# DOMAIN GENERALIZED FEW-SHOT IMAGE CLASSIFICATION VIA META REGULARIZATION NETWORK

*Min Zhang*[§†] *Siteng Huang*[†] *Donglin Wang*[⋆†]

[§] College of Information Science & Electronic Engineering, Zhejiang University, Hangzhou, China
[†] School of Engineering, Westlake University, Hangzhou, China

## ABSTRACT

In few-shot image classification scenarios, meta-learning methods aim to learn transferable feature representations extracted from seen domains (base classes) in the meta-training phase and quickly adapt to unseen domains (novel classes) in the meta-testing phase. However, when seen and unseen domains have a large discrepancy, existing approaches do not perform well due to the incapability of generalizing to unseen domains. In this paper, we investigate the challenging *domain generalized few-shot image classification* problem. We design an *Meta Regularization Network (MRN)* to learn a domain-invariant discriminative feature space, where a learning to learn update strategy is used to simulate domain shifts caused by seen and unseen domains. The simulation trains the model to learn to reorganize the feature knowledge acquired from seen domains to represent unseen domains. Extensive experiments and analysis show that our proposed MRN can significantly improve the generalization ability of various meta-learning methods to achieve state-of-the-art performance in domain generalized few-shot learning.

***Index Terms—*** Meta Learning, Domain Generalization, Few-shot Learning, Meta Regularization Network

## 1. INTRODUCTION

Deep learning has achieved great success with sufficient data [1], but in real-world applications, the demand for a large amount of data cannot be met commonly due to labor and time consumption. Few-shot image classification [2] instead aims to predict unlabeled query images with only a few labeled support images. Among various approaches for addressing the few-shot image classification problem, meta-learning methods have demonstrated their effectiveness and drawn considerable attention recently [3, 4, 5, 6, 7, 8]. In general, these approaches aim to learn a model that can extract transferable feature representations from seen domains in the meta-training phase and quickly adapt to unseen domains in the meta-testing phase with the same data distribution.

As illustrated in Figure 1, meta-learning models generally consist of a feature encoder $F_\theta$ and a classifier $C_\phi$ [9]. In the meta-training phase, the $F_\theta$ extracts image features and then the $C_\phi$ predicts the labels via an optimization procedure. If testing episodes are sampled from the same domain as training episodes (*e.g.*, the training and testing episodes are the same dataset in Figure 1 (a): 1.CUB $\rightarrow$ (b): 3.CUB), the updated model has a good performance. However, if testing episodes are given from a different domain like Figure 1 (a): 2.CUB $\rightarrow$ (b): 4.Flower, the learned model cannot quickly adapt to unseen domains, resulting in poor performance.

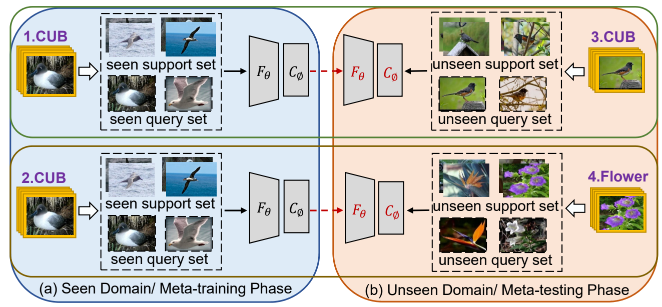



**Fig. 1. Problem formulation and motivation**

Recently, Chen et al. [10] have raised the issue that existing meta-learning approaches cannot generalize well to testing episodes from unseen domains that have different distributions as training episodes. To alleviate the domain shifts, domain adaptation (DA) [11] and domain generalization (DG) methods [12, 13, 14] have been proposed. Nevertheless, these methods assume that all domains share the same label space. However, few-shot learning needs to identify *novel* classes. Therefore, it is still an open challenge to address the DG for different label spaces between the training and testing phases.

In this paper, we tackle the *domain generalized few-shot image classification* problem. We propose to integrate a meta regularization network (MRN) to modulate the feature encoder, which assists the encoder to learn a *domain-invariant feature space* and improves the generalization ability to unseen domains. Furthermore, we present a *learning to learn* procedure to optimize the MRN. Specifically, we sample two episodes from non-overlapping seen domains to simulate domain shifts caused by seen and unseen domains during the meta-training phase. One episode is used to update the model based on the MRN, while the other tests the performance on the updated model, and the performance is used to update

⋆ Corresponding Author.

the MRN. Following this update strategy in the meta-training phase, the model can better adapt to meta-testing episodes from unseen domains. Our contributions can be summarized as follows: (1) We propose the MRN to help the encoder learn *domain-invariant discriminative features* with various distributions. The MRN is *method-agnostic* and can be applied to various meta-learning methods. (2) We use a *learning to learn* procedure to optimize the MRN. It can simulate the difference in domains, thereby improving the generalization ability of the trained model. (3) We conduct extensive experiments in domain generalized few-shot image classification to demonstrate that various meta-learning methods effectively improve the generalization ability by plugging the MRN.

## 2. METHODOLOGY

### 2.1. Preliminaries

**Meta Learning.** *In the meta-training phase*, the episodic training mechanism is used as an effective approach to learning the transferable feature representations from a large number of base classes $\mathcal{B}$ [4]. Specifically, each episode contains a support set $\mathcal{S}_\mathcal{B}$ with $n$ different classes and $k$ labeled images per class and a query set $\mathcal{Q}_\mathcal{B}$ with same classes and $q$ unlabeled images per class. We call this problem setting *$n$-way $k$-shot*. *In the meta-testing phase*, the novel classes $\mathcal{N}$ with a few classes and images are used to test the model performance. The support set $\mathcal{S}_\mathcal{N}$ are sampled from the $\mathcal{N}$, the model aims to classify each unlabeled query image $\mathcal{Q}_\mathcal{N}$.
**Domain Generalization.** In domain generalization (DG), there are $M$ source domains and $L$ target domains. The DG methods aim to learn domain-invariant feature representations from multiple source domains, either by moment matching or adversarial training, and generalize to unseen target domains [13, 15]. These methods assume that the marginal distribution P(X) changes while the conditional distribution P(Y|X) stays stable across domains, *i.e.*, the source and target domains have the same label space. However, the seen and unseen domains for meta-learning methods have different label spaces. To solve the domain generalized few-shot classification problem, we proposed the MRN to optimize the feature space so that the feature encoder $F_\theta$ is insensitive to different domains and then generalizes better to unseen domains. Furthermore, MRN improves the generalization performance of the model.

### 2.2. Description

In the meta-training phase, these episodes of each iteration are created as $\mathcal{T}_\mathcal{B} = \{\mathcal{S}_\mathcal{B}, \mathcal{Q}_\mathcal{B}\}$, where $\mathcal{S}_\mathcal{B} = (\mathcal{X}_\mathcal{B}^\mathcal{S}, \mathcal{Y}_\mathcal{B}^\mathcal{S})$ and $\mathcal{Q}_\mathcal{B} = (\mathcal{X}_\mathcal{B}^\mathcal{Q}, \mathcal{Y}_\mathcal{B}^\mathcal{Q})$. Generally, the model consists of a feature encoder $F_\theta$ and a classifier $C_\phi$. We optimize $\theta$ and $\phi$ by minimizing a cross-entropy loss and the loss is shown as:

$$L(\theta, \phi) = L_{ce}(\mathcal{Y}_\mathcal{B}^\mathcal{Q}, C_\phi(\mathcal{Y}_\mathcal{B}^\mathcal{S}, F_\theta(\mathcal{X}_\mathcal{B}^\mathcal{S}), F_\theta(\mathcal{X}_\mathcal{B}^\mathcal{Q}))), \quad (1)$$

where the $F_\theta$ first extracts the features from both the support and query images, the $C_\phi$ then predicts the categories of the

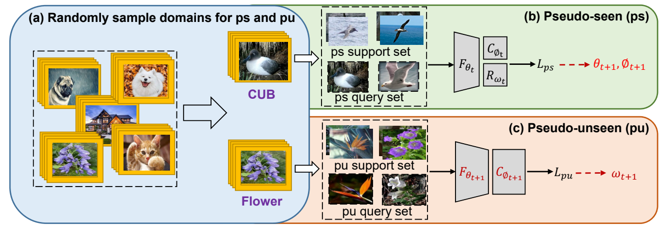



**Fig. 2. Framework of our method**:

query images according to the label of support images, the encoded support images and the encoded query image.

In the meta-testing phase, if testing episodes $\mathcal{T}_\mathcal{N}$ are sampled from the same domain as training episodes, the model has good performance. However, if test episodes are given from different domains, the model may overfit and fail to generalize to unseen domains. Because optimization through the cross-entropy loss solely, the $F_\theta$ cannot learn *domain-invariant* features. To alleviate the issue, we propose the following MRN, which uses a *learning to learn* update strategy.

### 2.3. Meta Regularization Network (MRN)

We proposed the MRN to adjust the $F_\theta$ to learn a *domain-invariant* feature space. Intuitively, the $F_\theta$ equipped with the MRN can produce more diverse feature distributions, which improves the generalization ability of the $C_\phi$. *How to integrate the MRN with the $F_\theta$?* Inspired by multi-tasks learning, we introduce an additional regularization loss $L_{mrn}$ generated by the MRN, combined with a few-shot learning loss $L_{ce}$ to optimize the model. So, the new loss becomes:

$$L^{new}(\theta, \phi, \omega) = L_{ce} + \beta L_{mrn}, \quad (2)$$

where $\beta$ is the trade-off hyperparameter. The $F_\theta$ is optimized using the above loss to produce more generalized feature representations. Specifically, we input these image features extracted by $F_\theta$ to the MRN, and then this network output a non-zero value, which is called $L_{mrn}$ and is shown as:

$$L_{mrn}(\theta, \omega) = R_\omega(F_\theta(\mathcal{X})), \quad (3)$$

where $R_\omega$ denotes the MRN network parameterized by $\omega$. The MRN can be applied to arbitrary meta-learning methods.

### 2.4. Learning to Learn with the MRN

We use a *learning to learn* update strategy to optimize the proposed MRN and illustrate the process in Figure 2. To the end, the MRN output an effective regular loss to help the $F_\theta$ to learn the domain-invariant discriminative feature space.

In each meta-training iteration, we randomly select a ***pseudo-seen*** domain $D_{ps}$ and a ***pseudo-unseen*** domain $D_{pu}$ from $M$ seen domains. Then, the episodes $\mathcal{T}_{ps} = \{(\mathcal{X}_{ps}^\mathcal{S}, \mathcal{Y}_{ps}^\mathcal{S}), (\mathcal{X}_{ps}^\mathcal{Q}, \mathcal{Y}_{ps}^\mathcal{Q})\}$ and $\mathcal{T}_{pu} = \{(\mathcal{X}_{pu}^\mathcal{S}, \mathcal{Y}_{pu}^\mathcal{S}), (\mathcal{X}_{pu}^\mathcal{Q}, \mathcal{Y}_{pu}^\mathcal{Q})\}$ are randomly sampled from $D_{ps}$ and $D_{pu}$ domains, respectively [16]. $\mathcal{T}_{ps}$ and $\mathcal{T}_{pu}$ aim to simulate the domain shifts in the meta-training and meta-testing phases (see Figure 2 (a)).

In iteration $t$ of the meta-training phase, we first integrate the MRN $R_{\omega_t}$ into the model. Then, the model with the $F_{\theta_t}$ and the $C_{\phi_t}$ is optimized using the loss $L_{mrn}$ and cross-entropy loss $L_{ce}$ to avoid the model overfitting in Eq. (4) and (5).

$$\begin{aligned} L^{ps}(\theta_t, \phi_t, \omega_t) =& L_{ce}(\mathcal{Y}^{\mathcal{Q}}_{ps}, C_{\phi_t}(\mathcal{Y}^{\mathcal{S}}_{ps}, F_{\theta_t}(\mathcal{X}^{\mathcal{S}}_{ps}), F_{\theta_t}(\mathcal{X}^{\mathcal{Q}}_{ps}))) \\ &+ \beta L_{mrn}(R_{\omega_t}([F_{\theta_t}(\mathcal{X}^{\mathcal{S}}_{ps}); F_{\theta_t}(\mathcal{X}^{\mathcal{Q}}_{ps})])), \end{aligned} \quad (4)$$

$$(\theta_{t+1}, \phi_{t+1}) = (\theta_t, \phi_t) - \alpha \nabla_{\theta_t, \phi_t} L^{ps}(\theta_t, \phi_t), \quad (5)$$

where $\alpha$ is the learning rate, $[;]$ denotes the concatenation (see Figure 2 (b)). We then measure the generalization ability of the updated model using $L_{mrn}$ by 1) removing the MRN from the framework and 2) computing the loss of the updated model on the *pseudo-unseen* episodes $\mathcal{T}_{pu}$ (see Figure 2 (c)):

$$L^{pu}(\theta_{t+1}, \phi_{t+1}) = L_{ce}(\mathcal{Y}^{\mathcal{Q}}_{pu}, C_{\phi_{t+1}}(\mathcal{Y}^{\mathcal{S}}_{pu}, F_{\theta_{t+1}}(\mathcal{X}^{\mathcal{S}}_{pu}), F_{\theta_{t+1}}(\mathcal{X}^{\mathcal{Q}}_{pu}))). \quad (6)$$

Finally, as the loss $L_{pu}$ reflects the effectiveness of the MRN to generate $L_{mrn}$, we optimize the parameters $\omega$ by

$$\omega_{t+1} = \omega_t - \alpha \nabla_{\omega_t} L^{pu}. \quad (7)$$

**Explanation.** *Why can the learning to learn update strategy learn the domain-invariant feature space*? Revisiting the adversarial learning DG methods [13], we found that the domain discriminator is optimized through a minimax game to learn a domain-invariance feature space. Different from such **explicit** adversarial learning above, the *learning to learn* is an **implicit** adversarial update strategy. Specifically, two non-overlapping episodes from different domains are randomly sampled in each meta-training iteration. One is used to update the model using the summation of $L_{ce}$ and $L_{mrn}$ in Eq.2. The other tests the performance of the updated model and then feeds it back to the MRN. Furthermore, the MRN is updated based on this feedback to improve the performance of the model in the next iteration. This double-adjustment strategy makes the model behave well: after each update from one domain, the performance gets improved for another domain.

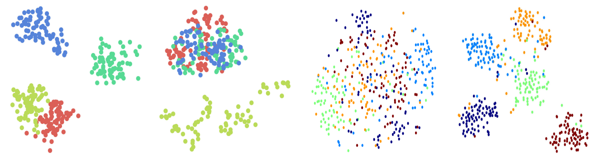
## 3. EXPERIMENTAL RESULTS

In this section, we conduct experiments to answer three main questions: **RQ1**: Does the MRN improves the performance of meta-learning methods? **RQ2**: How is its performance compared with state-of-the-art approaches? **RQ3**: How do different regularization, *i.e.*, L1 or L2, affect the performance?

### 3.1. Experimental details

**Datasets.** Four benchmark datasets are used to evaluate the model performance (*mini*ImageNet [4], CUB-200-2011 [17], *tiered*ImageNet [18] and CIFAR-FS [19]). The leave one-domain-out setting is adopted to select an unseen domain and we resize all images in the four datasets to 84 × 84 [20].

**Experimental setting.** We conduct experiments on 5-way 1-/5-shot settings, and each episode has 15 query images per



(a) MAML (b) MRN (c) MAML (d) MRN

**Fig. 3**. The t-SNE visualization of the embedding distributions learned by MAML without (a)/(c) or with (b)/(d) the MRN. The model is tested on *tiered*ImageNet datatset.

class in meta-training/-testing phases. We report the average accuracy (%) and the corresponding 95% confidence interval over the 10,000 episodes sampled from unseen domains.

**Implementation details.** For a fair comparison, we use the 4-layer ConvNet with 64 filters (Conv4-64F) as the backbone for all methods and do not adopt any data augmentation. In this paper, we get $L_{mrn}$ generated by MRN using an MLP with two linear layers and different options are shown in Section 3.4. Pytorch [21] is used to implement all experiments. All methods are trained via SGD with Adam [22], the initial learning rate is set to $1e^{-3}$ for $F_\theta / C_\phi$, and $1e^{-4}$ for $R_\omega$. For each method, we use the 60,000 (1-shot) or 40,000 (5-shot) episodes to train, and the best model on the validation set is used to evaluate final reporting performance on the test set.

### 3.2. RQ1. Meta-Learning Methods Equipped with MRN

To answer RQ1 and verify the effectiveness of the MRN, we embed it into two meta-learning methods: MAML [3] and Prototypical Network [6]. Table 1 shows the classification accuracies on different unseen domains, where the leave-one-domain-out setting is used to select one unseen domain for evaluation. It is observed that the performance of two simple meta-learning methods degrades in domain generalized few-shot image classification. However, for the two meta-learning methods, incorporating MRN leads to a significant improvement. An observation is that MRN boosts the performance of MAML nearly by 10% on 1-shot and 16% on 5-shot of unseen domain *mini*ImageNet, which is especially prominent when compared with other methods. We attribute this success to using the *learning to learn* update strategy, which divides non-overlapping domains from seen domains in the meta-training phase to simulate the data distributions of unseen domains in the meta-testing phase. Then, a domain-invariant discriminative feature space is learned to align the feature distributions of different domains and improve the model performance.

**Visualizing MRN.** To demonstrate that the MRN can assist feature encoder to learn a *domain-invariant discriminative feature space* and clustering spaces. We also apply t-SNE [23] to visualize the embedding distribution obtained with and without equipping with the MRN. In Figure 3, the (a) and (b) represent the image features extracted from different domains; the (c) and (d) mean the image features extracted from different classes. As shown in (a) and (b), the model equipped with the MRN has more compact features distance between different domains, indicating that the learned

**Table 1**. Classification average testing accuracy (%)

| 5-way | MRN | *mini*ImageNet | | *tiered*ImageNet | | CUB-200-2011 | | CIFAR-FS | |
|---|---|---|---|---|---|---|---|---|---|
| | | 1-shot | 5-shot | 1-shot | 5-shot | 1-shot | 5-shot | 1-shot | 5-shot |
| MAML [3] | | 35.18 | 45.82 | 30.83 | 45.91 | 30.96 | 42.28 | 35.62 | 37.28 |
| MAML [3] | ✓ | 45.30 (+10.12) | 62.33 (+16.51) | 42.59 (+11.76) | 59.83 (+13.92) | 35.42 (+4.46) | 47.63 (+5.35) | 39.27 (+3.65) | 48.61 (+11.33) |
| Prototypical Network [6] | | 47.72 | 64.83 | 43.50 | 58.04 | 39.47 | 55.36 | 38.87 | 54.41 |
| Prototypical Network [6] | ✓ | 53.77 (+6.05) | 68.05 (+3.22) | 46.67 (+3.17) | 64.23 (+6.19) | 42.32 (+2.85) | 60.83 (+5.47) | 41.33 (+2.46) | 58.91 (+4.50) |

**Table 2**. Average accuracy (%) comparison to state-of-the-arts.

| 5-way | Backbone | *mini*ImageNet | | *tiered*ImageNet | | CUB-200-2011 | | CIFAR-FS | |
|---|---|---|---|---|---|---|---|---|---|
| | | 1-shot | 5-shot | 1-shot | 5-shot | 1-shot | 5-shot | 1-shot | 5-shot |
| Relation Network LFT [16] | ResNet10 | 55.23±0.14 | 72.56±0.81 | 48.75±0.89 | 63.24±0.91 | 45.67±0.78 | 64.75±0.47 | 44.79±0.31 | 60.12±0.45 |
| Matching Network LFT [16] | ResNet10 | 56.01±0.31 | 73.45±0.65 | 49.31±0.21 | 65.41±0.45 | 45.12±0.65 | 65.14±0.74 | 45.98±0.31 | 59.12±0.34 |
| ProtoNet MRN (ours) | ResNet10 | **56.99±0.61** | **75.16±0.48** | **50.31±0.12** | **68.23±0.47** | **48.32±0.45** | **66.85±0.31** | **47.23±0.18** | **64.52±0.25** |

features are domain-invariant representations. The domain-invariant features decrease the domain shifts and improve the model generalization abilities. In addition, we visualize five classes distributions in Figure 3 (c) and (d) with good results.

### 3.3. RQ2. Comparison with State-of-the-Arts (SOTAs)

To answer RQ2 and prove that the meta-learning methods can surpass the SOTAs after equipping with the MRN, we report the experimental results on the four benchmark datasets. Note that few published approaches have been proposed to address the domain generalized few-shot image classification problems, and thus we choose the famous LFT baseline for comparison [16]. We display the results of Prototypical Network (ProtoNet) equipped with the MRN as our method. For a fair comparison, we use the ResNet10 backbone following [16].

Specifically, it is observed that our method significantly outperforms state-of-the-art LFT combined with Matching Network and Relation Network [16] in all dataset settings under ResNet10 backbone (see Table 2). Our method achieves 3% performance improvement for 5-way 1-shot and about 2% higher performance for 5-way 5-shot in the best performance of Matching Network or Relation Network equipped with LFT with CUB-200-2011 dataset. From Table 2, we can find that (1) our method significantly improves the performance and achieves the best results. (2) A new loss generated by the MRN is used to learn the domain-invariant discriminative feature space. (3) A *learning to learn* implicit adversarial update strategy can effectively align the feature distributions that they come from different training or testing domains.

### 3.4. RQ3. Ablation Study

**Influence of using different forms of L_mrn.** The loss $L_{mrn}$ generated by the MRN helps the feature encoder $F_\theta$ learn the *domain-invariant feature space* through a *learning to learn* update strategy. In Table 3, we consider four forms of $L_{mrn}$, where the first two are common $L1$ and $L2$ regularizations, and the other two represent different calculation modes for the MRN. The "MLP" means that the MRN ($R_\omega$) in Eq. (3) has two linear layers, *i.e.*, fully connected layers, $L_{mrn}^{MLP} = R_\omega(F_\theta(\mathcal{X}))$. The "Flatten" indicates using the linear flattened feature in Eq.3, *i.e.*, $L_{mrn}^{Flatten} = R_\omega((F_\theta(\mathcal{X}))^T F_\theta(\mathcal{X}))$, where $T$ is a transpose operation.

**Table 3**. Average testing accuracy (%).

| 5-way-1-shot | *mini*ImageNet | *tiered*ImageNet | CUB-200-2011 | CIFAR-FS |
|---|---|---|---|---|
| $L_1$ | 48.18±0.37 | 44.83±0.57 | 39.96±0.45 | 38.62±0.55 |
| $L_2$ | 48.63±0.35 | 45.95±0.38 | 40.06±0.42 | 38.79±0.58 |
| Flatten | 52.98±0.38 | 46.23±0.37 | **43.21±0.46** | **42.63±0.54** |
| MLP | **53.77±0.61** | **46.67±0.68** | 42.32±0.62 | 41.33±0.59 |
| **5-way-5-shot** | *mini*ImageNet | *tiered*ImageNet | CUB-200-2011 | CIFAR-FS |
| $L_1$ | 64.82±0.47 | 59.91±0.47 | 55.28±0.58 | 56.28±0.58 |
| $L_2$ | 65.23±0.37 | 58.98±0.37 | 56.32±0.42 | 55.78±0.48 |
| Flatten | 67.95±0.35 | 63.06±0.35 | 60.01±0.41 | 57.98±0.69 |
| MLP | **68.05±0.56** | **64.23±0.75** | **60.83±0.44** | **58.91±0.55** |

We display the results of four regularized forms based on Prototypical Network. Specifically, these two modes of MRN perform well, with the embedding of "MLP" performing slightly better than the covariance matrix embedding "Flatten". Meanwhile, compare with the best performance of $L_1$ and $L_2$, our method with "MLP" achieves 5% and 3% performance improvement for 1-shot and 5-shot on *mini*ImageNet, respectively. This indicates that the MRN using a *learn to learn* update strategy can help the $F_\theta$ learn better feature distribution from different domains than fixed loss $L_1$ or $L_2$.

## 4. CONCLUSION

We propose a meta-regularized network (MRN) to effectively enhance the generalization ability and significantly solve the domain generalized few-shot image classification problem. The novel MRN is presented to assist the feature encoder to learn a *domain-invariant feature space* through simulating various feature distributions extracted from the different domains. A *learning to learn* update strategy is used to optimize the parameters of the MRN. From extensive experiments, we demonstrate that our method can handle the domain generalized few-shot image classification problem, and shows an evident improvement over baselines to achieve new SOTA.

## 5. REFERENCES


[1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, “Deep residual learning for image recognition,” in *Conference on Computer Vision and Pattern Recognition CVPR*, 2016.

[2] Lake Brenden M, Salakhutdinov Ruslan, and Tenenbaum Joshua B, “Human-level concept learning through probabilistic program induction,” in *Science*, 2015.

[3] Finn Chelsea, Pieter Abbeel, and Sergey Levine, “Model-agnostic meta-learning for fast adaptation of deep networks,” in *International Conference on Machine Learning ICML*, 2017.

[4] Oriol Vinyals, Charles Blundell, Tim Lillicrap, Koray Kavukcuoglu, and Daan Wierstra, “Matching networks for one shot learning,” in *Conference on Neural Information Processing Systems NeurIPS*, 2016.

[5] 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 *Conference on Computer Vision and Pattern Recognition CVPR*, 2018.

[6] Jake Snell, Kevin Swersky, and Richard S. Zemel, “Prototypical networks for few-shot learning,” in *Conference on Neural Information Processing Systems NeurIPS*, 2017.

[7] Min Zhang, Donglin Wang, and Sibo Gai, “Knowledge distillation for model-agnostic meta-learning,” in *the 24th European Conference on Artificial Intelligence ECAI*, 2020.

[8] Siteng Huang, Min Zhang, Yachen Kang, and Donglin Wang, “Attributes-guided and pure-visual attention alignment for few-shot recognition,” in *Association for the Advancement of Artificial Intelligence AAAI*, 2021.

[9] Nichol Alex and Schulman John, “Reptile: a scalable metalearning algorithm,” in *arXiv*, 2018.

[10] Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Frank Wang, and Jia-Bin Huang, “A closer look at few-shot classification,” in *International Conference on Learning Representations ICLR*, 2019.

[11] Yuhua Chen, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool, “Domain adaptive faster R-CNN for object detection in the wild,” in *Conference on Computer Vision and Pattern Recognition CVPR*, 2018.

[12] Yiying Li, Yongxin Yang, Wei Zhou, and Timothy M. Hospedales, “Feature-critic networks for heterogeneous domain generalization,” in *ICML*, 2019.

[13] Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C. Kot, “Domain generalization with adversarial feature learning,” in *Conference on Computer Vision and Pattern Recognition CVPR*, 2018.

[14] Yogesh Balaji, Swami Sankaranarayanan, and Rama Chellappa, “Metareg: Towards domain generalization using meta-regularization,” in *Conference on Neural Information Processing Systems NeurIPS*, 2018.

[15] Foivos Ntelemis, Yaochu Jin, and Spencer A. Thomas, “Image clustering using an augmented generative adversarial network and information maximization,” in *IEEE Trans. Neural Networks Learn. Syst.*, 2020.

[16] Hung-Yu Tseng, Hsin-Ying Lee, Jia-Bin Huang, and Ming-Hsuan Yang, “Cross-domain few-shot classification via learned feature-wise transformation,” in *International Conference on Learning Representations ICLR*, 2020.

[17] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, “The Caltech-UCSD Birds-200-2011 Dataset,” in *California Institute of Technology*, 2011.

[18] Brenden M. Lake, Ruslan Salakhutdinov, and Joshua B. Tenenbaum, “The omniglot challenge: A 3-year progress report,” in *CoRR*, 2019.

[19] Luca Bertinetto, João F. Henriques, Philip H. S. Torr, and Andrea Vedaldi, “Meta-learning with differentiable closed-form solvers,” in *International Conference on Learning Representations ICLR*, 2019.

[20] Victor Garcia Satorras and Joan Bruna Estrach, “Few-shot learning with graph neural networks,” in *International Conference on Learning Representations ICLR*, 2018.

[21] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala, “Pytorch: An imperative style, high-performance deep learning library,” in *Conference on Neural Information Processing Systems NeurIPS*, 2019.

[22] Diederik P. Kingma and Jimmy Ba, “Adam: A method for stochastic optimization,” in *International Conference on Learning Representations ICLR*, 2015.

[23] Laurens, Van Der Maaten, and Hinton Geoffrey, “Visualizing data using t-sne,” in *Journal of Machine Learning Research*, 2008.