

# Binocular Mutual Learning for Improving Few-shot Classification

Ziqi Zhou<sup>1</sup> Xi Qiu<sup>1\*</sup> Jiangtao Xie<sup>2</sup> Jianan Wu<sup>1</sup> Chi Zhang<sup>1</sup>  
<sup>1</sup>Megvii Technology <sup>2</sup>Dalian University of Technology

{zhouziqi, qiuxi, wjn, zhangchi}@megvii.com jiangtaoxie@mail.dlut.edu.cn

## Abstract

Most of the few-shot learning methods learn to transfer knowledge from datasets with abundant labeled data (i.e., the base set). From the perspective of class space on base set, existing methods either focus on utilizing all classes under a global view by normal pretraining, or pay more attention to adopt an episodic manner to train meta-tasks within few classes in a local view. However, the interaction of the two views is rarely explored. As the two views capture complementary information, we naturally think of the compatibility of them for achieving further performance gains. Inspired by the mutual learning paradigm and binocular parallax, we propose a unified framework, namely **Binocular Mutual Learning (BML)**, which achieves the compatibility of the global view and the local view through both intra-view and cross-view modeling. Concretely, the global view learns in the whole class space to capture rich inter-class relationships. Meanwhile, the local view learns in the local class space within each episode, focusing on matching positive pairs correctly. In addition, cross-view mutual interaction further promotes the collaborative learning and the implicit exploration of useful knowledge from each other. During meta-test, binocular embeddings are aggregated together to support decision-making, which greatly improve the accuracy of classification. Extensive experiments conducted on multiple benchmarks including cross-domain validation confirm the effectiveness of our method<sup>1</sup>.

## 1. Introduction

Conventional classification methods heavily rely on massive labeled data [32] with diverse visual variations. However, in many realistic scenarios, only limited labeled data is available [43, 26], thereby giving rise to the investigation of few-shot classification (FSC), where only few available training data is given for the learning of new visual concepts. Such a setting makes FSC a challenging problem, since novel classes are unpredictable and the sampling of

Table 1. Comparison of BML with several representative methods from the view of base class space. Accuracy (Acc.) from *miniImageNet* [39]. Obviously, the unified perspective of BML is more effective.

	w/global	w/local	strategy	Acc.
MAML [10]			one-stage	63.1
MatchingNet [39]			one-stage	55.3
RelationNet [35]	✗	✓	one-stage	65.3
ProtoNet [33]			one-stage	68.2
Rethink [36]			one-stage	82.1
DC [24]	✓	✗	one-stage	79.0
CloserLook [6]			one-stage	75.7
DeepEMD [45]			two-stage	82.4
FEAT [44]	✓	✓	two-stage	82.1
Meta-Baseline [7]			two-stage	79.3
Neg-Cosine [25]			two-stage	81.6
<b>Our BML</b>	✓	✓	<b>one-stage</b>	<b>83.6</b>

few shots is also biased. In order to overcome those difficulties, many effective approaches have been proposed in recent years, which can be mainly summarized into two categories according to training strategies. The first category is fine-tuning based paradigms [29, 24, 6, 7, 36], which learn classifiers in the whole base class space with a straightforward purpose of maximize differences between classes. Since all base classes are visible under each iteration, we refer to this kind of methods as the *global* view. The other promising strategy is metric-based meta-training schemes [9, 39, 33, 35, 27, 44, 18, 13], which only tune on a few classes in each episode. The main idea comes from metric learning and the purpose is to match unlabeled query to its correct class with a small labeled support set. Because the visible range of base classes for each meta-task is limited, we oppositely call them the *local* view.

Considering that single view (whether *global* or *local*) is relatively weak, it is not enough to provide adequate knowledge for accurate classification. What’s more, the combination of dual views fits well with the characteristics of “people deepen their perception through two eyes”. To this end, we propose this new **Binocular Mutual Learning (BML)** paradigm, which equips the network with a global view

\*Xi Qiu (qiuxi@megvii.com) is the corresponding author.

<sup>1</sup><https://github.com/ZZQzzq/BML>

and a local view simultaneously. The combination of two complementary views works like a binocular system, and the mutual interaction [47] through two views further promotes their cooperation and calibrates the inappropriate expression caused by single “biased” view. Concretely, BML generates better expression through both intra-view and cross-view modeling. The intra-view training captures view-specific knowledge, where two balanced feature space are built, one focuses on inter-class relationship perception (global view) and the other pays attention on matching details (local view). Meanwhile, the cross-view mutual interaction facilitates the implicit knowledge transfer from each other. To balance “binocular parallax”, we enlarge the optimization difficulty of the local view, so that the global view can learn more useful knowledge from mutual interaction (For more details, please refer to Section.3.2.2).

Clearly, BML paradigm has two advantages: strong transferability and high time-efficiency, which are manifested from the following comparisons. Concretely, compared with single view based methods mentioned above, BML has two complementary views, so that more transferable and reliable knowledge can be learned. In contrast, single global training lacks additional constraints, making it easy to over-fit on base patterns [12]. Meanwhile, single local training is restricted by local perspectives, whose performance is heavily depend on the configuration of tasks, not to mention the complex structures. Moreover, compared with two-stage methods which firstly excute global training and then tune the embedding with local training, BML unifies the two views in a one-stage framework and enables the promotion of each other. By contrast, two-stage methods [34, 6, 45, 44, 25] are time-consuming. They focus on how to better learn embedding in the first stage to provide stronger features for the second stage of optimization [25], but ignore that local view and global view can promote each other in a unified manner.

We highlight the advantages of BML compared with several typical approaches in Table 1. As the first batch of methods to consider the combination of dual views, we propose an elegant compatibility strategy: binocular mutual learning, which is inspired by the fact that human beings usually perceive the world through two eyes (benefit from appropriate binocular parallax). The two complementary views simulate the binocular mode, and the mutual interaction calibrates deviation. Extensive experiments confirm the effectiveness of BML, which is mainly reflected on the stable performance under different granularity evaluation. No matter facing coarse-grained (e.g., *miniImageNet* [39]) or fine-grained (e.g. *CUB* [40]) situation, BML performs well. However, single view based methods cannot handle all the situations. This confirm that the unified framework does facilitate the mutual calibration of the two views. In summary, the contributions of this paper are as follows:

- We closely analyze the status quo of FSC and propose an efficient one-stage **Binocular Mutual Learning** paradigm: **BML**, which elegantly aggregate the global view and the local view through both intra-view and cross-view modeling.
- To enhance mutual learning, we propose an elastic loss to readjust the optimization difficulty of the local view, which promotes the bidirectional implicit knowledge transfer.
- Extensive experiments on multiple benchmarks including cross-domain validation verify the effectiveness of our framework.

## 2. Related Works

### 2.1. Fine-Tuning based Methods (*global view*)

Researches represented by [29, 34, 24, 6, 36, 25] pay more attention to global training by simply learning a class-specific embedding with fully-connected (FC) layers. Among them, FSIR [29] adapted the pretrained model to the new categories by directly predicting the parameters from activation, while DC [24] started from the perspective of spatial information mining and performed dense classification. MTL [34] further enhanced the pretrained model through a hard task meta-batch mechanism. And [6] proposed two different classification layers, including a conventional FC Layer (CloserLook) and a FC layer with feature normalization (CloserLook++). To alleviate overfitting on base patterns, The authors in [36] employed self-distillation strategy and data augmentation to constraint the learning process.

### 2.2. Meta-Training based Methods (*local view*)

One of the promising branch of meta-training is metric-based methods [39, 33, 35, 44, 27, 18, 20, 19, 42, 15, 13, 4, 28], which meta-learn an ideal metric space  $\mathcal{S}(\cdot, \cdot)$  to bring homogeneous samples closer while push heterogeneous samples away. The classification process is carried out under the guidance of the “*Nearest Neighbor*” principle. Specifically, Matching Networks [39] proposed a L-STM [14]-based encoding module to re-code the context of support feature, and computed attention scores using cosine distance. Similar variants include FEAT [44], which replaced LSTM with transformer [38]. Prototypical Networks [33] employed prototype to identify each class and calculated the Euclidean distance. Relation Networks [35] further adopted a learnable correlation calculation module to measure pairwise similarity. Subsequent follow-upers derived many variants with incorporating cross-modal information [18], introducing adversarial noise [20], employing local descriptors [22] or mutual information [13, 4], making pretext tasks [46], using attention mechanisms [8, 15] or learning task-relevant metrics [27, 19].

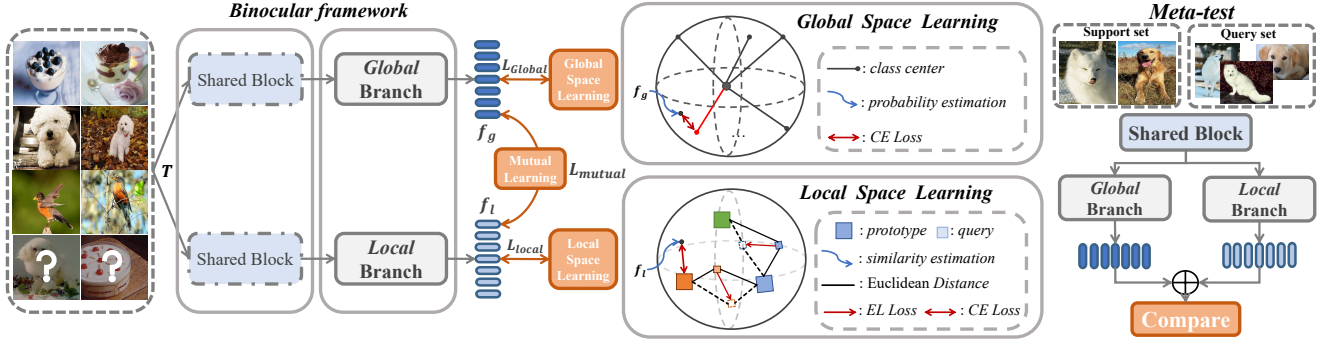


Figure 1. The framework of **BML**. For each task  $\mathcal{T}$ , BML minimizes the global classification loss  $\mathcal{L}_{global}$ , the local matching loss  $\mathcal{L}_{local}$  and the distribution consistency loss  $\mathcal{L}_{mutual}$  to capture discriminative expression. During test, the collaborative features from two views are combined to support the final decision.

Besides, some studies started from optimization [10, 23, 16, 30, 17]. A typical literature is MAML [10], in which, the parameters of a base learner were further optimized in a few iterations for quickly adaption for new tasks. Subsequent variants further developed MAML by designing more objectives [23], employing better classifiers [17] or dynamically adjusting the weights of different tasks [16]. More interestingly, the working mechanism of MAML was analyzed in detail in [30] and the experiments highlighted that “*Feature Reuse*” is the key to its role. In other words, the performance is majorly determined by the quality of the learned features, which is revalidated by experiments conducted in this paper.

We also found that some recent researches [25, 7, 45, 44] have realized the promotion role of two views. Some of them [25, 7, 45, 44] employed a two-stage training scheme. They firstly pretrained under global mode, and then tuned the parameters using local mode. Others [5, 15, 27] promoted local training by learning additional global classifiers. All of them focus on the promotion of global to local but ignore the bidirectional cross-view interaction (which plays an important role in BML).

### 2.3. Mutual Learning

Mutual learning [47] is a new distillation modal that has shined in many fields recently, which breaks the conventional “teacher-student” structure which has fixed direction of supervision. In mutual learning, a group of students are collaboratively learned from each other, which helps to obtain more general models without pretrained teachers. Similar ideas have been employed in person re-ID [11], but not effectively applied in FSC. Literatures like [36, 21] are close to this topic, but the training process is  $N$ -staged with fixed pretrained teachers.

Inspired by mutual learning [47] and binocular parallax, we propose this unified BML framework with two complementary views. Each of them can be regarded as a “*student*”. During training, besides their offline hard supervision,

two “*students*” also learn collaboratively and implicitly explore useful knowledge from each other.

## 3. Methodology

### 3.1. Preliminary

In the standard FSC scenario, we are given a base set with  $C_{base}$  classes and a novel set with  $C_{novel}$  classes, where  $C_{base} \cap C_{novel} = \emptyset$ . Training is usually performed on  $C_{base}$  classes and the optimization goal is to transfer the learned knowledge to new tasks built on  $C_{novel}$ . During meta-test, a family of tasks  $\{\mathcal{T}\}_1^n$  are constructed for evaluation. Concretely, each task  $\mathcal{T}$  has a support set  $\mathcal{S}$  and a query set  $\mathcal{Q}$ . The support set  $\mathcal{S}$  contains  $N$  classes and each class has  $K$  images (*i.e.*, the  $N$ -way  $K$ -shot setting). The query set  $\mathcal{Q}$  includes  $N \times Q$  unlabeled images. In most literature,  $N$  is set to 5 and  $K$  is set to 1 or 5, so do we.

### 3.2. Binocular Mutual Learning

As shown in Figure 1, BML has two complementary branches (views) based on shared blocks. In which, the *global* branch learns in the whole class space for inter-class relationship mining, and the *local* branch learns in each episode within few classes, aiming at matching each query sample to its support prototype. Besides, the two branches implicitly explore useful knowledge from each other by minimizing KL-Divergence based mimicry loss to match the feature distribution of its peers.

#### 3.2.1 Global Intra-view Training

For the *global* branch, the learned features are related to the whole class space, which explicitly contain rich inter-class relationships.

Specifically, given task  $\mathcal{T}$ , we learn a global learner  $\mathcal{A}_\phi^G$  to map each image  $\mathbf{x}_i$  in  $\mathcal{T}$  to a high-dimensional space,

and then a  $1 \times 1 \times C_{base}$  convolution layer  $\mathcal{W}$  is learned to classify each point of the feature to its corresponding class. Let  $w$  denotes the width of the feature and  $h$  denotes the height, the probability estimate of point  $(p, q)$  is formulated by:

$$P(y_i = y | \mathbf{x}_i^{(p,q)}) = \sigma(\mathcal{W}(\mathcal{A}_\phi^G(\mathbf{x}_i^{(p,q)}))) \quad (1)$$

where  $\sigma : \mathbb{R}^m \rightarrow \mathbb{R}^m$  is the *softmax function*,  $m$  represents the dimension size which is 640 here.

The negative logarithm of  $P(y_i = y | \mathbf{x}_i^{(p,q)})$  is calculated to represent the loss of current feature point  $(p, q)$  of input data  $x_i$ . The total global loss is the average value of all images in  $\mathcal{T}$ . Formulaically,

$$\mathcal{L}_{global} = \mathbb{E}_{(x_i, y_i) \in \mathcal{T}} \frac{1}{w \times h} \sum_{\substack{p=1 \\ q=1}}^{w, h} -y_i \log P(y_i = y | \mathbf{x}_i^{(p,q)}) \quad (2)$$

Meaningfully, classifying each feature point correctly is helpful to capture spatial structure information of the foreground. Since each point can be traced back to a local area of the foreground, multiple points represents multiple local areas that is equivalent to a memory-saving multi-crop operation [24]. For fair comparison, we also apply point-wise classification to the baselines we compared in this paper.

### 3.2.2 Local Intra-view Training

The *local* branch borrows the idea of metric learning and learns to match each query sample with support prototypes in embedding space.

For task  $\mathcal{T}$  with  $N$  classes, we divide it into a support set  $\mathbb{S} = \{(x_j, y_j) \mid j = 1, \dots, N \times S\}$  and a query set  $\mathbb{Q} = \{(x_j, y_j) \mid j = 1, \dots, N \times Q\}$ . A local learner  $\mathcal{A}_\phi^L$  is learned to map all samples  $\{x_j \mid x_j \in \mathbb{S} \cup \mathbb{Q}\}$  to an embedding space, and the matching process is guided by nearest neighbor strategy, where  $N$  prototypes are calculated, *i.e.*,  $C_i = \frac{1}{S} \sum_{j \in \mathbb{S}_i} \mathcal{A}_\phi^L(x_j)$ . To metric the similarity between query and each prototype, we simply employ Euclidean distance indicated by  $\mathcal{M}$ . The negative logarithm of all query samples are calculated to get the local loss.

$$P(C_i = C | \mathbf{x}_i) = \frac{\exp[-\mathcal{M}(C_i, \mathcal{A}_\phi^L(\mathbf{x}_i))]}{\sum_{j=1}^N \exp[-\mathcal{M}(C_j, \mathcal{A}_\phi^L(\mathbf{x}_i))]} \quad (3)$$

$$\mathcal{L}_{local} = \mathbb{E}_{(x_i, y_i) \in \mathbb{Q}} -C_i \log P(C_i = C | \mathbf{x}_i) \quad (4)$$

**Elastic constraint for magnifying optimization difficulty.** Moreover, we consider the following two points and further optimize the local loss by applying an elastic constraint.

On the one hand, each task  $\mathcal{T}$  is a collection of randomly sampled data, resulting in randomness of difficulty. Treating all tasks equally during cross-epoch training like [33]

will lead to an “unhealthy” phenomenon: further performance improvement is hard to obtain in the later training, because the model is dynamically growing while the difficulty is static [2]. On the other hand, the optimization difficulty of the local branch is relatively simple compared with the global one (since each query only has  $N - 1$  negative prototypes while global branch has  $C_{base} - 1$ ). The implementation of mutual promotion requires a seedbed, that is, the two views can provide valuable knowledge to each other.

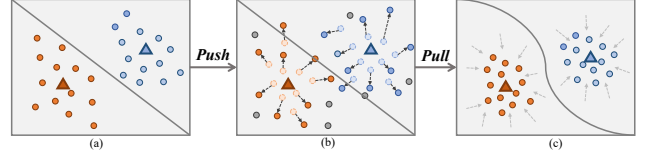


Figure 2. The process of elastic loss. All queries are pushed away from their clusters to magnify the optimization difficulty, and the network is forced to pull them back again.

Therefore, we propose this elastic loss to enhance the training difficulty of the local branch. The complete process is shown in Figure 2. Simply, we modify the position of each query sample  $\mathbf{x}_i$  in the embedding space according to the difference between  $\mathcal{M}(C_p, \mathcal{A}_\phi^L(\mathbf{x}_i))$  and  $\mathcal{M}(C_n, \mathcal{A}_\phi^L(\mathbf{x}_i))$ , where  $C_p$  is the positive prototype and  $C_n$  is the nearest negative prototype. The above operation can be described as “*push away*”. Then we demand the network to *pull* these pushed out samples *back* to the vicinity of their positive prototypes, thus to learn more implicit knowledge and get further performance improvement.

To better perceive the foreground from different patches and learn rich relationship between  $C_p$  and patch-specific  $C_n^{(p,q)}$ , similar to the global branch, instead of performing global average pooling, we calculate elastic constraint from different spatial position. Forcing the network to pull  $\mathbf{x}_i$  to its positive prototype under different  $(C_n^{(p,q)}, C_p)$  pairs helps to mine more hard samples, and avoid over-fitting on simple local tasks.

The updated local loss has a formulation as:

$$\hat{P}(C_i = C | \mathbf{x}_i^{(p,q)}) = \frac{\exp - \mathcal{M}(C_i, \mathcal{A}_\phi^L(\mathbf{x}_i^{(p,q)})) - d_{EL}^{i(p,q)}}{\sum_{j=1}^N \exp - \mathcal{M}(C_j, \mathcal{A}_\phi^L(\mathbf{x}_i^{(p,q)}))} \quad (5)$$

$$\hat{\mathcal{L}}_{local} = \mathbb{E}_{(x_i, y_i) \in \mathbb{Q}} \frac{1}{w \times h} \sum_{\substack{p=1 \\ q=1}}^{w, h} -C_i \log \hat{P}(C_i = C | \mathbf{x}_i^{(p,q)}) \quad (6)$$

where  $d_{EL}^{i(p,q)}$  is the patch-specific elastic constraint, whose specific detail is described in Algorithm 1.  $\alpha_1$  and  $\alpha_2$  are two scale factors,  $\alpha_1$  adjusts the push degree cross-epochs and  $\alpha_2$  adjusts the push degree cross-tasks.

---

**Algorithm 1:** Elastic Constraint  $d_{EL}^{(p,q)}$  on  $x_i^{(p,q)}$ .

---

**Input:** prototypes  $\{C_i^{(p,q)}\}_i^N$ ; query  $\mathcal{A}_\phi^l(x_i^{(p,q)})$ ;  
current epoch  $e$ ; total epoch  $E$ ;  $\alpha_1, \alpha_2$

**Output:** Elastic Constraint  $d_{EL}^{(p,q)}$

- 1 Get logits  $L^{(p,q)}$  of  $\mathcal{A}_\phi^l(x_i^{(p,q)})$  using Eq.3;
  - 2 Cal distance between  $\mathcal{A}_\phi^l(x_i^{(p,q)})$  and  $C_p$ :  
 $dis_P = \mathbb{I}_{c=C_p} L^{(p,q)}$ ;
  - 3 Cal distance between  $\mathcal{A}_\phi^l(x_i^{(p,q)})$  and  $C_n^{(p,q)}$ :  
 $dis_N = \mathbf{sort}(\mathbb{I}_{c \neq C_p} L^{(p,q)})[0]$ ;
  - 4 Cal  $\Delta = dis_P - dis_N$ ;
  - 5 Return  $d_{EL}^{(p,q)} = \frac{\alpha_1 \cdot e / E}{1 + \exp(-\alpha_2 \cdot \Delta)}$ ;
- 

### 3.2.3 Cross-view Mutual Learning

In addition to separate intra-view learning, the two views also promote each other through cross-view mutual interaction. Specifically, for each view, in addition to completing its own offline hard tasks, it is also forced to minimize the mimicry loss from another view (based on KL-divergence), which encourage the implicit knowledge transfer. Clearly, the mutual loss has two sub-items, which are formulated as:

$$\mathcal{L}_{mutual} = \mathcal{D}_{KL}(F_l || F_g) + \mathcal{D}_{KL}(F_g || F_l) \quad (7)$$

$$\mathcal{D}_{KL}(F_g || F_l) = F_g \log \frac{F_g}{F_l}, \mathcal{D}_{KL}(F_l || F_g) = F_l \log \frac{F_l}{F_g} \quad (8)$$

where  $F(\cdot)$  represents the feature distribution calculated by  $\sigma(\mathcal{A}_\phi(x))$ . We consider the issue of interaction from the perspective of feature distribution consistency, and only learn relative relationships rather than hard constraints such as mean squared error in Euclidean space. This is because that too strong supervision signal is harmful to retain the specificity of the two views.

In summary, the final loss has following formulation:

$$\mathcal{L}_{total} = \alpha \mathcal{L}_{global} + \beta \mathcal{L}_{local} + \gamma \mathcal{L}_{mutual} \quad (9)$$

where  $\alpha, \beta$  and  $\gamma$  are weighting factors.

### 3.2.4 Inference

During meta-test, we integrate the results of the global and local branches. We do not perform global average pooling but directly flatten the features and calculate the Euclidean distance (the same as ProtoNet [33]). The integration can be done on the feature level or on the logits level. In this paper, we simply integrate the global and local logits.

## 4. Experiments

In this section, we answer the following questions:

- How does our BML perform compared with SoTAs?
- Why is binocular learning better?
- How does the elastic loss work?
- Is BML less sensitive to image quality?
- Should the standards for validation be diverse?

### 4.1. Meta Datasets

We validate our BML on four commonly used benchmarks, including *miniImageNet* [39], *tieredImageNet* [31], CIFAR-FS [3] and CUB-200-2011 (CUB) [40]. Details of those datasets are summarized in Table 2. All the input images are resized to  $84 \times 84$  during comparison, and in the ablation part, we further analyze the stability of BML when enlarging the image size to  $224 \times 224$  or introducing other degraded components. In particular, for CUB, we crop the test images with the bounding box provided by [37] to make a fair comparison.

Table 2. Summarization of four benchmarks.

	Images	Classes	Split
<i>miniImageNet</i> [39]	60,000	100	64/16/20
<i>tieredImageNet</i> [31]	779,165	608	351/97/160
CIFAR-FS [3]	60,000	100	60/16/20
CUB-200-2011 [40]	11,788	200	100/50/50

### 4.2. Implementation Details

**Architecture.** Following previous works [27, 36, 17], we use ResNet12 as our backbone, which consists of 4 residual blocks. Each block has 3 convolutional layers with  $3 \times 3$  kernel and a  $2 \times 2$  max-pooling layer. We remove the last global average pooling layer to preserve spatial information. Similar to [17], we use Dropblock as a regularizer and the number of filters are set to (64, 160, 320, 640). Since BML have a binocular structure, we share the first three blocks and assign an independent block-4 to each *view*.

**Optimization setup.** We use SGD optimizer with a momentum of 0.9 and a weight decay of  $5e-4$ . The learning rate is initialized as 0.1. For *miniImageNet*, CIFAR-FS and CUB, we train 100 epochs. In the 50-*th* epoch, the learning rate is reduced to  $6e-3$ , and further reduced to  $1.2e-4$  in the 70-*th* epoch. For *tieredImageNet*, we train 150 epochs and decay the learning rate by 0.1 times per 40 epochs. In particular, when organizing data, in order to adapt to *binocular* requirements, we adopt a uniform sampling strategy. The ratio in the loss is set to  $\alpha:\beta:\gamma=4:2:1$ , and the two scale factors  $\alpha_1$  and  $\alpha_2$  in the elastic loss are experimentally set to 5.5 and 0.1, respectively. To ensure the stability of the evaluation results, for each benchmark, we test 2,000 episodes and report the average performance.

Table 3. Comparison on *miniImageNet*, *tieredImageNet* and CIFAR-FS. Results with  $\dagger$  are reported in [17].

Method	Backbone	<i>miniImageNet</i>		<i>tieredImageNet</i>		CIFAR-FS	
		<i>1-shot</i>	<i>5-shot</i>	<i>1-shot</i>	<i>5-shot</i>	<i>1-shot</i>	<i>5-shot</i>
MAML $\dagger$ [10]	<i>ConvNet</i>	48.70 $\pm$ 1.84	63.11 $\pm$ 0.92	51.67 $\pm$ 1.81	70.30 $\pm$ 1.75	58.90 $\pm$ 1.90	71.50 $\pm$ 1.00
TAML [16]	<i>ConvNet</i>	51.77 $\pm$ 1.86	65.60 $\pm$ 0.93	-	-	-	-
MetaOptNet [17]	<i>ResNet12</i>	64.09 $\pm$ 0.62	80.00 $\pm$ 0.45	65.81 $\pm$ 0.74	81.75 $\pm$ 0.53	72.00 $\pm$ 0.70	84.20 $\pm$ 0.50
ProtoNet $\dagger$ [33]	<i>ConvNet</i>	49.42 $\pm$ 0.78	68.20 $\pm$ 0.66	53.31 $\pm$ 0.89	72.69 $\pm$ 0.74	55.50 $\pm$ 0.70	72.00 $\pm$ 0.60
MatchingNet $\dagger$ [39]	<i>ConvNet</i>	43.56 $\pm$ 0.84	55.31 $\pm$ 0.73	-	-	-	-
RelationNet $\dagger$ [35]	<i>ConvNet</i>	50.44 $\pm$ 0.82	65.32 $\pm$ 0.70	54.48 $\pm$ 0.93	71.32 $\pm$ 0.78	55.00 $\pm$ 1.00	69.30 $\pm$ 0.80
DeepEMD [45]	<i>ResNet12</i>	65.91 $\pm$ 0.82	82.41 $\pm$ 0.56	71.16 $\pm$ 0.80	86.03 $\pm$ 0.58	46.47 $\pm$ 0.70	63.22 $\pm$ 0.71
FEAT [44]	<i>ResNet12</i>	66.78 $\pm$ 0.20	82.05 $\pm$ 0.14	70.80 $\pm$ 0.23	84.79 $\pm$ 0.16	-	-
TADAM [27]	<i>ResNet12</i>	58.50 $\pm$ 0.30	76.70 $\pm$ 0.30	-	-	-	-
CTM [19]	<i>ResNet18</i>	64.12 $\pm$ 0.82	80.51 $\pm$ 0.13	68.41 $\pm$ 0.39	84.28 $\pm$ 1.73	-	-
LR+ICI [41]	<i>ResNet12</i>	66.80 $\pm$ n/a	79.26 $\pm$ n/a	80.79 $\pm$ n/a	87.92 $\pm$ n/a	73.97 $\pm$ n/a	84.13 $\pm$ n/a
Rethink-Distill [36]	<i>ResNet12</i>	64.82 $\pm$ 0.60	82.14 $\pm$ 0.43	71.52 $\pm$ 0.69	86.03 $\pm$ 0.49	73.90 $\pm$ 0.80	86.90 $\pm$ 0.50
DC [24]	<i>ResNet12</i>	61.26 $\pm$ 0.20	79.01 $\pm$ 0.13	-	-	-	-
MTL [34]	<i>ResNet12</i>	61.20 $\pm$ 1.80	75.50 $\pm$ 0.80	-	-	-	-
CloserLook++ [6]	<i>ResNet18</i>	51.87 $\pm$ 0.77	75.68 $\pm$ 0.63	-	-	-	-
Meta-Baseline [7]	<i>ResNet12</i>	63.17 $\pm$ 0.23	79.26 $\pm$ 0.17	68.62 $\pm$ 0.27	83.29 $\pm$ 0.18	-	-
Neg-Cosine [25]	<i>ResNet12</i>	63.85 $\pm$ 0.81	81.57 $\pm$ 0.56	-	-	-	-
AFHN [20]	<i>ResNet18</i>	62.38 $\pm$ 0.72	78.16 $\pm$ 0.56	-	-	68.32 $\pm$ 0.93	81.45 $\pm$ 0.87
Centroid [1]	<i>ResNet18</i>	59.88 $\pm$ 0.67	80.35 $\pm$ 0.73	69.29 $\pm$ 0.56	85.97 $\pm$ 0.49	-	-
Baseline- <i>local</i>	<i>ResNet12</i>	58.96 $\pm$ 0.45	77.07 $\pm$ 0.34	64.46 $\pm$ 0.51	82.21 $\pm$ 0.36	67.60 $\pm$ 0.49	84.78 $\pm$ 0.34
Baseline- <i>global</i>	<i>ResNet12</i>	61.71 $\pm$ 0.48	81.21 $\pm$ 0.32	63.27 $\pm$ 0.52	82.22 $\pm$ 0.36	69.74 $\pm$ 0.49	87.37 $\pm$ 0.34
<b>BML</b>	<i>ResNet12</i>	<b>67.04<math>\pm</math>0.63</b>	<b>83.63<math>\pm</math>0.29</b>	<b>68.99<math>\pm</math>0.50</b>	<b>85.49<math>\pm</math>0.34</b>	<b>73.45<math>\pm</math>0.47</b>	<b>88.04<math>\pm</math>0.33</b>

### 4.3. Experimental Results

#### 4.3.1 Comparison on Coarse-grained Benchmark

Comparison to prior works are shown in Table 3, our results are highlighted in bold with gray background. Specific structural details of *ConvNet* are reported as follows (filter number of four blocks): MAML [10]: 32-32-32-32; TAML [16], ProtoNet [33], MatchingNet [39]: 64-64-64-64; RelationNet [35]: 64-96-128-256.

As is reported, compared with two baselines, the performance of BML is remarkable, even 9% higher in some cases (details is analyzed in Section.4.3.3). Compared with other competitors, BML achieves a new SoTA on *miniImageNet*, including the best metric-based method FEAT [44]. On *tieredImageNet*, we also achieve good performance by simply using the *nearest neighbor* principle. As for CIFAR-FS, we surpass all competitors and reach a new SoTA, including LR+ICI [41] which is based on the transductive strategy.

#### 4.3.2 Comparison on Fine-grained Benchmark

**Within domain evaluation** results is shown in Table 4 (Results with  $\dagger$  are reported in [6]). BML outperforms the runner-up DeepEMD [45] by 1.76% and 0.56% at 5-shot and 1-shot setting respectively. Although DeepEMD [45] adopts the similar idea of spatial information mining, but the task-dependent patch-wise matching is time-luxurious, while our BML shifts the attention to the bottom embed-

ding, which is time-efficient and valid.

Table 4. Within domain comparison on CUB-200-2011.

Method	Backbone	CUB-200-2011	
		<i>5-way 1-shot</i>	<i>5-way 5-shot</i>
MAML $\dagger$ [10]	<i>ResNet18</i>	68.42 $\pm$ 1.07	83.47 $\pm$ 0.62
ProtoNet $\dagger$ [33]	<i>ResNet18</i>	72.99 $\pm$ 0.88	86.64 $\pm$ 0.51
MatchingNet $\dagger$ [39]	<i>ResNet18</i>	73.49 $\pm$ 0.89	84.45 $\pm$ 0.58
RelationNet $\dagger$ [35]	<i>ResNet18</i>	68.58 $\pm$ 0.94	84.05 $\pm$ 0.56
DeepEMD [45]	<i>ResNet12</i>	75.65 $\pm$ 0.83	88.69 $\pm$ 0.50
CloserLook [6]	<i>ResNet18</i>	47.12 $\pm$ 0.74	64.16 $\pm$ 0.71
CloserLook++ [6]	<i>ResNet18</i>	60.53 $\pm$ 0.83	79.34 $\pm$ 0.61
Centroid [1]	<i>ResNet18</i>	74.22 $\pm$ 1.09	88.65 $\pm$ 0.55
Baseline- <i>local</i>	<i>ResNet12</i>	66.79 $\pm$ 0.49	86.55 $\pm$ 0.28
Baseline- <i>global</i>	<i>ResNet12</i>	60.13 $\pm$ 0.49	79.77 $\pm$ 0.36
<b>BML</b>	<i>ResNet12</i>	<b>76.21<math>\pm</math>0.63</b>	<b>90.45<math>\pm</math>0.36</b>

What’s more, comparing the performance of baseline-*global* reported in Tables 3-4 and Figure 3, we find that the performance of baseline-*global* is significantly reduced on fine-grained benchmark CUB, which shows that single global training loses its effect when the inter-class difference is relatively small. However, our BML performs well on all granularities, which proves that binocular framework is more robust against granularity change.

**Cross domain evaluation** results is reported in Table 5 (Results with  $\dagger$  are reported in [1]), BML outperforms all compared methods with a large margin. Among them, the method in [36] which also employs online soft label to improve the learning of inter-class relations, is 3.85% lower



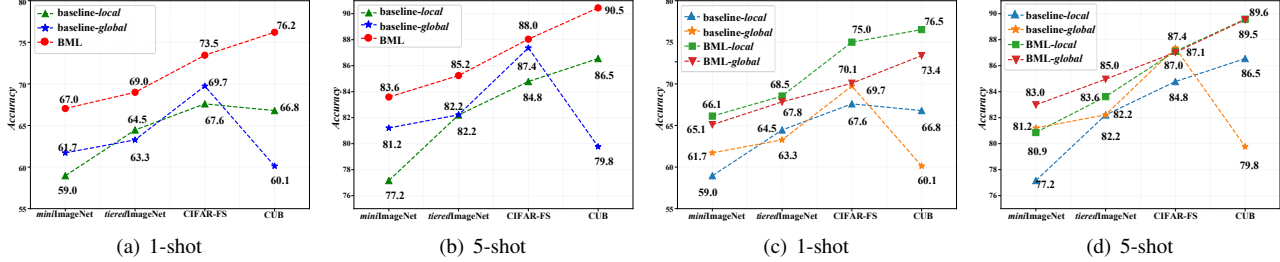


Figure 3. Effect of binocular learning on multiple benchmarks. The comparison result proves the superiority of our BML.

Table 5. Cross-domain comparison on CUB-200-2011.

Method	<i>miniImageNet</i> → CUB
MatchingNet [39]	53.07±0.74
ProtoNet <sup>†</sup> [33]	62.02±0.70
MAML <sup>†</sup> [10]	51.34±0.72
RelationNet <sup>†</sup> [35]	57.71±0.73
CloserLook [6]	65.57±0.70
CloserLook++ [6]	62.04±0.76
Centroid [1]	70.37±1.02
Rethink-distill [36]	68.57±0.39
<b>BML</b>	<b>72.42±0.54</b>

than our binocular mutual learning mechanism. This shows that bidirectional cross-view interaction is better than uni-directional same-view interaction.

### 4.3.3 Ablation Study

**Analyze of Binocular Learning:** Specifically, our baselines are two typical single-view methods: *baseline-global* and a stronger *ProtoNet* *baseline-local*. All the experimental configurations of the above two are the same as BML. According to the results shown in Figure 3, we highlight three observations:

1) Binocular mutual training can effectively integrate the advantages of the two views and obtain complementary integrated features. For instance, under 1-shot and 5-shot setting shown in subfigure.3(a) and 3(b), BML is significantly better than the two single-view training methods on both coarse-grained and fine-grained benchmarks.

2) Binocular mutual training makes the two views promote each other. As is shown in subfigure 3(c) and 3(d), BML-*local* and BML-*global* based on binocular training are better than *baseline-local* and *baseline-global* based on single-view mode, where BML-*local* is 1.4%-9.7% higher than *baseline-local*, and BML-*global* is 1.8%-13.3% higher than *baseline-global*. Since the only difference is whether the training is under binocular pattern, the results confirm the effectiveness of binocular training.

3) Binocular mutual training is less sensitive to shot count. By comparison, we find that BML-*local* and BML-*global* show different advantages under different settings.

The *global* branch performs better when shot count is greater than 1, indicating that *global* view captures richer expressions; while under *1-shot* setting, the *local* branch exhibits a higher advantage, which shows that *local* view is robust against sampling uncertainty, because local view minimizes the variation of features within a class according to the variation between classes [12]. Binocular aggregation avoids the instability of single view and absorbs the advantages of two branches.

**Analyze of Mutual Interaction ( $\mathcal{L}_{mutual}$ ):** Applying mutual interaction loss  $\mathcal{L}_{mutual}$  on the binocular framework, the performance is further improved by 0.79% as follows:

	w/o $\mathcal{L}_{mutual}$	w/ $\mathcal{L}_{mutual}$
BML	82.84	83.63(↑ 0.79)

Besides, we separately analyze the impact of  $\mathcal{L}_{mutual}$  on single view. Taking the local branch as example, the performance comparison with or without  $\mathcal{L}_{mutual}$  is as follows:

	w/o $\mathcal{L}_{mutual}$	w/ $\mathcal{L}_{mutual}$
BML- <i>Local</i>	79.29	80.95(↑ 1.66)

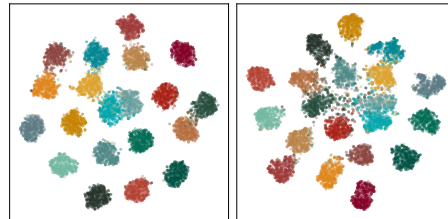


Figure 4. t-SNE results on *miniImageNet* base classes.

Together with qualitative results in Figure 4, we find that compared with the left one (w/o  $\mathcal{L}_{mutual}$ ), after introducing  $\mathcal{L}_{mutual}$  (right), the cluster structure on base classes is slightly broke but the performance on novel is significantly improved, which shows that in addition to the single hard label, minimizing the mimicry loss helps to alleviate the over-fitting problem while improving transferability.

**Analyze of Elastic Loss (ELoss):** We simply update the basic euclidean distance with the elastic loss, and keep the other settings unchanged. Here is the comparison result (all the experiments are conducted on *miniImageNet*):

	w/o ELloss	w/ ELloss
baseline-Local	77.07	77.77( $\uparrow$ 0.7)

Obviously, the introduction of ELloss has brought a great performance improvement. Next, we made a further detailed analysis:

**Effect of  $\alpha_1$  and  $\alpha_2$ :**  $\alpha_1$  and  $\alpha_2$  in Algorithm.1 are two factors controlling the push degree. Among them,  $\alpha_1 \in [4, 6]$ , this range is derived from the observation of normal training:  $\mathcal{M}(C_p, x_q)$  stably falls in [6, 12] after convergence. Therefore, we enlarge  $\mathcal{M}(C_p, x_q)$  about 50%. The range of  $\alpha_2$  is [0.05, 0.25], which controls push degree cross tasks. As shown in Figure 5, simply employing ELloss on baseline-local can obtain a significant improvement (best configuration  $\alpha_1 = 5.5, \alpha_2 = 0.1$ ), which shows that, compare to equally treats all the tasks, redefining the difficulty is beneficial to dig out more transferable knowledge.

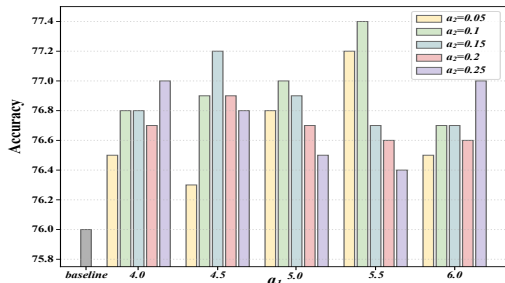


Figure 5. Effect of  $\alpha_1$  and  $\alpha_2$  (train 100 epochs with  $N=5$ ).

**Effect of Elastic Loss under different settings:** We further analyze ELloss in Figure 6, and the result shows a fact that local mode is indeed a relatively simple problem since the matching process only occurs within current episode, which naturally, makes the network too lazy to give satisfied solutions. The introduction of ELloss delays the appearance of performance saturation points, thereby alleviating the problem of sub-optimal solutions to some extent.

**Analyze of Stability:** A good model should perform well in any situation, especially when attacked by degraded input. The following stability test in Table 6 is carried out in several groups of degraded cases: size change; blur attack; noise attack and brightness attack. Results show that our BML performs well in any situation, which shows the stability of binocular learning strategy.

**Is Similarity Ranking important for a good model?** We randomly visualize a task (Figure 7) and find an interesting sidelight: while improving discriminability, BML also gets more accurate inter-class relationship and more accurate heatmap. This inspired us to think: for the current closed-set setup of FSC, should we pay attention to the similarity ranking besides the top-1 accuracy? Since semantical effective ranking is more practical and can further distinguish the advantages of existing methods.

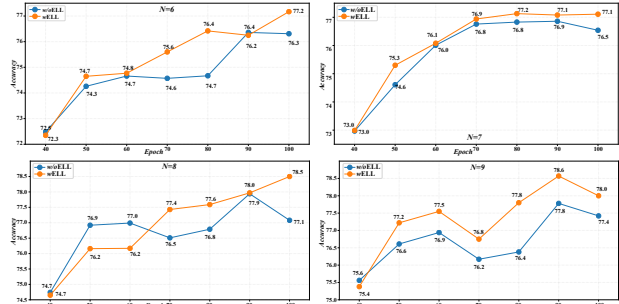


Figure 6. Performance trend of Elastic Loss with different  $N$ .

	$84 \times 84 \rightarrow 224 \times 224$
DeepEMD [45]	82.41 $\rightarrow$ 78.12
<b>BML</b>	<b>83.59 <math>\rightarrow</math> 81.57</b>
	+ GaussianBlur ( $\sigma \in [0.1, 2]$ )
Rethink-Distill [36]	82.14 $\rightarrow$ 49.30
<b>BML</b>	<b>83.59 <math>\rightarrow</math> 61.96</b>
	+ pepperNoise ( $r = 0.01$ )
Rethink-Distill [36]	82.14 $\rightarrow$ 63.97
<b>BML</b>	<b>83.59 <math>\rightarrow</math> 71.02</b>
	+ ColorJitter ( $B = 0.8$ )
Rethink-Distill [36]	82.14 $\rightarrow$ 81.05
<b>BML</b>	<b>83.59 <math>\rightarrow</math> 82.24</b>

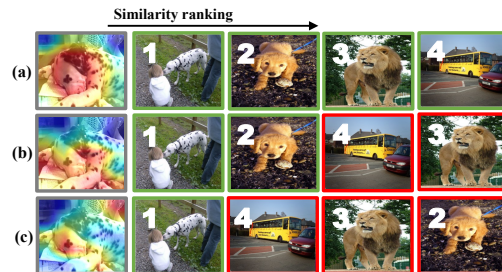


Figure 7. Similarity Ranking on a 4-way 1-shot task. (golden retriever, dalmatian, lion and bus). (a) Our BML, (b) single global view, (c) single local view. BML gives completely accurate ranking results while the left two failed to get inter-class relationship.

## 5. Conclusion

Inspired by mutual learning paradigm and binocular parallax, we propose a unified Binocular Mutual Learning (BML) framework, which achieves the compatibility of the global view and the local view through both intra-view and cross-view modeling. The effectiveness of BML has been fully demonstrated on both within-domain and cross-domain evaluations. The aggregated features are more robust than other competitors when dealing with degradation attacks. Besides, BML obtain accurate similarity ranking.

**Acknowledgement.** This work is supported by the National Key R&D Plan of the Ministry of Science and Technology (Project No. 2020AAA0104400).



## References

- [1] Arman Afrasiyabi, Jean-François Lalonde, and Christian Gagné. Associative alignment for few-shot image classification. In *European Conference on Computer Vision*, pages 18–35. Springer, 2020.
- [2] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48, 2009.
- [3] Luca Bertinetto, Joao F Henriques, Philip HS Torr, and Andrea Vedaldi. Meta-learning with differentiable closed-form solvers. *arXiv preprint arXiv:1805.08136*, 2018.
- [4] Malik Boudiaf, Ziko Imtiaz Masud, Jérôme Rony, José Dolz, Pablo Piantanida, and Ismail Ben Ayed. Transductive information maximization for few-shot learning. *arXiv preprint arXiv:2008.11297*, 2020.
- [5] Mengting Chen, Yuxin Fang, Xinggang Wang, Heng Luo, Yifeng Geng, Xinyu Zhang, Chang Huang, Wenyu Liu, and Bo Wang. Diversity transfer network for few-shot learning. In *AAAI*, pages 10559–10566, 2020.
- [6] Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Frank Wang, and Jia-Bin Huang. A closer look at few-shot classification. *arXiv preprint arXiv:1904.04232*, 2019.
- [7] Yinbo Chen, Xiaolong Wang, Zhuang Liu, Huijuan Xu, and Trevor Darrell. A new meta-baseline for few-shot learning. *arXiv preprint arXiv:2003.04390*, 2020.
- [8] Nanyi Fei, Zhiwu Lu, Tao Xiang, and Songfang Huang. {MELR}: Meta-learning via modeling episode-level relationships for few-shot learning. In *International Conference on Learning Representations*, 2021.
- [9] Li Fei-Fei, Rob Fergus, and Pietro Perona. One-shot learning of object categories. *IEEE transactions on pattern analysis and machine intelligence*, 28(4):594–611, 2006.
- [10] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, pages 1126–1135. PMLR, 2017.
- [11] Yixiao Ge, Dapeng Chen, and Hongsheng Li. Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification. *arXiv preprint arXiv:2001.01526*, 2020.
- [12] Micah Goldblum, Steven Reich, Liam Fowl, Renkun Ni, Valeriia Cherepanova, and Tom Goldstein. Unraveling meta-learning: Understanding feature representations for few-shot tasks. In *International Conference on Machine Learning*, pages 3607–3616. PMLR, 2020.
- [13] Yiluan Guo and Ngai-Man Cheung. Attentive weights generation for few shot learning via information maximization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13499–13508, 2020.
- [14] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.
- [15] Ruibing Hou, Hong Chang, Bingpeng Ma, Shiguang Shan, and Xilin Chen. Cross attention network for few-shot classification. *arXiv preprint arXiv:1910.07677*, 2019.
- [16] Muhammad Abdullah Jamal and Guo-Jun Qi. Task agnostic meta-learning for few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11719–11727, 2019.
- [17] Kwonjoon Lee, Subhansu Maji, Avinash Ravichandran, and Stefano Soatto. Meta-learning with differentiable convex optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10657–10665, 2019.
- [18] Aoxue Li, Weiran Huang, Xu Lan, Jiashi Feng, Zhenguo Li, and Liwei Wang. Boosting few-shot learning with adaptive margin loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12576–12584, 2020.
- [19] Hongyang Li, David Eigen, Samuel Dodge, Matthew Zeiler, and Xiaogang Wang. Finding task-relevant features for few-shot learning by category traversal. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1–10, 2019.
- [20] Kai Li, Yulun Zhang, Kunpeng Li, and Yun Fu. Adversarial feature hallucination networks for few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13470–13479, 2020.
- [21] Tianhong Li, Jianguo Li, Zhuang Liu, and Changshui Zhang. Few sample knowledge distillation for efficient network compression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14639–14647, 2020.
- [22] Wenbin Li, Lei Wang, Jinglin Xu, Jing Huo, Yang Gao, and Jiebo 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*, pages 7260–7268, 2019.
- [23] Zhenguo Li, Fengwei Zhou, Fei Chen, and Hang Li. Meta-sgd: Learning to learn quickly for few-shot learning. *arXiv preprint arXiv:1707.09835*, 2017.
- [24] Yann Lifchitz, Yannis Avrithis, Sylvaine Picard, and Andrei Bursuc. Dense classification and implanting for few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9258–9267, 2019.
- [25] Bin Liu, Yue Cao, Yutong Lin, Qi Li, Zheng Zhang, Mingsheng Long, and Han Hu. Negative margin matters: Understanding margin in few-shot classification. In *European Conference on Computer Vision*, pages 438–455. Springer, 2020.
- [26] Kushagra Mahajan, Monika Sharma, and Lovekesh Vig. Meta-dermdiagnosis: Few-shot skin disease identification using meta-learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 730–731, 2020.
- [27] Boris N Oreshkin, Pau Rodriguez, and Alexandre Lacoste. Tadam: Task dependent adaptive metric for improved few-shot learning. *arXiv preprint arXiv:1805.10123*, 2018.
- [28] Limeng Qiao, Yemin Shi, Jia Li, Yaowei Wang, Tiejun Huang, and Yonghong Tian. Transductive episodic-wise adaptive metric for few-shot learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3603–3612, 2019.

- [29] Siyuan Qiao, Chenxi Liu, Wei Shen, and Alan L Yuille. Few-shot image recognition by predicting parameters from activations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7229–7238, 2018.
- [30] Aniruddh Raghu, Maithra Raghu, Samy Bengio, and Oriol Vinyals. Rapid learning or feature reuse? towards understanding the effectiveness of maml. *arXiv preprint arXiv:1909.09157*, 2019.
- [31] Mengye Ren, Eleni Triantafillou, Sachin Ravi, Jake Snell, Kevin Swersky, Joshua B Tenenbaum, Hugo Larochelle, and Richard S Zemel. Meta-learning for semi-supervised few-shot classification. *arXiv preprint arXiv:1803.00676*, 2018.
- [32] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.
- [33] Jake Snell, Kevin Swersky, and Richard S Zemel. Prototypical networks for few-shot learning. *arXiv preprint arXiv:1703.05175*, 2017.
- [34] Qianru Sun, Yaoyao Liu, Tat-Seng Chua, and Bernt Schiele. Meta-transfer learning for few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 403–412, 2019.
- [35] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1199–1208, 2018.
- [36] Yonglong Tian, Yue Wang, Dilip Krishnan, Joshua B Tenenbaum, and Phillip Isola. Rethinking few-shot image classification: a good embedding is all you need? *arXiv preprint arXiv:2003.11539*, 2020.
- [37] Eleni Triantafillou, Richard Zemel, and Raquel Urtasun. Few-shot learning through an information retrieval lens. *arXiv preprint arXiv:1707.02610*, 2017.
- [38] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017.
- [39] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. *arXiv preprint arXiv:1606.04080*, 2016.
- [40] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.
- [41] Yikai Wang, Chengming Xu, Chen Liu, Li Zhang, and Yanwei Fu. Instance credibility inference for few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12836–12845, 2020.
- [42] Ling Yang, Liangliang Li, Zilun Zhang, Xinyu Zhou, Erjin Zhou, and Yu Liu. Dpgn: Distribution propagation graph network for few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13390–13399, 2020.
- [43] Huaxiu Yao, Yiding Liu, Ying Wei, Xianfeng Tang, and Zhenhui Li. Learning from multiple cities: A meta-learning approach for spatial-temporal prediction. In *The World Wide Web Conference*, pages 2181–2191, 2019.
- [44] Han-Jia Ye, Hexiang Hu, De-Chuan Zhan, and Fei Sha. Few-shot learning via embedding adaptation with set-to-set functions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8808–8817, 2020.
- [45] Chi Zhang, Yujun Cai, Guosheng Lin, and Chunhua Shen. Deepemd: Few-shot image classification with differentiable earth mover’s distance and structured classifiers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12203–12213, 2020.
- [46] Manli Zhang, Jianhong Zhang, Zhiwu Lu, Tao Xiang, Mingyu Ding, and Songfang Huang. {IEPT}: Instance-level and episode-level pretext tasks for few-shot learning. In *International Conference on Learning Representations*, 2021.
- [47] Ying Zhang, Tao Xiang, Timothy M Hospedales, and Huchuan Lu. Deep mutual learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4320–4328, 2018.