Semantic-based Selection, Synthesis, and Supervision for Few-shot Learning

Jinda Lu
University of Science and Technology of China
Hefei, Anhui, China
lujd@mail.ustc.edu.cn

Shuo Wang*
University of Science and Technology of China
Hefei, Anhui, China
shuowang.edu@gmail.com

Xinyu Zhang
University of Science and Technology of China
Hefei, Anhui, China
zhangxy21@mail.ustc.edu.cn

Yanbin Hao
University of Science and Technology of China
Hefei, Anhui, China
haoyanbin@hotmail.com

Xiangnan He*
University of Science and Technology of China
Hefei, Anhui, China
xiangnanhe@gmail.com

ABSTRACT

Few-shot learning (FSL) is designed to explore the distribution of novel categories from a few samples. It is a challenging task since the classifier is usually susceptible to over-fitting when learning from limited training samples. To alleviate this phenomenon, a common solution is to achieve more training samples using a generic generation strategy in visual space. However, there are some limitations to this solution. It is because a feature extractor trained on base samples (known knowledge) tends to focus on the textures and structures of the objects it learns, which is inadequate for describing novel samples. To solve these issues, we introduce semantics and propose a **S**emantic-based **S**election, **S**ynthesis, and **S**upervision (**4S**) method, where semantics provide more diverse and informative supervision for recognizing novel objects. Specifically, we first utilize semantic knowledge to explore the correlation of categories in the textual space and select base categories related to the given novel category. This process can improve the efficiency of subsequent operations (synthesis and supervision). Then, we analyze the semantic knowledge to hallucinate the training samples by selectively synthesizing the contents from base and support samples. This operation not only increases the number of training samples but also takes advantage of the contents of the base categories to enhance the description of support samples. Finally, we also employ semantic knowledge as both soft and hard supervision to enrich the supervision for the fine-tuning procedure. Empirical studies on four FSL benchmarks demonstrate the effectiveness of 4S.

CCS CONCEPTS

• **Computing methodologies** → **Transfer learning**.

*Corresponding authors are Shuo Wang and Xiangnan He.

MM '23, October 29–November 3, 2023, Ottawa, ON, Canada.

ACM ISBN 979-8-4007-0108-5/23/10…$15.00
https://doi.org/10.1145/3581783.3611784

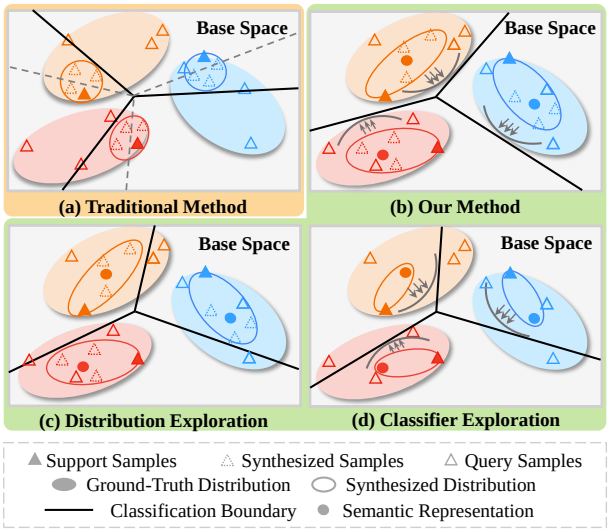

Figure 1: Given support samples and query samples from several novel categories (in different colors), the performance of classifiers trained with different strategies: (a) traditional FSL methods and generative methods, (b) our method (4S). And the different exploration strategies of our method: (c) distribution exploration, and (d) classifier exploration.

KEYWORDS

Data Synthesis, Semantic Supervision, Few-Shot Learning

ACM Reference Format:
Jinda Lu, Shuo Wang, Xinyu Zhang, Yanbin Hao, and Xiangnan He. 2023. Semantic-based Selection, Synthesis, and Supervision for Few-shot Learning. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23), October 29–November 3, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3581783.3611784

1 INTRODUCTION

In recent years, convolutional neural networks (CNNs) have demonstrated remarkable capabilities in various tasks [24, 25, 37, 38]. However, such data-driven networks require an enormous amount of

labeled training data to ensure their performance and capacity. And the process of collecting and annotating data is time-consuming and expensive. In contrast, humans can quickly recognize new objects with just a few examples owing to the immense quantity of prior knowledge that they have amassed. Therefore, few-shot learning (FSL) task is proposed to imitate this human ability [26, 40, 41, 53].

For this imitation, a popular solution is using a CNN trained on the base categories to extract the global features of novel objects directly [11, 43]. It aims to yield a transferable feature representation (textures and structures) to describe the novel categories. And then using these features to achieve classification. However, it is insufficient to represent the novel samples since their global features are insufficient to describe their category's distribution with the limited samples [40]. Thus, the few-shot classifier is usually susceptible to the over-fitting phenomenon. To alleviate this problem, recent methods devoted to designing different feature synthesis strategies and applying them in visual space to enrich training samples for classifier training [20, 28, 52]. However, these insufficient feature representations influence the synthesis procedures in two aspects: (1) the synthesized samples are finitely distributed around the given training samples, and (2) the contents that the CNN does not perceive will also be ignored in the synthesis process. Therefore, limited samples with insufficient description exacerbate the difficulty in perceiving novel distribution. As shown in Figure 1(a), given several support samples (solid triangles) and query samples (hollow triangles), the synthesized samples (dotted triangles) of traditional methods are around the support samples. And the performance of the classifiers (black lines) is slightly better than that of the baseline (dotted gray lines, which over-fit the support samples). Furthermore, we can find that the given support samples and the synthesized samples are distributed at the boundary of the actual distribution of the novel categories, which also shows that the trained CNN tends to focus on the base knowledge it learns (gray regions) and overlooks the contents of the novel objects. It also biases the classifier training.

To reduce the impact of insufficient feature description, recent training strategies have introduced semantic knowledge to constrain the classifier [9, 44, 58]. It is because textual knowledge provides potential contents which are already familiar with salient features of novel categories and helps the classifier capture the contents not in visual features [40]. Inspired by these strategies, we analyze semantic knowledge to explore its gains for the few-shot learning task, and propose a Semantic-based Selection, Synthesis, and Supervision method. In this method, we focus on two core issues: how to use semantics to (1) control the synthesized samples to explore the potential distribution of the novel categories, namely distribution exploration, and (2) control the classification process to help the classifier perceive the contents that are ignored in visual features, namely classifier exploration.

For the first issue, we devise semantic-based selection and synthesis strategies. Specifically, we first explore the potential knowledge from base categories (base knowledge) to describe the novel categories by calculating the correlation between these different categories, where semantics help us to filter out irrelevant information present in the base knowledge, and subsequently select related base samples from the corresponding visual space for further synthesis operations. Then, we design a semantic-based discriminator to select contents from both the support and the base features for synthesizing new training features, where these features for synthesis are prepared in advance by the pre-trained backbone. In this synthesis procedure, we aim to select the contents that may exist in the novel categories but have not been captured by the pre-trained backbone, where these contents can be approximated by using the contents from base categories. These strategies serve as a precise supplement to the support samples by leveraging base knowledge and semantic knowledge. Meanwhile, we can maximize the description of the novel categories in the data space. As shown in Figure 1(c), the semantically synthesized features complement the support features. They effectively extend the data space and prompt the classifier to concentrate on the enlarged data space formed by both the semantically synthesized and the support features.

For the second issue, we design semantic-based supervision training strategies in three aspects. First, we use the semantic-based discriminator to assign soft labels for the synthesized samples, which effectively expands the label space and aids in the development of a robust classifier. Second, we combine soft labels and hard labels to constrain the learning process of the visual classifier, where these labels help the visual classifier focus on different contents from the samples. Finally, we integrate textual information as hard supervision and propose a semantic-supervised classifier. This classifier takes both textual information and training samples as input while using textual information as supervision to direct the classifier's learning process for visual samples. As shown in Figure 1(d), semantic supervision guides the classifier to move toward semantic points. For instance, in the data space, all semantic representations are positioned in the clockwise direction relative to the support samples. This encourages the classifier to shift clockwise, leading to the formation of flexible classification boundaries.

Our method couples the aforementioned training strategies into one framework to train a powerful classifier. As depicted in Figure 1(b), benefiting from the aforementioned strategies, it not only expands the data space of novel categories through semantic selection and synthesis but also achieves flexible classification boundaries by semantic supervision. In summary, the main contributions of our method are threefold:

(1) We explore the effect of semantic knowledge in capturing the potential distribution of the novel categories and select related categories and contents from potential base knowledge to supplement the support samples, which helps the classifier to expand the perception of novel categories.
(2) We explore the benefit of semantic knowledge in the classifier learning process and introduce semantic supervision as both soft and hard supervision in training procedures to construct flexible classification boundaries.
(3) We align the synthesis and supervision within a unified framework. And our method outperforms existing approaches and achieves state-of-the-art performances on four popular FSL datasets under three settings, which boosts the recognition results of novel categories with limited samples.

2 RELATED WORK

In this section, we first briefly introduce common solutions for FSL and related knowledge exploration strategies, and then we list

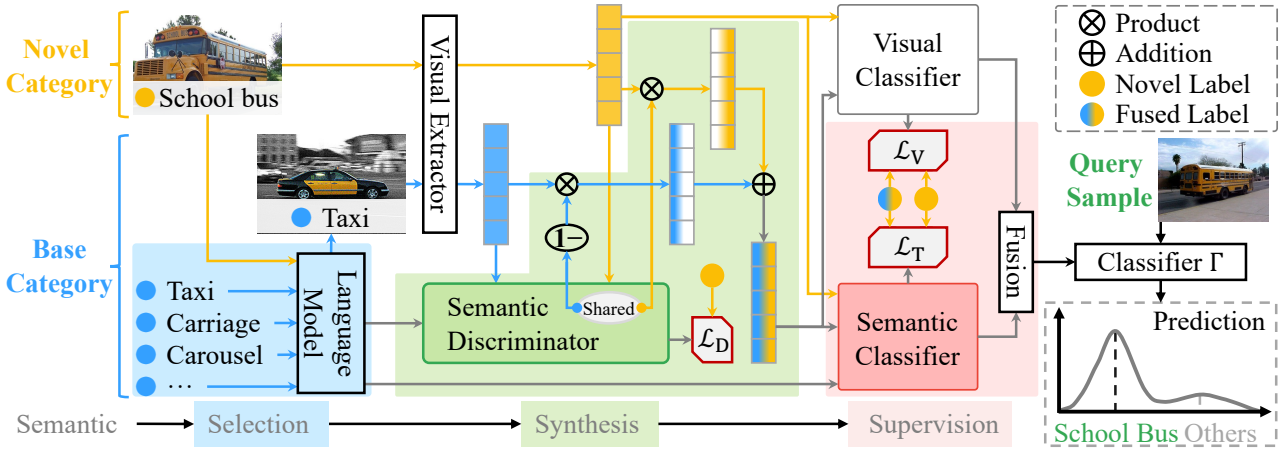

Figure 2: An overview of our semantic-based selection, synthesis, and supervision method.

recent data synthesis methods in FSL. Finally, we enumerate the differences between our methods and those of related methods.

2.1 Traditional Methods

The general solution to FSL is developing a robust feature extractor (backbone) on base data and then designing adaptation strategies to recognize novel objects [40, 41]. Fueled by the explosion of deep learning [23, 39, 42], Many deep-learning-based methods are proposed [22, 26]. Specifically, researchers have explored two different branches. One branch involves meta-learning. It trains a meta-learner on numerous FSL tasks (with base categories) to make the meta-learner learn how to solve FSL tasks, thus it can be easily adapted to new FSL tasks (with novel categories). Meta-learning methods can be divided into two groups: (1) Metric-based meta-learning, it aims at training a network that is capable of bringing samples from the same category closer together while pushing samples from different categories farther apart by mapping the given samples into feature space [31, 33, 35]. (2) Optimization-based meta-learning, it aims at training a network that is capable of generalizing to new FSL tasks with a small number of parameter updates by designing optimization strategies to learn a better initialization point or update direction [17, 29]. Another branch involves transfer learning; it trains a backbone on known (base) categories, aiming to produce a transferable feature representation (textures and structures) to describe novel categories. Specifically, [4, 32, 34] leverage features extracted by a pre-trained backbone to fine-tune a classifier for FSL task and show promising classification results.

2.2 Knowledge Exploration

Knowledge-based FSL methods have attracted attention for their strong performance. It is because knowledge from external sources can provide more information for novel categories. These methods focus primarily on the use of semantic information, such as exploring semantic relationships between different categories, bridging the gap between visual and semantic modalities, and using semantic information to enrich the supervision for classifier training. Specifically, Li *et al.* propose an adaptive margin loss using semantic similarity, which improves the supervision of the classifier and facilitates the separation of similar classes [18]. The work in [40] employs textual knowledge to generate soft labels, which serve as supervision information to help learn a robust classifier. Chen *et al.* introduce a knowledge graph transfer network (KGTN) that explores semantic relationships between categories using a graph and transfers similar classifier information from base categories to novel categories through these semantic relationships [3]. The work in [41] proposes a multi-directional knowledge transfer (MDKT), where knowledge from different modalities is fused through a bidirectional knowledge connection. Peng *et al.* put forward a knowledge transfer network (KTN) that incorporates a graph convolution network (GCN) to merge the visual and textual spaces, which assists in recognizing novel samples [26]. AM3 [49] combines textual features with visual features using an adaptive modality mixture mechanism, which helps to improve the performance of metric-based methods. Zhang *et al.* explore more detailed attributes or part information to complete visual prototypes [55].

2.3 Data Synthesis

Recent work has demonstrated that data synthesis can bring steady and effective improvement for few-shot learning. Synthesis-based FSL methods aim to design effective strategies to generate more training data to complete the novel categories. Thus, generative models are the direct choice for data synthesis, the work in [7, 20, 43] train generative adversarial networks(GAN) on the base data, and use the optimized generative model to synthesize novel samples. Similarly, Schwartz *et al.* train an auto-encoder (AE) on the base data and synthesize new samples for the novel category [30]. Since learning generative patterns from base data using generative models is time-consuming and expensive, [11, 12, 28, 52] propose to directly use base data for synthesis. Hariharan *et al.* learn variations between different base data from the same category and transfer variation patterns for novel data generation [11], The work in [12] assigns labels from the novel domain for base data by pseudo-labeling. Yang *et al.* use the statistics from base categories to calibrate the novel categories and sample features from the calibrated categories to complete the novel descriptions [52]. The work in [10, 54] can also be viewed as synthesis methods, Yue *et al.* split the novel features

into different equal parts, which enlarges the training samples, and they fuse the classification result of different parts as the final prediction[54]. The work in [10] takes few-shot learning as a binary ranking classification problem. It groups the novel samples into image triplets and takes the image triplets as input, which also enlarges the training samples.

Based on the aforementioned analysis, our method explores the semantic knowledge and synthesizes training samples for few-shot learning. The differences between the above methods and ours can be summarized in two aspects: (1) we utilize textual relations to filter base knowledge for further usage instead of directly using whole base knowledge, which avoids introducing irrelevant noise during the synthesis procedure. (2) We do not need to design complex exploration networks, such as the knowledge graph [3, 26], in the exploration of knowledge, which simplifies the calculation process and increases the efficiency of classifier training.

3 APPROACH

An overview of our framework is shown in Figure 2. Specifically, given a support sample and its label from the novel category, we first represent them into features by a pre-trained visual model and language model, respectively. Then we use the correlation between semantics to select the base sample close to the given novel category for the feature synthesis procedure. Meanwhile, we use semantics to distinguish the contents between different samples for synthesis. Finally, we enrich the supervision from semantic knowledge to train both semantic-supervised and visual classifiers. For inference, we fuse these classifiers to predict a given query sample. To illustrate the details of our method in this section, we first briefly revisit the preliminaries of the few-shot learning task. Then, we illustrate the operations of Semantic-Based Selection, Synthesis, and Supervision (4S). Finally, we describe the training and inference procedures.

3.1 Preliminaries

Given a dataset for the few-shot learning task, we follow the common solution to divide it into three parts: training set $\mathcal{D}_{\text{train}}$, support set $\mathcal{D}_{\text{support}}$, and testing set $\mathcal{D}_{\text{test}}$. Specifically, the training set $\mathcal{D}_{\text{train}}$ is used to pre-train the backbone, it has large-scale training samples (*e.g.*, about hundreds of samples in one category), and the categories of these samples are denoted as C_{base}. The support set $\mathcal{D}_{\text{support}}$ and the testing set $\mathcal{D}_{\text{test}}$ have the same categories, called C_{novel}, which are disjoint with that of the training set C_{base}. The goal of few-shot learning is to learn an image classification model that generalizes well to the N-way-K-shot task. Training samples for the N-way-K-shot task are sampled from $\mathcal{D}_{\text{support}}$ and the testing samples belong to $\mathcal{D}_{\text{test}}$, and a N-way-K-shot task identifies N novel categories, and each category has K support samples.

3.2 Selection

Our method is based on knowledge exploration. Therefore, we first introduce knowledge processing and illustrate the operation of semantic-based selection. Given several labels from base categories and novel categories, we use an available Word2Vec embedding method [19] to represent these labels into features as $\boldsymbol{t}^{C_{\text{base}}}$ and $\boldsymbol{t}^{C_{\text{novel}}}$, respectively. Then, we calculate the relations between the novel and base categories from these label features and select the base category for the synthesizing procedure. Specifically, given the label feature of n^{th} novel category as $\boldsymbol{t}_n^{C_{\text{novel}}}$, the relation with the b^{th} base category can be calculated by similarity scores:

$$r_{(n,b)} = \frac{\langle \boldsymbol{t}_n^{C_{\text{novel}}}, \boldsymbol{t}_b^{C_{\text{base}}} \rangle}{\|\boldsymbol{t}_n^{C_{\text{novel}}}\|_2 \cdot \|\boldsymbol{t}_b^{C_{\text{base}}}\|_2}, \tag{1}$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product between two features. We calculate the relations between n^{th} novel category and whole base categories as $\boldsymbol{R}_n = \{r_{(n,b)}\}_{b=1,\dots,|C_{\text{base}}|}$ and sort the relations to select the β^{th} related base category for synthesizing. For convenience, we denote the selected base category and its label feature as C_β and $\boldsymbol{t}_\beta$, respectively. Meanwhile, we use the superscript β to conveniently represent samples or visual features of the selected base category. The category selection brings semantic relevance into visual space, thus the irrelevant information in base categories can be filtered out for subsequent synthesis procedure.

3.3 Synthesis

For synthesis procedure, we first express the samples into features. Specifically, given a support sample of n^{th} novel category as $\boldsymbol{I}_n$ and its related base category C_β. We first randomly select a base sample from the related base category C_β as $\boldsymbol{I}_\beta$. Then we use a CNN Φ to extract the features of support sample $\boldsymbol{I}_n$ and selected base sample $\boldsymbol{I}_\beta$ as $\boldsymbol{f}_n$ and $\boldsymbol{f}_\beta$, respectively, where Φ is pre-trained on the base data $\mathcal{D}_{\text{train}}$. In the extraction procedure, we remove the last prediction layer of CNN Φ. Then, we design a semantic discriminator Ω to precisely discriminate the contents of the novel feature $\boldsymbol{f}_n$ and the selected base feature $\boldsymbol{f}_\beta$. The discriminator employs the textual features of the base categories $\boldsymbol{T}_B = \{\boldsymbol{t}_b\}_{b=1,\dots,|C_{\text{base}}|}$ and the novel categories $\boldsymbol{T}_N = \{\boldsymbol{t}_n\}_{n=1,\dots,|C_{\text{novel}}|}$ as the category descriptor, which can be formalized as $\boldsymbol{T} = \boldsymbol{T}_B \cup \boldsymbol{T}_N$. To connect the textual descriptor to the visual space, we use a projection head $\boldsymbol{W}_d \in \mathbb{R}^{d_v \times d_t}$, and the discriminator Ω can be formalized as follows:

$$\Omega = \boldsymbol{T}\boldsymbol{W}_d^\top. \tag{2}$$

Then, to better explore the potential knowledge of semantics, we use the novel feature to optimize the discriminator Ω in advance:

$$\mathcal{L}_{\text{D}} = \text{CE}((\boldsymbol{f}_n \Omega^\top), l_n), \tag{3}$$

where CE is the cross-entropy loss, and l_n is the one-hot label of the n^{th} novel category.

After this pre-optimization stage, we denote the optimized discriminator as $\bar{\Omega}$ and use it to distinguish the support feature $\boldsymbol{f}_n$ and the selected base feature $\boldsymbol{f}_\beta$ by classifying them into the same novel category $\boldsymbol{l}_n$, and the prediction scores are denoted as the content screening rate. Then, we calculate the fusion ratio for feature synthesis. Specifically, denoted content screening rate as γ_n and γ_β, respectively, the fusion ratio can be calculated as:

$$\alpha_{(n,\beta)} = \gamma_n - \gamma_\beta, \tag{4}$$

where $\alpha_{(n,\beta)}$ helps the synthesis procedure focus on the important contents of the novel category.

Finally, based on the fusion ratio $\alpha_{(n,\beta)}$, we fuse the support feature and the selected base feature to synthesize a new one. Specifically, we design two different fusion operations and set $\alpha_{(n,\beta)}$ as the weighting ratio and the thresholding ratio, respectively. For

the weighting synthesis, the synthesis of the new feature $\boldsymbol{f}_s$ can be described as:

$$\boldsymbol{f}_s = \alpha_{(n,\beta)}\boldsymbol{f}_\beta + (1-\alpha_{(n,\beta)})\boldsymbol{f}_n. \tag{5}$$

For the thresholding synthesis, we first sample a uniformly distributed random vector $\boldsymbol{V} \in (0,1)$ which is the same size as the feature $\boldsymbol{f}_n$, and then put $\alpha_{(n,\beta)}$ as the thresholding on the random vector $\boldsymbol{V}$. Thus, the mask vector $\bar{\boldsymbol{V}}$ can be formalized as:

$$\bar{v}_i = \begin{cases} 0, \text{if } v_i \geqslant \alpha_{(n,\beta)}. \\ 1, \text{if } v_i < \alpha_{(n,\beta)}. \end{cases} \tag{6}$$

where v_i is the i^{th} element of the random vector $\boldsymbol{V}$, and $\bar{v}_i$ denotes the i^{th} element of the mask vector $\bar{\boldsymbol{V}}$. Then, the synthesized feature of this operation can be calculated as:

$$\boldsymbol{f}_s = \bar{\boldsymbol{V}} \otimes \boldsymbol{f}_\beta + (1-\bar{\boldsymbol{V}}) \otimes \boldsymbol{f}_n, \tag{7}$$

where $\otimes$ is the hardmard product. The synthesis strategy provides precise supplementary contents for the given support sample, thus the potential space of the novel objects can be expanded.

3.4 Supervision

We design two format supervisions except for hard labels to train the classifier: (1) semantic supervision, and (2) soft labels. For semantic supervision, we design a semantic-supervised classifier, which transfers the knowledge containing textual descriptions of both the novel categories and the base categories from the textual space to the visual space and supervises the visual feature learning. Specifically, given the category descriptor $\boldsymbol{T}$, which contains the description of both base categories and novel categories in the textual space, the transfer operation can be described as follows:

$$\hat{\boldsymbol{T}} = \text{LeakyReLU}(\boldsymbol{TA}), \tag{8}$$

where $\boldsymbol{A}$ is a matrix that connects the textual space with the visual space, $\boldsymbol{A} \in \mathbb{R}^{d_t \times d_v}$, and LeakyReLU is the activation function. $\hat{\boldsymbol{T}}$ is the category descriptor transferred into the visual space, we then define the semantic-supervised classifier $\boldsymbol{\Gamma}_t$ as:

$$\boldsymbol{\Gamma}_{\boldsymbol{t}} = \hat{\boldsymbol{T}}\boldsymbol{W}_t, \tag{9}$$

where $\boldsymbol{W}_t$ is the visual learning matrix and $\boldsymbol{W}_t \in \mathbb{R}^{d_v \times d_v}$. Furthermore, we keep the vanilla visual classifier to boost the performance, and the visual classifier $\boldsymbol{\Gamma}_{\boldsymbol{v}}$ is described as:

$$\boldsymbol{\Gamma}_{\boldsymbol{v}} = \boldsymbol{W}_v, \tag{10}$$

where $\boldsymbol{W}_v \in \mathbb{R}^{(|C_{\text{base}}|+|C_{\text{novel}}|)\times d_v}$. Compared to the visual classifier that directly learns the mapping of the visual feature to the novel category, the learning process of the semantic classifier is constrained by the category descriptor $\hat{\boldsymbol{T}}$, and thus semantic supervision prevents the classifier from overfitting the visual feature and enlarges the classification boundaries.

To further identify the novel contents of the synthesized features, we assign soft labels for the synthesized features based on the feature synthesis. Specifically, given a synthesized feature $\boldsymbol{f}_s$, we use the fusion ratio $\alpha_{(n,\beta)}$ in Eq. (4) to control the label synthesis:

$$\boldsymbol{l}_s = (1-\alpha_{(n,\beta)})\boldsymbol{l}_n + \alpha_{(n,\beta)}\boldsymbol{l}_\beta, \tag{11}$$

where $\boldsymbol{l}_n$ and $\boldsymbol{l}_\beta$ is the label of n^{th} novel category and β^{th} base category, respectively, and $\boldsymbol{l}_s$ is a multi-label for synthesized feature $\boldsymbol{f}_s$. Both the original feature-label pairs $(\boldsymbol{f}_n, \boldsymbol{l}_n)$ and the synthesized feature-label pairs $(\boldsymbol{f}_s, \boldsymbol{l}_s)$ are handled for classifier learning to learn a powerful classifier. In summary, the selection and synthesis strategies enrich the data space of the novel category, and the supervision strategies enrich the supervision of the learning process. Our method benefits from these two strategies. As shown in Figure 2, given the query sample, these help the classifier to concentrate more on the novel category and make a precise prediction.

3.5 Training and Inference

For the training stage, given K support samples from the n^{th} novel category, we formalize the feature-label pairs as $\mathcal{N} = \left\{\boldsymbol{f}_n^i, \boldsymbol{l}_n^i\right\}_{i=1}^K$. We first optimize the discriminator Ω by $\mathcal{N}$, then we select the related base category and synthesize samples based on the optimized discriminator $\bar{\Omega}$, the synthesized feature-label pairs are denoted as $\mathcal{S} = \left\{\boldsymbol{f}_s^i, \boldsymbol{l}_s^i\right\}_{i=1}^K$. In our method, these synthesized samples are combined with support samples for classifier training. Specifically, we use the cross-entropy(CE) loss with the support feature-label pairs $\mathcal{N}$ and the multi-label cross-entropy(MCE) with the synthesized feature-label pairs $\mathcal{S}$ to train both the visual classifier $\boldsymbol{\Gamma}_v$ and the semantic-supervised classifier $\boldsymbol{\Gamma}_t$.

The training loss of the visual classifier $\boldsymbol{\Gamma}_v$ can be formalized by the following equation:

$$\mathcal{L}_{\text{V}} = \frac{1}{2K}\sum_{i=1}^{K} \text{CE}((\boldsymbol{f}_n^i\boldsymbol{\Gamma}_v^\top), \boldsymbol{l}_n^i) + \text{MCE}((\boldsymbol{f}_s^i\boldsymbol{\Gamma}_v^\top), \boldsymbol{l}_s^i). \tag{12}$$

The training loss of the semantic-supervised classifier $\boldsymbol{\Gamma}_t$ is:

$$\mathcal{L}_{\text{T}} = \frac{1}{2K}\sum_{i=1}^{K} \text{CE}((\boldsymbol{f}_n^i\boldsymbol{\Gamma}_t^\top), \boldsymbol{l}_n^i) + \text{MCE}((\boldsymbol{f}_s^i\boldsymbol{\Gamma}_t^\top), \boldsymbol{l}_s^i). \tag{13}$$

Thus, the total loss of training can be defined as follows:

$$\mathcal{L} = \mu_1\mathcal{L}_{\text{D}} + \mu_2\mathcal{L}_{\text{V}} + \mu_3\mathcal{L}_{\text{T}}, \tag{14}$$

where μ_1, μ_2, μ_3 are weighting factors.

For the inference stage, to boost the generalization and classification performance of the classifier, we define the classifier as the fusion of the semantic-supervised classifier and the visual classifier, which can be formalized as:

$$\boldsymbol{\Gamma} = \lambda\boldsymbol{\Gamma}_v + (1-\lambda)\boldsymbol{\Gamma}_t, \tag{15}$$

where λ is a hype parameter of the fusion operation, we use the classifier $\boldsymbol{\Gamma}$ to classify the query features into a specific novel category, and the query features are extracted from the last average pooling layer of the pre-trained CNN $\boldsymbol{\Phi}$.

4 EXPERIMENTS

In this section, we conduct experiments to evaluate the performance of our proposed method. We first introduce the experimental settings, then illustrate the ablation studies, and finally, we show the comparison with the state-of-the-art methods. Our experiments are intended to address the following research questions (**RQ**s):

RQ1: What are the influences of different semantic selection (β)?
RQ2: What are the effects of selected synthesis?
RQ3: What are the influences of fusion ratios between the visual and the semantic-supervised classifier (λ)?
RQ4: What are the influences of different feature fusion strategies

on the synthesis procedure?
RQ5: How does the performance comparison between our method and the state-of-the-art methods?

4.1 Experimental Settings

Datasets. We evaluate our method on four benchmark datasets, *i.e.*, Mini-ImageNet [35], CIFAR-FS [2], Caltech-UCSD Birds-200-2011 (CUB) [36], and ImageNet-FS [11]. Specifically, Mini-ImageNet and CIFAR-FS are the common few-shot classification datasets. Mini-ImageNet consists of 100 categories, and each category has 600 images. It is divided into three parts: 64 base categories for training, 16 novel categories for validation, and 20 novel categories for testing. CIFAR-FS is randomly sampled from CIFAR100 [15]. It has 100 categories with 60000 images, and it is divided into three parts: 64 base categories for training, 16 novel categories for validation, and 20 novel categories for testing. CUB is a fine-grained dataset, it contains 200 bird categories with 11788 images. Following the split strategy in [4], we divide this dataset into three parts, where 100 base categories for training, 50 novel categories for validation, 50 novel categories for testing. ImageNet-FS [11] is a large-scale few-shot classification dataset. it has 389 base categories and 611 novel categories, where 300 novel categories are used for validation, and the remaining 311 novel categories are used for testing.
Evaluation. For Mini-ImageNet, CIFAR-FS and CUB, the performance is evaluated on several N-way-K-shot classification tasks. In each task, N novel categories are sampled first, then K samples in each of the N categories are sampled for training, and the other 15 samples in each of the N categories are sampled for testing. To report the results, we sample 600 such tasks and report average accuracies with 95% confidence interval over all the tasks. In our experiments, $N = 5$ and $K = 1, 5$. For ImageNet-FS, the performance is evaluated under three settings with $K = 1$, 2, and 5 support samples per category. And we report the accuracy by recognizing the samples from the 311 testing novel categories. More details of the settings can be found in [11].
Implementation Details. We use the features extracted from the pre-trained CNNs for synthesis, then we use these original and synthesized features to train the classifier $\mathbf{\Gamma}$. The classifier is trained with loss $\mathcal{L}$ in Eq. (14) ($\mu_1 = 1, \mu_2 = 1, \mu_3 = 1$) for 800 epochs. We use the Adam optimizer [14] with the starting learning rate of 0.001 and the weight decay of 0.0001. The learning rate is divided by 10 every 200 epochs.

4.2 Ablation Studies

In the ablation study, we use Mini-ImageNet to evaluate the effectiveness of different parts of our method. Specifically, we set the 16 validation categories as the novel categories and the 64 base categories for feature selection and synthesis. All features are extracted with the available pre-trained backbone (ResNet-12) [5], all textual features are extracted with the available Word2Vec method [19], All experiments in this ablation study are conducted in 5-way-K-shot settings, where $K = 1$ or $K = 5$.

4.2.1 The effectiveness of the selection. **(RQ1)**

In this ablation, we conduct experiments under the $K = 1$ and $K = 5$ settings for all β-th related categories to evaluate the effectiveness of feature selection from different base categories. We mix two features according to Eq. 5 but replace the fusion ratio with a random number in this synthesis procedure. The results are shown in Figure 3. The ratio β ranges from 0 to 64 (whole base categories = 64), where $\beta = 0$ denotes that the visual classifier is trained only with novel features and $\beta = 64$ means that we select features from the least related base category. For comparison, we also plot the performance trained with whole base categories ("All"), which means that we randomly select features from all base categories and randomly synthesize them with the given novel features. The gray area refers to the difference between two curves. We can find that: (1) randomly mix novel samples with randomly selected base samples in feature space can slightly improve the performance of the classifier (compared "All" with $\beta = 0$), and (2) only a few categories above the "All" operation especially the $K = 5$ settings, and a smaller β brings positive effects, a larger β brings negative effects. This further indicates that irrelevant base knowledge has no effect on enhancing the description of novel categories, and shows the necessity that the synthesis operation should be constrained by the category relations. (3) The best performance is achieved with $\beta = 1$ in both the experimental settings $K = 1$ and $K = 5$, and in the $K = 1$ settings we achieve more than 2% accuracy improvement; thus we select $\beta = 1$ for subsequent operations.

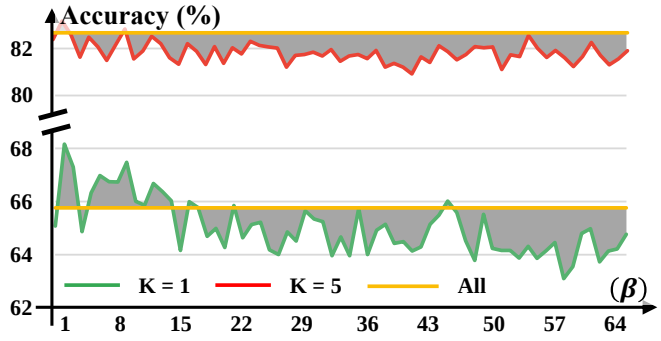

Figure 3: The accuracies(%) of the visual classifier trained with both novel features and randomly synthesized features, where the selected base category for synthesis is the β-th related category for the given novel category.

Table 1: The accuracies (%) of the visual classifier trained with different synthesis features.

Method	$K = 1$	$K = 5$
Baseline	65.08 ± 0.81%	82.45 ± 0.53%
$\mathbf{\Gamma}_v$ w/o $\mathcal{L}_\mathrm{D}$	68.17 ± 0.77%	82.76 ± 0.54%
$\mathbf{\Gamma}_v$ w/ $\mathcal{L}_\mathrm{D}$	**69.41** ± 0.76%	**83.07** ± 0.53%

4.2.2 The effects of the semantic discriminator. **(RQ2)**

To validate the effectiveness of the semantic discriminator in our synthesis procedures, we train two visual classifiers with different features. In this ablation, the samples used for synthesis are selected from the 1-st related base category. The results are shown in Table 1, where $\mathbf{\Gamma}_v$ is the visual classifier, "w/o $\mathcal{L}_\mathrm{D}$" means classifier trained with both novel features and randomly synthesized features, "w/

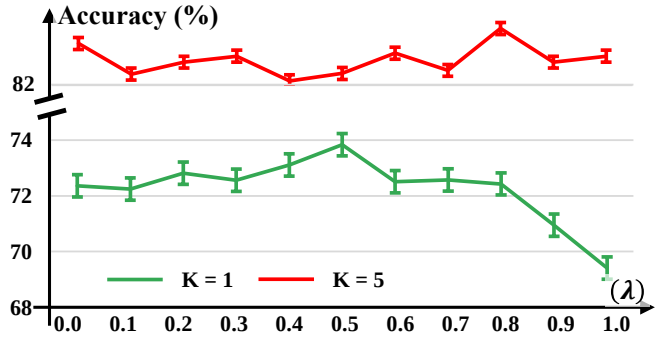

Figure 4: The accuracies(%) of the final classifier with different fusion ratios λ between the visual classifier and the semantic-supervised classifier.

Table 2: The accuracies(%) of the classifier trained with different feature fusion strategies.

Method	$K = 1$	$K = 5$
Baseline	65.08 ± 0.81%	82.45 ± 0.53%
Γ w/ $\mathcal{T}$	73.04 ± 0.77%	82.85 ± 0.54%
Γ w/ $\mathcal{W}$	**73.84** ± 0.75%	**83.28** ± 0.53%

$\mathcal{L}_D$" means classifier trained with both novel features and synthesized features by Eq. 5. For convenient comparison, we keep the "Baseline", which means the visual classifier trained with only novel features. Compared to the "Baseline", we can see that the selection and synthesis strategies make a significant improvement for the classifier. For example, it achieves 4.3% accuracy improvement in $K = 1$ settings. Meanwhile, we can also see that content selection in related features further improves the performance of the classifier. Specifically, compared to random fusion, the method with our semantic-based discriminator achieves 1.2% accuracy improvement in the $K = 1$ settings, which is significant.

4.2.3 The influences of the semantic supervision. **(RQ3)**

In this ablation, we conduct experiments with different fusion ratios in Eq. 15 to evaluate the influences of semantic supervision. In these experiments, we set the fusion ratio range from 0.0 to 1.0, where $\lambda = 0.0$ means that the final prediction is only determined by the semantic-supervised classifier, and $\lambda = 1.0$ implies that the final prediction is only influenced by the visual classifier. The total results are shown in Figure 4. We can see that: (1) semantic supervision makes a significant improvement for the classifier and achieves a performance improvement of more than 2% over the visual classifier. (2) Fusing the semantic-supervised classifier with the visual classifier further improves the performance of the classifier. The fused classifier achieves more than 4% performance improvements in the $K = 1$ settings. In detail, we can find that the classifier achieves the best performance with $\lambda = 0.5$ and $\lambda = 0.8$ for $K = 1$ and $K = 5$ experimental settings. Thus, we set $\lambda = 0.5$ for the $K = 1$ setting and $\lambda = 0.8$ for the $K = 5$ setting, respectively.

4.2.4 The influences of different feature fusion strategies. **(RQ4)**

Table 3: The accuracies (%) by different methods on the testing categories from Mini-ImageNet [35].

Method with ResNet-12	Mini-ImageNet	
	$K = 1$	$K = 5$
PN [31]	60.37 ± 0.83%	78.02 ± 0.57%
AM3 [49]	65.30 ± 0.49%	78.10 ± 0.36%
MetaOptNet [17]	62.64 ± 0.61%	78.63 ± 0.46%
DMF [50]	67.76 ± 0.46%	82.71 ± 0.31%
MixtFSL [1]	63.98 ± 0.79%	82.04 ± 0.49%
RENet [13]	67.60 ± 0.44%	82.58 ± 0.30%
DeepBDC [48]	67.34 ± 0.43%	84.46 ± 0.28%
Meta-Baseline [5]	63.17 ± 0.23%	79.26 ± 0.17%
Meta-Baseline + Ours	**71.27** ± 0.66%	**82.87** ± 0.54%
DeepEMD [56]	65.91 ± 0.82%	82.41 ± 0.56%
DeepEMD + Ours	**71.26** ± 0.70%	**83.50** ± 0.54%
FRN [46]	66.45 ± 0.19%	82.83 ± 0.13%
FRN + Ours	**72.66** ± 0.73%	**84.46** ± 0.50%
BML [59]	67.04 ± 0.63%	83.63 ± 0.29%
BML + Ours	**74.53** ± 0.68%	**85.78** ± 0.49%
FEAT [53]	66.78 ± 0.20%	82.05 ± 0.14%
FEAT + Ours	**72.64** ± 0.70%	**84.73** ± 0.50%
Method with WRN28-10		
PPA [27]	59.60 ± 0.41%	73.74 ± 0.19%
LEO [29]	61.76 ± 0.08%	77.59 ± 0.12%
wDAE-GNN [8]	61.07 ± 0.15%	76.75 ± 0.11%
IFSL [54]	64.12 ± 0.44%	80.97 ± 0.31%
FEAT [53]	65.10 ± 0.20%	81.11 ± 0.14%
FEAT + Ours	**71.92** ± 0.71%	**84.01** ± 0.51%
LRDC [52]	68.57 ± 0.55%	82.88 ± 0.42%
LRDC + Ours	**75.85** ± 0.69%	**87.37** ± 0.48%

In this ablation, we evaluate the influences of different feature fusion strategies (Eq.5 and Eq.7). The experimental results are shown in Table 2, where "Baseline" means the visual classifier trained with only novel features, "Γ" denotes the final classifier with the optimal fusion ratio between two classifiers, " w/ $\mathcal{W}$" means the features synthesized by Eq.5, and " w/ $\mathcal{T}$" means that the features synthesized by Eq.7. Specifically, we can see that both strategies achieve competitive results in $K = 1$ and $K = 5$ settings, which proves the effectiveness of our hypothesis. Meanwhile, we can see that the accuracy of the classifier trained with the weighting strategy is slightly better than that of the thresholding. Thus, we choose the weighting strategy for feature fusion for the best performance.

4.3 Comparison with other methods (RQ5)

We compare the performance of our method with the state-of-the-art methods under three few-shot classification settings, followings are detailed descriptions of the experimental results.

Traditional few-shot classification. For Mini-ImageNet, the results are shown in Table 3. The compared methods include PN [31], AM3 [49], MetaOptNet [17], DMF [50], MixtFSL [1], RENet [13], DeepBDC [48], PPA [27], LEO [29], wDAE-GNN [8], IFSL [54], and we apply our method on six popular methods, which are Meta-Baseline [5], DeepEMD [56], FRN [46], BML [59], FEAT [53],

Table 4: The accuracies (%) by different methods on the testing categories from CIFAR-FS [2].

Method	CIFAR-FS	
	$K = 1$	$K = 5$
ConstellationNet [51]	75.40 ± 0.20%	86.80 ± 0.20%
Meta Navigator [57]	74.63 ± 0.91%	86.45 ± 0.59%
NCA [16]	72.49 ± 0.12%	85.15 ± 0.09%
RENet [13]	74.51 ± 0.46%	86.60 ± 0.32%
TPMN [47]	75.50 ± 0.90%	87.20 ± 0.60%
DeepEMD [56]	74.58 ± 0.29%	86.92 ± 0.41%
DeepEMD + Ours	**79.66** ± 0.69%	**89.14** ± 0.51%
BML [59]	73.45 ± 0.47%	88.04 ± 0.33%
BML + Ours	**74.50** ± 0.84%	**88.76** ± 0.53%

Table 5: The accuracies (%) by different methods on the testing categories from CUB [36].

Method	CUB	
	$K = 1$	$K = 5$
PN [31]	72.99 ± 0.88%	86.64 ± 0.51%
CovNet [45]	80.76 ± 0.42%	92.05 ± 0.20%
ADM [21]	79.31 ± 0.43%	90.69 ± 0.21%
AFHN [20]	70.53 ± 1.01%	83.95 ± 0.63%
RENet [13]	74.51 ± 0.46%	86.60 ± 0.32%
LRDC [52]	79.56 ± 0.87%	90.67 ± 0.35%
LRDC + Ours	**83.86** ± 0.70%	**93.37** ± 0.34%
FRN [46]	83.55 ± 0.19%	92.92 ± 0.10%
FRN + Ours	**87.64** ± 0.64%	**94.24** ± 0.38%

LRDC [52]. From Table 3, we can see that our method significantly outperforms other methods, regardless of the methods and the pre-trained backbones. For the $K = 1$ setting, we achieve a performance improvement of more than 5% for all the methods applied, and the most improvement is more than 8% with Meta-Baseline. For ResNet-12, we achieve 74.5% accuracy with 7% improvement on BML, and for WRN28-10, we achieve 75.8% accuracy with 7% improvement on LRDC. For the $K = 5$ setting, we improve all applied methods with more than 1% improvement. For ResNet-12, we achieve 85.7% accuracy with 2% improvement on BML, for WRN28-10, we gain 87.3% accuracy with 4% improvement on LRDC. Both the improvements and performances are significant in few-shot learning, which further proves the effectiveness of our method.

For CIFAR-FS, we use the ResNet-12 as the feature extractor, and the classification accuracies are shown in Table 4. The compared methods include ConstellationNet[51], Meta Nevigator [57], NCA [16], RENet [13], TPMN [47], and we apply our method to DeepEMD [56] and BML [59]. Our method outperforms the applied methods, and it gains 5% improvement with DeepEMD for the $K = 1$ setting, 2% improvement with DeepEMD for the $K = 5$ setting. For $K = 1$, we achieve 79.6% accuracy with 5% improvement, and for $K = 5$, we achieve 89.1% accuracy with 2% improvement on DeepEMD.

Fine-grained few-shot classification. We evaluate our method with the fine-grained CUB dataset, and the results are shown in Table 5. We compare our method with PN [31], CovNet [45], ADM [21], AFHN [20], RENet [13], and we apply our method on LRDC [52] and FRN [46]. From Table 5, we can see that our method outperforms all applied methods with 4% improvement for $K = 1$ and 1% improvement for $K = 5$. For $K = 1$, we achieve 87.6% accuracy with 4% improvement on FRN, and for $K = 5$, we achieve 94.2% accuracy with 1% improvement on FRN.

Table 6: Top-5 accuracies (%) by different methods on the testing categories from ImageNet-FS [11].

Method	ResNet-10			ResNet-50		
	$K = 1$	$K = 2$	$K = 5$	$K = 1$	$K = 2$	$K = 5$
PN [31]	39.3	54.4	66.3	49.5	59.9	70.1
LR-H [11]	40.7	50.8	62.0	53.5	63.5	72.7
SGM-H [11]	44.3	56.0	69.7	52.8	64.4	77.3
IDeMe [6]	51.0	60.9	70.4	60.1	69.6	77.4
KTN [26]	54.7	61.7	70.4	61.9	68.7	76.4
MDKT [41]	55.2	63.2	**70.8**	62.6	70.1	**77.6**
Ours	**58.1**	**63.4**	69.4	**66.0**	**71.5**	76.8

Large-scale few-shot classification. We evaluate our method with the large-scale few-shot classification dataset ImageNet-FS, and we conduct experiments under three settings with ResNet-10 and ResNet-50 backbone, the performance is illustrated in Table 6. The compared methods include PN [31], SGM [11], IDeMe-Net [6], KTN [26], MDKT [41]. Our method outperforms others under $K = 1$, 2 settings. And we achieve a significant improvement in the $K = 1$ settings, with 3% improvement for both ResNet-10 and ResNet-50. Note that the competitive methods employ complex architectures, for example, [6, 11] use generative models to synthesize more training features, and [26, 41] incorporate graph models such as GNN, GCN *et al.* to explore semantic relations. But the architecture of our method is simple and we only synthesize 1 training sample for the given novel sample in each training step.

5 CONCLUSION

In this paper, we discuss the role of semantic knowledge in few-shot learning by introducing the Semantic-based Selection, Synthesis, and Supervision (4S) method. This innovative approach aligns synthesis and supervision within a unified framework, addressing the challenges posed by few-shot learning. The 4S method boasts several distinct features: (1) Our semantic-based selection and synthesis strategy not only expands the data space for novel categories but also capitalizes on the full potential of the base data. (2) Our semantic-based supervision strategy effectively constructs adaptive and flexible boundaries for novel categories. (3) To substantiate the effectiveness of our method, we conduct extensive experiments on four different FSL datasets. The results showcase the remarkable success of our approach, particularly in the context of 1-shot tasks.

We noticed that our method has certain fluctuations in different ablation studies. In the future, we will introduce more accurate feature representation, semantic representation, or robust models in our method to help the classifier obtain more stable gains.

ACKNOWLEDGMENTS

This work was supported by the National Key Research and Development Program of China under Grant 2020YFB1406703.

REFERENCES

[1] Arman Afrasiyabi, Jean-François Lalonde, and Christian Gagné. 2021. Mixture-based feature space learning for few-shot image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 9041–9051.

[2] L Bertinetto, J Henriques, P Torr, and A Vedaldi. 2019. Meta-learning with differentiable closed-form solvers. In *International Conference on Learning Representations (ICLR), 2019*. International Conference on Learning Representations.

[3] Riquan Chen, Tianshui Chen, Xiaolu Hui, Hefeng Wu, Guanbin Li, and Liang Lin. 2020. Knowledge graph transfer network for few-shot recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 10575–10582.

[4] Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Frank Wang, and Jia-Bin Huang. 2019. A Closer Look at Few-shot Classification. In *International Conference on Learning Representations*.

[5] Yinbo Chen, Zhuang Liu, Huijuan Xu, Trevor Darrell, and Xiaolong Wang. 2021. Meta-baseline: Exploring simple meta-learning for few-shot learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 9062–9071.

[6] Zitian Chen, Yanwei Fu, Yu-Xiong Wang, Lin Ma, Wei Liu, and Martial Hebert. 2019. Image deformation meta-networks for one-shot learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 8680–8689.

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

[8] Spyros Gidaris and Nikos Komodakis. 2019. Generating classification weights with gnn denoising autoencoders for few-shot learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 21–30.

[9] Jingcai Guo and Song Guo. 2020. A novel perspective to zero-shot learning: Towards an alignment of manifold structures via semantic feature expansion. *IEEE Transactions on Multimedia* 23 (2020), 524–537.

[10] Qianyu Guo, Gong Haotong, Xujun Wei, Yanwei Fu, Yizhou Yu, Wenqiang Zhang, and Weifeng Ge. 2023. RankDNN: Learning to Rank for Few-Shot Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 728–736.

[11] Bharath Hariharan and Ross Girshick. 2017. Low-shot visual recognition by shrinking and hallucinating features. In *Proceedings of the IEEE international conference on computer vision*. 3018–3027.

[12] Yiren Jian and Lorenzo Torresani. 2022. Label hallucination for few-shot classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 7005–7014.

[13] Dahyun Kang, Heeseung Kwon, Juhong Min, and Minsu Cho. 2021. Relational embedding for few-shot classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 8822–8833.

[14] Diederik P. Kingma and Jimmy Ba. 2015. Adam: a method for stochastic optimization. In *International Conference on Learning Representations*.

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

[16] Steinar Laenen and Luca Bertinetto. 2021. On episodes, prototypical networks, and few-shot learning. *Advances in Neural Information Processing Systems* 34 (2021), 24581–24592.

[17] Kwonjoon Lee, Subhransu Maji, Avinash Ravichandran, and Stefano Soatto. 2019. Meta-learning with differentiable convex optimization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10657–10665.

[18] Aoxue Li, Weiran Huang, Xu Lan, Jiashi Feng, Zhenguo Li, and Liwei Wang. 2020. Boosting few-shot learning with adaptive margin loss. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 12576–12584.

[19] Aoxue Li, Tiange Luo, Zhiwu Lu, Tao Xiang, and Liwei Wang. 2019. Large-scale few-shot learning: Knowledge transfer with class hierarchy. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*. 7212–7220.

[20] Kai Li, Yulun Zhang, Kunpeng Li, and Yun Fu. 2020. Adversarial feature hallucination networks for few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13470–13479.

[21] Wenbin Li, Lei Wang, Jing Huo, Yinghuan Shi, Yang Gao, and Jiebo Luo. 2021. Asymmetric distribution measure for few-shot learning. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*. 2957–2963.

[22] Chen Liu, Yanwei Fu, Chengming Xu, Siqian Yang, Jilin Li, Chengjie Wang, and Li Zhang. 2021. Learning a few-shot embedding model with contrastive learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 8635–8643.

[23] Zhenguang Liu, Runyang Feng, Haoming Chen, Shuang Wu, Yixing Gao, Yunjun Gao, and Xiang Wang. 2022. Temporal feature alignment and mutual information maximization for video-based human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 11006–11016.

[24] Zhenguang Liu, Shuang Wu, Shuyuan Jin, Shouling Ji, Qi Liu, Shijian Lu, and Li Cheng. 2022. Investigating pose representations and motion contexts modeling for 3D motion prediction. *IEEE transactions on pattern analysis and machine intelligence* 45, 1 (2022), 681–697.

[25] Zhenguang Liu, Shuang Wu, Shuyuan Jin, Qi Liu, Shijian Lu, Roger Zimmermann, and Li Cheng. 2019. Towards Natural and Accurate Future Motion Prediction of Humans and Animals. In *CVPR*. 10004–10012.

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

[27] Siyuan Qiao, Chenxi Liu, Wei Shen, and Alan L Yuille. 2018. Few-shot image recognition by predicting parameters from activations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7229–7238.

[28] Aniket Roy, Anshul Shah, Ketul Shah, Prithviraj Dhar, Anoop Cherian, and Rama Chellappa. 2022. FeLMi: few shot learning with hard mixup. *Advances in Neural Information Processing Systems* 35 (2022), 24474–24486.

[29] Andrei A Rusu, Dushyant Rao, Jakub Sygnowski, Oriol Vinyals, Razvan Pascanu, Simon Osindero, and Raia Hadsell. 2019. Meta-Learning with Latent Embedding Optimization. In *International Conference on Learning Representations*.

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

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

[32] Qianru Sun, Yaoyao Liu, Tat-Seng Chua, and Bernt Schiele. 2019. Meta-transfer learning for few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 403–412.

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

[34] Yonglong Tian, Yue Wang, Dilip Krishnan, Joshua B Tenenbaum, and Phillip Isola. 2020. 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*. Springer, 266–282.

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

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

[37] Shuo Wang, Huixia Ben, Yanbin Hao, Xiangnan He, and Meng Wang. 2023. Boosting Hyperspectral Image Classification with Dual Hierarchical Learning. *ACM Transactions on Multimedia Computing, Communications and Applications* 19, 1 (2023), 1–19.

[38] Shuo Wang, Dan Guo, Xin Xu, Li Zhuo, and Meng Wang. 2019. Cross-modality retrieval by joint correlation learning. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 15, 2s (2019), 1–16.

[39] Shuo Wang, Dan Guo, Wen-gang Zhou, Zheng-Jun Zha, and Meng Wang. 2018. Connectionist temporal fusion for sign language translation. In *Proceedings of the 26th ACM international conference on Multimedia*. 1483–1491.

[40] Shuo Wang, Jun Yue, Jianzhuang Liu, Qi Tian, and Meng Wang. 2020. Large-scale few-shot learning via multi-modal knowledge discovery. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part X 16*. Springer, 718–734.

[41] Shuo Wang, Xinyu Zhang, Yanbin Hao, Chengbing Wang, and Xiangnan He. 2022. Multi-directional Knowledge Transfer for Few-Shot Learning. In *Proceedings of the 30th ACM International Conference on Multimedia*. 3993–4002.

[42] Wenjie Wang, Xinyu Lin, Fuli Feng, Xiangnan He, and Tat-Seng Chua. 2023. Generative recommendation: Towards next-generation recommender paradigm. *arXiv preprint arXiv:2304.03516* (2023).

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

[44] Zhicai Wang, Yanbin Hao, Tingting Mu, Ouxiang Li, Shuo Wang, and Xiangnan He. 2023. Bi-directional Distribution Alignment for Transductive Zero-Shot Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 19893–19902.

[45] Davis Wertheimer and Bharath Hariharan. 2019. Few-shot learning with localization in realistic settings. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 6558–6567.

[46] Davis Wertheimer, Luming Tang, and Bharath Hariharan. 2021. Few-shot classification with feature map reconstruction networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8012–8021.

[47] Jiamin Wu, Tianzhu Zhang, Yongdong Zhang, and Feng Wu. 2021. Task-aware part mining network for few-shot learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 8433–8442.

[48] Jiangtao Xie, Fei Long, Jiaming Lv, Qilong Wang, and Peihua Li. 2022. Joint distribution matters: Deep brownian distance covariance for few-shot classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7972–7981.

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

[50] Chengming Xu, Yanwei Fu, Chen Liu, Chengjie Wang, Jilin Li, Feiyue Huang, Li Zhang, and Xiangyang Xue. 2021. Learning dynamic alignment via meta-filter for few-shot learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 5182–5191.

[51] Weijian Xu, Yifan Xu, Huaijin Wang, and Zhuowen Tu. 2021. Attentional constellation nets for few-shot learning. In *International Conference on Learning Representations*.

[52] Shuo Yang, Lu Liu, and Min Xu. 2021. Free Lunch for Few-shot Learning: Distribution Calibration. In *International Conference on Learning Representations*.

[53] Han-Jia Ye, Hexiang Hu, De-Chuan Zhan, and Fei Sha. 2020. Few-shot learning via embedding adaptation with set-to-set functions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 8808–8817.

[54] Zhongqi Yue, Hanwang Zhang, Qianru Sun, and Xian-Sheng Hua. 2020. Interventional few-shot learning. *Advances in neural information processing systems* 33 (2020), 2734–2746.

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

[56] Chi Zhang, Yujun Cai, Guosheng Lin, and Chunhua Shen. 2020. 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*. 12203–12213.

[57] Chi Zhang, Henghui Ding, Guosheng Lin, Ruibo Li, Changhu Wang, and Chunhua Shen. 2021. Meta navigator: Search for a good adaptation policy for few-shot learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 9435–9444.

[58] Li Zhang, Tao Xiang, and Shaogang Gong. 2017. Learning a deep embedding model for zero-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2021–2030.

[59] Ziqi Zhou, Xi Qiu, Jiangtao Xie, Jianan Wu, and Chi Zhang. 2021. Binocular mutual learning for improving few-shot classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 8402–8411.