Semantic-Aligned Attention with Refining Feature Embedding for Few-Shot Image Classification

Xianda Xu, Xing Xu, Fumin Shen, Yujie Li

***Abstract*—Autonomous driving relies on trusty visual recognition of surrounding objects. Few-shot image classification is used in autonomous driving to help recognize objects that are rarely seen. Successful embedding and metric-learning approaches to this task normally learn a feature comparison framework between an unseen image and the labeled images. However, these approaches usually have problems with ambiguous feature embedding because they tend to ignore important local visual and semantic information when extracting intra-class common features from the images. In this paper, we introduce a Semantic-Aligned Attention (SAA) mechanism to refine feature embedding and it can be applied to most of the existing embedding and metric-learning approaches. The mechanism highlights pivotal local visual information with attention mechanism and aligns the attentive map with semantic information to refine the extracted features. Incorporating the proposed mechanism into the prototypical network, evaluation results reveal competitive improvements in both few-shot and zero-shot classification tasks on various benchmark datasets.**

***Index Terms*—Autonomous Driving, Few-Shot Image Classification, Zero-Shot Image Classification, Attention Mechanism, Visual-Semantic Alignment**

I. Introduction

In autonomous driving, it is required to recognize objects on road in real-time, including lanes, vehicles, pedestrians, animals, trees, traffic signs, and so on. Deep neural networks (DNNs) are often used in these visual recognition tasks to help autonomous vehicles run normally [1], [2], [3]. However, these conventional methods have two major drawbacks: (1) they usually require large-scale datasets, which adds burden to the cost of collecting data and limits their adaptability to new classes with rare examples like new vehicle modes, new trafiic signs, animals, and building in new places, etc; (2) they usually have complex network architecture and low computational efficiency, which limits their use in real-time applications. To tackle these challenges, *few-shot image classification* is proposed.

Few-shot image classification aims to recognize new classes with little supervised data. Specifically, it classifies unseen images into a set of new classes with a limited number of labeled images in each new class. The shortage of labeled data leads to overfitting and generalization issues in conventional deep learning methods so new approaches to this problem are being studied. Recently, an increase of interest emerges in the embedding and metric-learning approaches.

The embedding and metric-learning approaches [4], [5], [6], [7], [8], [9] produce feature embeddings for comparison between each unlabeled sample embedding and the class embeddings [7]. Specifically, they roughly follow three steps: (1) using a feature embedding network to map all the labeled (support) and unlabeled (query) samples into an embedding space; (2) representing each class in the embedding space based on the embeddings of its support samples; (3) measuring the distances of each query sample embedding to all the class embeddings and assigning each query sample to the nearest class. The first step, also known as feature embedding, is decisive in class embedding. To alleviate overfitting, most embedding and metric-learning approaches [4], [5], [6] adopt a simple 4-layer convolutional network in feature embedding, which suffers from the following two major limitations.

The first limitation is that the feature embedding and measuring network tends to extract global features, which makes it prone to bring noisy information such as background distraction. The second limitation is that the feature embedding and measuring network ignores the use of semantic information. When human beings recognize an object, they not only focus on visual information but also consider some prior knowledge like semantic descriptions of the object. It indicates that there is a latent correlation between visual information and semantic knowledge in the human's brain [10], [11]. Semantic information accessible to few-shot image classification is usually the label embeddings learned from large unsupervised text corpora [12]. Exploring the relationship between visual information and the semantic knowledge could assist the process of feature embedding and class embedding [13], [14].

Motivated by the above two limitations in the existing feature embedding network, we propose a semantic-aligned attention (SAA) mechanism for few-shot image classification. Specifically, we leverage the attention mechanism to highlight pivotal local information and alleviate noisy global distractions. We learn the visual-and-semantic correlation and align the visual attentive maps with it to further refine the extracted local features. We further extend our method to zero-shot image classification. Rather than a few labeled examples in few-shot image classification, each new class in zero-shot image classification comes with side information, usually attribute vectors. Like in few-shot image classification, we improve class embedding in zero-shot image classification by incorporating the class label embeddings into the network.

Proven by the experimental results, the semantic-aligned attention network has achieved a significant improvement in

X. Xu, X. Xu, F. Shen, are with the Center for Future Multimedia and School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China (Email: bryce_xu@163.com, xing.xu@uestc.edu.cn, fumin.shen@gmail.com)(Corresponding author: Xing Xu)

Y. Li is with School of Information Engineering, Yangzhou University, Yangzhou, China (Email: yzyjli@gmail.com)

image classification tasks with rare examples. At the same time, it shows high computational efficiency due to its superior network structure. Therefore, we believe that our method has great potential in real-time visual recognition applications like autonomous driving.

To sum up, our contributions in this paper include:

- We introduce a semantic-aligned attention (SAA) mechanism for few-shot image classification to refine feature embedding. The mechanism refines feature embedding with important local visual information aligned with semantic information to improve class embedding.
- We extend our method to zero-shot image classification. The network improves class embedding by absorbing knowledge from a semantic vector concatenated with attributes and label embeddings.
- Experimental results show that our method brings a considerable performance boost to the existing embedding and metric-learning approaches such as the prototypical network in both image classification tasks.

The remainder of this paper is organized as follows. In Section II, we have a review about the related work in few-shot learning and zero-shot learning respectively. Section III contains the elaboration of the proposed approach. Extensive experimental studies on few-shot image classification and zero-shot image classification are presented in Section IV. The conclusion is drawn in Section V.

II. Related Work

In this section, we conduct a review of the related work in few-shot learning and zero-shot learning. We begin by briefly discussing about deep learning in visual recognition for autonomous driving. Then, we introduce three paradigms in few-shot learning. Next, we discuss the research progress in zero-shot learning.

Deep Learning in Visual Recognition for Autonomous Driving: Visual recognition in autonomous driving applications usually trains a deep convolutional neural network architecture to detect objects in general [15] or objects in a certain category like traffic signs [1] and vehicle logos [2]. In recent years, much work has been drawn to use few-shot learning or zero-shot learning in autonomous driving. For example, Kim *et al.* [16] uses a variational prototyping-encoder to learn the prototypes of traffic signs and logos in the few-shot learning scenario. Wong *et al.* [17] created a category-agnostic embedding space to identify unknown instances on road. It indicates that few-shot learning and zero-shot learning are achieving steady progress in visual recognition for autonomous driving.

Few-shot Learning: Much attention has been drawn to this task in recent years. Existing approaches to few-shot learning can be roughly divided into three paradigms: (i) data augmentation (ii) optimization (iii) embedding and metric-learning.

Data augmentation approaches [18], [19], [20] are straightforward as they solve the few-shot learning task by augmenting the number of labeled samples in each novel class. One category focuses on transforming existing datasets. The idea in the early works is to copy original samples and modify the new samples with learned transformation. Schwartz *et al.* [19] applies auto-encoders that add the new samples with the intra-class variances learned from a similar class. Tsutsui *et al.* [20] generates new samples with GANs and learns to synthesize them with original samples. The other category explores transforming external datasets. For example, Chen *et al.* [18] combines blocks from the existing labeled images and the unlabeled gallery images in a self-training scheme. These approaches alleviate overfitting by expanding the training dataset to meet the sufficient sample complexity, but they do not solve the problem in essence.

Optimization approaches [21], [22], [23], [24] meta-learn an algorithm to fine-tune the model in the case of few-shot learning. In early works, Finn *et al.* [21] proposes MAML that learns a good initialization so that the classifiers for novel classes can be learned with a small number of gradient update steps. Li *et al.* [23] presents Meta-SGD which learns a good optimizer for the new task. These methods have been improved in recent years. Based on MAML, Jamal *et al.* [22] presents an entropy-based meta-learning approach to alleviate the bias towards existing tasks in the initial model and Oh *et al.* [25] proposes a novel meta-learning algorithm which updates only the extractor while freezes the classifier. These approaches learn to fine-tune the model from an abundant data regime, but they do not check the relatedness of the two tasks, thus being affected by the previous work.

Embedding and metric-learning approaches train a neural network to embed all the labeled and unlabeled samples into an embedding space so that the similar pairs and dissimilar pairs can be identified using a metric. Inductive embedding and metric-learning approaches [4], [5], [6] are traditional and only support samples are used in the model training. The prototypical network [4] uses the mean of support embeddings as the class representation (or the prototype) and the Euclidean distance as the metric. The relation network [5] learns the metric by evaluating each query-support pair with a neural network. The multi-level metric learning network [6] further considers part-level, pixel-level, and distribution-level similarities simultaneously and adopts multiple metric methods. Transductive embedding and metric-learning approaches [8], [9] are less ambitious and the entire query set is also involved in the model training. For example, Oreshkin *et al.* [8] learns a task-dependent metric in the network and Liu *et al.* [9] learns to embed a visual subgraph to a task-oriented cross-modal graph. In this paper, we only consider the inductive task setting. It is challenging yet more generalized because, in real-time applications, objects are always unpredictable.

In the inductive task setting, several networks with attention mechanism [26], [27], [28], [29] have achieved good results in few-shot image classification. Approaches include [26] and [27] localize relevant regions by applying feature-level and instance-level attention schemes. Hao *et al.* [28] and Li *et al.* [29] further take advantage of semantic information to select attentive region pairs. Inspired by their work, we introduce a novel semantic-aligned attention mechanism. It combines spatial attention mechanism and semantic alignment mechanism in a different way: semantic alignment guides visual attention in the embedding network to better focus on

regional intra-class commonality.

Zero-shot Learning: Our method is designed for few-shot learning, but it can be elegantly extended to zero-shot learning. Like few-shot learning, zero-shot learning aims to recognize new classes whose instances are unavailable during training. But in zero-shot learning, each new class comes with side information, usually semantic knowledge, rather than a few labeled samples.

Early works in zero-shot learning [30], [31], [32] directly learn a mapping from an image feature space to a semantic space. For example, Akata *et al.* [30] uses ranking loss to learn a bilinear compatibility function between the image feature space and the semantic space. The compatibility function is further optimized by [31] with SVM loss and [32] with a regularizer. Bucher *et al.* [33] improves the consistency of the semantic embedding with metric-learning in the semantic space.

Recent advances [34], [35], [36], [37], [38], [39] focus on learning deep multi-modal embeddings. Socher *et al.* [38] use a neural network to project image features into a semantic space. Zheng *et al.* [39] argues that the image feature space is more discriminative so they map semantic features into the image feature space instead with a deep embedding model. Reed *et al.* [34] learns deep multi-modal embeddings for both image features and semantic knowledge, which is further improved in the following works [35], [36], [37].

Our work uses a deep neural network to generate both image feature embeddings and semantic embeddings. Following [4], we obtain class embeddings by transforming the semantic attribute vectors in each class. We improve the process by introducing semantic alignment to the transformation network and we again benefit from gaining additional semantic knowledge.

III. The Proposed Approach

This section presents our proposed approach to dealing with few-shot image classification and zero-shot image classification. We begin by offering a problem definition of few-shot image classification and zero-shot image classification. Next, we introduce our Semantic-Aligned Attention (SAA) mechanism and show how our method is implemented in the feature embedding network of few-shot image classification. At last, we extend our work to zero-shot image classification and demonstrate how to improve the task with the label embeddings.

A. Problem Definition

Few-shot image classification deals with a supervised learning task given a dataset $D = \left\{D^{\text{train}}, D^{\text{val}}, D^{\text{test}}\right\}$ which is comprised of a training set, a validation set and a test set. The three sub-datasets have different label spaces, which means that a certain class in one of the three sub-datasets is unseen in the other two sub-datasets.

In traditional image classification, the three sub-datasets share the same label space and each class has a massive number of samples. That is why we could easily train a classifier to assign each sample a class label in the test set. However, since the three sub-datasets in few-shot image classification are disjoint with each other and each class is given a few number of samples, such a classifier does not work well in such a circumstance. Therefore, meta-learning is normally performed on the training set, to learn transferrable knowledge about how to classify from N new classes with K examples in each class. It allows us to perform better few-shot image classification on the test set.

Most successful few-shot image classification approaches follow an *Episodic Training* paradigm proposed by [40] which simulates the few-shot image classification task during training. Under this paradigm, a N-way K-shot *episode* contains two sets: (1) the *support* set $\mathcal{S} = \{(s_i, y_i)\}_{i=1}^{N\times K}$ with K samples from each of the N classes (2) the *query* set $\mathcal{Q} = \{(q_j, y_j)\}_{j=1}^{N\times Q}$ with Q samples from each of the N classes. The *support* set and the *query* set are disjoint $(S \cap Q = \varnothing)$ with each other. A certain image can never co-exist in both sets. All the training, validation and test processes are implemented on *episodes* which are randomly sampled from each dataset. In each training iteration, the model is updated in an *episode* and in each validation or test iteration, one *episode* is validated or tested.

Zero-shot image classification is similar to few-shot image classification but with K kept to 0, which means labeled samples are not available to each test class. Instead, each class comes with side information, usually class attribute vectors $V = \{v_k\}_{k=1}^{N}$ describing the semantic attributes of each class. $X = \{(x_m, y_m)\}^{N\times M}$ is the set of samples with M samples from each of the N classes. The zero-shot task follows the normal zero-shot setting where three sub-datasets are disjoint with each other.

B. Semantic-Aligned Attention (SAA) Mechanism

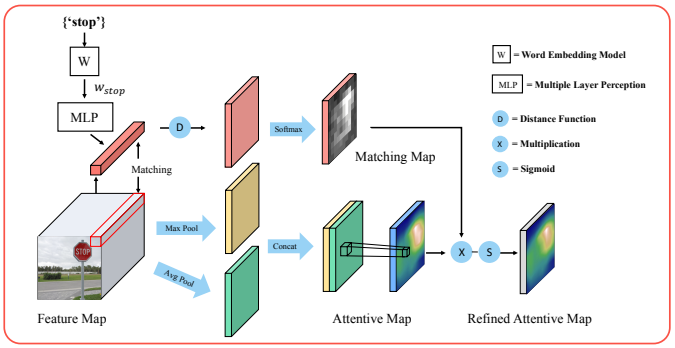

Fig. 1. The overall architecture of our Semantic-Aligned Attention (SAA) mechanism. It generates an visual attentive map with the spatial attention mechanism and a visual-semantic matching map with the semantic alignment mechanism. The refined visual attentive map is produced by combining and then activating these two maps.

The Semantic-Aligned Attention (SAA) mechanism is illustrated in Fig. 1, which consists of two branches, i.e., the Spatial Attention Mechanism and the Semantic Alignment Mechanism. The first branch is to extract more important local features. The second branch is to align the local features with semantic information. It generates a semantic-aligned attentive map $M_r \in \mathbf{R}^{1\times H\times W}$ based on the input feature

map $F \in \mathbf{R}^{C\times H\times W}$. The output feature map $F' \in \mathbf{R}^{C\times H\times W}$ would be:

$$\mathbf{F}' = \mathbf{F} \otimes \mathbf{M}_r. \tag{1}$$

where $\otimes$ is element-wise multiplication. Through multiplication, the attention values can be broadcasted accordingly.

1) Spatial Attention Mechanism:

Inspired by [41], [42], the spatial attention mechanism in our method is designed to explore the inter-spatial relationship of features in the input feature map. In other words, it helps decide "where" to focus on in the feature map.

Specifically, the spatial attention mechanism firstly applies average-pooling and max-pooling operations along the channel dimension to generate the following two visual maps: $F_{avg} \in \mathbb{R}^{1\times H\times W}$ and $F_{max} \in \mathbb{R}^{1\times H\times W}$. Then, it concatenates these two visual maps and obtains the attentive map $M_a \in \mathbb{R}^{1\times H\times W}$ by applying a convolution operation on the concatenated visual map to determine where to emphasize or suppress. To sum up, M_a is obtained as:

$$\begin{aligned}\mathbf{M}_a &= Conv\left(\left[f_{ap}(\mathbf{F})\,;\, f_{mp}(\mathbf{F})\right]\right)\\ &= Conv\left(\left[\mathbf{F_{avg}}\,;\, \mathbf{F_{max}}\right]\right).\end{aligned} \tag{2}$$

where f_{ap} is the average-pooling operation, f_{mp} is the max-pooling operation and $Conv$ is the convolution operation with the filter size of 7×7 in our few-shot image classification experiments.

2) Semantic Alignment Mechanism:

The semantic alignment mechanism in our method is used to associate visual local regions with class label embeddings. By exploring the relationship between visual information and the pre-learned semantic knowledge (the label embeddings), it helps locate informative visual regions and refine the visual attentive map. Note that it is only applied in feature embedding of the support samples since the label embeddings are not accessible to query samples.

We obtain the label embeddings of all the classes in each set using a word embedding model W. The word embedding model in this work is *GloVe* [12]. It has been pre-trained on large text corpora and proved reliable on word representation. The label embeddings are then transformed via a multiple layer perceptron [43] to match the channel dimension of the input feature map.

Each episode contains N label embeddings belonging to N different classes and for each support sample, it has 1 positive label embedding w^+ and $N-1$ negative label embeddings $\{w_n^-\}_{n=1}^{N-1}$. A visual-semantic matching map M_w can be generated using w^+ based on the input feature map with M_{wi} representing the relevance between w^+ and the local region F_i:

$$\mathbf{M}_{wi} = \frac{\exp\left(f_s\left(\mathbf{w}^+, \mathbf{F}_i\right)\right)}{\sum_j \exp\left(f_s\left(\mathbf{w}^+, \mathbf{F}_j\right)\right)}. \tag{3}$$

where f_s is the similarity function measuring the relevance between two embeddings.

Therefore, the refined attentive map M_r can be achieved by multiplying the attentive map and the matching map and then activating with the sigmoid function [44]:

$$\mathbf{M}_r = \sigma(\mathbf{M}_w \otimes \mathbf{M}_a). \tag{4}$$

In order to link visual local regions and class label embeddings, $loss_w$ is introduced here. The loss function is inspired from [45], [46]. $loss_w$ needs to be computed in each episode, which is:

$$loss_w = \sum_{n=1}^{N-1} \sum_{\mathbf{F}_i \in \mathbf{F}} \left[\mathbf{M}_{wi} \cdot |\alpha + f_s(\mathbf{w}_n^-, \mathbf{F}_i) - f_s(\mathbf{w}^+, \mathbf{F}_i)|_+\right]. \tag{5}$$

where α is a manually-set margin and M_{wi} here is for reinforcing the relevance between w^+ and F_i. Fig. 2 well illustrates the computing process of $loss_w$ in one episode.

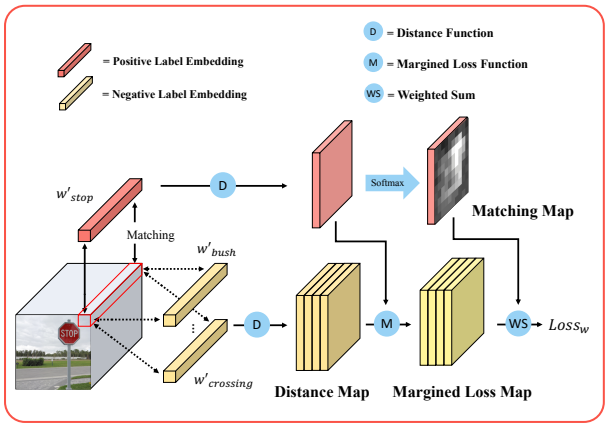

Fig. 2. The illustration of computing $loss_w$ in the 5-way scenario. In this case, the positive class is "Poodle" and the rest four classes are negative classes.

C. Network Architecture

Following most existing embedding and metric-learning approaches, we choose *ConvNet* as the neural network architecture for feature embedding in few-shot image classification, as shown in Fig. 3.

The neural network is comprised of four convolutional blocks. The first three blocks contain a convolution layer with 64 3×3 filters, a batch normalization layer, a spatial attention mechanism layer, and a ReLU nonlinearity layer. The last block replaces the spatial attention mechanism layer with the semantic-aligned attention mechanism layer for the support samples while keeps the same for the query samples. A 2×2 max-pooling layer is placed between every two convolutional blocks. The output size of the feature embedding network is $D = 64 * 5 * 5 = 1600$ for both the miniImageNet dataset and the tieredImageNet dataset.

The neural network generates 1600-dimensional feature embeddings of both the support samples and the query samples in the episode, which are used in the following metric-learning process. The support feature embeddings are utilized to form class embeddings and the query feature embeddings are measured to every class embedding to determine which class they belong to.

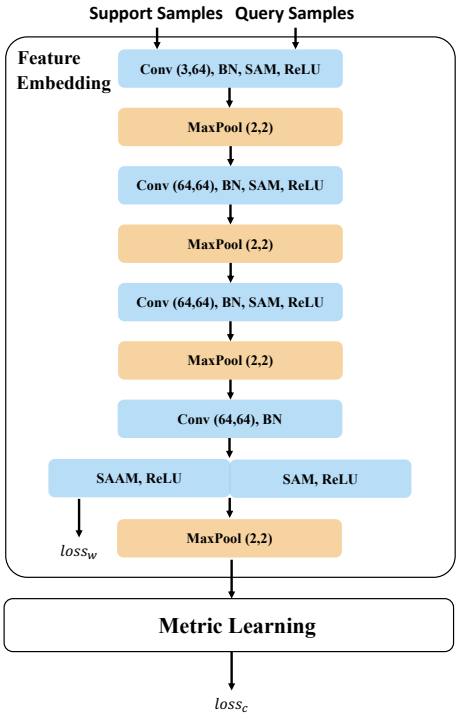

Fig. 3. The architecture of the feature embedding neural network in few-shot image classification. It chooses $ConvNet$ as the basic feature embedding network.

There are two losses in the whole workflow. One is $loss_w$ used to update the visual-semantic network in the semantic alignment mechanism. The other one is $loss_c$ (seen in Eq. 9) used to update the overall network. A hyper-parameter λ is introduced in calculating the overall network loss. Therefore, the network loss would be:

$$loss = loss_c + \lambda \cdot loss_w. \tag{6}$$

D. Zero-shot Image Classification

Zero-shot image classification differs from few-shot image classification in that it produces class embeddings based on side information rather than feature embeddings from labeled samples. Most embedding and metric-learning approaches choose class attribute vectors as side information. These vectors are usually provided by the dataset, describing semantic attributes of each class in the dataset. Here, like our approach in few-shot image classification, we improve class embedding by introducing additional semantic information, the class label embeddings to the network. We achieve so by concatenating the class attribute vectors and the class label embedding and then sending the concatenated vector to the following class embedding network.

We follow the architectures of the class embedding neural network and the feature embedding neural network in most embedding and metric-learning approaches to zero-shot image classification, as shown in Fig. 4.

The class embedding network learns a linear network to create a 1024-dimensional embedding space. Each linear block is comprised of a fully-connected layer, a batch normalization layer, and a ReLU nonlinearity layer. We use the default class attribute vectors offered in the dataset and obtain the class label embeddings using *GloVe* like in few-shot image classification.

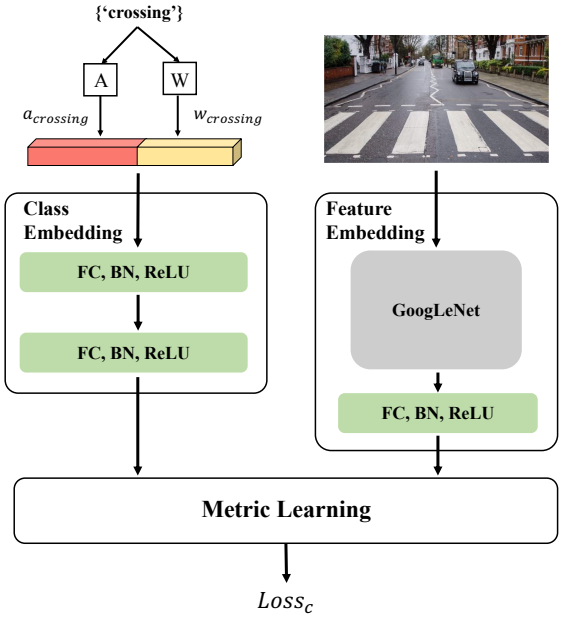

Fig. 4. The architectures of the class embedding neural network and the feature embedding neural network in zero-shot image classification

The feature embedding network utilizes a pre-trained GoogLeNet [47] as the main feature embedding model. In order to match the dimension of the embedding space, we replace its last linear block with a new linear block so that the output size of the feature embedding network can be reshaped to 1024.

During the process of metric learning, the feature embedding of each sample is measured to the class embeddings in the embedding space and then assigned to the nearest one. $loss_c$ is used to update the overall network.

IV. Experiments

We test our approach on one effective embedding and metric-learning approach, namely the prototypical network [4]. In the few-shot image classification task, the improved prototypical network is evaluated on the two widely used datasets: the miniImageNet dataset [40] and the tieredImageNet dataset [48]. In the zero-shot image classification task, the improved prototypical network is evaluated on the CUB dataset [49] and the SUN dataset [50]. All the experiments are implemented using the PyTorch framework [51].

A. Prototypical Network

In this paper, we evaluate our approach on the prototypical network [4]. As a matter of fact, our approach is plug-and-play for most existing embedding and metric-learning approaches. We choose the prototypical network in this paper due to its simplicity.

1) Few-shot Image Classification:

During the process of feature embedding, the prototypical network maps all the support samples and the query samples into a D-dimensional embedding space using the feature embedding neural network. The feature embedding neural network $f : \mathbb{R}^V \rightarrow \mathbb{R}^D$, parameterized by θ_f, is used to learn the embedding space where samples of the same class are clustered closely and samples of different classes are separated widely.

Then, during the process of metric learning, the prototypical network firstly computes the class embedding (or the prototype) for each class in the current episode by averaging its support embeddings. For example, the prototype of class c is computed as:

$$\mathbf{p}_c = \frac{1}{|S_c|} \sum_{(s_i, y_i) \in \mathcal{S}_c} f\left(\mathbf{s}_i\right). \tag{7}$$

Next, the prototypical network produces a distribution over the N classes within the episode for each query feature embedding $f(q_i)$ based on a softmax over the Euclidean distances to the class embeddings (or the prototypes) in the embedding space:

$$\mathbf{p}(y = c|q_j) = \frac{\exp\left(-d\left(f(\mathbf{q}_j), \mathbf{p}_c\right)\right)}{\sum_{c'} \exp\left(-d\left(f(\mathbf{q}_j), \mathbf{p}_{c'}\right)\right)}. \tag{8}$$

In each training iteration, the loss used to update the feature embedding neural network is calculated as:

$$loss_c = \sum_{c=1}^{N} \sum_{j=1}^{Q} \left[d(f(\mathbf{q_j}), \mathbf{p}_c)) + \log \sum_{c'} \exp(-d(f(\mathbf{q_j}), \mathbf{p}_{c'})) \right]. \tag{9}$$

2) Zero-shot Image Classification:

The class embedding process in zero-shot image classification maps the class attribute vectors into a D'-dimensional embedding space via the class embedding neural network $g : \mathbb{R}^A \rightarrow \mathbb{R}^{D'}$ parameterized by θ_g. For example, the prototype of class c is computed as:

$$\mathbf{p}_c = g\left(\mathbf{v}_k\right). \tag{10}$$

Meanwhile, the feature embedding process produces the feature embedding of each sample x_m using the pre-trained deep neural network $f' : \mathbb{R}^{V'} \rightarrow \mathbb{R}^{D'}$. The distance-based distribution for each sample x_m over the N classes and the computation of $loss_c$ follow the Eq. 8 and the Eq. 9 respectively.

B. Datasets

The few-shot image classification task is evaluated on the miniImageNet dataset and the tieredImageNet dataset while the zero-shot image classification task is tested on the CUB dataset and the SUN dataset.

The miniImageNet dataset [40] is actually a subset of the ImageNet-2012 dataset [52], which consists of 100 classes with 600 images in each class. We follow the standard splitting rule given by Vinvals *et al.* [40]: training with 64 classes, validating with 16 classes and testing with 20 classes.

The tieredImageNet dataset [48] is also but a larger subset of the ImageNet-2012 dataset [52], which comprises totally 608 classes. Following the standard splitting rule provided by Ren *et al.* [48], we manage the dataset in the following way: training with 351 classes, validating with 97 classes and testing with 160 classes.

The CUB dataset [49] is a fine-grained image dataset which contains images from 200 bird classes. Each class is annotated with 312 attributes encoding characteristics like color and shape. Following the splitting rule provided by Snell *et al.* [4], we divide the classes into: training with 100 classes, validating with 50 classes, and testing with 50 classes.

The SUN dataset [50] is also a fine-grained image dataset which contains images from 717 types of scenes. Each class is marked with 102 attributes. Following [53], we manage the dataset as: training with 580 classes, validating with 65 classes, and testing with 72 classes.

C. Implementation Details

Few-shot image classification experiments are implemented under 5-way 1-shot and 5-way 5-shot settings. The input image size is 84 $\times$ 84 and no data augmentation is involved during training. The model is trained for 120 epochs, with 4 episodes in each mini-batch and 200 mini-batches in each epoch. We use Adam [54] as the optimizer with an initial learning rate of 10^{-3}. The weight decay is 5×10^{-5}. For the miniImageNet dataset, the learning rate is dropped by half every 12,000 episodes. For the tieredImageNet dataset, the learning rate is dropped by half every 18,000 episodes. We use *GloVe* [12] as the word embedding model and generate all the label embeddings from it. All the class labels are available in *GloVe* and we suggest using a word embedding model that has been pre-trained on large text corpus and has been proved stable on word representation like *GloVe* and *BERT*. Each label embedding has a length of 300 in dimension. The manually-set margin α in computing $loss_w$ is set to be 0.3 under the 5-way 1-shot setting and 0.4 under the 5-way 5-shot setting. The hyper-parameter λ in the overall network loss is set to be 1.0. The overall accuracy is counted as the average testing accuracy of the 16,000 episodes sampled from the testing set. Each episode contains 15 query images for each class in all the training, validation, and test procedures. The number of query images in each episode is constrained by the GPU memory.

Zero-shot image classification experiments are implemented on the two datasets. The input image is resized to 256 $\times$ 256 and there is no data augmentation during training. The feature embedding network generates and reshapes the features to 1024 dimensions. A linear class embedding network produces the 1024-dimensional prototypes in the embedding space based on the concatenated vectors between the class attribute vectors and the class label embeddings. The prototypes are formalized to the unit length. The class attribute vectors provided by the datasets are 312-dimensional in the CUB dataset and are 102-dimensional in the SUN dataset. The class label embeddings generated with the pre-trained word

TABLE I
FEW-SHOT IMAGE CLASSIFICATION ACCURACIES ON THE MINIIMAGENET DATASET. "∼" MEANS "NO REPORTED". ALL THE INCLUDED APPROACHES USE *ConvNet* AS THE BASIC FEATURE EMBEDDING NETWORK.

Paradigm	Model	5-way 1-shot	5-way 5-shot
Memory-based	MetaNet [55]	49.21 ± 0.96	∼
	MM-Net [56]	**53.57 ± 0.48**	66.97 ± 0.35
Optimization-based	MAML [21]	48.70 ± 1.84	63.11 ± 0.92
	Meta-SGD [23]	50.47 ± 1.87	64.03 ± 0.94
	REPTILE [24]	49.97 ± 0.32	65.99 ± 0.58
	BOIL [25]	49.61 ± 0.16	66.45 ± 0.37
Embedding-and-metric-learning-based	MatchingNet [40]	43.44 ± 0.77	60.60 ± 0.71
	Subspace [57]	∼	68.12 ± 0.67
	ProtoNet [4]	49.42 ± 0.78	68.20 ± 0.66
	RelationNet [5]	50.44 ± 0.82	65.32 ± 0.70
	DSN [58]	51.24 ± 0.74	71.02 ± 0.64
	Neg-Margin [59]	52.68 ± 0.76	70.41 ± 0.66
	SSA (Ours)	53.18 ± 0.74	**70.86 ± 0.69**

TABLE II
FEW-SHOT IMAGE CLASSIFICATION ACCURACIES ON THE TIEREDIMAGENET DATASET. "∼" MEANS "NO REPORTED". ALL THE INCLUDED APPROACHES USE *ConvNet* AS THE BASIC FEATURE EMBEDDING NETWORK.

Paradigm	Model	5-way 1-shot	5-way 5-shot
Optimization-based	MAML [21]	49.0 ± 1.8	66.5 ± 0.9
	MAML++ [60]	51.5 ± 0.5	70.6 ± 0.5
Embedding-and-metric-learning-based	Soft k-means [48]	52.39 ± 0.44	69.88 ± 0.20
	Subspace [57]	∼	71.15 ± 0.67
	ProtoNet [4]	53.31 ± 0.89	72.69 ± 0.74
	RelationNet [5]	54.48 ± 0.93	71.32 ± 0.78
	CovaMNet [61]	54.98 ± 0.90	71.51 ± 0.75
	SSA (Ours)	**56.04 ± 0.83**	**74.28 ± 0.72**

embedding model *GloVe* [12] are 300-dimensional during the CUB experiment and are 100-dimensional during the SUN experiment. For both datasets, the networks are trained for 50 epochs, with 10 episodes in each epoch. We use Adam [54] as the optimizer with an initial learning rate of 10^{-4}. The weight decay is 10^{-5}. Each episode contains 5 query images for each class, which is also constrained by the GPU memory.

D. Results

1) Few-shot Image Classification: We evaluate our proposed approach on the prototypical network in few-shot image classification. We then compare the improved version with other existing methods on the miniImageNet dataset shown in Table I and the tieredImageNet dataset shown in Table II.

For fair comparisons, all the methods in the table have the following common ground: (i) data augmentation is not involved during training; (ii) basic feature embedding network is the *ConvNet*; (iii) they are all inductive methods. They can roughly be divided into three categories. As shown in Table I, the first block of the methods contains memory-based methods, the second contains optimization-based methods and the last contains embedding and metric-learning methods.

It is easy to observe the effectiveness of our SAA approach. With SAA, the improved version of the prototypical network performs much better in few-shot image classification on both of the datasets. For example, the performance is 68.20% *v.s.* 70.86%, and 72.69% *v.s.* 74.28% under the 5-way 5-shot setting on the miniImageNet dataset and the tieredImageNet dataset respectively. Furthermore, it compares favorably against most of the other methods coming from the three categories by a large margin.

TABLE III
ZERO-SHOT IMAGE CLASSIFICATION ACCURACIES ON THE CUB DATASET AND THE SUN DATASET. "∼" MEANS "NO REPORTED".

Model	CUB	SUN
ALE [30]	26.9	∼
ESZSL [32]	∼	65.8
SSE [62]	30.4	82.5
SC [63]	44.3	∼
LATEM [64]	45.5	∼
SJE [31]	50.1	∼
DS-SJE [34]	50.4	∼
AS-SJE [34]	50.9	∼
PN [4]	54.6	85.8
SSA (Ours)	**57.8**	**88.6**

We also compare the validation accuracy between the networks with our method and the networks without our method on the miniImageNet dataset, as shown in Fig. 5. In Fig. 6, it shows a fast convergence in our method under both settings of 5-way 1-shot and 5-way 5-shot. We believe that this is because our method could find a better class embedding in a short time.

2) Zero-shot Image Classification: We also evaluate our proposed approach on the prototypical network in zero-shot image classification. Table III shows the comparison between the improved version and other existing methods which have evaluation records in the same setting.

Like in few-shot image classification experiments, there is no data augmentation involved during training and all the methods in the table are inductive methods.

The results show that our proposed approach in zero-shot image classification achieves a significant performance boost compared with other existing methods. It further demonstrates the effectiveness of introducing semantic alignment to the network in zero-shot image classification.

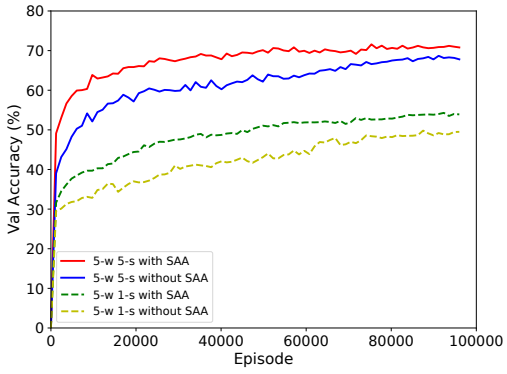

Fig. 5. The validation accuracy on the miniImageNet dataset.

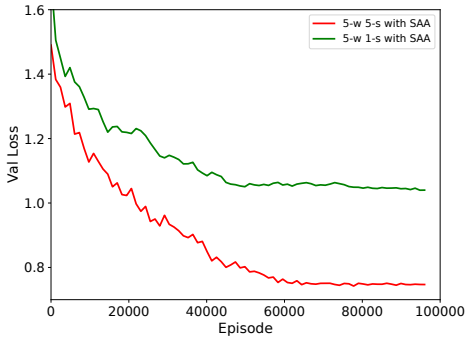

Fig. 6. The validation loss on the miniImageNet dataset.

E. Ablation Study

We study the impacts of sub-modules, the manually-set margin α in $loss_w$, the hyper-parameter λ in $loss$ and the dimension of the label embeddings. Experiments are conducted on the miniImageNet dataset.

1) Impacts of sub-modules: The semantic-aligned attention mechanism consists of two sub-modules and we study their impacts on the performance, as shown in Table IV. Here "AM" refers to the attention mechanism in Section and "SM" refers to the semantic-aligned mechanism described in Section III-B. Additionally, "(1)" means each convolutional block in the network contains both the attention mechanism and the semantic-aligned mechanism. "(2)" means only the last convolutional blocks contain both mechanisms while the first three convolutional blocks only have the attention mechanism, which is used in our network described in Section III-C.

The first thing we could see is that our semantic-aligned attention mechanism, the combination of the attention mechanism and the semantic-aligned mechanism surpasses the attention mechanism or the semantic-aligned mechanism alone in improving feature embedding of the prototypical network. The attention mechanism alone improves performances by 2.14% and 1.28% under the 5-way 1-shot setting and the 5-way 5-shot setting respectively. It indicates the importance of local features in forming accurate class embeddings. The semantic-aligned mechanism alone, however, deteriorates the model. We believe this is because it is more difficult to learn the relationship between the semantic description of the object and the global visual features. With the semantic-aligned attention mechanism, the local visual features are generated so that visual-semantic alignment can be well learned to refine local feature embedding. In this case, performances

TABLE IV
ABLATION STUDY RESULTS ON THE MINIIMAGENET DATASET.

Model	5-way 1-shot	5-way 5-shot
ProtoNet	49.42	68.20
ProtoNet+AM	51.56	69.48
ProtoNet+SM	47.87	67.12
ProtoNet+SAA$^{(1)}$	52.79	70.54
ProtoNet+SAA$^{(2)}$	**53.18**	**70.86**

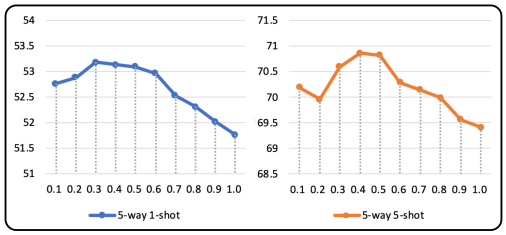

Fig. 7. The impact of α in computing $loss_w$ on few-shot image classification on the miniImageNet dataset.

are improved under both the 5-way 1-shot and 5-way 5-shot settings compared with the network with the attention mechanism alone. Therefore, our semantic-aligned attention mechanism is more effective in feature embedding.

The second thing we could obverse is that introducing semantic alignment into the last layer alone is better than into every single layer of the network. As a matter of fact, many works related to visual-semantic embedding apply semantic alignment in the last layer. It verifies the effectiveness of our network architecture selection in Section 3.3.

2) Impacts of α: α is a hyper-parameter in computing $loss_w$ in the network, so we study the impact of its value on few-shot image classification using the miniImageNet dataset, as shown in Fig. 7. It is easy to observe that the 5-way 1-shot performance is best when α is set to be 0.3 and the optimal α is 0.4 in the 5-way 5-shot performance. We set the hyper-parameter α according to this in our experiments.

As a matter of fact, α controls the margin in exploring the relationship between visual local regions and class label embeddings in a common embedding space. A higher margin might lead to the overfitting problem. This is why there is a performance degradation when α becomes bigger.

3) Impacts of λ: λ is a hyper-parameter in calculating $loss$ in the overall network, and we study its impact on few-shot image classification using the miniImageNet dataset, as shown in Fig. 8. It can be observed that the optimal value of λ is within the range of 0.6 to 1.2 under both circumstances of 5-way 1-shot and 5-way 5-shot.

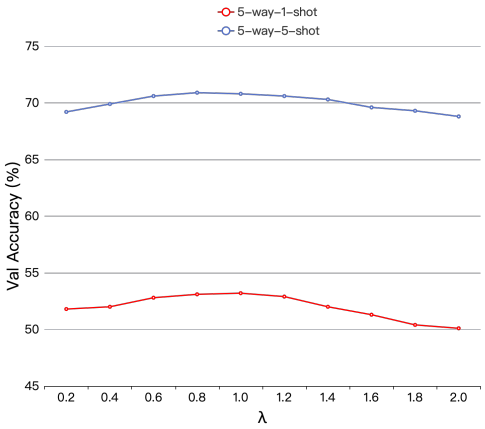

Fig. 8. The impact of λ in computing the overall network loss on few-shot image classification on the miniImageNet dataset.

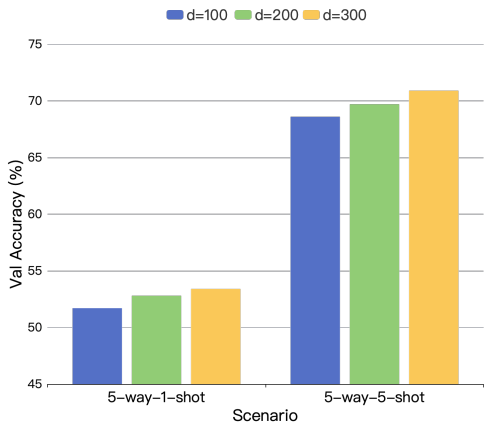

Fig. 9. The impact of the dimension of the label embeddings on few-shot image classification on the miniImageNet dataset.

In fact, λ can be seen as a coefficient controlling the role the semantic-aligned mechanism plays in updating the network. We choose 1.0 in our experiments.

4) Impacts of the dimension of the label embeddings: The pre-trained word embedding model *GloVe* [12] offers three kinds of word embeddings in three different dimensions: 100, 200, and 300. A higher dimension means richer prior semantic knowledge in the word embeddings. We study the impact of the dimension of the label embeddings in our method on few-shot image classification in Fig. 9.

When the dimension becomes bigger, our method performs better in few-shot image classification under both circumstances of 5-way 1-shot and 5-way 5-shot. We choose 300 as the dimension of the label embeddings in our experiments.

F. Computational Complexity

Computational efficiency for the recognition process is essential to real-time applications like autonomous driving. In terms of forwarding time per image for recognition, we conduct a computational comparison between our approach and other applied algorithms in recognition. The experiments are performed on a single GeForce RTX 2080 Ti GPU and the results are shown in Fig. 10. It takes about 18ms for our approach to recognize and respond, which is faster than other selected algorithms. This is because the deep architecture in our approach to deal with few-shot problems is simpler and thus easier to use in cases requiring a high response rate.

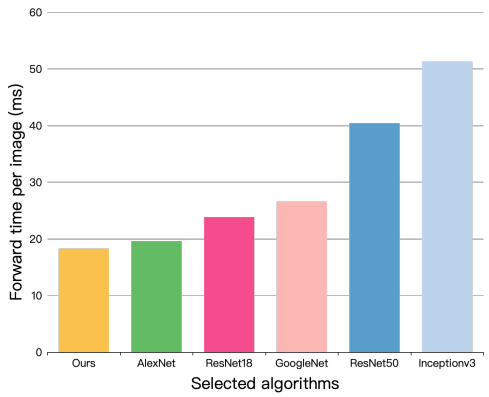

Fig. 10. An analysis of average forward time per image between our approach and other applied algorithms on our machine.

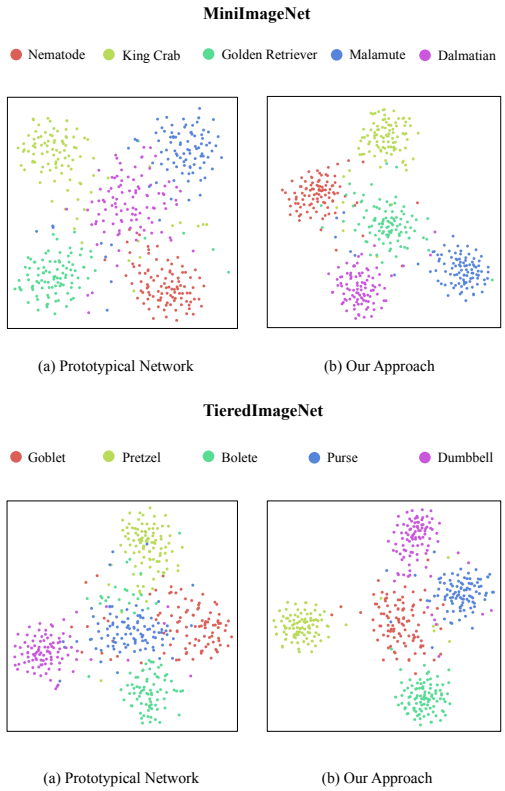

Fig. 11. The t-SNE visualization of feature embedding on the miniImageNet dataset and the tieredImageNet dataset. It is under the 5-way 5-shot setting and 100 support samples are included in each class for a better view. We can see that the effect of introducing semantic-aligned attention mechanism to feature embedding is obvious: making clusters more compact and discriminative from each other.

G. Visualization

Visualization of feature embedding in Fig. 11 using t-SNE [65] perfectly proves the effectiveness of the proposed SAA approach.

As shown in Fig. 11, the features are computed under the 5-way 5-shot setting. The 5 classes selected from the miniImageNet test dataset are “Nematode”, “King Crab”, “Golden Retriever”, “Malamute” and “Dalmatian”. The 5 classes selected from the tieredImageNet test dataset are “Goblet”, “Pretzel”, “Bolete”, “Purse” and “Dumbbell”. To have a better view, 100 support samples are visualized for each class. Fig. 11(a) corresponds to the prototypical network and Fig. 11(b) corresponds to the improved version with our proposed semantic-aligned attention mechanism. As can be seen clearly, the improved version of the prototypical network with SAA has more compact clusters and features in one class are discriminative from ones in other classes. It indicates that our proposed SAA helps find intra-class commonality and improves feature embedding greatly in the model.

V. Conclusion

In this paper, we introduced a Semantic-Aligned Attention (SAA) mechanism for few-shot image classification that can be used in real-time visual recognition applications like autonomous driving. To be more specific, the mechanism highlights pivotal local information with attention mechanism and refines extracted local features with semantic alignment. We further extended the idea to zero-shot image classification by introducing important semantic information in generating the class embedding. Extensive experimental results show that our approach in few-shot image classification lifts the performance on the miniImageNet dataset and the tieredImageNet dataset and our approach in zero-shot image classification boosts the performance on the CUB dataset and the SUN dataset. The t-SNE visualization shows more compact and discriminative clusters in our method and the class activation map visualization shows a refined attentive map offered by our method. They further verify the effectiveness of the proposed approach.

Acknowledgment

We sincerely thank the anonymous reviewers for their kind review of this paper. This work was supported in part by the National Natural Science Foundation of China under Grants (61976049 and U20B2063); the Fundamental Research Funds for the Central Universities under Project (ZYGX2019Z015) and the Sichuan Science and Technology Program, China (2019ZDZX0008 and 2019YFG0533).

References

[1] J. Jin, K. Fu, and C. Zhang, "Traffic sign recognition with hinge loss trained convolutional neural networks," *IEEE Transactions on Intelligent Transportation Systems*, vol. 15, no. 5, pp. 1991–2000, 2014.

[2] H. Peng, X. Wang, H. Wang, and W. Yang, "Recognition of low-resolution logos in vehicle images based on statistical random sparse distribution," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 2, pp. 681–691, 2015.

[3] M. Rezaei and M. Shahidi, "Zero-shot learning and its applications from autonomous vehicles to covid-19 diagnosis: A review," *arXiv preprint arXiv:2004.14143*, 2020.

[4] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," in *Advances in neural information processing systems*, 2017, pp. 4077–4087.

[5] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. Torr, and T. M. Hospedales, "Learning to compare: Relation network for few-shot learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1199–1208.

[6] H. Chen, H. Li, Y. Li, and C. Chen, "Multi-level metric learning for few-shot image recognition," *arXiv preprint arXiv:2103.11383*, 2021.

[7] Y. Zhu, W. Min, and S. Jiang, "Attribute-guided feature learning for few-shot image recognition," *IEEE Transactions on Multimedia*, pp. 1–1, 2020.

[8] B. Oreshkin, P. R. López, and A. Lacoste, "Tadam: Task dependent adaptive metric for improved few-shot learning," in *Advances in Neural Information Processing Systems*, 2018, pp. 721–731.

[9] X. Liu, Z. Ji, Y. Pang, J. Han, and X. Li, "Dgig-net: Dynamic graph-in-graph networks for few-shot human-object interaction," *IEEE Transactions on Cybernetics*, 2021.

[10] H. Xu, G. Qi, J. Li, M. Wang, K. Xu, and H. Gao, "Fine-grained image classification by visual-semantic embedding." in *IJCAI*, 2018, pp. 1043–1049.

[11] B. Wang, Y. Yang, X. Xu, A. Hanjalic, and H. T. Shen, "Adversarial cross-modal retrieval," in *Proceedings of the 2017 ACM on Multimedia Conference*, 2017, pp. 154–162.

[12] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[13] Z. Ji, Y. Fu, J. Guo, Y. Pang, Z. M. Zhang *et al.*, "Stacked semantics-guided attention model for fine-grained zero-shot learning," in *Advances in Neural Information Processing Systems*, 2018, pp. 5995–6004.

[14] H. T. Shen, L. Liu, Y. Yang, X. Xu, Z. Huang, F. Shen, and R. Hong, "Exploiting subspace relation in semantic labels for cross-modal hashing," *IEEE Transactions on Knowledge and Data Engineering*, p. 10.1109/TKDE.2020.2970050, 2020.

[15] A. Uçar, Y. Demir, and C. Güzeliş, "Object recognition and detection with deep learning for autonomous driving applications," *Simulation*, vol. 93, no. 9, pp. 759–769, 2017.

[16] J. Kim, T.-H. Oh, S. Lee, F. Pan, and I. S. Kweon, "Variational prototyping-encoder: One-shot learning with prototypical images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9462–9470.

[17] K. Wong, S. Wang, M. Ren, M. Liang, and R. Urtasun, "Identifying unknown instances for autonomous driving," in *Conference on Robot Learning*. PMLR, 2020, pp. 384–393.

[18] Z. Chen, Y. Fu, K. Chen, and Y.-G. Jiang, "Image block augmentation for one-shot learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 3379–3386.

[19] E. Schwartz, L. Karlinsky, J. Shtok, S. Harary, M. Marder, A. Kumar, R. Feris, R. Giryes, and A. Bronstein, "Delta-encoder: an effective sample synthesis method for few-shot object recognition," in *Advances in Neural Information Processing Systems*, 2018, pp. 2845–2855.

[20] S. Tsutsui, Y. Fu, and D. Crandall, "Meta-reinforced synthetic data for one-shot fine-grained visual recognition," in *Advances in Neural Information Processing Systems*, 2019, pp. 3057–3066.

[21] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017, pp. 1126–1135.

[22] M. A. Jamal and G.-J. Qi, "Task agnostic meta-learning for few-shot learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 11 719–11 727.

[23] Z. Li, F. Zhou, F. Chen, and H. Li, "Meta-sgd: Learning to learn quickly for few-shot learning," *arXiv preprint arXiv:1707.09835*, 2017.

[24] A. Nichol, J. Achiam, and J. Schulman, "On first-order meta-learning algorithms," *arXiv preprint arXiv:1803.02999*, 2018.

[25] J. Oh, H. Yoo, C. Kim, and S.-Y. Yun, "Boil: Towards representation change for few-shot learning," in *Proc. Int. Conf. Learn. Represent.(ICLR)*, 2021.

[26] T. Gao, X. Han, Z. Liu, and M. Sun, "Hybrid attention-based prototypical networks for noisy few-shot relation classification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 6407–6414.

[27] M. Ren, R. Liao, E. Fetaya, and R. Zemel, "Incremental few-shot learning with attention attractor networks," in *Advances in Neural Information Processing Systems*, 2019, pp. 5276–5286.

[28] F. Hao, F. He, J. Cheng, L. Wang, J. Cao, and D. Tao, "Collect and select: Semantic alignment metric learning for few-shot learning," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 8460–8469.

[29] H. Li, R. Tao, J. Li, H. Qin, Y. Ding, S. Wang, and X. Liu, "Multi-pretext attention network for few-shot learning with self-supervision," *arXiv preprint arXiv:2103.05985*, 2021.

[30] Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid, "Label-embedding for attribute-based classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 819–826.

[31] Z. Akata, S. Reed, D. Walter, H. Lee, and B. Schiele, "Evaluation of output embeddings for fine-grained image classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 2927–2936.

[32] B. Romera-Paredes and P. Torr, "An embarrassingly simple approach to zero-shot learning," in *International Conference on Machine Learning*, 2015, pp. 2152–2161.

[33] M. Bucher, S. Herbin, and F. Jurie, "Improving semantic embedding consistency by metric learning for zero-shot classiffication," in *European Conference on Computer Vision*. Springer, 2016, pp. 730–746.

[34] S. Reed, Z. Akata, H. Lee, and B. Schiele, "Learning deep representations of fine-grained visual descriptions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 49–58.

[35] Y. Yu, Z. Ji, J. Han, and Z. Zhang, "Episode-based prototype generating network for zero-shot learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 14 035–14 044.

[36] X. Xu, H. Lu, J. Song, Y. Yang, H. T. Shen, and X. Li, "Ternary adversarial networks with self-supervision for zero-shot cross-modal retrieval," vol. 50, no. 6, pp. 2400–2413, 2020.

[37] S. Rahman, S. Khan, and N. Barnes, "Deep0tag: Deep multiple instance learning for zero-shot image tagging," *IEEE Transactions on Multimedia*, vol. 22, no. 1, pp. 242–255, 2020.

[38] R. Socher, M. Ganjoo, C. D. Manning, and A. Ng, "Zero-shot learning through cross-modal transfer," in *Advances in neural information processing systems*, 2013, pp. 935–943.

[39] L. Zhang, T. Xiang, and S. Gong, "Learning a deep embedding model for zero-shot learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2021–2030.

[40] O. Vinyals, C. Blundell, T. Lillicrap, D. Wierstra *et al.*, "Matching networks for one shot learning," in *Advances in neural information processing systems*, 2016, pp. 3630–3638.

[41] S. Woo, J. Park, J.-Y. Lee, and I. So Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 3–19.

[42] S. Zagoruyko and N. Komodakis, "Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer," *arXiv preprint arXiv:1612.03928*, 2016.

[43] U. Seiffert, "Multiple layer perceptron training using genetic algorithms." in *ESANN*. Citeseer, 2001, pp. 159–164.

[44] D. J. Finney, *Probit analysis: a statistical treatment of the sigmoid response curve*. Cambridge university press, cambridge, 1952.

[45] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, M. Ranzato, and T. Mikolov, "Devise: A deep visual-semantic embedding model," in *Advances in neural information processing systems*, 2013, pp. 2121–2129.

[46] J. Weston, S. Bengio, and N. Usunier, "Large scale image annotation: learning to rank with joint word-image embeddings," *Machine learning*, vol. 81, no. 1, pp. 21–35, 2010.

[47] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[48] M. Ren, E. Triantafillou, S. Ravi, J. Snell, K. Swersky, J. B. Tenenbaum, H. Larochelle, and R. S. Zemel, "Meta-learning for semi-supervised few-shot classification," *arXiv preprint arXiv:1803.00676*, 2018.

[49] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The caltech-ucsd birds-200-2011 dataset," 2011.

[50] G. Patterson and J. Hays, "Sun attribute database: Discovering, annotating, and recognizing scene attributes," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 2751–2758.

[51] A. Paszke, S. Gross, S. Chintala, and G. Chanan, "Pytorch: Tensors and dynamic neural networks in python with strong gpu acceleration," *PyTorch: Tensors and dynamic neural networks in Python with strong GPU acceleration*, vol. 6, 2017.

[52] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, no. 3, pp. 211–252, 2015.

[53] C. H. Lampert, H. Nickisch, and S. Harmeling, "Attribute-based classification for zero-shot visual object categorization," *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 3, pp. 453–465, 2013.

[54] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[55] T. Munkhdalai and H. Yu, "Meta networks," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017, pp. 2554–2563.

[56] Q. Cai, Y. Pan, T. Yao, C. Yan, and T. Mei, "Memory matching networks for one-shot image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4080–4088.

[57] C. Simon, P. Koniusz, and M. Harandi, "Projective subspace networks for few-shot learning," 2018.

[58] W. Li, L. Wang, J. Xu, J. Huo, Y. Gao, and J. Luo, "Revisiting local descriptor based image-to-class measure for few-shot learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7260–7268.

[59] B. Liu, Y. Cao, Y. Lin, Q. Li, Z. Zhang, M. Long, and H. Hu, "Negative margin matters: Understanding margin in few-shot classification," in *European Conference on Computer Vision*. Springer, 2020, pp. 438–455.

[60] A. Antoniou, H. Edwards, and A. Storkey, "How to train your maml," 2018.

[61] W. Li, J. Xu, J. Huo, L. Wang, Y. Gao, and J. Luo, "Distribution consistency based covariance metric networks for few-shot learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 8642–8649.

[62] Z. Zhang and V. Saligrama, "Zero-shot learning via semantic similarity embedding," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 4166–4174.

[63] R. Liao, A. Schwing, R. Zemel, and R. Urtasun, "Learning deep parsimonious representations," in *Advances in Neural Information Processing Systems*, 2016, pp. 5076–5084.

[64] Y. Xian, Z. Akata, G. Sharma, Q. Nguyen, M. Hein, and B. Schiele, "Latent embeddings for zero-shot classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 69–77.

[65] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.