anchor). It is designed to minimize the distance between the anchor and the positive example while simultaneously increasing the distance between the anchor and the negative example. (d) **hardest negative triplet loss:** In the computation of the triplet loss, only the samples most sim-

setting	miniImageNet		tieredImageNet		CIFAR-FS	
	5Way-1Shot	5Way-5Shot	5Way-1Shot	5Way-5Shot	5Way-1Shot	5Way-5Shot
ProtoNet	**62.42 ± 0.26**	77.16 ± 0.18	67.53 ± 0.31	83.32 ± 0.21	71.43 ± 0.29	85.25 ± 0.20
ProtoNet + SEVPro	61.66 ± 0.26	**79.06 ± 0.18**	**67.95 ± 0.31**	**83.58 ± 0.21**	**74.01 ± 0.29**	**85.86 ± 0.20**
SEGA	69.04 ± 0.26	**79.03 ± 0.18**	72.18 ± 0.30	**84.28 ± 0.21**	78.45 ± 0.24	86.00 ± 0.20
SEGA + SEVPro	**71.81 ± 0.22**	78.88 ± 0.18	**72.77 ± 0.30**	84.04 ± 0.21	**80.36 ± 0.24**	**86.12 ± 0.20**

Table 3: **Ablation study.** Average classification accuracies (%) over different combinations.



(a) ProtoNet (b) ProtoNet + SEVPro (c) SEGA (d) SEGA + SEVPro

Figure 4: **t-SNE Visualization.** Test on CIFAR-FS in 5-Way 1-Shot

L_{nce}	L_{cos}	L_{tri}	L^h_{tri}	CIFAR-FS	
				1Shot	5Shot
✓				**80.36 ± 0.24**	**86.12 ± 0.20**
	✓			69.99 ± 0.29	85.06 ± 0.20
		✓		67.47 ± 0.29	83.27 ± 0.20
			✓	70.06 ± 0.29	84.60 ± 0.60

Table 4: **Ablation study with different constrastive loss.** Four common contrastive losses are employed for comparison.

ilar to the anchor in the batch are considered, as opposed to all negative samples. Table **??** presents the experimental results, highlighting that the infoNCE loss outperforms other variants on the CIFAR-FS dataset for both 1-shot and 5-shot tasks. This can be attributed to the effectiveness of the infoNCE loss in contrastive learning scenarios, where the objective is to encourage the proximity of similar samples and the separation of dissimilar ones. Notably, this approach is well-suited for scenarios with limited labeled data, as it relies on the relative relationships between samples.

t-SNE Visualization

To gain deeper insights into the role of semantic knowledge during the pre-training stage, we conducted t-SNE visualization. Figure 4 depicts the transformation of prototypes before and after the incorporation of semantics in pre-training, specifically in the context of the 5-Way 1-Shot scenario. Referencing Table 3, the four verification settings are labeled as ProtoNet, ProtoNet + SEVPro, SEGA, and SEGA + SEVPro, respectively. As observed, the generated prototypes exhibit significant instability in the absence of semantic knowledge utilization, as illustrated in ProtoNet. However, after applying contrastive learning with both visual and word embeddings, the final prototypes demonstrate increased stability, as seen in ProtoNet + ours. Going a step further, as illustrated in SEGA, our SEGA baseline serves as guidance in the few-shot learning phase, resulting in improved separability among different classes, although some classes remain intertwined. Ultimately, as depicted in SEGA+ours, our comprehensive method, which integrates semantic knowledge in pre-training and few-shot learning phases, achieves the highest level of separability, evident in the apparent gap between prototypes of different classes. This implies that prototypes from the same class draw closer together while those from other classes move farther apart. These observations shed light on why our model excels with incorporating semantic knowledge.

6 Conclusion

In this work, we study the problem of few-shot image classification. We focus on one limitation of existing metric-based approaches, i.e., they ignore learning a perfect embedding space in the pre-training stage and only emphasize the importance of fine-tuning the prototype during the few-shot earning stage. We introduce SEVPro, a plug-and-play framework based on contrastive learning. SEVPro employs semantic knowledge in the learning of feature extractor, which can better separate visually confusing classes. With more discriminative embedding space, novel classes can accept better knowledge transferred from base classes. The performance of semantic-based methods can also be further improved, especially in the 1-Shot setting. Experiments demonstrate the effectiveness of our framework, which achieves comparable performance on three popular few-shot classification benchmarks. In future work, we can explore the effect of richer semantic knowledge, such as descriptions of categories.

Acknowledgments

This work was partially supported by the National Science Foundation of China under Grant 62106164 and 62376175, the 111 Project under Grant B21044, and the Sichuan Science and Technology Program under Grants 2021ZDZX0011.

References

[Bertinetto *et al.*, 2018] Luca Bertinetto, Joao F Henriques, Philip HS Torr, and Andrea Vedaldi. Meta-learning with differentiable closed-form solvers. *arXiv preprint arXiv:1805.08136*, 2018.

[Chen *et al.*, 2019] Zitian Chen, Yanwei Fu, Yinda Zhang, Yu-Gang Jiang, Xiangyang Xue, and Leonid Sigal. Multi-level semantic feature augmentation for one-shot learning. *IEEE Transactions on Image Processing*, 28(9):4594–4605, 2019.

[Chen *et al.*, 2020a] Jiaxin Chen, Li-Ming Zhan, Xiao-Ming Wu, and Fu-lai Chung. Variational metric scaling for metric-based meta-learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 3478–3485, 2020.

[Chen *et al.*, 2020b] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, pages 1597–1607. PMLR, 2020.

[Chen *et al.*, 2021] Yinbo Chen, Zhuang Liu, Huijuan Xu, Trevor Darrell, and Xiaolong Wang. Meta-baseline: Exploring simple meta-learning for few-shot learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9062–9071, 2021.

[Deng *et al.*, 2009] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on Computer Vision and Pattern Recognition*, pages 248–255. Ieee, 2009.

[Dong *et al.*, 2021] Shi Dong, Ping Wang, and Khushnood Abbas. A survey on deep learning and its applications. *Computer Science Review*, 40:100379, 2021.

[Fei-Fei *et al.*, 2006] Li Fei-Fei, Robert Fergus, and Pietro Perona. One-shot learning of object categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(4):594–611, 2006.

[Finn *et al.*, 2017] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, pages 1126–1135. PMLR, 2017.

[Garcia and Bruna, 2017] Victor Garcia and Joan Bruna. Few-shot learning with graph neural networks. *arXiv preprint arXiv:1711.04043*, 2017.

[Gidaris and Komodakis, 2018] Spyros Gidaris and Nikos Komodakis. Dynamic few-shot visual learning without forgetting. In *Proceedings of the IEEE Conference on computer vision and pattern recognition*, pages 4367–4375, 2018.

[Gidaris and Komodakis, 2019] Spyros Gidaris and Nikos Komodakis. Generating classification weights with gnn denoising autoencoders for few-shot learning. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 21–30, 2019.

[Guan *et al.*, 2020] Jiechao Guan, Zhiwu Lu, Tao Xiang, Aoxue Li, An Zhao, and Ji-Rong Wen. Zero and few shot learning with semantic feature synthesis and competitive learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(7):2510–2523, 2020.

[Han *et al.*, 2021] Zongyan Han, Zhenyong Fu, Shuo Chen, and Jian Yang. Contrastive embedding for generalized zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2371–2381, 2021.

[He *et al.*, 2020] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020.

[Huang *et al.*, 2022] Lian Huang, Shaosheng Dai, and Ziqiang He. Few-shot object detection with semantic enhancement and semantic prototype contrastive learning. *Knowledge-Based Systems*, 252:109411, 2022.

[Khosla *et al.*, 2020] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33:18661–18673, 2020.

[Koch *et al.*, 2015] Gregory Koch, Richard Zemel, Ruslan Salakhutdinov, et al. Siamese neural networks for one-shot image recognition. In *ICML deep learning workshop*, volume 2. Lille, 2015.

[Krizhevsky *et al.*, 2010] Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Cifar-10 (canadian institute for advanced research). *URL http://www. cs. toronto. edu/kriz/cifar. html*, 5(4):1, 2010.

[LeCun *et al.*, 1998] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[Lee *et al.*, 2019] Kwonjoon Lee, Subhransu Maji, Avinash Ravichandran, and Stefano Soatto. Meta-learning with differentiable convex optimization. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 10657–10665, 2019.

[Li *et al.*, 2020] Aoxue Li, Weiran Huang, Xu Lan, Jiashi Feng, Zhenguo Li, and Liwei Wang. Boosting few-shot learning with adaptive margin loss. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 12576–12584, 2020.

[Liu *et al.*, 2020] Bin Liu, Yue Cao, Yutong Lin, Qi Li, Zheng Zhang, Mingsheng Long, and Han Hu. Negative margin matters: Understanding margin in few-shot classification. In *Computer Vision–ECCV 2020: 16th European*

Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16, pages 438–455. Springer, 2020.

[Miller *et al.*, 2000] Erik G Miller, Nicholas E Matsakis, and Paul A Viola. Learning from one example through shared densities on transforms. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2000 (Cat. No. PR00662)*, volume 1, pages 464–471. IEEE, 2000.

[Oord *et al.*, 2018] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[Park *et al.*, 2020] Taesung Park, Alexei A Efros, Richard Zhang, and Jun-Yan Zhu. Contrastive learning for unpaired image-to-image translation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*, pages 319–345. Springer, 2020.

[Peng *et al.*, 2019] Zhimao Peng, Zechao Li, Junge Zhang, Yan Li, Guo-Jun Qi, and Jinhui Tang. Few-shot image recognition with knowledge transfer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 441–449, 2019.

[Qi *et al.*, 2018] Hang Qi, Matthew Brown, and David G Lowe. Low-shot learning with imprinted weights. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5822–5830, 2018.

[Qiao *et al.*, 2019] Limeng Qiao, Yemin Shi, Jia Li, Yaowei Wang, Tiejun Huang, and Yonghong Tian. Transductive episodic-wise adaptive metric for few-shot learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3603–3612, 2019.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[Ravichandran *et al.*, 2019] Avinash Ravichandran, Rahul Bhotika, and Stefano Soatto. Few-shot learning with embedded class models and shot-free meta training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 331–339, 2019.

[Ren *et al.*, 2018] Mengye Ren, Eleni Triantafillou, Sachin Ravi, Jake Snell, Kevin Swersky, Joshua B Tenenbaum, Hugo Larochelle, and Richard S Zemel. Meta-learning for semi-supervised few-shot classification. *arXiv preprint arXiv:1803.00676*, 2018.

[Satorras and Estrach, 2018] Victor Garcia Satorras and Joan Bruna Estrach. Few-shot learning with graph neural networks. In *International Conference on Learning Representations*, 2018.

[Schwartz *et al.*, 2022] Eli Schwartz, Leonid Karlinsky, Rogerio Feris, Raja Giryes, and Alex Bronstein. Baby steps towards few-shot learning with multiple semantics. *Pattern Recognition Letters*, 160:142–147, 2022.

[Snell *et al.*, 2017] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in Neural Information Processing Systems*, 30, 2017.

[Sung *et al.*, 2018] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1199–1208, 2018.

[Tian *et al.*, 2020] Yonglong Tian, Yue Wang, Dilip Krishnan, Joshua B Tenenbaum, and Phillip Isola. Rethinking few-shot image classification: a good embedding is all you need? In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, pages 266–282. Springer, 2020.

[Vinyals *et al.*, 2016] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. *Advances in Neural Information Processing Systems*, 29, 2016.

[Wu *et al.*, 2018] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 3733–3742, 2018.

[Xing *et al.*, 2019] Chen Xing, Negar Rostamzadeh, Boris Oreshkin, and Pedro O O Pinheiro. Adaptive cross-modal few-shot learning. *Advances in Neural Information Processing Systems*, 32, 2019.

[Yang *et al.*, 2022] Fengyuan Yang, Ruiping Wang, and Xilin Chen. Sega: Semantic guided attention on visual prototype for few-shot learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1056–1066, 2022.

[Zhang *et al.*, 2020] Chi Zhang, Yujun Cai, Guosheng Lin, and Chunhua Shen. Deepemd: Few-shot image classification with differentiable earth mover's distance and structured classifiers. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 12203–12213, 2020.

[Zhang *et al.*, 2021] Baoquan Zhang, Xutao Li, Yunming Ye, Zhichao Huang, and Lisai Zhang. Prototype completion with primitive knowledge for few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3754–3762, 2021.