SEMANTICALLY PROPORTIONAL PATCHMIX FOR FEW-SHOT LEARNING

Jingquan Wang[†], Jing Xu[†], Yu Pan[†], Zenglin Xu[†,‡,]*

[†]Harbin Institute of Technology (Shenzhen), Shenzhen, Guangdong, China
[‡]PengCheng Laboratory, Shenzhen, Guangdong, China

ABSTRACT

Few-shot learning aims to classify unseen classes with only a limited number of labeled data. Recent works have demonstrated that training models with a simple transfer learning strategy can achieve competitive results in few-shot classification. Although excelling at distinguishing training data, these models are not well generalized to unseen data, probably due to insufficient feature representations on evaluation. To tackle this issue, we propose Semantically Proportional Patchmix (**SePPMix**), in which patches are cut and pasted among training images and the ground truth labels are mixed proportionally to the semantic information of the patches. In this way, we can improve the generalization ability of the model by regional dropout effect without introducing severe label noise. To learn more robust representations of data, we further take rotate transformation on the mixed images and predict rotations as a rule-based regularizer. Extensive experiments on prevalent few-shot benchmarks have shown the effectiveness of our proposed method.

Index Terms— few-shot learning, image classification, data augmentation, generalization

1. INTRODUCTION

Deep learning has achieved significant progress in computer vision tasks. This success can be partly attributed to the availability of large-scale data [1]. However, acquiring enough labeled data is infeasible in some situations due to event scarcity or expensive labor costs. Inspired by the human visual system that can learn new classes with only a few samples, few-shot learning (FSL) emerges as a promising method that aims to equip a learner with fast adaptability to novel concepts. For FSL, there are usually a base set and a novel set that share no class overlap. Then a model is trained on the base set with sufficient labeled data, and is further adapted to categorize the unseen classes in the novel set with scarce labeled data. In particular, the performance of the few-shot classification highly relies on the model's generalization ability.

Recent works [2, 3] have shown that a model trained on the base set with standard supervision can achieve impressive results on the novel set. However, we argue that the learned model tends to be overly discriminative to the training data in the base set, leading to sub-optimal performance when evaluated on unseen classes. This is because that deep networks are likely to memorize some specific training statistics, resulting in overconfident prediction and poor generalization ability [4]. To solve this issue and make the model produce sufficient discriminative representations of unseen test data, it is essential to introduce uncertainty to the input data and to regularise the model in the training process.

Data mixing methods introduce uncertainty and reduce the risk of overly memorizing input data by blending original images and labels, which have proved valid in training a more general and transferable model for image classification recently [5–7]. In this paper, we propose a simple yet effective regularization method called Semantically Proportional Patchmix (**SePPMix**) to improve the performance of FSL. In SePPMix, we divide the input images into N×N patches. The patches are randomly cut from one image and pasted onto another image at their respective original locations, which prevents the model from learning over specific structures of training data. Furthermore, rather than directly using the area of patches, the label of the mixed image is generated by the semantic proportion of patches which is estimated by the class activation maps (CAMs) [8]. To increase the number of mixed samples and learn more robust representations of data, we rotate the mixed images and predict the rotated angle of the images as an auxiliary task. We conduct extensive experiments on two widely-used datasets, the empirical results show that the proposed method can significantly improve few-shot image classification performance over corresponding baseline methods. Remarkably, under the same experiment setting, the proposed method improves 1-shot and 5-shot tasks by nearly **6%** and **5%** on MiniImageNet and CUB, respectively.

2. RELATED WORKS

Few-shot learning aims to build a model using available training data of base classes that can classify unseen novel classes with few labeled examples in each class. One popular paradigm to tackle few-shot classification is meta-learning,

which can be roughly divided into two streams, optimized-based methods [9–11] and metric-based methods [12–15]. The first class of methods, i.e., learning to learn methods, aiming to learn a suitable model initialization, so that the model can achieve fast adaption with a few labeled examples. The second category is learning to compare methods, targeting at projecting input samples into a specific embedding space where data from different classes can be distinguished by using distance metrics. Apart from these, recent works [2, 3] adopt the transfer learning strategy to train the model on all the base classes, then use it to extract the feature representation of input data on the novel classes, showing competitive performance. Our work follows this school of thought and further improves the generalization of the model.

Data augmentation has been widely used to train deep learning models. Regional-erasing methods [16, 17] erasing part of the training images, aiming to encourage the networks better utilize the entire context of the data. Mixup [18] generates images by linearly combining data and fusing their labels using the same coefficients and shows its generalization ability in large datasets; Manifold Mixup [19] applies Mixup to the intermediate layer, leading to smoother decision boundaries which benefit feature representation; Cutmix [6] combines the advantages of Cutout [17] and Mixup, which produces a new image by cutting out one random image patch and pasting it to another image. Different from Cutmix, Snapmix [7] generates the target label for the mixed image by estimating its intrinsic semantic composition, which achieves top-level performance in fine-grained data. However, these data mixing methods are not specifically designed for limited data, which may result in suboptimal performance in FSL tasks.

3. METHOD

3.1. Problem Formulation

Few-shot learning usually consists of two disjoint datasets $\mathcal{D}_b$ and $\mathcal{D}_n$. A model is firstly trained on the sufficient labeled dataset $\mathcal{D}_b$, and then performs the FSL task based on the novel dataset $\mathcal{D}_n$. The common way to build the FSL task is called N-way K-shot that should classify N classes sampled from $\mathcal{D}_n$ with K labeled data in each class. The few labeled data is called support set $\mathcal{S} = \{x_n^j, y_n^j\}_{j=1}^{N \times K}$ that contains N classes with each class K examples. The performance is evaluated on the query set $\mathcal{Q} = \{x_n^j, y_n^j\}_{j=N \times K+1}^{N \times K+H}$, where the H denotes the number of unlabeled examples and the data in $\mathcal{Q}$ is sampled from the same N classes in each episode. Most meta-learning methods [9, 12, 13] adopt the same episode training scheme as evaluation when trained on $\mathcal{D}_b$, making training and evaluation stage in the same situation. However, recent works [2, 3, 20] show that training the model on the whole $\mathcal{D}_b$ in a supervised manner is efficient for the FSL task. We follow the same idea and utilize our proposed method to train a
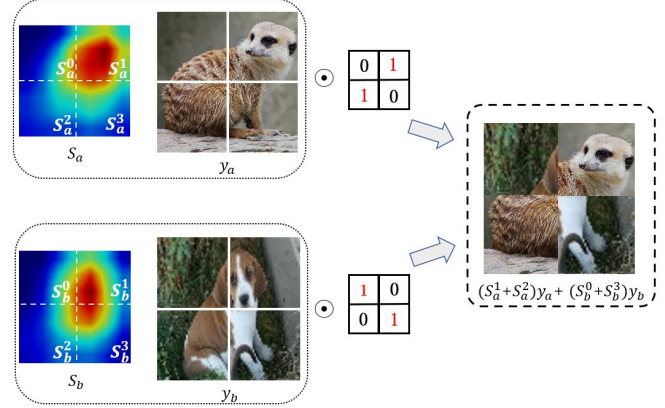

Fig. 1. An overview of SePPMix. S_a and S_b are the semantic information maps of input image x_a and x_b respectively.

more generalized embedding network on $\mathcal{D}_b$, which performs better than baseline when tested on $\mathcal{D}_n$.

3.2. Framework Overview

In this work, we consider the simple but effective procedure as [3], during the first stage, we train the embedding network f_ϕ on $\mathcal{D}_b$,

$$\phi = \arg\min_\phi \mathcal{L}_{base}(\mathcal{D}_b; \phi), \tag{1}$$

where $\mathcal{L}_{base}$ is the loss function, ϕ is the parameters of the embedding model. Then the pre-trained f_ϕ is fixed and used as the feature extractor in D_n. A linear classifier g_θ is firstly trained on the extracted features of $\mathcal{S}$,

$$\theta = \arg\min_\theta \mathcal{L}_{base}(\mathcal{S}; \theta, \phi) + R(\theta), \tag{2}$$

where θ contains weight and bias terms and R is the regulation term. Then g_θ can be leveraged as predictor on the features of $\mathcal{Q}$.

3.3. Sample Generation via SePPMix

In few-shot learning, SePPMix can better improve the model generalization than other data mixing methods. On the one hand, it increases the uncertainty and diversity of the training data; on the other hand, it generates more accurate labels of mixed data without introducing severe noise. Fig. 1 illustrates the procedure of our proposed method.

Formally, let $x \in \mathbb{R}^{3 \times h \times w}$ denotes the training image and y is its corresponding label. SePPMix create a new training sample $(\hat{x}, \hat{y})$ given two distinct samples (x_a, y_a) and (x_b, y_b). The generation of mixed image is defined as

$$\hat{x} = M \odot x_a + (\mathbf{1} - M) \odot x_b, \tag{3}$$

where $M \in \{0, 1\}^{N \times N}$ denotes a binary mask indicating which patches belong to either of the two images. The $\mathbf{1}$ is a

binary mask filled with ones and the probability of its appearing is 0.5, and $\odot$ denotes the element-wise multiplication.

When using the original data of the base set for training, the different patches in the image contain varying levels of information that are valid for the classification task. To generate a more accurate label, we calculate the class activation maps of images as the indicators of semantic information and use them to estimate how a patch correlates with its label, which has been proved useful to measure the semantic relatedness of each original image patch to the corresponding label [7]. Given an image x, the class activation map can be calculated by

$$CAM(x) = \Phi(\sum_{l=1}^{c} w_y^l F_l(x)), \qquad (4)$$

where $\Phi(\cdot)$ is the operation that upsamples a feature map to match the dimensions of the input image. $F(\cdot)$ denotes the feature extractor and $F_l(x)$ denotes the lth feature map with input x, c is the number of feature maps. The w_y is the classifier weight of class y, and for simplicity, the bias term is ignored. We obtain the semantic information map $S(x)$ by normalizing the $CAM(x)$ to sum to one. The semantic information map of x, $S(x)$, is defined as

$$S(x) = \frac{CAM(x)}{sum(CAM(x))}. \qquad (5)$$

Finally, we can calculate the corresponding semantic label of the image produced using Eq. 3,

$$\rho_a = sum(M \odot S(x_a)), \qquad (6)$$
$$\rho_b = sum((\mathbf{1} - M) \odot S(x_b)), \qquad (7)$$
$$\hat{y} = \rho_a y_a + \rho_b y_b, \qquad (8)$$

where ρ_a and ρ_b are the weights corresponding to label y_a and y_b respectively. The newly generated image label in Fig 1 depicts this process. In this way, not only can the model learn rich visual features, but also prevent introducing heavy noise in the augmented data, especially when in the few-shot situation.

3.4. Training SePPMix

Our SePPMix is applied in the training stage to learn better representations with generalization ability. Given a new sample $(\hat{x}, \hat{y})$ generated by SePPMix, we rotate it and predict the angles as an auxiliary task to learn more generalizable features. $\hat{x}^r$ denotes the mixed image rotated by r degrees, r $\in C_R = \{0°, 90°, 180°, 270°\}$. The loss $\mathcal{L}_m$ for training the image classification task using mixed images and the auxiliary loss $\mathcal{L}_r$ for predicting the rotation angle are defined as

$$\mathcal{L}_m = \sum_{x \in D_b} \sum_{r \in C_R} \mathcal{L}_{ce}[f_\phi(\hat{x}^r), \hat{y}], \qquad (9)$$

$$\mathcal{L}_r = \frac{1}{|C_R|} \sum_{x \in D_b} \sum_{r \in C_R} L_{ce}[g_r(f_\phi(\hat{x}^r)), r], \qquad (10)$$

Table 1. Average 5-way few-shot classification accuracies (%) with 95% confidence intervals on miniImageNet dataset.

Method	Backbone	1-shot	5-shot
MAML [9]	32-32-32-32	48.70±1.84	63.11±0.92
MatchingNet [12]	64-64-64-64	43.56±0.84	55.31±0.73
ProtoNet [13]	64-64-64-64	49.42±0.78	68.20±0.66
TADAM [23]	ResNet-12	58.50±0.30	76.70±0.30
MTL [24]	ResNet-12	61.20±1.80	75.50±0.80
MetaOptNet [25]	ResNet-12	62.64±0.61	78.63±0.64
Fine-tuning [20]	WRN-28-10	57.73±0.62	78.17±0.49
VLCL [26]	WRN-28-10	61.75±0.43	76.32±0.49
S2M2 [5]	WRN-28-10	64.93±0.18	83.18±0.11
FEAT [27]	ResNet-12	66.78±0.20	82.05±0.14
DeepEMD [15]	ResNet-12	65.91±0.82	82.41±0.56
RFS-distill [3]	ResNet-12	64.82±0.60	82.14±0.43
RFS-simple [3]	ResNet-12	62.02±0.63	79.64±0.44
Ours	ResNet-12	**66.98±0.81**	**83.88±0.54**

where $\mathcal{L}_{ce}$ is standard cross-entropy loss function and $g_r(\cdot)$ is a 4-way linear classifier. $|C_R|$ denotes the cardinality of C_R. Then the overall loss in training stage is

$$\mathcal{L}_{base} = \alpha \mathcal{L}_m + \beta \mathcal{L}_r. \qquad (11)$$

where α and β are the weighting factors. For evaluation, the embedding network f_ϕ is fixed and tested on FSL tasks consists of $\mathcal{S}$ and $\mathcal{Q}$ sampled from $\mathcal{D}_n$ which plays a role as feature extractor, and we obtain embeddings of the images in $\mathcal{S}$ and $\mathcal{Q}$. We follow the implementation of [3], which trains a logistic regression classifier based on the embeddings of images in $\mathcal{S}$ and their corresponding labels, and uses the classifier to predict the labels of images in $\mathcal{Q}$.

4. EXPERIMENTS

4.1. Datasets

We perform our experiments on two widely used datasets in FSL, i.e., miniImageNet [21] and CUB [22]. The miniImageNet is a subset of ILSVRC-12 [1], including 100 distinct classes, each of which contains 600 images of size 84 × 84. We adopt the common setup introduced by [21], which split the categories into 64,16,20 classes for training, validation and evaluation respectively. The CUB dataset consisting of 11,788 images of size 84 × 84 from 200 bird classes. We follow the splits from [2], where 100 classes are used for training, 50 for validation, and 50 for testing.

4.2. Implementation Details

Following [3], we use a ResNet-12 network as our backbone, which consists of 4 residual blocks with Dropblock as a regularizer. To generate the initial CAMs of the images, we train the network from scratch in D_b at the beginning. In all experiments, we use Stochastic Gradient Descent (SGD) with a

Table 2. Average 5-way few-shot classification accuracies (%) with 95% confidence intervals on CUB. RFS-simple* indicates the model re-implemented by ours on CUB dataset.

Method	Backbone	1-shot	5-shot
Baseline [2]	ResNet-18	65.51±0.87	82.85±0.55
Baseline++ [2]	ResNet-18	67.02±0.90	83.58±0.54
MAML [9]	ResNet-18	68.42±1.07	83.47±0.62
MatchingNet [12]	ResNet-12	71.87±0.85	85.08±0.57
ProtoNet [13]	ResNet-12	66.09±0.92	82.50±0.58
Negative-Cosine [28]	ResNet-18	72.66±0.85	89.40±0.43
S2M2 [5]	ResNet-18	71.43±0.28	85.55±0.52
DeepEMD [15]	ResNet-12	75.65±0.83	88.69±0.50
RFS-simple* [3]	ResNet-12	72.78±0.86	87.24±0.50
Ours	ResNet-12	**78.55±0.77**	**91.81±0.41**

Table 3. Performance comparison between different mix methods and ablation study of our model, r indicates rotation.

Method	miniImageNet	
	1-shot	5-shot
baseline	61.06	78.83
Mixup	59.12 (-1.94)	79.01 (+0.18)
Cutmix	61.61 (+0.55)	80.35 (+1.52)
Snapmix	62.52 (+1.46)	80.90 (+2.07)
Patchmix	62.75 (+1.69)	80.69 (+1.86)
SePPMix (w/o r)	64.09 (**+3.03**)	81.21 (**+2.38**)
SePPMix (w/ r)	66.98 (**+5.92**)	83.88 (**+5.05**)

learning rate of 0.05, a momentum of 0.9, and a weight decay of $5e^{-4}$. The training epoch is 65 epochs and the learning rate is decreased with a factor of 0.1 at 30, 45 and 60 epoch. The batch size is 64 and 16 on miniImageNet and CUB respectively. We empirically set the number of patches N=2, the coefficients $\alpha = 1$ and $\beta = 0.5$. We randomly sample 600 episodes to report the accuracies on both datasets.

4.3. Results

We present the results of SePPMix on two representative benchmarks of the few-shot learning task. As detailed in Table 1, our method outperforms all previous approaches on miniImageNet. Specifically, our approach achieves 4.96% and 4.24% improvement over our baseline RFS-simple [3] for 5-way 1-shot and 5-way 5-shot tasks respectively. Table 2 illustrates the experimental results on CUB dataset. We re-implement the baseline RFS-simple [3] on CUB. In the 5-way tasks, our method improves the accuracies by 5.77% and 4.57% over the baseline on 1/5-shot respectively. Besides, our approach outperforms all the strong competitors and sets new records on 5-way 1/5-shot tasks on CUB.
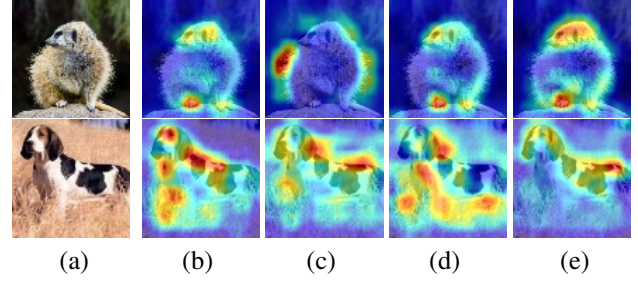

Fig. 2. Class activation maps for different methods: (a) input, (b) baseline, (c) Mixup, (d) Cutmix, and (e) SePPMix. Here the warmer color indicates higher value.

4.4. Ablations and Analysis

We conduct an ablation study to compare SePPMix with other data mixing methods and assess the effects of different components of SePPMix. We report the results on both 5-way 1/5-shot tasks. In Table 3, we train a simple baseline model without using data mixing augmentation. Notably, Mixup leads to worse performance in 5-way 1-shot task than the baseline, which indicates that linearly combining images and labels cannot improve the transfer and generalization abilities of the model in the FSL scenario. For a fair comparison, in the auxiliary task "without rotation" (w/o r), our proposed SePP-Mix (w/o r) outperforms the baseline by 3.03% and 2.38% on 1/5-shot tasks respectively, which significantly surpasses other mix-based methods. Fig. 2 gives some class activation maps of samples obtained using different methods. The proposed SePPMix pays more attention on the object while less on the background, which intuitively demonstrates that our approach yields better results.

To show that every component in our model has a valid contribution, we use Patchmix to denote that the label is determined proportionally to the area of pixels like Cutmix, while the image is generated the same as ours. The Patchmix performs better than Cutmix, which proves the effectiveness of our patch mix strategy. After combining with our semantic label, SePPMix(w/o r) without rotation improves more than 1% accuracy compared with Patchmix. Besides, with the auxiliary rotation task, SePPMix(w r) further improves the accuracies to 66.98% and 83.88% on 1/5-shot task, respectively.

5. CONCLUSION

In this paper, we aim at better generalizing the model to unseen classes in few-shot learning. To achieve this, we propose a simple but effective mix method called SePPMix, which generates new training data with semantically proportional labels. Additionally, we rotate the generated samples and predict the angles as auxiliary signals. Extensive experiments on the miniImageNet and CUB datasets verify the effectiveness of our method.

6. REFERENCES

[1] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Fei-Fei Li, "Imagenet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, 2015.

[2] Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Frank Wang, and Jia-Bin Huang, "A closer look at few-shot classification," in *ICLR (Poster)*. 2019, OpenReview.net.

[3] Yonglong Tian, Yue Wang, Dilip Krishnan, Joshua B. Tenenbaum, and Phillip Isola, "Rethinking few-shot image classification: A good embedding is all you need?," in *ECCV (14)*. 2020, vol. 12359 of *Lecture Notes in Computer Science*, pp. 266–282, Springer.

[4] Sunil Thulasidasan, Gopinath Chennupati, Jeff A. Bilmes, Tanmoy Bhattacharya, and Sarah Michalak, "On mixup training: Improved calibration and predictive uncertainty for deep neural networks," in *NeurIPS*, 2019, pp. 13888–13899.

[5] Puneet Mangla, Mayank Singh, Abhishek Sinha, Nupur Kumari, Vineeth N. Balasubramanian, and Balaji Krishnamurthy, "Charting the right manifold: Manifold mixup for few-shot learning," in *WACV*. 2020, pp. 2207–2216, IEEE.

[6] Sangdoo Yun, Dongyoon Han, Sanghyuk Chun, Seong Joon Oh, Youngjoon Yoo, and Junsuk Choe, "Cutmix: Regularization strategy to train strong classifiers with localizable features," in *ICCV*. 2019, pp. 6022–6031, IEEE.

[7] Shaoli Huang, Xinchao Wang, and Dacheng Tao, "Snapmix: Semantically proportional mixing for augmenting fine-grained data," in *AAAI*. 2021, pp. 1628–1636, AAAI Press.

[8] Bolei Zhou, Aditya Khosla, Àgata Lapedriza, Aude Oliva, and Antonio Torralba, "Learning deep features for discriminative localization," in *CVPR*. 2016, pp. 2921–2929, IEEE Computer Society.

[9] Chelsea Finn, Pieter Abbeel, and Sergey Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *ICML*. 2017, vol. 70 of *Proceedings of Machine Learning Research*, pp. 1126–1135, PMLR.

[10] Alex Nichol, Joshua Achiam, and John Schulman, "On first-order meta-learning algorithms," *CoRR*, vol. abs/1803.02999, 2018.

[11] Andrei A. Rusu, Dushyant Rao, Jakub Sygnowski, Oriol Vinyals, Razvan Pascanu, Simon Osindero, and Raia Hadsell, "Meta-learning with latent embedding optimization," in *ICLR (Poster)*. 2019, OpenReview.net.

[12] Oriol Vinyals, Charles Blundell, Tim Lillicrap, Koray Kavukcuoglu, and Daan Wierstra, "Matching networks for one shot learning," in *NIPS*, 2016, pp. 3630–3638.

[13] Jake Snell, Kevin Swersky, and Richard S. Zemel, "Prototypical networks for few-shot learning," in *NIPS*, 2017, pp. 4077–4087.

[14] Xu Luo, Longhui Wei, Liangjian Wen, Jinrong Yang, Lingxi Xie, Zenglin Xu, and Qi Tian, "Rectifying the shortcut learning of background for few-shot learning," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[15] Chi Zhang, Yujun Cai, Guosheng Lin, and Chunhua Shen, "Deepemd: Few-shot image classification with differentiable earth mover's distance and structured classifiers," in *CVPR*. 2020, pp. 12200–12210, Computer Vision Foundation / IEEE.

[16] Krishna Kumar Singh and Yong Jae Lee, "Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization," in *ICCV*. 2017, pp. 3544–3553, IEEE Computer Society.

[17] Terrance Devries and Graham W. Taylor, "Improved regularization of convolutional neural networks with cutout," *CoRR*, vol. abs/1708.04552, 2017.

[18] Hongyi Zhang, Moustapha Cissé, Yann N. Dauphin, and David Lopez-Paz, "mixup: Beyond empirical risk minimization," in *ICLR (Poster)*. 2018, OpenReview.net.

[19] Vikas Verma, Alex Lamb, Christopher Beckham, Amir Najafi, Ioannis Mitliagkas, David Lopez-Paz, and Yoshua Bengio, "Manifold mixup: Better representations by interpolating hidden states," in *ICML*. 2019, vol. 97 of *Proceedings of Machine Learning Research*, pp. 6438–6447, PMLR.

[20] Guneet Singh Dhillon, Pratik Chaudhari, Avinash Ravichandran, and Stefano Soatto, "A baseline for few-shot image classification," in *ICLR*. 2020, OpenReview.net.

[21] Sachin Ravi and Hugo Larochelle, "Optimization as a model for few-shot learning," in *ICLR*. 2017, OpenReview.net.

[22] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona, "Caltech-UCSD Birds 200," Tech. Rep. CNS-TR-2010-001, California Institute of Technology, 2010.

[23] Boris N. Oreshkin, Pau Rodríguez López, and Alexandre Lacoste, "TADAM: task dependent adaptive metric for improved few-shot learning," in *NeurIPS*, 2018, pp. 719–729.

[24] Qianru Sun, Yaoyao Liu, Tat-Seng Chua, and Bernt Schiele, "Meta-transfer learning for few-shot learning," in *CVPR*. 2019, pp. 403–412, Computer Vision Foundation / IEEE.

[25] Kwonjoon Lee, Subhransu Maji, Avinash Ravichandran, and Stefano Soatto, "Meta-learning with differentiable convex optimization," in *CVPR*. 2019, pp. 10657–10665, Computer Vision Foundation / IEEE.

[26] Xu Luo, Yuxuan Chen, Liangjian Wen, Lili Pan, and Zenglin Xu, "Boosting few-shot classification with view-learnable contrastive learning," in *2021 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2021, pp. 1–6.

[27] Han-Jia Ye, Hexiang Hu, De-Chuan Zhan, and Fei Sha, "Few-shot learning via embedding adaptation with set-to-set functions," in *CVPR*. 2020, pp. 8805–8814, Computer Vision Foundation / IEEE.

[28] Bin Liu, Yue Cao, Yutong Lin, Qi Li, Zheng Zhang, Mingsheng Long, and Han Hu, "Negative margin matters: Understanding margin in few-shot classification," in *ECCV (4)*. 2020, vol. 12349 of *Lecture Notes in Computer Science*, pp. 438–455, Springer.