This ICCV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

# Collect and Select: Semantic Alignment Metric Learning for Few-Shot Learning

Fusheng Hao[1,2], Fengxiang He[3], Jun Cheng[1,2]*, Lei Wang[1,2], Jianzhong Cao[4], and Dacheng Tao[3]

[1]CAS Key Laboratory of Human-Machine Intelligence-Synergy Systems
Shenzhen Institutes of Advanced Technology, CAS, China
[2]The Chinese University of Hong Kong, Hong Kong, China
[3]UBTECH Sydney AI Centre, School of Computer Science, Faculty of Engineering
The University of Sydney, Darlington, NSW 2008, Australia
[4]Xi'an Institute of Optics and Precision Mechanics, CAS, China

{fs.hao, jun.cheng, lei.wang1}@siat.ac.cn, {fengxiang.he, dacheng.tao}@sydney.edu.au,
cjz@opt.ac.cn

## Abstract

*Few-shot learning aims to learn latent patterns from few training examples and has shown promises in practice. However, directly calculating the distances between the query image and support image in existing methods may cause ambiguity because dominant objects can locate anywhere on images. To address this issue, this paper proposes a Semantic Alignment Metric Learning (SAML) method for few-shot learning that aligns the semantically relevant dominant objects through a "collect-and-select" strategy. Specifically, we first calculate a relation matrix (RM) to "collect" the distances of each local region pairs of the 3D tensor extracted from a query image and the mean tensor of the support images. Then, the attention technique is adapted to "select" the semantically relevant pairs and put more weights on them. Afterwards, a multi-layer perceptron (MLP) is utilized to map the reweighted RMs to their corresponding similarity scores. Theoretical analysis demonstrates the generalization ability of SAML and gives a theoretical guarantee. Empirical results demonstrate that semantic alignment is achieved. Extensive experiments on benchmark datasets validate the strengths of the proposed approach and demonstrate that SAML significantly outperforms the current state-of-the-art methods. The source code is available at https://github.com/haofusheng/SAML.*

## 1. Introduction

Few-shot learning aims to learn knowledge from few training examples [21], in contrast to conventional methods which usually need large-scale datasets (such as ImageNet [7]). It addresses the problem that collecting such large amounts of data is extremely time-consuming and sometimes unrealistic in practice [37].

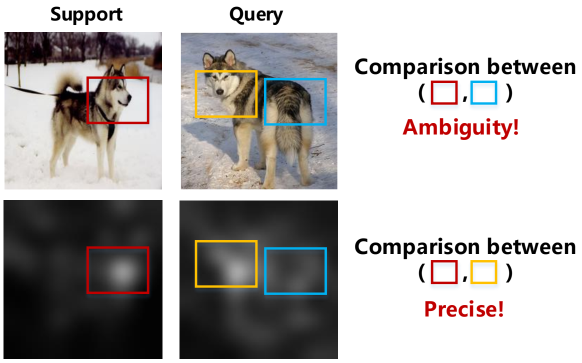

Figure 1. The two images shown belong to the same category, while the key objects (dogs) appear in different locations. Directly calculating the distance between the two images according to the spatial indices introduces ambiguity that pairs between the dog's head (red boxes) and the dog's tail (blue boxes). The proposed method SAML aligns the local regions with the same semantic information (see the comparison between red and yellow boxes).

Recently, the features extracted from images by neural networks have demonstrated profound representation ability in many computer vision tasks [42, 20, 23, 14]. Based on the 3D tensors extracted from the images, metric learning methods have significantly prompted the frontier of few-shot learning. Specifically, metric learning first calculates the distance between the 3D tensors respectively extracted from the query and support images and then learn a classifier based on the distances.

However, most existing methods [48, 37, 16] calculate the distance metric between each tensor pair directly ac-

*Corresponding author.

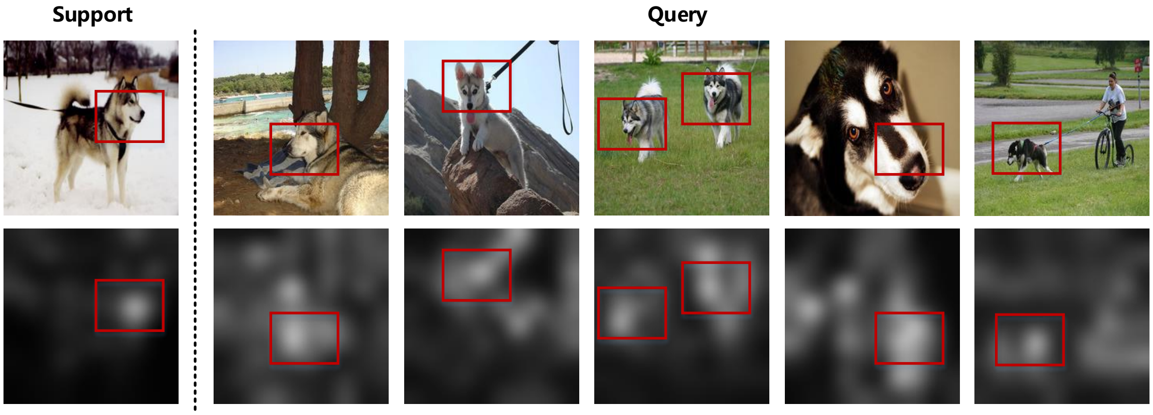

Figure 2. Illustrations of semantically relevant local regions. They demonstrate that the semantic alignment is realized by SAML.

cording to the element index. This direct compare may introduce severe ambiguity because dominant objects can locate anywhere on images. Therefore, the dominant object in one image is probably compared with the semantically irrelevant local region of the other image (please see the comparisons of red and blue boxes in Figure 1).

To address this issue, this paper proposes a Semantic Alignment Metric Learning (SAML) method to align the semantically relevant local regions on images through a "collect-and-select" strategy. Specifically, SAML first "collects" the distances of all local region pairs from query and support images in a relation matrix (RM). Each local region is represented by a vector in the 3D tensor extracted from the corresponding image by a convolutional neural network (CNN). Afterwards, SAML "selects" the semantically relevant local region pairs and reweighs them according to the relevance by employing the attention technique. The relevance-reweighted RM is then processed by a multi-layer perceptron (MLP) to calculate a similarity score in order to determine whether the query image is from the support class.

Theoretical analysis evaluates the generalization ability of SAML and provides an $\mathcal{O}(1/\sqrt{N})$ generalization bound with no explicit dependence on the parameter sizes of the embedding network and MLP ($N$ is the number of episodes in the training sample set). Extensive experiments on two standard benchmark datasets CUB [40] and miniImageNet [34, 31] demonstrate that SAML significantly outperforms the state-of-the-art methods.

## 2. Related Works

Existing works for few-shot learning are mainly from the following four categories: metric learning, meta-learning, hallucination, and attention-based.

**Metric learning:** Metric learning-based approaches share the same paradigm: (1) map all images (including both support and query ones) into a representation space by embedding networks and compute the representation of each support category; (2) calculate the distances of each query image to all support classes; and (3) assign each query instance to the support class with the closest distance to itself.

Existing approaches mainly focus on one of the first two steps, since the third step is relatively well-developed: (1) the design of embedding networks has evolved from early siamese neural networks [19] to rapid adaptation with conditionally shifted neurons [27] and memory matching networks [5]. Recently, to better capture Geometrical information, 3D tensors are introduced by [48, 37]. The computation of class representations can be dated back to [16] which uses the mean of embedding deep features for each support category as its representation. Recently, Qiao et al. present to predict the representation of each novel support class from the activations in a pre-trained neural network [33]; and (2) the design of distance metrics to perform classification stems from cosine distance [31] and Euclidean distance [16] to more advanced distance measurements, such as the one calculated by Graph Neural Networks [39].

A significant problem for most existing works is that the distances are obtained by calculating straight according to the element indices [48, 37] and often introduce severe ambiguity that compares semantically irrelevant parts.

**Meta-learning:** Meta-learning-based methods learn the learning algorithm itself. Ravi et al. present an LSTM-based meta-learner, which learns the exact optimization algorithm in order to train a neural network classifier in the few-shot regime [34]. Finn et al. design MAML to train a meta-learner that provides good parameter initialization such that only a small number of updates can lead to fast learning on novel tasks [8]. Meta-SGD adjusts the update direction and the learning rate for fast adaption to new tasks [22]. Nevertheless, these methods often need costly higher-order gradients which may lead to fail-

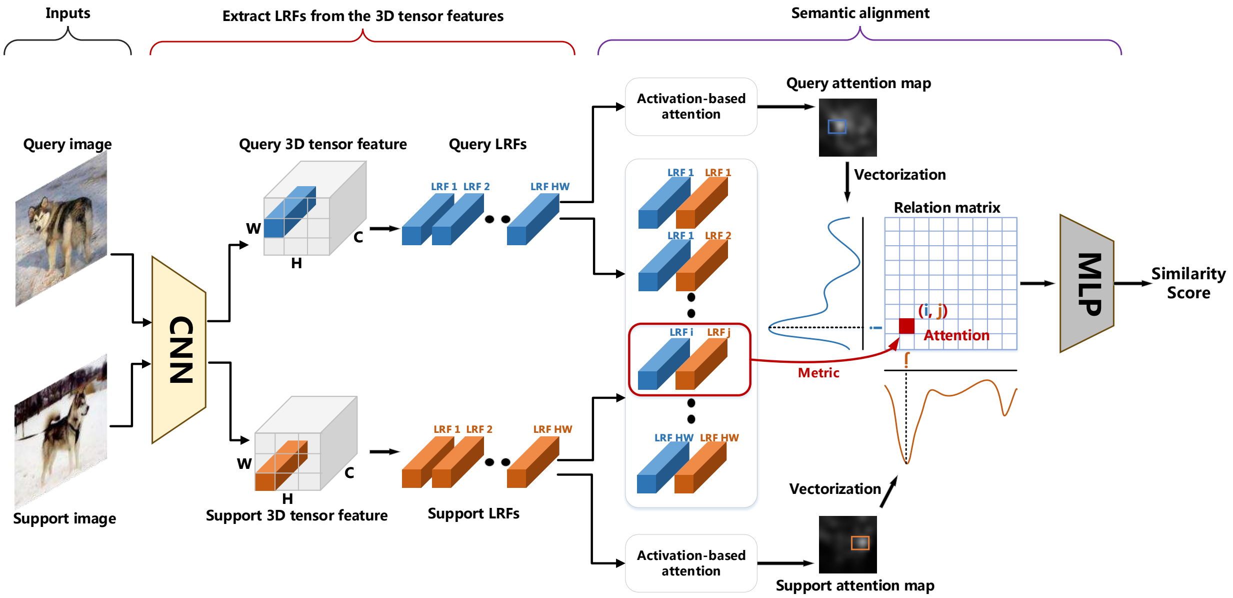

Figure 3. Flowchart of the proposed method SAML.

ure when scales to deeper network architectures as shown in [26]. Correspondingly, Mishra et al. combine temporal convolutions and soft attention to propose generic meta-learner architectures that scale to deeper network architectures [26]. In addition, first-order optimization [29] and latent embedding optimization [36] techniques have been proposed to solve the issues.

**Hallucination:** Hallucination-based methods tackle few-shot learning by increasing the number of labeled instances in each novel category via two different directions. An early work by Bharath et al. applies category-independent transformations to generate as many new instances for each novel category as possible [13]. Its subsequent development [43] exploits recent progress in meta-learning that jointly optimizes a meta-learner and a hallucinator to produce high-quality training instances. Zhang et al. propose MetaGAN to generate samples indistinguishable from the true data sampled from a specific task [47] and Akshay et al. regard the generated fake instances as a strong regularizer [24].

**Attention-based:** Recently, the attention mechanism has been introduced into the few-shot regime. For example, Yan et al. [44] utilize the spatial attention to localize relevant object regions and the task attention to select similar training data for label prediction, and thus present a dual-attention network based on the two attention mechanisms. Ren et al. [35] propose to regularize the learning of novel classes by an Attention Attractor Networks. Hu et al. [15] propose an Attention-based Multi-Context Guiding (A-MCG) network, which integrates multi-scale context features between support and query branches, enforcing better guidance from the support set. By contrast, we adopt the attention mechanism to "select" semantically relevant regions.

## 3. Semantic Alignment Metric Learning

This section presents our proposed method SAML. We start by reviewing the problem definition for few-shot learning, before describing the image embedding. Then, we describe the "collect-and-select" for the semantic alignment. Finally, two instantiations of metrics are provided.

### 3.1. Problem Definition

In this work, we focus on the $M$-way $K$-shot problem, where $M$ is the number of categories and $K$ is the number of the examples in each categorise ($K$ is usually a small integer, such as 1 or 5).

Few-shot learning datasets are constituted by three parts: training set, validation set, and test set, whose label spaces are disjoint with each other (e.g., a category seen during training is not seen during validation or test). Generally, every set contains abundant categories and examples that are significantly larger than $M$ and $K$. Since proposed by [31], the three sets are usually divided into many episodes, each of which contains a *support set* $S = \{(x_i, y_i)|i = 1 \ldots MK, y_i \in \{1, \ldots, M\}\}$ and a *query set* $Q = \{(\tilde{x}_j, \tilde{x}_j)|j = 1 \ldots MT, y_j \in \{1, \ldots, M\}\}$. Both the *support set* and the *query set* are randomly drawn from the training/validation/test set. Additionally, $S$ and $Q$ are disjoint ($S \cap Q = \varnothing$), while share the same label space.

To simulate the real-world scenarios of few-shot learning, all training, validation, and test procedures are implemented on episodes. For example, in each training iteration, an episode is randomly sampled from the training set to update the learnable parameters. This procedure repeats many times until the model converges to a stable state. The validation and test on episodes are similar.

### 3.2. Image Embedding

By convolving each image $x_i$ through a neural network, we can obtain a 3D tensor $f_\Theta(x_i) \in \Omega \subset \mathbb{R}^{C\times H\times W}$ to represent this image, where $f_\Theta$ is the hypothesis function learned by the neural network, $\Theta$ is the parameter of the neural network, $\Omega$ is the representation space formed by all 3D tensors, and $C$, $H$, and $W$ are respectively the lengths of the three dimensions of the tensor. By this way, we embed all images to a representation space. There are $H \times W$ $C$-dimensional cells in each 3D tensor, each of which is a *local region feature* (LRF) of a region in the corresponding image (it is also the receptive field). Compared with 1D [16, 31] or features with other dimensions, 3D tensors can better capture geometrical information, and thus a common choice in metric learning-based few-shot learning methods. Image embedding can be realized through many neural networks. For the details of our embedding network, please refer to Section 5.1.

There are $K$ images from each support class in an episode. When $K > 1$, an important task is to calculate a representation for the support class from the 3D tensors of the $K$ single images. In this paper, we utilize the empirical mean of the $K$ 3D tensors to be the representation of the corresponding support class:

$$c_m = \frac{1}{|S_m|} \sum_{(x_i,y_i)\in S_m} f_\Theta(x_i), \tag{3.1}$$

where $c_m$ is the class representation of the $m$-th support class, $S_m$ is the support set of class $m$ in the episode, and $|S_m|$ is the number of examples in $S_m$. The support class representation $c_m$ also locates in the representation space $\Omega$: $c_m \in \Omega \subset \mathbb{R}^{C\times H\times W}$. Similar with the representations of single images, each class representation $c_m$ is also constituted by $H \times W$ $C$-dimensional features as LRFs.

For the convenience of explanations, the $H \times W$ LRFs of the representation $c_m$ is denoted as $\left\{o_m^1, \ldots, o_m^{HW}\right\}$. Similarly, the LRFs of the 3D representation $f_\Theta\left(\tilde{x}_n\right)$ of each query image $\tilde{x}_n$ are $\left\{\hat{o}^1, \ldots, \hat{o}^{HW}\right\}$.

### 3.3. Collect and Select for Semantic Alignment

The local regions in an image that determine its class can locate anywhere. For example, for dog images, the determinant local regions that contain dogs may locate in the top right corner in one image and in the central area in another. Therefore, directly calculating the distance between them according to the location indices may pair semantically irrelevant local regions and may lead to severe ambiguity [48, 37] (see Figure 1). Thus semantically aligning the representations is an important task to calculate the distances.

To address this issue, we propose a "collect-and-select" strategy to realize the semantic alignment. Specifically, we calculate the metric distances of all LRF pairs constituted by one LRF from the query image and one from the support class. All distance values are collected in a relation matrix (RM) and are located according to the spatial indices. For example, the $(i, j)$-element $r_{ij}$ of the RM $R$ is the distance defined by some metric $g$ between the $i$-th LRF $\hat{o}^i$ of the query image $\tilde{x}_n$ and the $j$-th LRF $o_m^j$ of the support class representation:

$$\begin{aligned} r_{ij} = g(\hat{o}^i, o_m^j) \in \mathbb{R}, \\ \forall i, j \in \{1, \ldots, HW\}, \end{aligned} \tag{3.2}$$

where $r_{ij}$ is a scalar reflecting the similarity of two LRFs. The metric can vary according to various scenarios. Discussions regarding the instances of the metric is provided in Section 3.4. The RM carries all information of the similarity between the query image and the support class. The process of obtaining the RM is called the "collect" phase of the strategy.

The RM contains the distances of semantically irrelevant local regions. The attention technique strengthens key objects, while suppresses the background [46, 4, 17]. By employing the attention technique, we can pay more attention to semantically relevant LRF pairs. In this paper, we chose the activation-based attention [46], where the norm of each LRF is defined as the attention value $a$:

$$\begin{aligned} a(\hat{o}^i) = \|\hat{o}^i\|, a(o_m^j) = \|o_m^j\|, \\ \forall i, j \in \{1, \ldots, HW\}. \end{aligned} \tag{3.3}$$

We reweight the distance $r_{ij}$ by

$$r'_{ij} = a(\hat{o}^i) a(o_m^j) r_{ij}. \tag{3.4}$$

By this way, the distances of semantically irrelevant local regions are suppressed; meanwhile, the distances of semantically relevant local regions are enhanced. Thus, we realize the semantic alignment. The process of adopting the attention technique to find the semantically relevant local regions is called the "select" phase of the strategy. Afterwards, the reweighted RM $R'$ is fed to an MLP to calculate a similarity score (a factor) to perform further classification:

$$s_m = MLP_\Phi(R'), \tag{3.5}$$

where $\Phi$ represents the learnable parameters of the MLP.

For each query instance, there are $M$ similarity score $s_m$ that respectively express the similarity of the query instance with all support class. Link all $M$ similarity score to constitute a vector, we form a discriminative function. To perform the final classification, we employ the softmax function to calculate probability $p_m$ of the test example $\tilde{x}_n$ assigned to the $m$-th class:

$$p_m = \frac{e^{s_m}}{\sum_{i=1}^{M} e^{s_i}}. \tag{3.6}$$

Based on the probabilities, we further define the loss function:

$$L = \frac{-1}{MT} \sum_{n=1}^{MT} \sum_{m=1}^{M} I(\tilde{y}_n = m) ln(p_m), \tag{3.7}$$

where $I(\cdot)$ is the indicator function that equals one if its arguments is true and zero otherwise, $\tilde{y}_n$ is the label of $\tilde{x}_n$, and $T$ is the number of the instances in each query set.

### 3.4. Instantiations of Metrics

The metric function in the previous subsection has many choices. This paper implements experiments with two simple metrics: cosine metric and Gaussian metric. The experiments show that simple metrics are enough to perform well.

**Cosine metric:** Cosine distance is defined as the cosine of the angle between two features:

$$g(\hat{o}^i, o^j_m) = \cos(\theta_{ij}) = \frac{< \hat{o}^i, o^j_m >}{\|\hat{o}^i\| \cdot \|o^j_m\|}, \tag{3.8}$$

$$\forall i, j \in \{1, \ldots, HW\}, \tag{3.9}$$

where $\theta_{ij}$ is the angle between $\hat{o}^i$ and $o^j_m$, $< \cdot, \cdot >$ is the inner product, and $\| \cdot \|$ is the norm. It is effective for the face verification [41] and image classification [12].

**Gaussian metric:** Gaussian function can also be taken as a choice of $g$:

$$g(\hat{o}^i, o^j_m) = e^{\hat{o}^i \cdot o^j_m}, \quad \forall i, j \in \{1, \ldots, HW\}. \tag{3.10}$$

## 4. Theoretical Analysis

This section studies the generalization abilities of our method in term of the size of the training samples from the theoretical perspective. We first present an upper bound for the covering number (covering bound) of the proposed model. The covering bound controls the magnitude of the complexity of the hypothesis space induced by our proposed method. We then obtain an upper bound on the generalization error (generalization bound) of the proposed method. The generalization bound provides a theoretical guarantee for our method.

Few-shot learning can be modelled as a binary classification problem. Specifically, each episode is an example; the query and support images are instances and the label is whether they are from the same class. In this section, we use theories for the binary classification to evaluate our method. Thereby, we give a theoretical guarantee that how many episodes in the training sample set is enough.

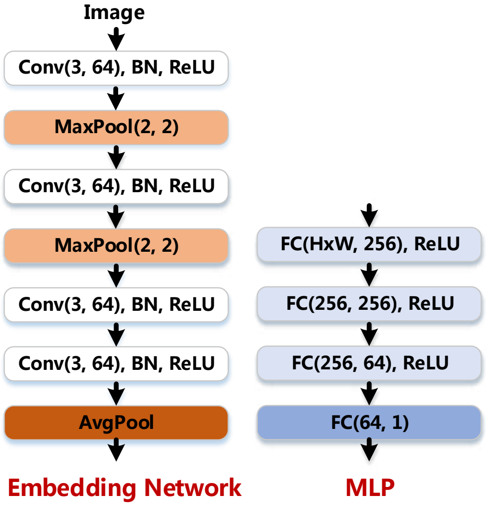

Figure 4. The network architectures of the embedding network and the MLP.

As shown in Figure 4, our model involves two neural networks, the embedding network and the MLP. They are connected by a fixed operation which does not influence the hypothesis complexity. Specifically, the operation calculated the representation of each support class and calculate the relation matrix induced from the query instance and the class representation. Suppose the input to the proposed model is $X$. The embedding network is constituted by four convolutional layers, two max-pooling layers, and an average pooling layer. We denote them respectively by weight matrices $A_1$, $A_2$, $A_3$, and $A_4$ and nonlinearities $\sigma_1$, $\sigma_2$, $\sigma_3$, and $\sigma'_4$. Correspondingly, the output of the embedding network can be expressed as

$$F^E(X) = \sigma'_4 \left(A_4\sigma_3 \left(A_3\sigma_2 \left(A_2\sigma_1 \left(A_1X\right)\right)\right)\right). \tag{4.1}$$

The collection of relation matrix and the attention technique is a fixed nonlinear operation. Here, we denote it as $\sigma_f$. Additionally, we express the MLP by weight matrices $A_5$, $A_6$, $A_7$, and $A_8$ and nonlinearities $\sigma_5$, $\sigma_6$, and $\sigma_7$. The final output of our proposed algorithm is thus:

$$F(X) = A_8\sigma_7 \left(A_7\sigma_6 \left(A_6\sigma_5 \left(A_5 O_f F^E(X)\right)\right)\right). \tag{4.2}$$

For the brevity in the following theorem, we define $\sigma_4 = \sigma_f\sigma'_4$. Suppose the hypothesis space of the output classifiers of our model is $\mathcal{H}$. Then, we can obtain the following theorem.

**Theorem 1** (Covering bound)**.** *Suppose the Lipschitz constant of the $i$-th nonlinearity $\sigma_i$ is $\rho_i$ and the Lipschitz constant of the operation $\sigma_f$ is $\rho_f$. Suppose the spectral norm of* *each weight matrix is bounded: $\|A_i\|_\sigma \leq s_i$, $i = 1, \dots, 8$. Also, suppose each weight matrix $A_i$ has a reference matrix $M_i$, which is satisfied that $\|A_i - M_i\|_\sigma \leq b_i$, $i = 1, \dots, 8$. Then, the $\varepsilon$-covering number satisfies that*

$$\begin{aligned} &\log \mathcal{N}\left(\mathcal{N}, \varepsilon, \|\cdot\|_2\right) \\ \leq &\frac{\log\left(2W^2\right)\|X\|_2^2}{\varepsilon^2}\left(s_8 \prod_{i=1}^{7} s_i \rho_i\right)^2 \left(\sum_{i=1}^{8} \frac{b_i^{2/3}}{s_i^{2/3}}\right)^3 . \end{aligned} \quad (4.3)$$

*and $W$ is the largest dimension of the feature maps throughout the algorithm.*

A detailed proof is omitted here and given in the Appendix. Based on the covering bound, we can obtain the following theorem. For the brevity, we denote the right-hand-side (RHS) of Eq. (4.3) as $\frac{R^2}{\varepsilon^2}$. Also, we define the expected risk and empirical risk respectively as

$$\mathcal{R}(F) = \mathbb{E}_{X,Y} l(F(X), Y), \quad (4.4)$$

$$\hat{\mathcal{R}}(F) = \frac{1}{N} \sum_{n=1}^{N} l(F(X_n), Y_n), \quad (4.5)$$

where $(X, Y)$ is a feature-label pair, $N$ is the training sample size, and $l$ is the loss function.

**Theorem 2** (Generalization Bound)**.** *For any real $\delta \in (0, 1)$, with probability at least $1 - \delta$, the following inequality holds for any hypothesis $F_\theta \in \mathcal{N}$:*

$$\begin{aligned} \mathcal{R}(F_\theta) \leq &\hat{\mathcal{R}}(F_\theta) + \frac{24R}{N}\left[1 + \log\left(\frac{N}{3R}\right)\right] \\ &+ 3\sqrt{\frac{\log \frac{2}{\delta}}{2N}}, \end{aligned} \quad (4.6)$$

*where $N$ is the training sample size.*

Theorem 2 can be directly obtained by applying Theorem 1 to two classic results in learning theory which are omitted here but provided in the Appendix. A detailed proof is also provided in the Appendix. Eq. (4.6) gives an $O\left(\frac{1}{\sqrt{N}}\right)$ generalization bound for our proposed algorithm. It provides a theoretical guarantee for our proposed method.

## 5. Experiments

This section introduces the experimental settings, ablation studies, and comparisons with the state-of-the-art methods.

### 5.1. Experimental Settings

**Datasets:** The miniImageNet dataset is a subset of ImageNet [7] comprised of 100 categories, each of which contains 600 labeled instances. We adopt the common split to get 64, 16, and 20 categories for training, validation and test, respectively. The CUB dataset was initially designed for fine-grained classification and is comprised of $11,788$ instances of birds over 200 species. We randomly split the dataset into 100 training, 50 validation, and 50 test categories. For both miniImageNet and CUB, the images are resized to $224 \times 224$, and no data augmentations are adopted. For a more intuitive understanding of the two datasets, some images are shown in the Appendix.

| Models | Image Size | 5way-1shot | 5way-5shot |
|---|---|---|---|
| RelationNet [37] | 84 × 84 | 50.44% | 65.32% |
|  | 224 × 224 | 50.16% | 65.98% |
| **SAML** (ours) | 84 × 84 | 52.22% | 66.49% |
|  | 224 × 224 | **56.68%** | **71.34%** |

Table 1. The effect of image size on the performance of few-shot classification. Experiments are conducted on miniImageNet.

| Metric Functions | 5way-1shot | 5way-5shot |
|---|---|---|
| Gaussian | 52.35±0.40% | 68.54±0.46% |
| Cosine | 49.52±0.42% | 62.82±0.45% |
| Guassian + Attention | 56.40±0.48% | 71.28±0.39% |
| **Cosine + Attention** | **56.68±0.40%** | **71.34±0.41%** |

Table 2. The effect of different metric functions on the few-shot classification accuracies. Experiments are conducted on miniImageNet.

**Networks:** The details of our embedding network and MLP are illustrated in Figure 4. Since the embedding network is our focus and to perform fair comparisons, our embedding network shares a similar backbone to [31, 37], while is still with some minor changes to obtain enough LRFs. For example, only the first two max pooling layers are preserved, and the last max pooling layer is replaced by the average pooling layer. The stride of the average pooling layer for miniImageNet and CUB is set to $5$, resulting in $100$ LRFs. In addition, the similarity score is often limited within the range $0$ to $1$, which is realized by adding the sigmoid function after the last fully connected layer. Here, we omit the commonly used sigmoid function for the MLP.

**Implementation details:** We take the cosine metric as an example to introduce the implementation details of SAML. The overall flowchart is shown in Figure 3. After embedding all the support images and query images, we merge their spatial dimensions. SAML can be directly implemented by performing matrix product on the reshaped LRFs, as shown in Figure 8 in the Appendix. All experiments are conducted under the PyTorch framework [1]. We use Adam [18] with an initial learning rate of $10^{-3}$, which is halved every $2,000$ episodes. The total number of training episodes is $20,000$. Note that, $T = 15$ query images per class are tested in every test episode.

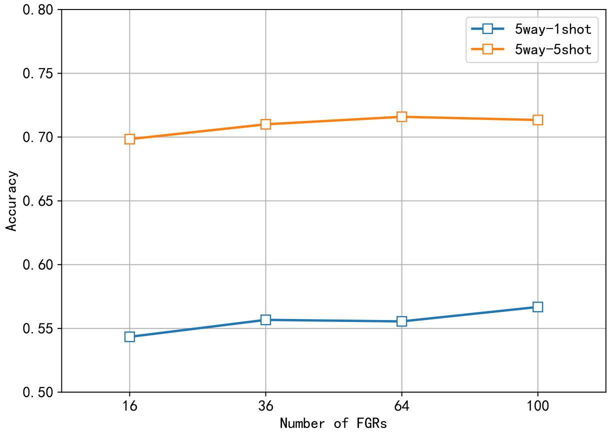

Figure 5. The effect of the number of LRFs on the performance of few-shot classification on miniImageNet.

## 5.2. Ablation Study

We study the impacts of metric functions, the number of LRFs, image size, and attention method. Also, time complexity is discussed. Experiments are conducted on miniImageNet.

**Impact of image size:** To better control the number of LRFs, we adopt a larger image size of 224 × 224 instead of 84 × 84. For the fair comparison, we evaluate the impact of image size on few-shot classification performance, and the results are shown in Table 1. We also conduct similar comparison using RelationNet [37]. Our approach benefits more from the increase in image size than RelationNet. The larger the image size, the smaller the LRF overlap, and the more independent the individual LRF is. Intuitively, the reason why the performance increases with the size of the image can be attributed to the smaller ratio of the receptive field to the image size.

**Impact of metric functions:** We study the choice of metric function on the performance of few-shot learning, as shown in Table 2. The cosine and Gaussian without attention perform much worse than their attentional version, which can be attributed to their equal treatment of the comparison result of all LRF pairs. For the same category, the pair of the same semantic LRFs is obviously more important than the pair of the semantically irrelevant LRFs as the latter can increase the difference. The attentional cosine and attentional Gaussian suppress the comparisons of the semantically irrelevant LRFs via the attention method. For different categories, the key regions show significant differences, which are also captured by SAML.

**Impact of the number of LRFs:** The number of LRFs can be adjusted by setting the stride of the final average pooling layer of the embedding network. We evaluate the impact of the number of LRFs on few-shot learning performance, and the results are shown in Figure 5. Both 5-way 1-shot and 5-way 5-shot classification accuracies improve as the number of LRFs increases, and they reach saturation when the number of LRFs is 100. Based on this observation, we set the number of LRFs to 100 in the following experiments by default, and the setting works well for both miniImageNet and CUB.

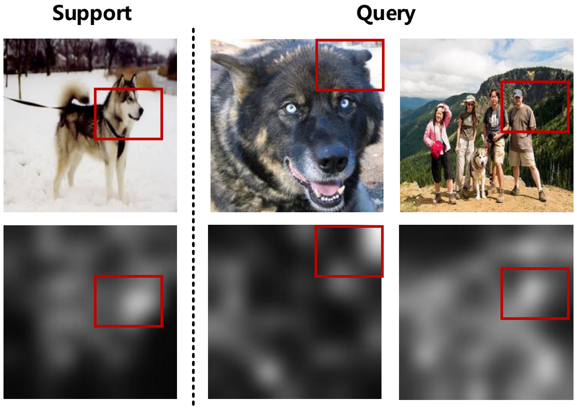

Figure 6. SAML fails when the scale of the dominant object in the query image (middle) differs greatly from that of the support image or the key object (right) in the query image is not salient in a complex background.

**Impact of attention method:** We study the impact of the attention method. Some correctly classified images and their corresponding attention maps are shown in Figure 2. The key regions are marked with red rectangles, and all these regions correspond to the same semantic concept, namely dog. The effect of attention maps is to strengthen objects while suppressing backgrounds. The combination of the attention method and the metric function reduces the ambiguity introduced by the comparison between the key object and the semantically irrelevant part. However, our approach suffers when dominant objects of different scales exist in a complex background, and some failure examples are shown in Figure 6. We further optimize our method to address this issue. Specifically, we introduce an inception [38] operator ([2 × 2 max pooling, 3 × 3 convolutions, 5 × 5 convolutions, 7 × 7 convolutions]) after the embedding network (see Figure 4) to extract features for the objects with different scales. The results are discussed in Section 5.3.

**Time complexity:** The computations of SAML are matrix multiplications (see Figure 8 in the Appendix), which have been well optimized in popular deep learning platforms, e.g., PyTorch [1] and TensorFlow [2], and is no longer a major computing bottleneck. We compared the time costs of different methods, and the results are shown

| Model | Training / Test (ms/episode) | |
|---|---|---|
| | 5way-1shot | 5way-5shot |
| MAML [8] | 61.92 / 31.04 | 72.64 / 38.38 |
| Prototypical Net [16] | 14.55 / 4.51 | **15.46** / 5.22 |
| Matching Net [31] | **6.89 / 2.88** | 19.10 / 6.83 |
| RelationNet [37] | 20.78 / 4.13 | 22.84 / 5.26 |
| **SAML(Ours)** | 16.30 / 3.95 | 19.59 / **5.17** |

Table 3. Training/test time costs per episode of different methods on miniImageNet.

in Table 3. Our method is competitive regarding time cost (2nd fast for 5way-1shot and 1st for 5way-5shot during test).

### 5.3. Comparisons with the State-of-the-art

We compare SAML with the state-of-the-art methods. Here, the cosine metric is employed. More empirical results, including results with more complex metrics, can be found in the Appendix.

**Results on miniImageNet:** To perform fair comparisons, two common tasks are conducted by way of evaluation, namely, 5-way 1-shot and 5-way 5-shot classification. We randomly sample 600 episodes from the miniImageNet test set, and then report the few-shot classification accuracies with 95% confidence intervals. We also repeat the test process 10 times and report the variance. The results are shown in Table 4. Our approach achieves much better performance than the-state-of-the-art methods for both 5-way 1-shot and 5-way 5-shot classification, especially on the 5-way 5-shot task ($> 2.4\%$). The scales of key objects in miniImageNet vary greatly. When adopting scale-invariant features, the improvement over naive SAML for 5-way 1-shot and 5-way 5-shot classification are 1.01% and 1.69%, respectively.

**Results on CUB:** CUB is a fine-grained image classification dataset comprised of birds of different species. Compared with miniImageNet collected for generic recognition, CUB is simple, as the dominant objects are always birds and the backgrounds are relatively clean. However, the birds still show great variability in position. Two tasks are conducted on CUB, namely, 5way-1shot and 5way-5shot classification, and the experimental results are shown in Table 5. Our approach performs better than existing methods. Specifically, the increments on fine-grained 5way-1shot classification and 5way-5shot classification tasks are 6.88% and 2.22%, respectively, which are surprising and impressive performance boosts. As we crop all images with given bounding box for CUB, the scales of dominant objects in CUB are roughly the same. Thus, adopting scale-invariant features has little effect on performance.

## 6. Conclusions

| Model | 5way-1shot | 5way-5shot |
|---|---|---|
| Prototypical Net [16] | 49.42±0.78% | 68.20±0.66% |
| Matching Net [31] | 43.56±0.84% | 55.31±0.73% |
| M-L LSTM [34] | 43.44±0.77% | 60.60±0.71% |
| MAML [8] | 48.70±1.84% | 63.11±0.92% |
| RelationNet [37] | 50.44±0.82% | 65.32±0.70% |
| Meta-SGD [22] | 50.47±1.87% | 64.03±0.94% |
| LLAMA [11] | 49.40±1.83% | - |
| REPTILE [29] | 49.97±0.32% | 65.99±0.58% |
| MM-Net [5] | 53.37±0.48% | 66.97±0.35% |
| PLATIPUS [9] | 50.13±1.86% | - |
| **SAML** (ours) | **56.68±0.40%** | **71.34±0.41%** |
| **SAML*** (ours) | **57.69±0.20%** | **73.03±0.16%** |

Table 4. Few-shot classification accuracies on miniImageNet. "-" means "not reported". "*" means "adopting inception operator".

| Method | 5way-1shot | 5way-5shot |
|---|---|---|
| Baseline [6] | 47.12±0.74% | 64.16±0.71% |
| Baseline++ [6] | 60.53±0.83% | 79.34±0.61% |
| Matching Net [31] | 61.16±0.89% | 72.86±0.70% |
| Prototypical Net [16] | 51.31±0.91% | 70.77±0.69% |
| MAML [8] | 55.92±0.95% | 72.09±0.76% |
| RelationNet [37] | 62.45±0.98% | 76.11±0.69% |
| **SAML** (ours) | **69.33±0.22%** | **81.56±0.15%** |
| **SAML*** (ours) | **69.35±0.22%** | **81.37±0.15%** |

Table 5. Few-shot classification accuracies on CUB. "*" means "adopting inception operator".

Dominant objects may appear in any part of an image. Thus, directly calculating the distance between the features extracted from images according to the indices may lead to serious ambiguity, because we probably compare semantically irrelevant local regions. To this end, we present a Semantic Alignment Metric Learning (SAML) method that aligns the semantically relevant local regions through the "collect-and-select" strategy. Specifically, we define a relation matrix (RM) to "collect" all distances of local regions pair of query instances and support class means, and then utilize the attention technique to "select" and pay more attention to semantically relevant local region pairs. Empirical results demonstrate that the semantic alignment is achieved. Theoretical analysis of the generalization bound proves SAML's feasibility on unseen data is guaranteed. Extensive experiments on standard benchmark datasets demonstrate the superiority of SAML by comparing with the state-of-the-art few-shot learning methods.

## Acknowledgments

This work was supported in part by National Natural Science Foundation of China (61772508, U1713213), Shenzhen Technology Project (JCYJ20170413152535587, JCYJ20180507182610734), CAS Key Technology Talent Program, and Australian Research Council (FL-170100117, DP-180103424).

# References

[1] https://pytorch.org/.

[2] https://www.tensorflow.org/.

[3] Matthias Bauer, Mateo Rojas-Carulla, Jakub Bartlomiej Swiatkowski, Bernhard Scholkopf, and Richard E. Turner. Discriminative k-shot learning using probabilistic models. arXiv:1706.00326, 2017.

[4] Ali Borji, Ming-Ming Cheng, Huaizu Jiang, and Jia Li. Salient object detection: A benchmark. *IEEE Trans. Image Processing*, 24(12):5706–5722, December 2015.

[5] Qi Cai, Yingwei Pan, Ting Yao, Chenggang Yan, and Tao Mei. Memory matching networks for one-shot image recognition. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 4080–4088, 2018.

[6] Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Frank Wang, and Jia-Bin Huang. A closer look at few-shot classification. In *Proc. Int. Conf. Learn. Represent.*, 2019.

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. Imagenet: A large-scale hierarchical image database. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 248–255, 2009.

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

[9] Chelsea Finn, Kelvin Xu, and Sergey Levine. Probabilistic model-agnostic meta-learning. In *Proc. Adv. Neural Inf. Process. Syst.*, pages 9537–9548, 2018.

[10] Spyros Gidaris and Nikos Komodakis. Dynamic few-shot visual learning without forgetting. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 4367–4375, 2018.

[11] Erin Grant, Chelsea Finn, Sergey Levine, Trevor Darrell, and Thomas Griffiths. Recasting gradient-based meta-learning as hierarchical bayes. In *Proc. Int. Conf. Learn. Represent.*, 2018.

[12] Fusheng Hao, Jun Cheng, Lei Wang, Xinchao Wang, Jianzhong Cao, Xiping Hu, and Dapeng Tao. Anchor-based nearest class mean loss for convolutional neural networks. arXiv: 1804.08087, 2018.

[13] Bharath Hariharan and Ross Girshick. Low-shot visual recognition by shrinking and hallucinating features. In *Proc. IEEE Conf. Int. Comput. Vis.*, pages 3037–3046, 2017.

[14] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Low-shot visual recognition by shrinking and hallucinating features. In *Proc. IEEE Conf. Int. Comput. Vis.*, pages 3037–3046, 2017.

[15] Tao Hu, Pengwan Yang, Chiliang Zhang, Gang Yu, Yadong Mu, and Cees G.M. Snoek. Attention-based multi-context guiding for few-shot semantic segmentation. In *Proc. Conf. AAAI Artif. Intell.*, pages 8441–8448, 2019.

[16] Snell Jake, Swersky Kevin, and Zemel Richard. Prototypical networks for few-shot learning. In *Proc. Adv. Neural Inf. Process. Syst.*, pages 4077–4087, 2017.

[17] Saumya Jetley, Nicholas A. Lord, Namhoon Lee, and Philip H.S. Torr. Learn to pay attention. In *Proc. Int. Conf. Learn. Represent.*, 2018.

[18] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proc. Int. Conf. Learn. Represent.*, 2015.

[19] Gregory Koch. Siamese neural networks for one-shot image recognition. 2015.

[20] Alex Krizhevsky, Sutskever Ilya, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proc. Adv. Neural Inf. Process. Syst.*, pages 1097–1105, 2012.

[21] Fei-Fei Li, Rob Fergus, and Pietro Perona. One-shot learning of object categories. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28(4):594–611, April 2006.

[22] Zhenguo Li, Fengwei Zhou, Fei Chen, and Hang Li. Meta-sgd: Learning to learn quickly for few-shot learning. arXiv:1707.09835, 2017.

[23] Tsung-Yu Lin and Subhransu Maji. Improved bilinear pooling with cnns. In *British Mach. Vis. Conf.*, 2017.

[24] Akshay Mehrotra and Ambedkar Dukkipati. Generative adversarial residual pairwise networks for one shot learning. arXiv:1703.08033, 2017.

[25] Mohri Mehryar, Rostamizadeh Afshin, and Talwalkar Ameet. *Foundations of machine learning*. MIT press, 2012.

[26] Nikhil Mishra, Mostafa Rohaninejad, Xi Chen, and Pieter Abbeel. A simple neural attentive meta-learner. In *Proc. Int. Conf. Learn. Represent.*, 2018.

[27] Tsendsuren Munkhdalai, Xingdi Yuan, Soroush Mehri, and Adam Trischler. Rapid adaptation with conditionally shifted neurons. In *Proc. Int. Conf. Mach. Learn.*, pages 3664–3673, 2018.

[28] Tsendsuren Munkhdalai, Xingdi Yuan, Soroush Mehri, Tong Wang, and Adam Trischler. Learning rapid-temporal adaptations. arXiv:1712.09926, 2017.

[29] Alex Nichol, Joshua Achiam, and John Schulman. On first-order meta-learning algorithms. arXiv:1803.02999, 2018.

[30] Boris Oreshkin, Pau Rodriguez Lopez, and Alexandre Lacoste. Tadam: Task dependent adaptive metric for improved few-shot learning. In *Proc. Adv. Neural Inf. Process. Syst.*, pages 719–729, 2018.

[31] Vinyals Oriol, Blundell Charles, Lillicrap Tim, kavukcuoglu koray, and Wierstra Daan. Matching networks for one shot learning. In *Proc. Adv. Neural Inf. Process. Syst.*, pages 3630–3638, 2016.

[32] Bartlett Peter, Foster Dylan J, and Telgarsky Matus. Spectrally-normalized margin bounds for neural networks. In *Proc. Adv. Neural Inf. Process. Syst.*, pages 6240–6249, 2017.

[33] Siyuan Qiao, Chenxi Liu, Wei Shen, and Alan Yuille. Few-shot image recognition by predicting parameters from activations. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 7229–7238, 2018.

[34] Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. In *Proc. Int. Conf. Learn. Represent.*, 2017.

[35] Mengye Ren, Renjie Liao, Ethan Fetaya, and Richard S. Zemel. Incremental few-shot learning with attention attractor networks. arXiv:1810.07218, 2018.

[36] Andrei A. Rusu, Dushyant Rao, Jakub Sygnowski, Oriol Vinyals, Razvan Pascanu, Simon Osindero, and Raia Hadsell. Meta-learning with latent embedding optimization. In *Proc. Int. Conf. Learn. Represent.*, 2019.

[37] 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 *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 1199–1208, 2018.

[38] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 1–9, 2015.

[39] Garcia Victor and Bruna Joan. Few-shot learning with graph neural networks. In *Proc. Int. Conf. Learn. Represent.*, 2018.

[40] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. In *Computation & Neural Systems Technical Report, CNS-TR-2011-001*, 2011.

[41] Feng Wang, Weiyang Liu, Haijun Liu, and Jian Cheng. Additive margin softmax for face verification. arXiv:1801.05599, 2018.

[42] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 7794–7803, 2018.

[43] Yu-Xiong Wang, Ross Girshick, Martial Hebert, and Bharath Hariharan. Low-shot learning from imaginary data. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 7278–7286, 2018.

[44] Shipeng Yan, Songyang Zhang, and Xuming He. A dual attention network with semantic embedding for few-shot learning. In *Proc. Conf. AAAI Artif. Intell.*, pages 9079–9086, 2019.

[45] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *British Mach. Vis. Conf.*, 2016.

[46] Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In *Proc. Int. Conf. Learn. Represent.*, 2017.

[47] Ruixiang Zhang, Tong Che, Zoubin Ghahramani, Yoshua Bengio, and Yangqiu Song. Metagan: An adversarial approach to few-shot learning. In *Proc. Adv. Neural Inf. Process. Syst.*, pages 2371–2380, 2018.

[48] Xueting Zhang, Flood Sung, Yuting Qiang, Yongxin Yang, and Timothy M. Hospedales. Deep comparison: Relation columns for few-shot learning. arXiv:1811.07100, 2018.