Multi-Attention Based Visual-Semantic Interaction for Few-Shot Learning

Peng Zhao[1] , **Yin Wang**[1] , **Wei Wang**[2*] , **Jie Mu**[3] , **Huiting Liu**[1] ,
Cong Wang[2,4] and **Xiaochun Cao**[2]
[1]School of Computer Science and Technology, Anhui University
[2]School of Cyber Science and Technology, Shenzhen Campus of Sun Yat-sen University
[3]School of Data Science and Artificial Intelligence, Dongbei University of Finance and Economics
[4]Department of Computing, The Hong Kong Polytechnic University
{zhaopeng_ad,htliu}@ahu.edu.cn, {wyin5630,supercong94}@gmail.com,
{wangwei29,caoxiaochun}@mail.sysu.edu.cn, jiemu@dufe.edu.cn

Abstract

Few-Shot Learning (FSL) aims to train a model that can generalize to recognize new classes, with each new class having only very limited training samples. Since extracting discriminative features for new classes with few samples is challenging, existing FSL methods leverage visual and semantic prior knowledge to guide discriminative feature learning. However, for meta-learning purposes, the semantic knowledge of the query set is unavailable, so their features lack discriminability. To address this problem, we propose a novel Multi-Attention based Visual-Semantic Interaction (MAVSI) approach for FSL. Specifically, we utilize spatial and channel attention mechanisms to effectively select discriminative visual features for the support set based on its ground-truth semantics while using all the support set semantics for each query set sample. Then, a relation module with class prototypes of the support set is employed to supervise and select discriminative visual features for the query set. To further enhance the discriminability of the support set, we introduce a visual-semantic contrastive learning module to promote the similarity between visual features and their corresponding semantic features. Extensive experiments on four benchmark datasets demonstrate that our proposed MAVSI could outperform existing state-of-the-art FSL methods.

1 Introduction

With a large number of annotated data, deep learning methods have achieved tremendous success in various computer vision tasks [Chen *et al.*, 2024; Huang *et al.*, 2021; Huang *et al.*, 2022a; Huang *et al.*, 2022b; Huang *et al.*, 2023; Huang *et al.*, 2024], such as image recognition, object detection, *etc*. However, in many real-world scenarios, collecting a large number of labeled samples for each class can be impractical or expensive [Wu *et al.*, 2023a; Wu *et al.*, 2023b; Wu *et al.*, 2024; Wang *et al.*, 2023a; Wang *et al.*, 2022; Wang *et al.*, 2023b; Wang *et al.*, 2021]. Therefore, Few-Shot Learning (FSL), which requires only a small number of samples to recognize new classes, has been receiving increasing attention recently [Liu *et al.*, 2022; Xu *et al.*, 2023; Cao *et al.*, 2021]. Existing FSL approaches usually adopt the meta-learning strategy, which simulates FSL tasks using an episodic training approach on a fully annotated base dataset. The base dataset does not share any classes with the test dataset. Each episode consists of a labeled support set and an unlabeled query set, with the query set sharing the same label space as the support set.

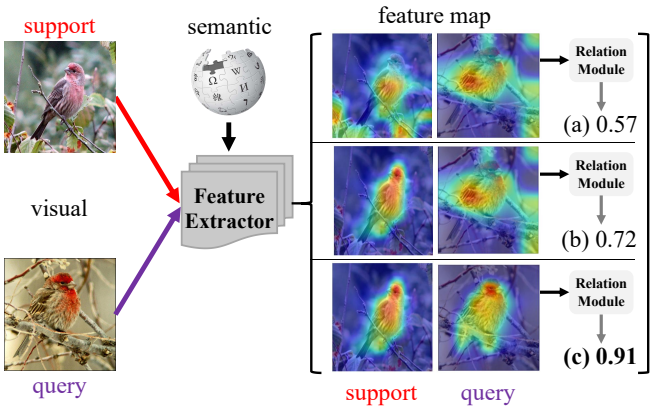

Figure 1: Visual-semantic interaction for FSL. (a) w/o semantic; (b) w/ semantic in support set; (c) w/ semantic in both sets.

It could be observed that humans can quickly recognize a novel category, as they could utilize semantic prior knowledge to efficiently extract discriminative visual features from a small number of samples. Therefore, existing FSL methods have started using semantic knowledge to assist in learning discriminative visual features [Chen *et al.*, 2019; Xing *et al.*, 2019; Peng *et al.*, 2019]. Although these methods have achieved promising performance, the discriminability of visual features on the query set cannot be well guaranteed because semantic knowledge of the query set is unavailable. As shown in Figure 1, when semantic knowledge is not considered, the feature extractor may capture features of both the target of interest (American Redstart) and the back-

*Corresponding author: Wei Wang.

ground (branches). In the right part of Figure 1, the first row shows the case where neither the support set nor the query set considers semantic knowledge, the second row considers the semantic knowledge only on the support set, and the third row considers semantic knowledge on both sets. It can be observed that the similarity scores continuously increase when semantic knowledge is considered. Therefore, considering semantic knowledge in both sets is crucial for improving the performance of FSL.

To address the problem mentioned above, we propose a Multi-Attention based Visual-Semantic Interaction (MAVSI) approach for FSL. As shown in Figure 2, we first obtain the visual features of the support set, query set, and the semantic features of the support set, through the feature encoder and semantic encoder. Next, the proposed MAVSI employs Spatial Cross-Attention (SCAVSI) and Channel Attention (CAVSI) based Visual-Semantic Interaction modules to select discriminative visual features. As there is no semantic knowledge of the query set, we use all the semantic features from the support set for each query sample to generate visual features for the query set, with the number of features equal to the number of semantic features. Then, based on the support set, we calculate the class prototypes for each category and introduce a relation module to compute the relationship scores between each visual feature of the query sample and the corresponding class prototype. By promoting the relationship scores corresponding to the query sample's class label, we can extract discriminative visual features for the query set.

To further enhance the discriminability of the support set, we design a Visual-Semantic Contrastive Learning (VSCL) module. As shown in Figure 2, we perform a concatenation operation between the visual and semantic features in the support set. We calculate the visual-semantic similarity through a comparison network and aim to prompt the similarity between the visual features and their ground-truth semantic features. Our main contributions are summarized as follows,

- We propose a Multi-Attention based Visual-Semantic Interaction (MAVSI) approach for FSL, which contains two distinct modules, SCAVSI and CAVSI, and prompts discriminative visual features from both spatial and channel perspectives.
- We propose to utilize the semantic knowledge from the support set to guide the visual-semantic interaction on the query set with a relation module and introduce a VSCL module further to enhance the feature discriminability of the support set.
- Extensive experiments on four FSL benchmark datasets demonstrate that the proposed MAVSI outperforms state-of-the-art FSL approaches.

2 Related Work

FSL learns new concepts with only one or a few annotated samples and has gained extensive attention in recent years. Existing FSL methods mostly employ the meta-learning strategy, which aims to transfer knowledge from an auxiliary base dataset to new tasks. Meta-learning-based FSL methods can generally be categorized into the following three types: optimization-based methods [Finn *et al.*, 2017; Ravi and Larochelle, 2017; Rusu *et al.*, 2019; Baik *et al.*, 2020], metric-based methods [Vinyals *et al.*, 2016; Snell *et al.*, 2017], and data augmentation-based methods [Wang *et al.*, 2018; Zhang *et al.*, 2019; Wang *et al.*, 2020]. The proposed approach belongs to a metric-based FSL.

In recent years, semantic information has played a crucial role in tasks such as image-text matching [Ji *et al.*, 2021] and zero-shot learning [Zhao *et al.*, 2023]. Particularly, in scenarios with scarce visual samples, semantic information is a preferable choice to assist model training, and existing FSL methods that utilize semantic information could be summarized into three categories.

The combination of visual features and semantic features. For instance, Xing *et al.* propose a mechanism that can adaptively combine information from both modalities (visual and semantic) according to novel classes to be learned. Furthermore, Schwartz *et al.* combine visual features with multiple and richer semantic information such as category labels, attributes, and natural language descriptions. Peng *et al.* propose a semantic-visual mapping network to conduct knowledge inference for novel categories from base categories.

Employing semantic information to synthesize visual features, to mitigate the deficiencies in visual features. For example, Chen *et al.* propose directly synthesizing instance features by leveraging semantics using a novel auto-encoder network. Xu and Le propose using a conditional variational autoencoder model to generate visual samples based on semantic embeddings.

Utilizing semantic information to refine visual features, to enhance discriminability of visual features. For instance, Yang *et al.* use the semantic knowledge to guide the visual perception about what visual features should be paid attention to when distinguishing a category from the others. As such, the feature embedding of the novel class, even with few samples, can be more discriminative. Afham *et al.* force the prototypes to retain semantic information about the class description, improving their generalization to novel classes at inference. Xu *et al.* propose to highlight pivotal local visual information with attention mechanism and align the attentive map with semantic information to refine the extracted visual features. Li *et al.* develop a class-relevant additive margin loss and a task-relevant additive margin loss with semantic information to prompt discriminative visual features.

The proposed approach belongs to the third category. Given the effectiveness of attention mechanisms in cross-modal tasks, many FSL methods employ them to facilitate interactions between semantic and visual features, thereby enhancing the discriminability of visual features, such as [Yang *et al.*, 2023], [Yang *et al.*, 2022], [Xu *et al.*, 2022], and [Wang *et al.*, 2020]. However, these methods only consider interactions between semantic and visual features within the support set. In contrast, this paper simultaneously considers semantic-visual interactions in both the support and query sets. Moreover, we introduce a multi-attention mechanism that operates in spatial and channel dimensions to enhance discriminative visual features more effectively.

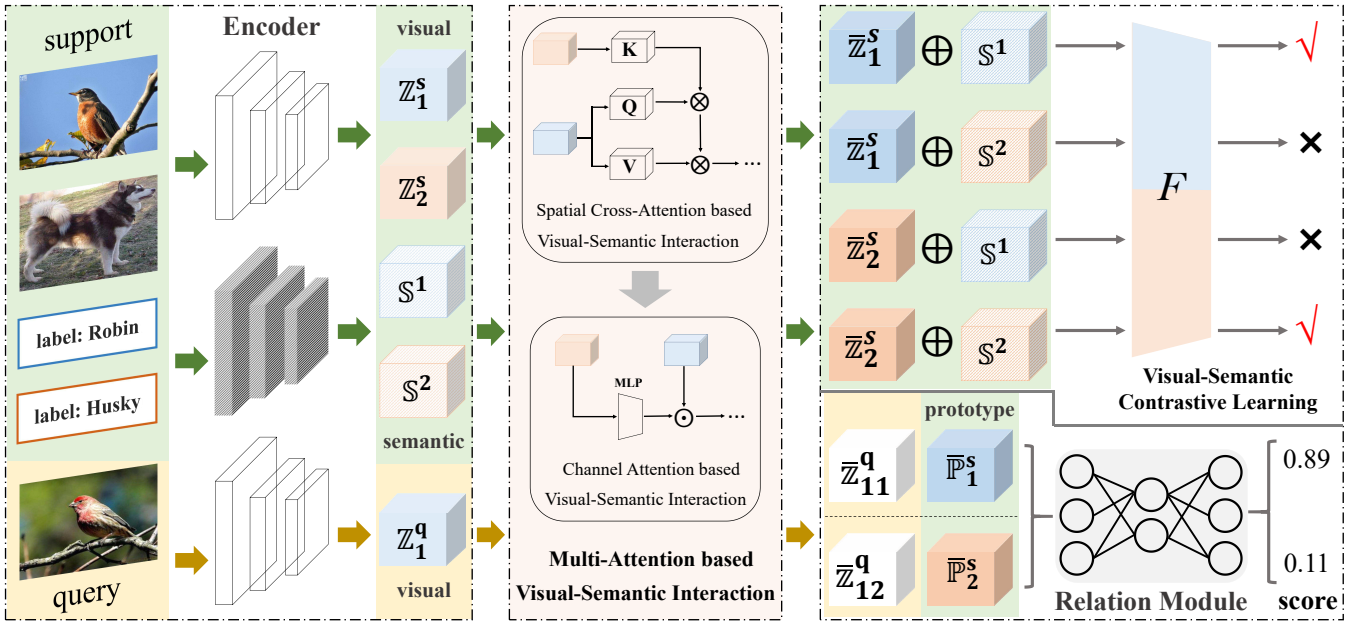

Figure 2: The whole pipeline of MAVSI. 1) We employ visual and semantic encoders to extract features from the support and query sets. 2) We utilize the MAVSI module, which consists of both spatial (SCAVSI) and channel (CAVSI) attention mechanisms, to select discriminative visual features. 3) We employ the relation module, utilizing class prototypes from the support set to guide the selection of discriminative visual features for the query set. 4) We further enhance the discriminability of the support set by introducing the VSCL module.

3 Proposed Approach

In this section, we first provide problem formulation for FSL. Then, we detail the proposed multi-attention based visual-semantic interaction and visual-semantic contrastive learning modules. Finally, we introduce the relation network and our overall training loss.

3.1 Problem Formulation

During the training phase, we generally have a base dataset for meta-learning purposes, and each class has abundant samples. Then, N novel classes do not overlap with the base dataset during testing. Additionally, each novel class has only K labeled samples (support set), resulting in a total of $N \times K$ samples, referred to as the N-way K-shot setting.

Following [Xing *et al.*, 2019], we adopt the episodic training with the base dataset, performed with episodes that simulate the testing setup. Generally, our model is trained in an N-way K-shot setting, where each episode consists of N classes from the meta-training set, divided into two parts: K labeled samples in the support set and several unlabeled samples in the query set. The support set contains $N \times K$ labeled samples with semantic knowledge.

In one episode, we have a support set $\mathbb{X}^s$ and a query set $\mathbb{X}^q$. Then, we use a Convolutional Neural Network (CNN) to embed them into a visual space: $\mathbb{Z} = \mathcal{G}_\theta(\mathbb{X}) \in \mathbf{R}^{C\times H\times W}$, where θ represents a learnable parameter set. Moreover, C, H, and W are the channel, height, and width, respectively. Notably, we can regard $r = H \times W$ as the spatial resolution of an image. Each C-dimensional vector corresponds to a position and represents a local region of an image, where some of these local regions may contain important features of the target object of interest.

3.2 Multi-Attention Based Visual-Semantic Interaction

Inspired by the self-attention mechanism, we propose a Multi-Attention based Visual-Semantic Interaction (MAVSI) approach that operates at both spatial and channel levels. With semantic knowledge, MAVSI could mitigate the negative impact of background noise by selecting discriminative visual features. Specifically, MAVSI consists of a Spatial Cross-Attention (SCAVSI) and a Channel Attention (CAVSI) based Visual-Semantic Interaction modules.

Spatial Cross-Attention Visual-Semantic Interaction

SCAVSI aims to explore the relationship between local regions and semantics with labels of support set. It selects local regions highly correlated with semantic knowledge, avoiding introducing irrelevant background noise at the spatial level. Due to the limited exploration of the relationship between semantics and local features by a single-head attention mechanism, we choose multi-head attention to find more highly correlated local features. For simplicity, we illustrate our proposed SCAVSI with only one attention head in Figure 3.

Let $\boldsymbol{s}_0^i \in \mathbf{R}^{d_0}$ denotes the semantic from class i by semantic extractor. We transform the semantic to the same dimension as the visual feature through a semantic encoder. Specifically, we first pass $\boldsymbol{s}_0^i$ through a Multilayer Perceptron (MLP) layer to obtain $\boldsymbol{s}_1^i = MLP(\boldsymbol{s}_0^i) \in \mathbf{R}^C$. Then, we duplicate $\boldsymbol{s}_1^i$ in the spatial dimension to obtain $\mathbb{S}^i = repeat(\boldsymbol{s}_1^i) \in \mathbf{R}^{C\times H\times W}$, where $repeat$ is a replication operation.

As shown in Figure 3, we use three 1×1 convolution kernels ω_1, ω_2, and ω_3 with learnable parameter sets to linearly

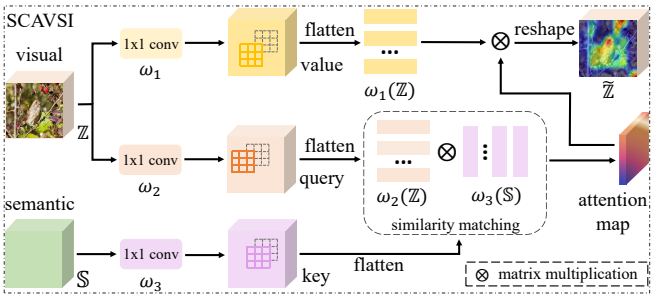

Figure 3: The proposed SCAVSI. We pass visual and semantic features through three 1×1 convolution kernels, followed by a series of flattening and replication operations. Next, we obtain an attention map through visual and semantic information interaction, from which the spatially corrected visual feature $\widetilde{\mathbb{Z}}$ is derived.

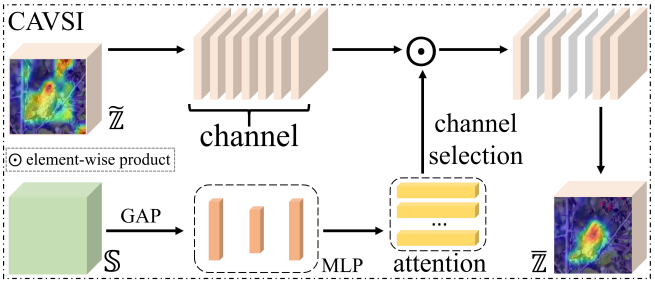

Figure 4: The proposed CAVSI. We pass the semantic feature through GAP and MLP, and perform channel selection on the spatially corrected visual feature to obtain the final modified visual feature of $\widetilde{\mathbb{Z}}$ in channel level.

transform the visual feature $\mathbb{Z}_j^{s,i}$ of the j-th sample from class i in support set, and the semantic $\mathbb{S}^i$ of class i. We denote the transformed semantic as a key $\omega_3(\mathbb{S}^i)$, and denote the transformed visual feature as a query $\omega_2(\mathbb{Z}_j^{s,i})$ and a value $\omega_1(\mathbb{Z}_j^{s,i})$. As mentioned before, each C-dimensional vector corresponds to a position and represents a local region of an image. Thus, we reshape them into the size of $\mathbf{R}^{C\times(HW)}$. Then, the transpose of $\omega_3(\mathbb{S}^i)$ is multiplied with $\omega_2(\mathbb{Z}_j^{s,i})$. This matrix multiplication calculates the similarity between the spatial feature of each semantic channel in $\omega_3(\mathbb{S}^i)$ and all visual channel features in $\omega_2(\mathbb{Z}_j^{s,i})$, establishing the relationship between semantic information and visual features in the spatial level.

Next, we apply the softmax function on the resulting matrix to perform spatial normalization, obtaining the visual-semantic attention map $\boldsymbol{A}_1^{s,i} \in \mathbf{R}^{HW\times HW}$:

$$\boldsymbol{A}_1^{s,i} = softmax(\omega_3(\mathbb{S}^i)^\top \cdot \omega_2(\mathbb{Z}_j^{s,i})/\sqrt{\gamma}), \tag{1}$$

where γ is a scaling factor.

After obtaining the semantic-visual attention map, the next step is to perform a matrix multiplication between the value $\omega_1(\mathbb{Z}_j^{s,i})$ and the transpose of the attention map. The resulting matrix is then reshaped back to its original size, *i.e.*, $\mathbf{R}^{C\times H\times W}$. Finally, the spatially modified feature is obtained by element-wise addition of the spatially corrected visual feature with the original input feature:

$$\widetilde{\mathbb{Z}}_j^{s,i} = \omega_1(\mathbb{Z}_j^{s,i}) \cdot \boldsymbol{A}_1^{s,i} + \mathbb{Z}_j^{s,i}, \tag{2}$$

where $\widetilde{\mathbb{Z}}_j^i$ denotes the modified features in spatial level. As we have no semantic knowledge of the query, we obtain its modified features with all semantics of the support set, with the number of features being equal to the number of semantics. Then, we use the relation module introduced in the following subsection, to select the modified feature with the highest score for each query sample.

Channel Attention Visual-Semantic Interaction

As some background noise may also affect discriminative information in the channel dimension, we propose a CAVSI module as a complement to SCAVSI, further enhancing the discriminative power of visual features in the channel level, as shown in Figure 4.

Given the spatially modified feature $\widetilde{\mathbb{Z}}_j^{s,i} \in \mathbf{R}^{C\times H\times W}$ of the j-th sample from i-th class and the semantic $\boldsymbol{s}^i$ for the i-th class, where we use the Global Average Pooling (GAP) layer to transform $\mathbb{S}^i \in \mathbf{R}^{C\times H\times W}$ to $\boldsymbol{s}^i \in \mathbf{R}^C$. Then, we use a two-layer MLP network as the channel attention generator to produce the attention weights $\boldsymbol{a}_2^i \in \mathbf{R}^C$. The final layer of the MLP is a sigmoid function, ensuring that $\boldsymbol{a}_2^i$ consists of values in the range [0,1]:

$$\boldsymbol{a}_2^i = \sigma(\boldsymbol{W}_2\sigma(\boldsymbol{W}_1\boldsymbol{s}^i + \boldsymbol{b}_1) + \boldsymbol{b}_2). \tag{3}$$

We duplicate $\boldsymbol{a}_2^i$ in the spatial dimension to enable it have the same dimension as the visual feature (*i.e.*, $\boldsymbol{A}_2^i$), and conduct element-wise product with the spatially modified feature $\widetilde{\mathbb{Z}}_j^{s,i}$. Then, we obtain modified features of $\overline{\mathbb{Z}}_j^{s,i}$ in channel level as the following equation,

$$\overline{\mathbb{Z}}_j^{s,i} = \boldsymbol{A}_2^i \odot \widetilde{\mathbb{Z}}_j^{s,i}, \tag{4}$$

where $\boldsymbol{W}_1$, $\boldsymbol{W}_2$, $\boldsymbol{b}_1$, and $\boldsymbol{b}_2$ are the parameters of the MLP network, and σ is the activation function sigmoid. Similar to the modified features at the spatial level, we could obtain modified features for each query sample at the channel level.

3.3 Visual-Semantic Contrastive Learning

To further enhance the feature discriminability of the support set, inspired by contrastive learning techniques [Badamdorj *et al.*, 2022], we propose a Visual-Semantic Contrastive Learning (VSCL) module. Since the visual and semantic features are not embedded in the same space, it is impossible to compute the similarity between them directly. Therefore, we introduce a comparison network F to deal with this problem, which aims to minimize the following equation,

$$\mathcal{L}_{ce}(\overline{\mathbb{Z}}^s, \mathbb{S}) = \frac{1}{NK}\sum_{j=1}^{NK} -\log \frac{e^{(F(\overline{\mathbb{Z}}_j^s, \mathbb{S}^+)/\tau)}}{\sum_{i=1}^{N} e^{(F(\overline{\mathbb{Z}}_j^s, \mathbb{S}^i)/\tau)}}, \tag{5}$$

where $\overline{\mathbb{Z}}^s$ is the modified feature of support set by the proposed MAVSI. N and K are the class number and sample number of each class in one episode. $\mathbb{S}^+$ is the semantic corresponding to the ground-truth of $\overline{\mathbb{Z}}_j^s$, and $\tau > 0$ is a temperature parameter. Equation (5) enables the similarity between

Method	Semantic	Backbone	miniImageNet		tieredImageNet	
			1-shot	**5-shot**	**1-shot**	**5-shot**
MAML ([Finn *et al.*, 2017])	No	ConvNet	48.70±1.84	63.11±0.92	51.67±1.81	70.30±1.75
MatchingNet ([Vinyals *et al.*, 2016])	No	ConvNet	43.56±0.84	55.31±0.73	-	-
RelationNet ([Sung *et al.*, 2018])	No	ConvNet	50.44±0.82	65.32±0.70	54.48±0.93	71.32±0.78
ProtoNet ([Snell *et al.*, 2017])	No	ConvNet	49.42±0.78	68.20±0.66	53.31±0.89	72.69±0.74
Dynamic-FSL ([Gidaris and Komodakis, 2018])	No	ResNet12	62.81±0.27	78.97±0.18	78.97±0.18	83.09±0.12
DeepEMD ([Zhang *et al.*, 2020])	No	ResNet12	66.50±0.80	**82.41±0.56**	72.65±0.31	86.03±0.58
Neg-Cosine ([Liu *et al.*, 2020])	No	ResNet12	63.85±0.81	81.57±0.56	-	-
MBSS ([Cheng *et al.*, 2023])	No	ResNet12	65.79±0.20	81.90±0.14	71.74±0.23	86.34±0.15
KTN ([Peng *et al.*, 2019])	Yes	ConvNet	64.42±0.72	74.16±0.56	63.43±0.21	74.32±0.58
TriNet ([Chen *et al.*, 2019])	Yes	ResNet18	58.12±1.37	76.92±0.69	-	-
AM3 ([Xing *et al.*, 2019])	Yes	ResNet12	65.30±0.49	78.10±0.36	69.08±0.47	82.58±0.31
SEGA ([Yang *et al.*, 2022])	Yes	ResNet12	69.04±0.26	79.03±0.18	72.18±0.30	84.28±0.21
MultiSem ([Schwartz *et al.*, 2022])	Yes	ResNet12	67.30	82.10	62.18±0.18	78.54±0.27
LPE ([Yang *et al.*, 2023])	Yes	ResNet12	68.28±0.43	78.88±0.33	72.03±0.49	83.76±0.37
MAVSI (**Ours**)	Yes	ResNet12	**69.74±0.21**	82.23±0.41	**74.61±0.18**	**87.45±0.46**

Table 1: Average accuracy (%) comparison on miniImageNet and tieredImageNet in 5-way 1-shot and 5-way 5-shot settings.

a visual feature and its ground-truth semantic larger with the comparison network F, so discriminative visual features of the support set are further promoted.

3.4 Relation Network Module for Classification

Following [Sung *et al.*, 2018], the visual prototype $\overline{\mathbb{P}}^i$ for class i could be obtained by $\overline{\mathbb{Z}}^{s,i}$ in support set as below,

$$\overline{\mathbb{P}}^i = \frac{1}{K}\sum_{j=1}^{K}\overline{\mathbb{Z}}_j^{s,i}. \tag{6}$$

Then, we concatenate each class prototype with each query set and calculate the similarity score as below,

$$r_{i,j} = \mathcal{R}(\overline{\mathbb{P}}^i, \overline{\mathbb{Z}}_j^q), \tag{7}$$

where $\mathcal{R}$ consists of two convolutional layers and one MLP. Moreover, we utilize the loss of Mean Square Error (MSE) to train the model,

$$\mathcal{L}_{MSE} = \frac{1}{N\times m}\sum_{i=1}^{N}\sum_{j=1}^{m}(r_{i,j} - \mathbf{1}(i == y_j^q))^2, \tag{8}$$

where m denotes the sample number of the query set. Different from [Sung *et al.*, 2018], we compute the relation scores between the prototypes of the support set and query samples.

After introducing the modules in our proposed model, we could obtain the overall loss function as below,

$$\mathcal{L}_{overall} = \mathcal{L}_{MSE} + \lambda\mathcal{L}_{ce}, \tag{9}$$

where λ is a hyper-parameter used to balance the importance between the two losses.

4 Experiments

In this section, we validate the effectiveness of our proposed model on four benchmark datasets for FSL.

4.1 Datasets

miniImageNet [Vinyals *et al.*, 2016] consists of 100 classes. These classes are divided into 64, 16, and 20 for training, validation, and testing. **tiredImageNet** [Ren *et al.*, 2018] contains 608 classes, split into 351, 97, and 160 for training, validation, and testing. **CIFAR-FS** [Bertinetto *et al.*, 2019] consists of 100 classes. These classes are divided into 64, 16, and 20 for training, validation, and testing. **CUB-200-2011** [Wah *et al.*, 2011] contains images from 200 bird species, where 200 species are divided into 100, 50, and 50 for training, validation, and testing, respectively.

4.2 Implementation Details

Similar to previous works [Xing *et al.*, 2019; Schwartz *et al.*, 2022; Yang *et al.*, 2022], we utilize ResNet-12 as the backbone network, and modify the number of convolutional filters from [64, 128, 256, 512] to [64, 160, 320, 640]. In all cases, the comparison network F is the MLP with a LeakyReLU-activated hidden layer, and the relation network consists of convolutional layers and the MLP. We use Glove [Pennington *et al.*, 2014] as the semantic extractor, which is pre-trained on a large corpus. Our experiments are implemented under 5-way 1-shot and 5-way 5-shot settings. The input image size is 84×84. Following [Peng *et al.*, 2019], we train the model for 150 epochs, with 800 episodes in each epoch. We use the Adam optimizer with a learning rate of 5e-3 and weight decay of 5e-6. The learning rate is dropped by half every 6,000 episodes, and other parameters such as λ, γ, and the temperature parameter τ are adjusted during end-to-end training. We conduct 5-way 5-shot and 5-way 1-shot classification tasks on each dataset during the testing phase. The final classification accuracy results are obtained by averaging over 10,000 episodes and we report it with a 95% confidence interval.

4.3 Main Results

To evaluate the effectiveness of our proposed model, we conduct extensive experiments on miniImageNet, tieredImageNet, CUB, and CIFAR-FS. The comparison results with

Method	Semantic	Backbone	CUB	
			1-shot	**5-shot**
MAML ([Finn *et al.*, 2017])	No	ConvNet	54.73±0.97	75.75±0.75
MatchingNet ([Vinyals *et al.*, 2016])	No	ConvNet	60.52±0.88	75.29±0.75
RelationNet ([Sung *et al.*, 2018])	No	ConvNet	62.34±0.94	77.84±0.68
ProtoNet ([Snell *et al.*, 2017])	No	ConvNet	50.46±0.88	76.39±0.64
FEAT ([Ye *et al.*, 2020])	No	ResNet12	68.87±0.22	82.90±0.15
DeepEMD ([Zhang *et al.*, 2020])	No	ResNet12	75.65±0.83	88.69±0.50
MBSS ([Cheng *et al.*, 2023])	No	ResNet12	73.29±0.69	87.49±0.40
TriNet ([Chen *et al.*, 2019])	Yes	ResNet18	69.61±0.46	84.10±0.35
AM3 ([Xing *et al.*, 2019])	Yes	ResNet12	73.60	79.90
SEGA ([Yang *et al.*, 2022])	Yes	ResNet12	84.57±0.22	90.85±0.16
MultiSem ([Schwartz *et al.*, 2022])	Yes	ResNet12	76.10	82.90
LPE ([Yang *et al.*, 2023])	Yes	ResNet12	80.76±0.40	88.98±0.26
MAVSI (**Ours**)	Yes	ResNet12	**85.21±0.42**	**91.56±0.25**

Table 2: Average accuracy (%) comparison on CUB in 5-way 1-shot and 5-way 5-shot settings.

Method	Semantic	Backbone	CIFAR-FS	
			1-shot	**5-shot**
MAML ([Finn *et al.*, 2017])	No	ConvNet	58.9±1.9	71.5±1.0
ProtoNet ([Snell *et al.*, 2017])	No	ConvNet	55.5±0.7	72.0±0.6
MetaOptNet ([Lee *et al.*, 2019])	No	ResNet12	72.0±0.7	84.2±0.5
RFS ([Tian *et al.*, 2020])	No	ResNet12	73.9±0.8	86.9±0.5
SEGA ([Yang *et al.*, 2022])	Yes	ResNet12	78.45±0.24	86.00±0.20
LPE ([Yang *et al.*, 2023])	Yes	ResNet12	74.88±0.45	85.30±0.35
MAVSI (**Ours**)	Yes	ResNet12	**80.12±0.21**	**87.13±0.14**

Table 3: Average accuracy (%) comparison on CIFAR-FS in 5-way 1-shot and 5-way 5-shot settings.

state-of-the-art methods are shown in Tables 1, 2, and 3. Compared to previous methods that utilize semantic information to address FSL problem (*e.g.*, KTN [Peng *et al.*, 2019], TriNet [Chen *et al.*, 2019], AM3 [Xing *et al.*, 2019], SEGA [Yang *et al.*, 2022], MultiSem [Schwartz *et al.*, 2022]), our model achieves the best performance in both 1-shot and 5-shot settings on all benchmark datasets. Particularly, as shown in Table 1, in the 1-shot and 5-shot settings of tieredImageNet, our method outperforms SEGA by 2.43% and 3.17% improvements, respectively. This is because we conduct visual-semantic interaction on both the support and query sets, allowing us to capture more discriminative visual features. Additionally, our proposed approach consistently achieves the best results across almost all datasets and settings compared to all methods. Due to our aim at maintaining consistent parameters across all datasets and all FSL settings, we do not surpass DeepEMD on the miniImageNet with the 5-shot setting. These results could verify the effectiveness and superiority of our proposed FSL model.

4.4 Empirical Analysis

This section provides a deeper analysis and discussion of the proposed model.

Ablation Study. In Table 4, we record the accuracy results of our model using three proposed modules, including SCAVSI, CAVSI, and VSCL, as well as their combinations. It can be observed that SCAVSI, CAVSI, and VSCL are all highly effective, with average 1-shot accuracy improvements of 4.91%, 3.89%, and 2.58% on the three datasets, respectively. Moreover, when combining them together, further improvements in 1-shot learning accuracy are achieved. These results indicate that the three proposed modules all play important roles in the final performance of the model.

Analysis for Hyper-Parameter λ. λ is used to embody the importance of our proposed VSCL module. We conduct experiments on validation of miniImageNet and CIFAR-FS to analyze its impact on the model's performance. As shown in Figure 5 (a), we observe that, in 5-way 1-shot setting, the optimal values for λ are all within the range of 0.3 to 0.6. In comparison results, we set the value of λ as 0.5.

Analysis for Number of Attention Heads. By using different numbers of attention heads and their corresponding scaling factors [Vaswani *et al.*, 2017], we evaluate their impact on the performance of the miniImageNet dataset. As shown in Table 5, the model's performance shows a trend of initially increasing and then decreasing. We speculate that when the number of attention heads is too small, the model's feature learning capacity is insufficient. Moreover, when the number is too large, it introduces significant redundancy.

Analysis for Temperature Parameter τ. There exists a temperature parameter τ in the proposed VSCL module. We select different values of τ and assess their impact on the performance of the tiredImageNet validation. As shown in Figure 5 (b), the accuracy curves remain relatively stable with increasing values of τ in both the 1-shot and 5-shot settings.

SCAVSI	CAVSI	VSCL	miniImageNet	tieredImageNet	CUB	Average
✗	✗	✗	62.81	68.55	67.78	66.38
✓	✗	✗	66.23	70.21	77.43	71.29
✗	✓	✗	65.34	69.34	76.12	70.27
✗	✗	✓	64.52	69.12	73.25	68.96
✓	✗	✓	67.32	72.25	78.21	72.26
✓	✓	✗	67.55	72.12	78.45	72.71
✗	✓	✓	66.54	71.67	79.46	72.56
✓	✓	✓	**69.74**	**74.61**	**80.12**	**74.82**

Table 4: Ablation study on three different datasets under 5-way 1-shot setting.

Heads	Scaling factor	miniImageNet	
		1-shot	**5-shot**
1	640	67.23	78.23
4	160	68.57	80.24
8	**80**	**69.74**	**82.23**
16	40	68.12	79.23
32	20	66.23	77.10

Table 5: Analysis for the number of attention heads and corresponding scaling factors on miniImageNet dataset.

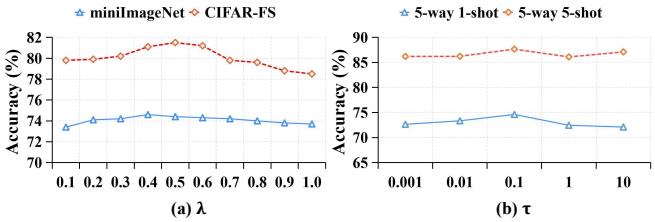

Figure 5: Analysis for the parameter sensitivity.

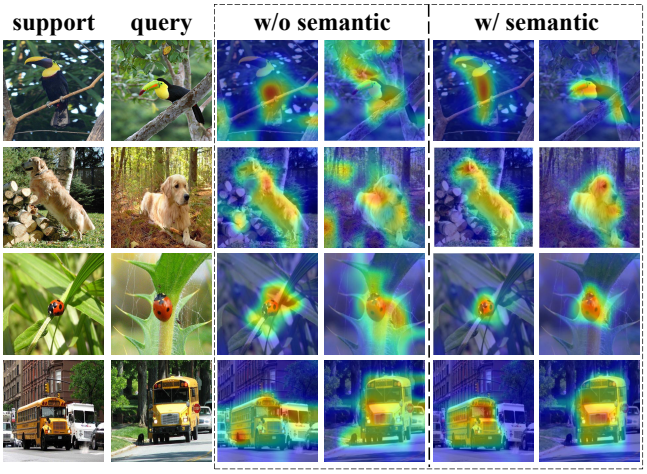

Figure 6: Visualization results on miniImageNet dataset.

These results indicate that our proposed model is robust to the choice of τ. In both settings, the highest accuracy results are achieved when τ is set to 0.1.

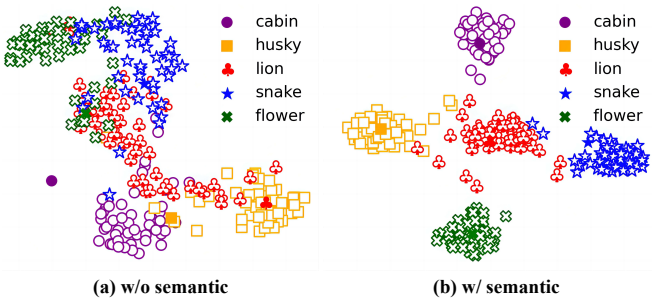

Figure 7: Feature distributions on miniImageNet dataset.

4.5 Visualization Results

We visualize four feature maps from support and query sets, respectively. As shown in Figure 6, the visual features without visual-semantic interaction exhibit some background noise, which makes the visual features less discriminative. In contrast, background noise has been almost completely eliminated, demonstrating that our proposed approach could effectively enhance discriminative visual features. Additionally, we visualize feature distributions on the miniImageNet dataset, as illustrated in Figure 7. Through visual-semantic interaction, the dots representing different classes in the query set become more distinguishable and discriminative.

5 Conclusion

We propose a novel FSL approach named MAVSI, where two different attention modules are designed to enhance discriminative visual features at spatial and channel levels. To conduct MAVSI on both the support and query sets, we introduce a relation module and utilize the semantic knowledge from the support set to guide the visual-semantic interaction on the query set. Moreover, we propose a visual-semantic contrastive learning module to promote discriminative visual features of the support set. Extensive experiments on four benchmark datasets demonstrate that the proposed model could outperform existing mainstream FSL approaches. In future work, we will explore higher-quality semantic extractors, such as vision-language large models, and investigate the cross-domain FSL problem where a distribution shift exists between base and unseen classes.

Acknowledgments

This work was supported by the Key Natural Science Project of Anhui Provincial Education Department (No.KJ2021A0043) and the National Natural Science Foundation of China (No.61602004, No.62025604, No.62306343), Shenzhen Science and Technology Program (No.KQTD20221101093559018).

References

[Afham *et al.*, 2021] Mohamed Afham, Salman Khan, Muhammad Haris Khan, Muzammal Naseer, and Fahad Shahbaz Khan. Rich semantics improve few-shot learning. In *BMVC*, page 152, 2021.

[Badamdorj *et al.*, 2022] Taivanbat Badamdorj, Mrigank Rochan, Yang Wang, and Li Cheng. Contrastive learning for unsupervised video highlight detection. In *IEEE CVPR*, pages 14022–14032, 2022.

[Baik *et al.*, 2020] Sungyong Baik, Seokil Hong, and Kyoung Mu Lee. Learning to forget for meta-learning. In *IEEE CVPR*, pages 2376–2384, 2020.

[Bertinetto *et al.*, 2019] Luca Bertinetto, João F. Henriques, Philip H. S. Torr, and Andrea Vedaldi. Meta-learning with differentiable closed-form solvers. In *ICLR*, 2019.

[Cao *et al.*, 2021] Kaidi Cao, Maria Brbic, and Jure Leskovec. Concept learners for few-shot learning. In *ICLR*, 2021.

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

[Chen *et al.*, 2024] Ruoyu Chen, Hua Zhang, Siyuan Liang, Jingzhi Li, and Xiaochun Cao. Less is more: Fewer interpretable region via submodular subset selection. In *ICLR*, 2024.

[Cheng *et al.*, 2023] Jun Cheng, Fusheng Hao, Fengxiang He, Liu Liu, and Qieshi Zhang. Mixer-based semantic spread for few-shot learning. *IEEE TMM*, 25:191–202, 2023.

[Finn *et al.*, 2017] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *ICML*, volume 70, pages 1126–1135, 2017.

[Gidaris and Komodakis, 2018] Spyros Gidaris and Nikos Komodakis. Dynamic few-shot visual learning without forgetting. In *IEEE CVPR*, pages 4367–4375, 2018.

[Huang *et al.*, 2021] Chao Huang, Zongju Peng, Yong Xu, Fen Chen, Qiuping Jiang, Yun Zhang, Gangyi Jiang, and Yo-Sung Ho. Online learning-based multi-stage complexity control for live video coding. *IEEE TIP*, 30:641–656, 2021.

[Huang *et al.*, 2022a] Chao Huang, Zhihao Wu, Jie Wen, Yong Xu, Qiuping Jiang, and Yaowei Wang. Abnormal event detection using deep contrastive learning for intelligent video surveillance system. *IEEE TII*, 18(8):5171–5179, 2022.

[Huang *et al.*, 2022b] Chao Huang, Zehua Yang, Jie Wen, Yong Xu, Qiuping Jiang, Jian Yang, and Yaowei Wang. Self-supervision-augmented deep autoencoder for unsupervised visual anomaly detection. *IEEE TCYB*, 52(12):13834–13847, 2022.

[Huang *et al.*, 2023] Chao Huang, Jie Wen, Yong Xu, Qiuping Jiang, Jian Yang, Yaowei Wang, and David Zhang. Self-supervised attentive generative adversarial networks for video anomaly detection. *IEEE TNNLS*, 34(11):9389–9403, 2023.

[Huang *et al.*, 2024] Chao Huang, Chengliang Liu, Jie Wen, Lian Wu, Yong Xu, Qiuping Jiang, and Yaowei Wang. Weakly supervised video anomaly detection via self-guided temporal discriminative transformer. *IEEE TCYB*, 54(5):3197–3210, 2024.

[Ji *et al.*, 2021] Zhong Ji, Xiyao Liu, Yanwei Pang, Wangli Ouyang, and Xuelong Li. Few-shot human-object interaction recognition with semantic-guided attentive prototypes network. *IEEE TIP*, 30:1648–1661, 2021.

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

[Li *et al.*, 2020] Aoxue Li, Weiran Huang, Xu Lan, Jiashi Feng, Zhenguo Li, and Liwei Wang. Boosting few-shot learning with adaptive margin loss. In *IEEE CVPR*, pages 12573–12581, 2020.

[Liu *et al.*, 2020] Bin Liu, Yue Cao, Yutong Lin, Qi Li, Zheng Zhang, Mingsheng Long, and Han Hu. Negative margin matters: Understanding margin in few-shot classification. In *ECCV*, volume 12349, pages 438–455, 2020.

[Liu *et al.*, 2022] Yang Liu, Tu Zheng, Jie Song, Deng Cai, and Xiaofei He. DMN4: few-shot learning via discriminative mutual nearest neighbor neural network. In *AAAI*, pages 1828–1836, 2022.

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

[Pennington *et al.*, 2014] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *EMNLP*, pages 1532–1543, 2014.

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

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

[Rusu *et al.*, 2019] Andrei A. Rusu, Dushyant Rao, Jakub Sygnowski, Oriol Vinyals, Razvan Pascanu, Simon Osindero, and Raia Hadsell. Meta-learning with latent embedding optimization. In *ICLR*, 2019.

[Schwartz *et al.*, 2022] Eli Schwartz, Leonid Karlinsky, Rogério Feris, Raja Giryes, and Alexander M. Bronstein. Baby steps towards few-shot learning with multiple semantics. *Pattern Recognit. Lett.*, 160:142–147, 2022.

[Snell *et al.*, 2017] Jake Snell, Kevin Swersky, and Richard S. Zemel. Prototypical networks for few-shot learning. In *NeurIPS*, pages 4077–4087, 2017.

[Sung *et al.*, 2018] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip H. S. Torr, and Timothy M. Hospedales. Learning to compare: Relation network for few-shot learning. In *IEEE CVPR*, pages 1199–1208, 2018.

[Tian *et al.*, 2020] Yonglong Tian, Yue Wang, Dilip Krishnan, Joshua B. Tenenbaum, and Phillip Isola. Rethinking few-shot image classification: A good embedding is all you need? In *ECCV*, volume 12359, pages 266–282, 2020.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, pages 5998–6008, 2017.

[Vinyals *et al.*, 2016] Oriol Vinyals, Charles Blundell, Tim Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. In *NeurIPS*, pages 3630–3638, 2016.

[Wah *et al.*, 2011] Catherine Wah, Steve Branson, Peter Welinder, Pietro Per-ona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. In *Technical Report CNS-TR-2011-001, California Institute of Technology*, 2011.

[Wang *et al.*, 2018] Yu-Xiong Wang, Ross B. Girshick, Martial Hebert, and Bharath Hariharan. Low-shot learning from imaginary data. In *IEEE CVPR*, pages 7278–7286, 2018.

[Wang *et al.*, 2020] Haoran Wang, Ying Zhang, Zhong Ji, Yanwei Pang, and Lin Ma. Consensus-aware visual-semantic embedding for image-text matching. In *ECCV*, volume 12369, pages 18–34, 2020.

[Wang *et al.*, 2021] Wei Wang, Shenglun Chen, Yuankai Xiang, Jing Sun, Haojie Li, Zhihui Wang, Fuming Sun, Zhengming Ding, and Baopu Li. Sparsely-labeled source assisted domain adaptation. *Pattern Recognit.*, 112:107803, 2021.

[Wang *et al.*, 2022] Wei Wang, Baopu Li, Mengzhu Wang, Feiping Nie, Zhihui Wang, and Haojie Li. Confidence regularized label propagation based domain adaptation. *IEEE TCSVT*, 32(6):3319–3333, 2022.

[Wang *et al.*, 2023a] Wei Wang, Haojie Li, Zhengming Ding, Feiping Nie, Junyang Chen, Xiao Dong, and Zhihui Wang. Rethinking maximum mean discrepancy for visual domain adaptation. *IEEE TNNLS*, 34(1):264–277, 2023.

[Wang *et al.*, 2023b] Wei Wang, Mengzhu Wang, Xiao Dong, Long Lan, Quannan Zu, Xiang Zhang, and Cong Wang. Class-specific and self-learning local manifold structure for domain adaptation. *Pattern Recognit.*, 142:109654, 2023.

[Wu *et al.*, 2023a] Yanan Wu, Zhixiang Chi, Yang Wang, and Songhe Feng. Metagcd: Learning to continually learn in generalized category discovery. In *IEEE ICCV*, pages 1655–1665, 2023.

[Wu *et al.*, 2023b] Yanan Wu, Tengfei Liang, Songhe Feng, Yi Jin, Gengyu Lyu, Haojun Fei, and Yang Wang. Metazscil: A meta-learning approach for generalized zero-shot class incremental learning. In *AAAI*, pages 10408–10416, 2023.

[Wu *et al.*, 2024] Yanan Wu, Zhixiang Chi, Yang Wang, Konstantinos N. Plataniotis, and Songhe Feng. Test-time domain adaptation by learning domain-aware batch normalization. In *AAAI*, pages 15961–15969, 2024.

[Xing *et al.*, 2019] Chen Xing, Negar Rostamzadeh, Boris N. Oreshkin, and Pedro O. Pinheiro. Adaptive cross-modal few-shot learning. In *NeurIPS*, pages 4848–4858, 2019.

[Xu and Le, 2022] Jingyi Xu and Hieu Le. Generating representative samples for few-shot classification. In *IEEE CVPR*, pages 8993–9003, 2022.

[Xu *et al.*, 2022] Xianda Xu, Xing Xu, Fumin Shen, and Yujie Li. Semantic-aligned attention with refining feature embedding for few-shot image classification. *IEEE TITS*, 23(12):25458–25468, 2022.

[Xu *et al.*, 2023] Chengming Xu, Chen Liu, Xinwei Sun, Siqian Yang, Yabiao Wang, Chengjie Wang, and Yanwei Fu. Patchmix augmentation to identify causal features in few-shot learning. *IEEE TPAMI*, 45(6):7639–7653, 2023.

[Yang *et al.*, 2022] Fengyuan Yang, Ruiping Wang, and Xilin Chen. SEGA: semantic guided attention on visual prototype for few-shot learning. In *IEEE WACV*, pages 1586–1596, 2022.

[Yang *et al.*, 2023] Fengyuan Yang, Ruiping Wang, and Xilin Chen. Semantic guided latent parts embedding for few-shot learning. In *IEEE WACV*, pages 5436–5446, 2023.

[Ye *et al.*, 2020] Han-Jia Ye, Hexiang Hu, De-Chuan Zhan, and Fei Sha. Few-shot learning via embedding adaptation with set-to-set functions. In *IEEE CVPR*, pages 8805–8814, 2020.

[Zhang *et al.*, 2019] Hongguang Zhang, Jing Zhang, and Piotr Koniusz. Few-shot learning via saliency-guided hallucination of samples. In *IEEE CVPR*, pages 2770–2779, 2019.

[Zhang *et al.*, 2020] Chi Zhang, Yujun Cai, Guosheng Lin, and Chunhua Shen. Deepemd: Few-shot image classification with differentiable earth mover's distance and structured classifiers. In *IEEE CVPR*, pages 12200–12210, 2020.

[Zhao *et al.*, 2023] Peng Zhao, Huihui Xue, Xia Ji, Huiting Liu, and Li Han. Zero-shot learning via visual feature enhancement and dual classifier learning for image recognition. *Information Sciences.*, 642:119161, 2023.