Simple Semantic-Aided Few-Shot Learning

Hai Zhang[1*] Junzhe Xu[1,2*‡] Shanlin Jiang[3] Zhenan He[1†]
[1]College of Computer Science, Sichuan University
[2]Banma Network Technology, Alibaba Group
[3]Naveen Jindal School of Management, University of Texas at Dallas
zhanghi@stu.scu.edu.cn, xujunzhe.xjz@alibaba-inc.com
shanlin.jiang@utdallas.edu, zhenan@scu.edu.cn

Abstract

Learning from a limited amount of data, namely Few-Shot Learning, stands out as a challenging computer vision task. Several works exploit semantics and design complicated semantic fusion mechanisms to compensate for rare representative features within restricted data. However, relying on naive semantics such as class names introduces biases due to their brevity, while acquiring extensive semantics from external knowledge takes a huge time and effort. This limitation severely constrains the potential of semantics in Few-Shot Learning. In this paper, we design an automatic way called Semantic Evolution to generate high-quality semantics. The incorporation of high-quality semantics alleviates the need for complex network structures and learning algorithms used in previous works. Hence, we employ a simple two-layer network termed Semantic Alignment Network to transform semantics and visual features into robust class prototypes with rich discriminative features for few-shot classification. The experimental results show our framework outperforms all previous methods on six benchmarks, demonstrating a simple network with high-quality semantics can beat intricate multi-modal modules on few-shot classification tasks. Code is available at https://github.com/zhangdoudou123/SemFew.

1. Introduction

Deep learning models have witnessed substantial advancements by leveraging extensive annotated data [8]. Nevertheless, in numerous real-world scenarios, the shortage of labeled data restricts the applicability of conventional deep learning approaches. In contrast, humans possess a remarkable cognitive ability to learn new concepts and recognize categories from just a few examples [3]. Motivated by this, Few-Shot Learning (FSL) [46] is proposed to mimic the ability of humans to learn from a few labeled samples.

In a typical setting of FSL [46], the support set, which consists of N novel classes and each class contains K samples, is provided for the model. The model learned on the support set is required to accurately classify test samples, which are labeled as the query set, into one of the N categories. A conventional classification approach involves projecting both support and query sets onto a pre-established metric space. Then, query images are classified by finding the nearest support image. However, few support samples may not contain sufficient discriminative features for recognition. These samples are typically located on the periphery of the sample cluster in the metric space, resulting in unstable distance evaluations and inaccurate classification.

To address this issue, researchers start from visual-based methods [7, 11, 13, 16, 19, 29, 32, 37, 46, 49, 53, 60], aiming at extracting class-related features from periphery image, reducing the intra-class variation among samples, and thus constructing robust features for classification. Although visual-based methods have undergone extensive investigation and achieved great success, this type of approach still struggles to handle situations when there is only one hard periphery sample with minimal semantic features per class, *i.e.*, the one-shot learning task [25, 53]. Hence, some researchers turned to applying semantics as auxiliary information to help the model better understand periphery images, leveraging the synergies between language and vision modalities.

From this perspective, semantic-based methods [5, 28, 38, 44, 50, 56–59, 61] dedicated to exploring different types of semantics to enhance periphery samples. One kind of method [57–59] applies the class name as the semantic source. However, class name is not the best way to understand a novel class. For instance, If a person never saw a zebra, it would be easier to identify a zebra by the definition, *i.e.*, a horse with stripes, than by the name zebra.

[*] Equal Contribution. [†] Corresponding Author.
[‡] Work done during master's degree at Sichuan University.

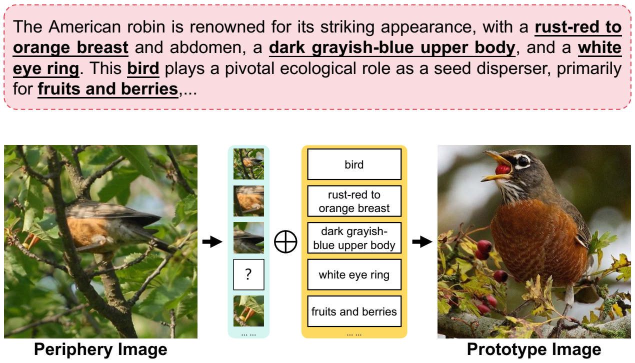

The American robin is renowned for its striking appearance, with a **rust-red to orange breast** and abdomen, a **dark grayish-blue upper body**, and a **white eye ring**. This **bird** plays a pivotal ecological role as a seed disperser, primarily for **fruits and berries**,...

Figure 1. The introduction of how high-quality semantics reconstruct the class prototype through complementarity between modalities. Periphery image is an image with fewer discriminative features and prototype image represents images with concrete and enriched representative features.

Meanwhile, the class name contains ambiguity in some scenarios. For example, the class *ear* is easily misidentified as the human organ, but images in MiniImageNet [53] belong to this class are actually photos of corn. Another kind of semantic exploits discrete attribute labels as the substitution of mere class names and achieves good performance [61]. Nevertheless, the collection process is time-consuming and expert assistance requiring. Moreover, discrete attribute labels lack robust representation and struggle to benefit from context. Therefore, effectively collecting and leveraging high-quality semantics becomes an urgent research need.

In this paper, we propose a simple framework called **Sem**antic-Aided **Few**-Shot Learning (SemFew). To begin with, we identify the vital role of high-quality semantics in FSL and innovatively propose an automatic step-by-step **Sem**antic **Evo**lution (SemEvo) process to acquire detailed and accurate semantics. It first converts the class name into a short description that matches the image content of each class. Next, to further enrich recognizable characteristics, we expand and paraphrase short descriptions. The paraphrased descriptions contain more class-related knowledge than class names and short descriptions, making them better encompass a vast array of visual features. To illustrate, Fig. 1 showcases an instance that the paraphrased semantic contains a large number of detailed descriptions matching visual features *e.g*. organ characteristics and eating habitat. With this information, the model can better reconstruct an image with periphery feature (left image), into an image with prototype feature (right image), which benefits few-shot classification. Then, to fully exploit the advantage of paraphrased semantics, we design a network called **Sem**antic **Align**ment Network (SemAlign). Different from previous methods, SemAlign does not adopt complex semantic understanding modules but utilizes a basic two-layer network. After accepting semantics and visual features as input, it reconstructs prototype features for subsequent classification. Experimental results on six widely adopted benchmarks demonstrate the effectiveness of our method, providing a simple insight into future FSL research. Our contributions can be summarized as three points:

- To the best of our knowledge, we are the first to consider the automatic way of collecting high-quality semantics and applying them in FSL.
- We design a simple and efficient way to translate high-quality semantics and visual features into prototypes, without any intricate semantic understanding modules.
- Our approach achieves state-of-the-art performance across six benchmarks in FSL research, underscoring that a basic network can obtain excellent performance when supported by high-quality semantics.

2. Related Work

The significant issue in FSL is how to effectively and accurately extract class-related features from limited data. Targeting this issue, existing research explored two different types of approaches, *i.e*., visual-based methods and semantic-based methods.

Visual-Based Methods focus on extracting class-related features from images for classification. There are two main types of this approach. The approaches with the optimization principle behind [1, 13, 40] concentrated on acquiring a set of initial model parameters capable of rapid adaptation to novel classes. However, updating the entire model with a limited amount of labeled data may result in meta-overfitting [12, 22, 35, 42, 65]. The approaches considered metric design aiming at training a metric space, where inter-class distances are maximized, while intra-class distances are minimized. Within this type of approach, one attracting method focuses on integrating self-supervised loss to pretrain the backbone network [10, 15, 19, 32, 49]. Other methods are dedicated to improving the metric, and previous research spanning from initially manually designed metrics [2, 25, 46, 62] to subsequently model-based metrics [48, 53], all of which have yielded encouraging outcomes.

Semantic-Based Methods sought to enhance visual recognition performance by fusing complementary information from both visual and textual modalities [5, 28, 57, 58, 61], and related research introducing intricate network frameworks to effectively utilize textual information. Chen *et al*. [57] introduced an adaptive fusion mechanism to integrate visual prototypes with semantic prototypes obtained through word embeddings of class labels. Zhang *et al*. [61] designed a framework for prototype completion, leveraging rare attribute-level information from labels to construct representative class prototypes. Xu and Le [58] proposed the use of the Conditional Variational Autoencoder (CVAE) [47] model to generate visual features based on semantic embeddings for further classification. Li *et al*. [28] presented a task-relevant adaptive margin loss based on the semantic similarity between class names. Chen *et al.* [5] devised complementary mechanisms for inserting semantic prompts into the feature extractor, allowing the feature extractor to better focus on class-specific features.

Previous research emphasized the exploitation of class names and designed a complicated network to integrate semantics with visual information. In contrast, our primary concern lies in the inherent quality of semantics, and how it can relieve the burden of designing intricate network structures and learning algorithms. Remarkably, we exhibit that high-quality semantics information has the ability to yield superior outcomes even within a simple network.

3. Methodology

This section first revisits the background of FSL, then introduces two components of SemFew: explaining how semantics are gradually improved through SemEvo, and demonstrating the details of our proposed SemAlign.

3.1. Background

The dataset of FSL includes two components: a base set $\mathcal{D}_{base} = \{(x, y)|x \in \mathcal{X}_{base}, y \in \mathcal{C}_{base}\}$ for training a metric space and a novel set $\mathcal{D}_{novel} = \{(x, y)|x \in \mathcal{X}_{novel}, y \in \mathcal{C}_{novel}\}$ for testing, where x denotes the image and y represents the label. It is important to note that the label space for both sets is disjoint, *i.e.*, $\mathcal{C}_{base} \cap \mathcal{C}_{novel} = \emptyset$. During the testing process, the support set $\mathcal{S} = \{(x_i, y_i)\}_{i=0}^{N\times K}$ is randomly sampled from $\mathcal{D}_{novel}$, which contains N classes with each class consisting of K samples. Then, the model is required to correctly classify images in the query set $\mathcal{Q} = \{(x_i, y_i)\}_{i=0}^{N\times M}$ into one of N classes in the support set $\mathcal{S}$, where M is the number of query samples in each class. Generally, this type of classification task is named as N-way K-shot task.

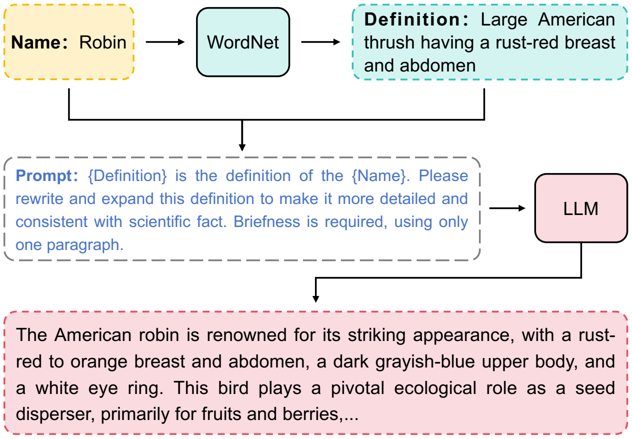
A simple but effective FSL method is to calculate the prototype of each support class, and then classify query samples by finding the nearest prototype in the metric space. To be specific, the prototype of class t, labeled as p_t, is the mean vector of all samples in the support set belonging to class t:

$$p_t = \frac{1}{|\mathcal{S}_t|} \sum_{x_i \in \mathcal{S}_t} f(x_i), \tag{1}$$

where $\mathcal{S}_t$ is a set of all samples belonging to class t in the support set $\mathcal{S}$ and f denotes the backbone network. Then, the probability of a query sample q belonging to class t is related to the distance between q and prototype of class t [46], which is calculated as follows:

$$P(y = t) = \frac{\exp d(q, p_t)}{\sum_{i=1}^{N} \exp d(q, p_i)}, \tag{2}$$

where $d(\cdot, \cdot)$ means distance metric, *e.g.*, cosine distance, and N denotes the number of classes.


Name: Robin
WordNet
Definition: Large American thrush having a rust-red breast and abdomen
LLM

Prompt: {Definition} is the definition of the {Name}. Please rewrite and expand this definition to make it more detailed and consistent with scientific fact. Briefness is required, using only one paragraph.

The American robin is renowned for its striking appearance, with a rust-red to orange breast and abdomen, a dark grayish-blue upper body, and a white eye ring. This bird plays a pivotal ecological role as a seed disperser, primarily for fruits and berries,...

Figure 2. The illustration of how Semantic Evolution converts the simple name and the definition into the high-quality description.

3.2. Semantic Evolution

This section explains how we automatically generate high-quality semantics for few-shot classification by the designed process termed Semantic Evolution. This process first converts class names to short descriptions and then paraphrases them into high-quality semantics.

As we illustrated before, adopting class names as semantics [5, 39, 58] unavoidably introduces ambiguity. Therefore, we first retrieve the definition of each class name from WordNet [33] that matches image content, where definitions are more detailed and concrete than mere class names. However, even though definitions are able to precisely describe a class to some extent, brief descriptions may still overlook key visual information, due to the fact that visual features are highly variable.

To comprehensively conclude variable visual features, we leverage the extensive knowledge stored in the pretrained Large Language Models (LLMs) [4] to augment definitions. The prompt we used is: *{WordNet Definition} is the definition of the {Class Name}. Please rewrite and expand this definition to make it more detailed and consistent with scientific fact. Briefness is required, using only one paragraph.* The illustration of the Semantic Evolution process is shown in Fig. 2.

The paraphrased class description generated by LLMs precisely describes the corresponding class itself, which does not apply to other categories. This characteristic ensures that semantic descriptions do not leak information to new categories, which follows the paradigm of FSL.

3.3. Semantic Alignment Network

After obtaining paraphrased semantics in the previous section, we devise the Semantic Alignment Network (SemAlign), aiming at translating hard periphery samples into robust class prototypes.

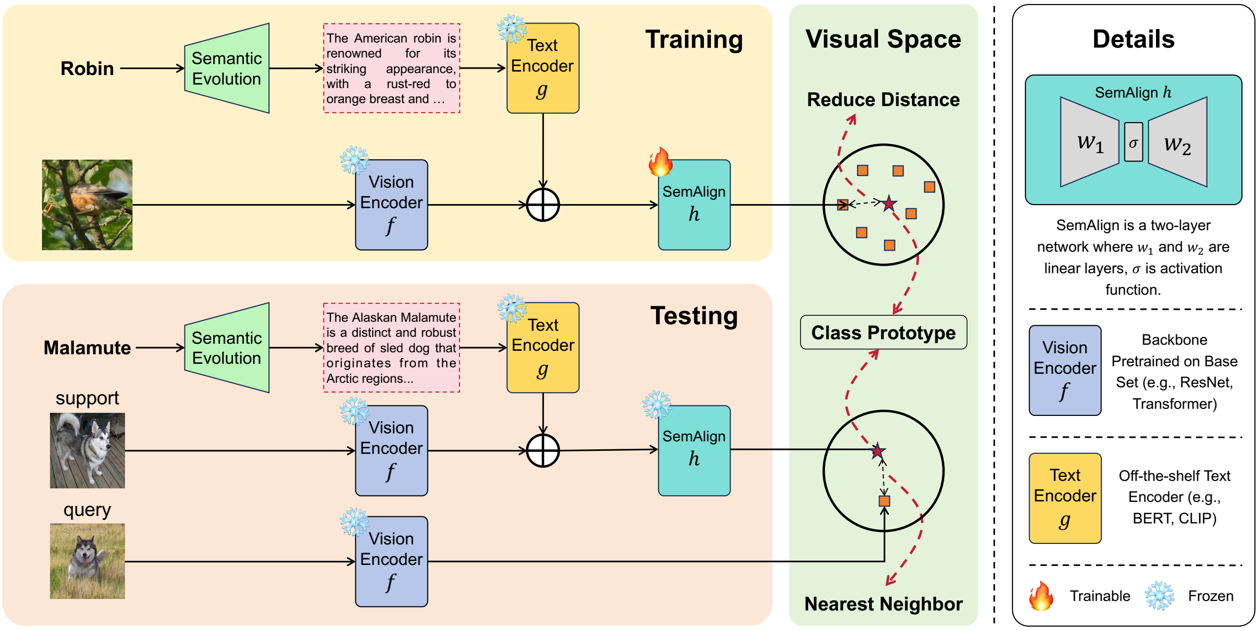

Training
Robin
Semantic Evolution
The American robin is renowned for its striking appearance, with a rust-red to orange breast and …
Text Encoder g
Vision Encoder f
SemAlign h
Visual Space
Reduce Distance
Class Prototype
Testing
Malamute
Semantic Evolution
The Alaskan Malamute is a distinct and robust breed of sled dog that originates from the Arctic regions...
Text Encoder g
support
Vision Encoder f
SemAlign h
query
Vision Encoder f
Nearest Neighbor
Details
SemAlign h
w_1
σ
w_2
SemAlign is a two-layer network where w_1 and w_2 are linear layers, σ is activation function.
Vision Encoder f
Backbone Pretrained on Base Set (e.g., ResNet, Transformer)
Text Encoder g
Off-the-shelf Text Encoder (e.g., BERT, CLIP)
Trainable
Frozen

Figure 3. The framework of our proposed SemFew. During the training stage, images and paraphrased semantics are encoded and fed into SemAlign h, with the objective of reducing the distance between the output of h and the class prototype in the visual space. During the testing stage, images in the support set are transformed into class prototypes by h, and query images are classified by identifying the nearest prototype. The symbol $\oplus$ denotes a concatenation operation.

For a cluster of samples, the cluster center best encapsulates the characteristic of a category and has been regarded as a prototype in previous work [46]. Suppose S represents semantic features, and C stands for the center of each class, the simplest implementation of SemAlign is to learn the alignment between semantics S and centers C:

$$\min_{W_1, W_2} L(\sigma(S^\top W_1)W_2, C), \tag{3}$$

where W_1 and W_2 are parameters of network, σ is an activation function.

Considering that each modality views things from a different and complementary perspective, this phenomenon can be reasonably utilized to obtain better prototypes. Therefore, we fuse visual features with high-quality semantics, and then align multi-modal output with the cluster center, the details of the whole pipeline can be found in Fig. 3. Specifically, suppose the i-th image in the dataset is represented as x_i, its label and semantic are denoted as y_i and s_i, respectively. First, x_i is encoded by a vision encoder f, which is pre-trained on the base set. Also, semantic description s_i is encoded by an off-the-shelf text encoder g. Then, the encoded visual and text features are concatenated together and fed into SemAlign h to reconstruct the prototype. The training loss supervises the distance between the reconstructed prototype towards the cluster center:

$$\min_{W_1, W_2} \mathbb{E}[||h(f(x_i), g(s_i)) - c_y||_1], \tag{4}$$

where $h(f(x_i), g(s_i)) = \sigma([f(x_i) \cdot g(s_i)]^\top W_1)W_2$ is SemAlign with learnable weights W_1 and W_2, $[\cdot]$ means concatenation operation.

During the testing process, a N-way K-shot support set is randomly sampled in the novel set. For class t, we obtain its paraphrased description s_t by Semantic Evolution proposed in Sec. 3.2. Then, support image x and description s_t are encoded by f and g, respectively, and fed into h to get the reconstructed prototype:

$$r_t = \frac{1}{K} \sum_{i=1}^{K} h(f(x_i), g(s_t)), \tag{5}$$

where r_t is the reconstructed prototype. To ensure the reconstructed prototype is further aligned with the ground-truth class prototype, it will be fused with support images in a convex combination manner [57]:

$$p_t = kr_t + (1-k)u_t, \tag{6}$$

where p_t means the classification prototype of class t, $u_t = \frac{1}{K}\sum_{i=1}^{K} f(x_i)$ is the mean vector of support samples, r_t is the reconstructed prototype produced by SemAlign, and $k \in [0,1]$ is a manually controlled fusion factor. Then, Eq. (2) is utilized to assign label for each query sample, where $d(\cdot,\cdot)$ is set as cosine distance.

	Method	**Venue**	**Backbone**	**MiniImageNet**		**TieredImageNet**	
				5-way 1-shot	**5-way 5-shot**	**5-way 1-shot**	**5-way 5-shot**
Visual-Based	MatchNet [53]	NeurIPS'16	ResNet-12	65.64 ± 0.20	78.72 ± 0.15	68.50 ± 0.92	80.60 ± 0.71
	ProtoNet [46]	NeurIPS'17	ResNet-12	62.39 ± 0.21	80.53 ± 0.14	68.23 ± 0.23	84.03 ± 0.16
	MAML [13]	ICML'17	ResNet-12	49.24 ± 0.21	58.05 ± 0.10	67.92 ± 0.17	72.41 ± 0.20
	TADAM [37]	NeurIPS'18	ResNet-12	58.50 ± 0.30	76.70 ± 0.30	62.13 ± 0.31	81.92 ± 0.30
	CAN [20]	NeurIPS'19	ResNet-12	63.85 ± 0.48	79.44 ± 0.34	69.89 ± 0.51	84.23 ± 0.37
	CTM [29]	CVPR'19	ResNet-18	64.12 ± 0.82	80.51 ± 0.13	68.41 ± 0.39	84.28 ± 1.73
	RFS [49]	ECCV'20	ResNet-12	62.02 ± 0.63	79.64 ± 0.44	69.74 ± 0.72	84.41 ± 0.55
	FEAT [60]	CVPR'20	ResNet-12	66.78 ± 0.20	82.05 ± 0.14	70.80 ± 0.23	84.79 ± 0.16
	Meta-Baseline [6]	ICCV'21	ResNet-12	63.17 ± 0.23	79.26 ± 0.17	68.62 ± 0.27	83.29 ± 0.18
	SUN [11]	ECCV'22	ViT-S	67.80 ± 0.45	83.25 ± 0.30	72.99 ± 0.50	86.74 ± 0.33
	FewTURE [19]	NeurIPS'22	Swin-T	72.40 ± 0.78	86.38 ± 0.49	76.32 ± 0.87	**89.96 ± 0.55**
	FGFL [7]	ICCV'23	ResNet-12	69.14 ± 0.80	86.01 ± 0.62	73.21 ± 0.88	87.21 ± 0.61
	Meta-AdaM [45]	NeurIPS'23	ResNet-12	59.89 ± 0.49	77.92 ± 0.43	65.31 ± 0.48	85.24 ± 0.35
Semantic-Based	KTN [38]	ICCV'19	Conv-128	64.42 ± 0.72	74.16 ± 0.56	74.16 ± 0.56	-
	AM3 [57]	NeurIPS'19	ResNet-12	65.30 ± 0.49	78.10 ± 0.36	69.08 ± 0.47	82.58 ± 0.31
	TRAML[28]	CVPR'20	ResNet-12	67.10 ± 0.52	79.54 ± 0.60	-	-
	AM3-BERT [59]	ICMR'21	ResNet-12	68.42 ± 0.51	81.29 ± 0.59	77.03 ± 0.85	87.20 ± 0.70
	SVAE-Proto [58]	CVPR'22	ResNet-12	74.84 ± 0.23	83.28 ± 0.40	76.98 ± 0.65	85.77 ± 0.50
	SP-CLIP [5]	CVPR'23	Visformer-T	72.31 ± 0.40	83.42 ± 0.30	78.03 ± 0.46	88.55 ± 0.32
	SemFew	Ours	ResNet-12	77.63 ± 0.63	83.04 ± 0.48	78.96 ± 0.80	85.88 ± 0.62
	SemFew-Trans	Ours	Swin-T	**78.94 ± 0.66**	**86.49 ± 0.50**	**82.37 ± 0.77**	**89.89 ± 0.52**

Table 1. Results (%) on MiniImageNet and TieredImageNet. The ± shows 95% confidence intervals. The best results are shown in **bold**.

4. Experiments

We exhibit experimental results in this section, including benchmarks introduction, implementation details, and few-shot classification performance.

4.1. Datasets

The study evaluates the proposed method on six established FSL datasets: MiniImageNet [53], TieredImageNet [41], CIFAR-FS [27], FC100 [37], Places [63], and CUB [54]. **MiniImageNet** and **TieredImageNet** are both subsets of ImageNet [8]. MiniImageNet consists of 64 training classes, 16 validating classes, and 20 testing classes. TieredImageNet encompasses 351 training classes, 97 validating classes, and 160 testing classes. **CIFAR-FS** and **FC100** are derived from the CIFAR-100 [26]. CIFAR-FS employs a random partitioning strategy with 64 training classes, 16 validating classes, and 20 testing classes. FC100 introduces a unique superclass partitioning approach, where the training set comprises 12 superclasses, namely 60 classes, while both the validation and test sets include four superclasses, totaling 20 classes. **Places** and **Caltech-UCSD Birds-200-2011 (CUB)** are datasets for cross-domain scenario testing. In a cross-domain evaluation [51], the model is initially trained on MiniImageNet's base split and tested on the corresponding test split.

4.2. Implementation Details

Architecture. We employ a pre-trained ResNet-12 [18] and Swin-T [30] as the vision encoders. Additionally, we apply two types of text encoder: the text encoder from ViT-B/16 CLIP [39] where output dimension is 512, and the BERT-Base text encoder [9] which has 768 output dimension. The network h that combines visual and textual features consists of two fully connected layers and a LeakyReLU [31] activation function. The hidden layer has a dimension of 4,096. In the context of Semantic Evolution, the utilized Large Language Model is GPT-3.5-turbo [36].

Training Details. A two-step training procedure is performed in experiments. In step one, two kinds of vision encoder are pre-trained, the ResNet-12 [18] follows the same pretraining settings as the FEAT [60], while the Swin-T [30] adopts the identical pretraining strategy as FewTURE [19]. Next, the network h is trained 50 epochs. This training is conducted with a batch size of 128 and employs an Adam optimizer [24] with a learning rate of 1e-4.

Evaluation protocol. The proposed method is evaluated under 5-way 1-shot and 5-way 5-shot settings. For each setting, 600 classification tasks are uniformly sampled from the novel set. In each task, there are 15 query samples per class. The mean and 95% confidence interval of the accuracy are reported.

Method	**Venue**	**Backbone**	**CIFAR-FS**		**FC100**	
			5-way 1-shot	**5-way 5-shot**	**5-way 1-shot**	**5-way 5-shot**
ProtoNet [46]	NeurIPS'17	ResNet-12	72.20 ± 0.70	83.50 ± 0.50	41.54 ± 0.76	57.08 ± 0.76
TADAM [37]	NeurIPS'18	ResNet-12	-	-	40.10 ± 0.40	56.10 ± 0.40
MetaOptNet [27]	CVPR'19	ResNet-12	72.80 ± 0.70	84.30 ± 0.50	47.20 ± 0.60	55.50 ± 0.60
MABAS [23]	ECCV'20	ResNet-12	73.51 ± 0.92	85.65 ± 0.65	42.31 ± 0.75	58.16 ± 0.78
RFS [49]	ECCV'20	ResNet-12	71.50 ± 0.80	86.00 ± 0.50	42.60 ± 0.70	59.10 ± 0.60
SUN [11]	ECCV'22	ViT-S	78.37 ± 0.46	88.84 ± 0.32	-	-
FewTURE [19]	NeurIPS'22	Swin-T	77.76 ± 0.81	88.90 ± 0.59	47.68 ± 0.78	63.81 ± 0.75
Meta-AdaM [45]	NeurIPS'23	ResNet-12	-	-	41.12 ± 0.49	56.14 ± 0.49
SP-CLIP [5]	CVPR'23	Visformer-T	82.18 ± 0.40	88.24 ± 0.32	48.53 ± 0.38	61.55 ± 0.41
SemFew	Ours	ResNet-12	83.65 ± 0.70	87.66 ± 0.60	**54.36 ± 0.71**	62.79 ± 0.74
SemFew-Trans	Ours	Swin-T	**84.34 ± 0.67**	**89.11 ± 0.54**	54.27 ± 0.77	**65.02 ± 0.72**

Table 2. Results (%) on CIFAR-FS and FC100. The ± shows 95% confidence intervals. The best results are shown in **bold**.

Method	**Venue**	**CUB**		**Places**	
		1-shot	**5-shot**	**1-shot**	**5-shot**
GNN [43]	ICLR'18	45.69	62.25	53.10	70.84
S2M2 [32]	WACV'20	48.24	70.44	-	-
FT [51]	ICLR'20	47.47	66.98	55.77	73.94
ATA [55]	IJCAI'21	45.00	66.22	53.57	75.48
AFA [21]	ECCV'22	46.86	68.25	54.04	76.21
StyleAdv [14]	CVPR'23	48.49	68.72	58.58	**77.73**
LDP-net [64]	CVPR'23	49.82	70.39	53.82	72.90
SemFew-Name	Ours	57.58	72.26	63.22	74.54
SemFew	Ours	**59.07**	**72.47**	**64.01**	74.70

Table 3. Average results (%) on cross-domain scenarios. SemFew-Name denotes that semantics are class names.

4.3. Experimental Results

4.3.1 Few-Shot Classification

The comparative results of our proposed approach against previous methods across four benchmarks are shown in Tabs. 1 and 2. Specifically, in the 1-shot task, SemFew exhibits notable superiority, surpassing SP-CLIP [5] **0.93%-5.83%** on four benchmarks. Recent studies [5, 11, 19] have explored different Transformer architectures as backbones. Consequently, we conduct experiments with a Transformer backbone, resulting in SemFew-Trans. This variant further elevates performance, outperforming SP-CLIP by **2.16%-6.63%** in the 1-shot setting. In summary, SemFew-Trans achieves state-of-the-art results compared to both previous visual-based and semantic-based approaches across all 1-shot and 5-shot tasks.

4.3.2 Cross-Domain Classification

To evaluate whether semantics serve as an effective bridge across novel tasks, we test SemFew on completely novel datasets. Such a task is also known as Cross-Domain Few-Shot Learning (CDFSL) [51], which involves extending the capacity to recognize novel categories to domains distinct from the training domain. All the results are reported in Tab. 3, where the experimental setup is the same as FT [51]. Tab. 3 indicates that SemFew significantly surpassing the state-of-the-art methods, where it outperforms best compared methods by **9.25**% (1-shot) and **1.87**% (5-shot) on CUB. It also surpasses the lastest method [64] **10.19**% (1-shot) and **1.80**% (5-shot) on Places. Conversely, SemFew-Name, which adopts a template (*i.e*., *The Photo of a bird called* {*Class Name*}) as the semantic, performs worse than SemFew. This observation highlights the inadequacy of using solely names as semantics.

4.4. Ablation Studies

We conduct ablation experiments on three datasets, *i.e*., MiniImageNet, CIFAR-FS, and FC100, and all experiments are performed using the ResNet-12 network architecture.

4.4.1 Alignment Source

Given that our framework is primarily responsible for translating visual samples V and semantics descriptions S into ground-truth class prototype (*i.e*., $V+S \Rightarrow C$), we conduct ablation studies by removing S or V. The results can be found in Tab. 4. First, we remove semantics and force pure visual samples to reconstruct prototypes (*i.e*., $V \Rightarrow C$). The result shows that it gets the worst performance, indicating that reconstructing a prototype from pure visual features is challenging. Then, we abandon visual features to implement our proposed principle in Sec. 3.3 (*i.e*., $S \Rightarrow C$). As we can see, it achieves substantial performance gains, reflecting the rich class representative features within semantics. However, our method combining visual and semantic acquires the best performance, which indicates the complementarity between visual and semantic modalities does en-

Method	5-way 1-shot MiniImageNet	 CIFAR-FS	 FC100	5-way 5-shot MiniImageNet	 CIFAR-FS	 FC100
$V \Rightarrow C$	66.27 ± 0.81	74.43 ± 0.90	44.72 ± 0.75	81.70 ± 0.55	86.74 ± 0.61	60.58 ± 0.76
$S \Rightarrow C$	76.77 ± 0.66	82.27 ± 0.71	52.12 ± 0.73	82.92 ± 0.48	87.42 ± 0.60	62.22 ± 0.75
$V + S \Rightarrow C$	**77.63 ± 0.63**	**83.65 ± 0.70**	**54.36 ± 0.71**	**83.04 ± 0.48**	**87.66 ± 0.60**	**62.79 ± 0.74**

Table 4. Results (%) on different alignment sources. V means visual features, S represents semantic features, C stands for class prototypes.

Center	MiniImageNet		TieredImageNet		CIFAR-FS		FC100	
	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
Cluster	77.16	82.94	78.36	85.19	83.39	87.61	53.30	62.22
Mean	77.63	83.04	78.96	85.88	83.65	87.66	54.36	62.79

Table 5. Average results (%) on different prototypes. **Mean** represents the mean vector of sample clusters. **Cluster** means the prototype is determined by Clustering Algorithm.

classifier	MiniImageNet		TieredImageNet		CIFAR-FS		FC100	
	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
LR	77.55	82.70	78.66	84.72	83.08	86.70	51.53	57.49
EU	77.56	83.15	78.65	85.26	82.83	86.87	54.14	62.54
CO	77.63	83.04	78.96	85.88	83.65	87.66	54.36	62.79

Table 6. Average results (%) on different classifiers. **LR**: linear logistic regression classifier. **EU**: euclidean-distance nearest prototype classifier. **CO**: cosine-distance nearest prototype classifier.

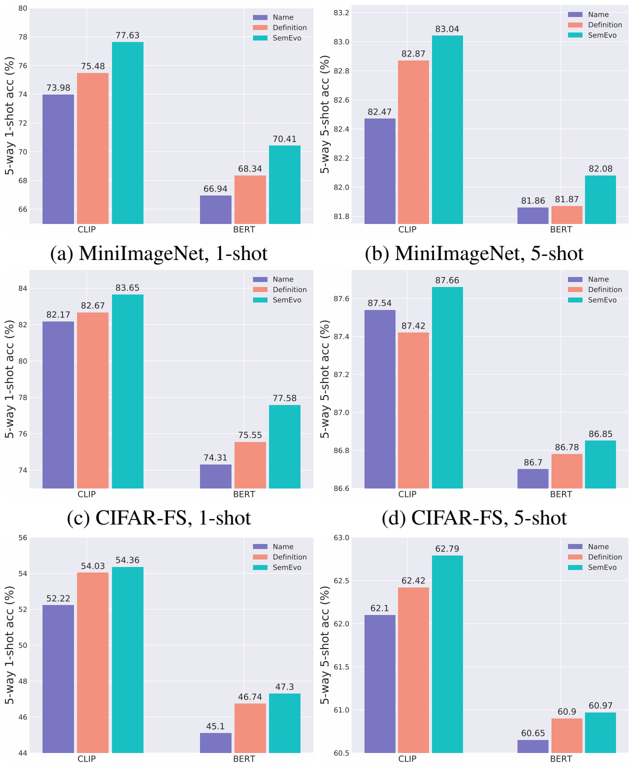
hance the robustness of the reconstructed prototype.

4.4.2 Prototype and Classifier Selection Strategy

Tab. 5 displays the performance associated with various learning targets. Results reveal that employing the mean vector as a prototype yields slightly better results compared to prototypes identified by the clustering algorithm [17]. This suggests that the mean vector consolidates representative visual information, making it a highly effective class prototype. As for the classification head design, we present the results in Tab. 6, which indicates the use of cosine distance as the metric to achieve nearest neighbor classification gets the best results, euclidean distance gets similar but lower outcomes, and linear logistic classifier gets the worst performance.

4.4.3 Semantic Evolution

We use different kinds of semantics and text encoders to evaluate the influence of semantics, where results can be found in Fig. 4. In most scenarios, our model attains optimal results when aided by paraphrased semantics, and the model trained on simple definitions secures the second-best performance. This trend is reasonable since paraphrased descriptions contain more knowledge than simple definitions and class names. Between the comparison of different text encoders, the model trained on CLIP encoded semantics greatly outperforms the model trained on BERT encoded semantics. This is because CLIP is trained in a multi-modal manner, rendering it better capable of understanding images. Although a multi-modal text encoder has been widely adopted in FSL research, our model obtains competitive results on BERT, demonstrating that uni-modal semantics can also achieve excellent results.



(a) MiniImageNet, 1-shot

(b) MiniImageNet, 5-shot

(c) CIFAR-FS, 1-shot

(d) CIFAR-FS, 5-shot

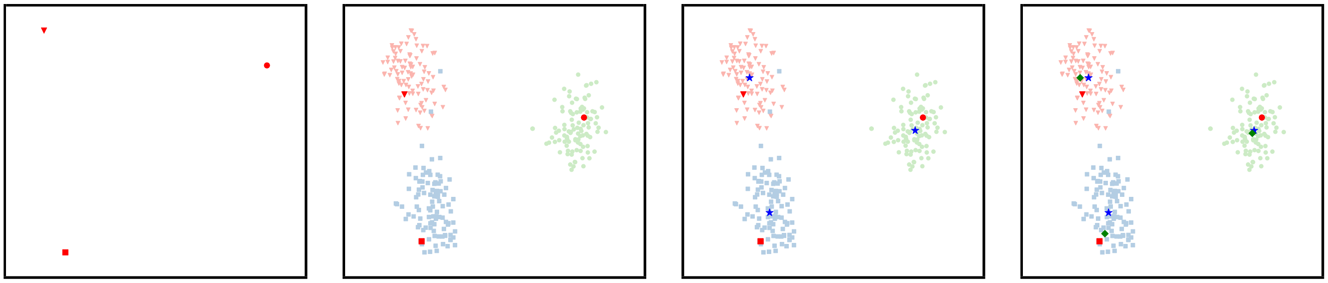
(e) FC100, 1-shot

(f) FC100, 5-shot

Figure 4. Average results (%) on different semantics.

4.5. Visualization Analysis

We visualize visual space features through the t-SNE algorithm [52], where the results are presented in Fig. 5. We select three novel classes on the MiniImageNet dataset, and then randomly sampled three support samples and one hundred query samples, where the distribution of support and query samples in visual space can be found in Figs. 5a



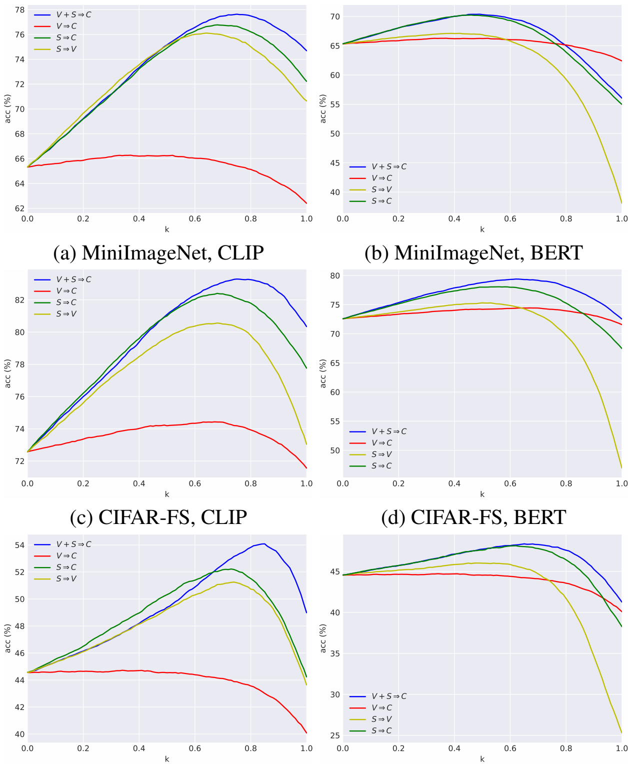
(a) Support

(b) Query

(c) Real Prototype

(d) Reconstructed Prototype

Figure 5. Visualization results on the MiniImageNet dataset. Different colors or shapes represent different classes. The ⋆ represents the class prototypes, and the ♢ denotes the prototypes reconstructed by our method.



(a) MiniImageNet, CLIP

(b) MiniImageNet, BERT

(c) CIFAR-FS, CLIP

(d) CIFAR-FS, BERT

(e) FC100, CLIP

(f) FC100, BERT

Figure 6. Average results (%) on different fusion factor k.

and 5b. In Fig. 5c, we put support samples, *i.e.*, red dots, mean vector of each class, *i.e.*, blue stars, and query samples in the same picture. As we can see, support samples are close to the cluster edge rather than the mean center, which leads to difficulties in correctly classifying most samples in the query set. However, after being transformed by SemFew, the reconstructed prototypes are closer to the mean center of each cluster compared to the support set samples. This phenomenon demonstrates that our method can reasonably transform visual samples into class prototypes corresponding to their respective categories.

4.6. Fusion Factor Analysis

The fusion factor, denoted as k in Eq. (6), significantly influences performance. To demonstrate, we systematically vary the value of k between 0 and 1, using a step size of 0.01, and present the accuracy trends across three benchmarks in Fig. 6. Notably, our model, *i.e.*, blue line, gets the highest accuracy when k is large. Given that k represents the percentage of the reconstructed prototype r_t during fusion in Eq. (6), the experiment highlights that our model produces high-quality prototypes. This observation is further supported by the accuracy comparison when $k = 1$. In such cases, our model consistently outperforms others, providing strong evidence for the effectiveness of our reconstructed prototypes.

Additionally, we add $S \Rightarrow V$ experiment for comparison, which generates visual samples rather than prototypes through a Conditional Generative Adversarial Network [34]. This experiment allows us to explore whether we can build robust prototypes indirectly by expanding the sample of the support set. As we can see, $S \Rightarrow V$, *i.e.*, yellow line, gets relatively low performance in all settings. The experiments indicate that producing the prototype by expanding then averaging the support set is not as effective as directly generating class prototypes.

5. Conclusion

To unleash the potential of semantics in FSL, we propose the Semantic Evolution to automatically generate high-quality semantics from simple class names. Then, we design a simple Semantic Alignment Network to transform the concatenated visual features and high-quality semantics features into robust class prototypes for classification. The experimental results show that with high-quality semantics, the basic network can easily achieve greater performance compared to other state-of-the-art methods.

Acknowledgement. This work was supported by 2023YFF1204901, NSFC-62076172, 2021DQ02-0903, 2023YFG0116.

References

[1] Marcin Andrychowicz, Misha Denil, Sergio Gomez, Matthew W Hoffman, David Pfau, Tom Schaul, Brendan Shillingford, and Nando De Freitas. Learning to learn by gradient descent by gradient descent. In *NeurIPS*, 2016. 2

[2] Peyman Bateni, Raghav Goyal, Vaden Masrani, Frank Wood, and Leonid Sigal. Improved few-shot visual classification. In *CVPR*, pages 14493–14502, 2020. 2

[3] Irving Biederman. Recognition-by-components: a theory of human image understanding. *Psychol Rev*, 94(2):115, 1987. 1

[4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *NeurIPS*, 33:1877–1901, 2020. 3

[5] Wentao Chen, Chenyang Si, Zhang Zhang, Liang Wang, Zilei Wang, and Tieniu Tan. Semantic prompt for few-shot image recognition. In *CVPR*, pages 23581–23591, 2023. 1, 2, 3, 5, 6

[6] Yinbo Chen, Zhuang Liu, Huijuan Xu, Trevor Darrell, and Xiaolong Wang. Meta-baseline: Exploring simple meta-learning for few-shot learning. In *ICCV*, pages 9062–9071, 2021. 5

[7] Hao Cheng, Siyuan Yang, Joey Tianyi Zhou, Lanqing Guo, and Bihan Wen. Frequency guidance matters in few-shot learning. In *ICCV*, pages 11814–11824, 2023. 1, 5

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. Ieee, 2009. 1, 5

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 5

[10] Carl Doersch, Ankush Gupta, and Andrew Zisserman. Crosstransformers: spatially-aware few-shot transfer. In *NeurIPS*, pages 21981–21993, 2020. 2

[11] Bowen Dong, Pan Zhou, Shuicheng Yan, and Wangmeng Zuo. Self-promoted supervision for few-shot transformer. In *ECCV*, pages 329–347. Springer, 2022. 1, 5, 6

[12] Thomas Elsken, Benedikt Staffler, Jan Hendrik Metzen, and Frank Hutter. Meta-learning of neural architectures for few-shot learning. In *CVPR*, pages 12365–12375, 2020. 2

[13] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *Int. Conf. Mach. Learn.*, pages 1126–1135. PMLR, 2017. 1, 2, 5

[14] Yuqian Fu, Yu Xie, Yanwei Fu, and Yu-Gang Jiang. Styleadv: Meta style adversarial training for cross-domain few-shot learning. In *CVPR*, pages 24575–24584, 2023. 6

[15] Spyros Gidaris, Andrei Bursuc, Nikos Komodakis, Patrick Pérez, and Matthieu Cord. Boosting few-shot visual learning with self-supervision. In *ICCV*, pages 8059–8068, 2019. 2

[16] Fusheng Hao, Fengxiang He, Liu Liu, Fuxiang Wu, Dacheng Tao, and Jun Cheng. Class-aware patch embedding adaptation for few-shot image classification. In *ICCV*, pages 18905–18915, 2023. 1

[17] John A Hartigan and Manchek A Wong. Algorithm as 136: A k-means clustering algorithm. *J R Stat Soc Ser C Appl Stat*, 28(1):100–108, 1979. 7

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 5

[19] Markus Hiller, Rongkai Ma, Mehrtash Harandi, and Tom Drummond. Rethinking generalization in few-shot classification. *NeurIPS*, 35:3582–3595, 2022. 1, 2, 5, 6

[20] Ruibing Hou, Hong Chang, Bingpeng Ma, Shiguang Shan, and Xilin Chen. Cross attention network for few-shot classification. In *NeurIPS*, 2019. 5

[21] Yanxu Hu and Andy J Ma. Adversarial feature augmentation for cross-domain few-shot classification. In *ECCV*, pages 20–37. Springer, 2022. 6

[22] Muhammad Abdullah Jamal and Guo-Jun Qi. Task agnostic meta-learning for few-shot learning. In *CVPR*, pages 11719–11727, 2019. 2

[23] Jaekyeom Kim, Hyoungseok Kim, and Gunhee Kim. Model-agnostic boundary-adversarial sampling for test-time generalization in few-shot learning. In *ECCV*, pages 599–617. Springer, 2020. 6

[24] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[25] Gregory Koch, Richard Zemel, Ruslan Salakhutdinov, et al. Siamese neural networks for one-shot image recognition. In *ICML Deep Learn. workshop*. Lille, 2015. 1, 2

[26] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 5

[27] Kwonjoon Lee, Subhransu Maji, Avinash Ravichandran, and Stefano Soatto. Meta-learning with differentiable convex optimization. In *CVPR*, pages 10657–10665, 2019. 5, 6

[28] Aoxue Li, Weiran Huang, Xu Lan, Jiashi Feng, Zhenguo Li, and Liwei Wang. Boosting few-shot learning with adaptive margin loss. In *CVPR*, pages 12576–12584, 2020. 1, 2, 5

[29] Hongyang Li, David Eigen, Samuel Dodge, Matthew Zeiler, and Xiaogang Wang. Finding task-relevant features for few-shot learning by category traversal. In *CVPR*, pages 1–10, 2019. 1, 5

[30] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, pages 10012–10022, 2021. 5

[31] Andrew L Maas, Awni Y Hannun, Andrew Y Ng, et al. Rectifier nonlinearities improve neural network acoustic models. In *Proc. icml*, page 3. Atlanta, Georgia, USA, 2013. 5

[32] Puneet Mangla, Nupur Kumari, Abhishek Sinha, Mayank Singh, Balaji Krishnamurthy, and Vineeth N Balasubramanian. Charting the right manifold: Manifold mixup for few-shot learning. In *IEEE Win. Conf. Appl. Comput. Vis.*, pages 2218–2227, 2020. 1, 2, 6

[33] George A Miller. Wordnet: a lexical database for english. *Commun. ACM*, 38(11):39–41, 1995. 3

[34] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014. 8

[35] Nikhil Mishra, Mostafa Rohaninejad, Xi Chen, and Pieter Abbeel. A simple neural attentive meta-learner. *arXiv preprint arXiv:1707.03141*, 2017. 2

[36] OpenAI. Chatgpt. `https://chat.openai.com`, 2023. 5

[37] Boris Oreshkin, Pau Rodríguez López, and Alexandre Lacoste. Tadam: Task dependent adaptive metric for improved few-shot learning. In *NeurIPS*, 2018. 1, 5, 6

[38] Zhimao Peng, Zechao Li, Junge Zhang, Yan Li, Guo-Jun Qi, and Jinhui Tang. Few-shot image recognition with knowledge transfer. In *ICCV*, pages 441–449, 2019. 1, 5

[39] 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 *Int. Conf. Mach. Learn.*, pages 8748–8763. PMLR, 2021. 3, 5

[40] Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. In *ICLR*, 2017. 2

[41] 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. 5

[42] Andrei A Rusu, Dushyant Rao, Jakub Sygnowski, Oriol Vinyals, Razvan Pascanu, Simon Osindero, and Raia Hadsell. Meta-learning with latent embedding optimization. *arXiv preprint arXiv:1807.05960*, 2018. 2

[43] Victor Garcia Satorras and Joan Bruna Estrach. Few-shot learning with graph neural networks. In *ICLR*, 2018. 6

[44] Edgar Schonfeld, Sayna Ebrahimi, Samarth Sinha, Trevor Darrell, and Zeynep Akata. Generalized zero-and few-shot learning via aligned variational autoencoders. In *CVPR*, pages 8247–8255, 2019. 1

[45] Hongyang Gao Siyuan Sun. Meta-adam: A meta-learned adaptive optimizer with momentum for few-shot learning. In *NeurIPS*, 2023. 5, 6

[46] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *NeurIPS*, 2017. 1, 2, 3, 4, 5, 6

[47] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. In *NeurIPS*, 2015. 2

[48] 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 *CVPR*, pages 1199–1208, 2018. 2

[49] 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 *ECCV*, pages 266–282. Springer, 2020. 1, 2, 5, 6

[50] Pavel Tokmakov, Yu-Xiong Wang, and Martial Hebert. Learning compositional representations for few-shot recognition. In *ICCV*, pages 6372–6381, 2019. 1

[51] Hung-Yu Tseng, Hsin-Ying Lee, Jia-Bin Huang, and Ming-Hsuan Yang. Cross-domain few-shot classification via learned feature-wise transformation. In *ICLR*, 2020. 5, 6

[52] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *J Mach Learn Res*, 9(11), 2008. 7

[53] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. In *NeurIPS*, 2016. 1, 2, 5

[54] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 5

[55] Haoqing Wang and Zhi-Hong Deng. Cross-domain few-shot classification via adversarial task augmentation. In *IJCAI*, pages 1075–1081. International Joint Conferences on Artificial Intelligence Organization, 2021. 6

[56] Xin Wang, Fisher Yu, Ruth Wang, Trevor Darrell, and Joseph E Gonzalez. Tafe-net: Task-aware feature embeddings for low shot learning. In *CVPR*, pages 1831–1840, 2019. 1

[57] Chen Xing, Negar Rostamzadeh, Boris Oreshkin, and Pedro O O Pinheiro. Adaptive cross-modal few-shot learning. In *NeurIPS*, 2019. 1, 2, 4, 5

[58] Jingyi Xu and Hieu Le. Generating representative samples for few-shot classification. In *CVPR*, pages 9003–9013, 2022. 2, 3, 5

[59] Kun Yan, Zied Bouraoui, Ping Wang, Shoaib Jameel, and Steven Schockaert. Aligning visual prototypes with bert embeddings for few-shot learning. In *ICMR*, pages 367–375, 2021. 1, 5

[60] Han-Jia Ye, Hexiang Hu, De-Chuan Zhan, and Fei Sha. Few-shot learning via embedding adaptation with set-to-set functions. In *CVPR*, pages 8808–8817, 2020. 1, 5

[61] Baoquan Zhang, Xutao Li, Yunming Ye, Zhichao Huang, and Lisai Zhang. Prototype completion with primitive knowledge for few-shot learning. In *CVPR*, pages 3754–3762, 2021. 1, 2

[62] Chi Zhang, Yujun Cai, Guosheng Lin, and Chunhua Shen. Deepemd: Differentiable earth mover's distance for few-shot learning. *IEEE TPAMI*, 2022. 2

[63] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6):1452–1464, 2017. 5

[64] Fei Zhou, Peng Wang, Lei Zhang, Wei Wei, and Yanning Zhang. Revisiting prototypical network for cross domain few-shot learning. In *CVPR*, pages 20061–20070, 2023. 6

[65] Luisa Zintgraf, Kyriacos Shiarli, Vitaly Kurin, Katja Hofmann, and Shimon Whiteson. Fast context adaptation via meta-learning. In *Int. Conf. Mach. Learn.*, pages 7693–7702. PMLR, 2019. 2