FS-BAN: Born-Again Networks for Domain Generalization Few-Shot Classification

Yunqing Zhao Ngai-Man Cheung

***Abstract*—Conventional Few-shot classification (FSC) aims to recognize samples from novel classes given limited labeled data. Recently, domain generalization FSC (DG-FSC) has been proposed with the goal to recognize novel class samples from unseen domains. DG-FSC poses considerable challenges to many models due to the domain shift between base classes (used in training) and novel classes (encountered in evaluation). In this work, we make two novel contributions to tackle DG-FSC. Our first contribution is to propose Born-Again Network (BAN) episodic training and comprehensively investigate its effectiveness for DG-FSC. As a specific form of knowledge distillation, BAN has been shown to achieve improved generalization in conventional supervised classification with a closed-set setup. This improved generalization motivates us to study BAN for DG-FSC, and we show that BAN is promising to address the domain shift encountered in DG-FSC. Building on the encouraging findings, our second (major) contribution is to propose Few-Shot BAN (FS-BAN), a novel BAN approach for DG-FSC. Our proposed FS-BAN includes novel multi-task learning objectives: Mutual Regularization, Mismatched Teacher, and Meta-Control Temperature, each of these is specifically designed to overcome central and unique challenges in DG-FSC, namely overfitting and domain discrepancy. We analyze different design choices of these techniques. We conduct comprehensive quantitative and qualitative analysis and evaluation over six datasets and three baseline models. The results suggest that our proposed FS-BAN consistently improves the generalization performance of baseline models and achieves state-of-the-art accuracy for DG-FSC. Project Page: yunqing-me.github.io/Born-Again-FS/.**

***Index Terms*—Few-shot classification, domain generalization, born-again network, episodic training, meta-learning.**

I. Introduction

WHILE modern deep learning models achieve superior performance in many visual recognition tasks, *e.g.*, image classification [8] and object detection [46], they require a large number of labeled data during training [50]. In contrast, in few-shot classification (FSC) [9], [49], [53], [10], the models are required to classify samples from *novel* categories given only *a few* labeled data from each category.

A. Domain Generalization FSC

Recently, meta-learning based FSC [53], [60], [10], [29] has achieved outstanding performance in the *single* domain setup, where the base classes for training and the novel classes for evaluation are from the same domain. However, in real-world applications, the deployed models are often required to classify objects from domains that are unseen during training, given limited labeled data (*e.g.*, recognize rare bird species in a fine-grained setup [57]). In particular, our work addresses this challenging *domain generalization* (DG) FSC: to recognize samples from novel classes of unseen domains given only a few labeled data of each class. We follow recent DG-FSC works [57], [51] and assume to have several seen domains during training; however, we do not have access to samples from the unseen domains which will be encountered during evaluation. DG-FSC has attracted a fair amount of attention recently [6], [57], [51], [39]. Due to the significant discrepancy between the seen domains used in training and the unseen domains encountered in evaluation, existing FSC models designed only for the single domain setup often perform poorly [6]. Therefore, DG-FSC still has much room for improvement in generalization under the domain shift setup.

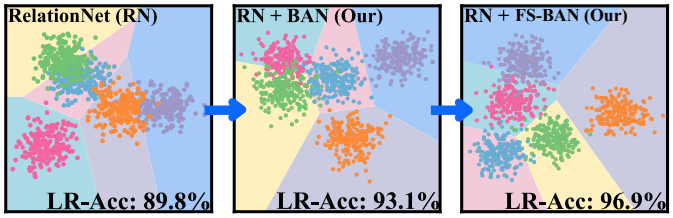

Fig. 1: In this visualization, we select 5 novel classes with 200 query samples per class from an unseen domain (Places [73]). Each point indicates the feature representation of RelationNet (RN) [53] with backbone network (ResNet-10 [20]), projected by LDA [17], [40]. We use the linear regression prediction accuracy ("LR-Acc") to demonstrate the improved decision boundaries: baseline RN (left), BAN episodic training applied to RN (mid), and our further proposed FS-BAN applied to RN (right). See numerical results and comparisons in Sec. VI.

B. Born-Again Networks (BANs)

In their pioneer work, Breiman and Shang [3] proposed *born-again trees*. Given a complex predictor, *e.g.*, a model with multiple trees ensemble, they train a single tree which outputs (decisions) match that of the complex predictor. This single born-again tree is simple and more interpretable compared to the complex predictor while it still maintains a decent decision performance [59]. More recently, [12] investigated the knowledge transfer [22] from one model (the teacher) to another model (the student). Focusing on the conventional image classification tasks, they first train the teacher network to convergence using the standard cross-entropy loss; then, they train the student network with the dual goals of prediction of correct label and matching of teacher's probability prediction.

Yunqing Zhao and Ngai-Man Cheung are with the Information Systems Technology and Design Pillar, Singapore University of Technology and Design, Singapore 487372 (email: yunqing_zhao@mymail.sutd.edu.sg, ngaiman_cheung@sutd.edu.sg). Correspondence to: Ngai-Man Cheung.

Surprisingly, despite that the teacher and the student models have *identical network structure*, and the *same training data* is used for teacher training and knowledge transfer process, they reported that with this BAN approach, the student outperforms the teacher network accuracy consistently in various conventional image classification setups, *e.g.*, DenseNets [24] on CIFAR-10 and CIFAR-100 [28]. The student models were found to have *better generalization*. This is attributed to the distillation of *dark knowledge*, *i.e.*, teacher's prediction on the wrong outputs, and *importance weighting*, *i.e.*, teacher's confidence on the correct outputs. Recently, Zhu and Li [2] presented a rigorous analysis on this improved generalization. From the perspective of *multi-view* data structure, they argue that the BAN approach can be viewed as a combination of implicit ensemble and knowledge distillation [22], enabling the student model to learn multi-view features and eventually achieve better generalization compared to the teacher models which have identical structures. Besides the conventional image classification, BAN has been applied in other areas, *e.g.*, multi-task natural language processing [7].

C. Motivation and Our Contributions

This work is motivated by the empirical results and theoretical analysis presented by [12] and [2]. Both works suggested BANs can achieve improved generalization without modification to the network structure, which could be extremely useful for existing FSC models, especially, under domain shift. In particular, **our first contribution** is to propose BAN episodic training for DG-FSC. Note that previous work has focused on applying BAN in conventional supervised training [12], [2], [55], and our work on applying BAN in *episodic training* is novel. In Sec.IV, we discuss the subtleties in BAN episodic training, perform a rigorous study to show that BAN can lead to models with improved generalization on novel tasks sampled from an unseen domain. Furthermore, we also validate that BAN enables learning of more compact features with a lower intra-class to inter-class variance ratio which is useful for few-shot learning as discussed in [17] (sSee Linear Discriminant Analysis (LDA) [40] of features in Figure 1).

Based on the encouraging results investigated in Sec. IV, **our second contribution** is to propose Few-Shot BAN (FS-BAN) that addresses the unique issues in DG-FSC. Specifically, different from conventional image classification, DG-FSC poses unique challenges that inhibit the improvement of BAN episodic training: (i) Because of limited labeled data in FSC, the teacher model in BAN training may suffer from overfitting, and this degrades the knowledge transferring to the student model; (ii) In DG-FSC, the student model needs to handle unseen domains during the evaluation stage.

To address the above challenges in BAN for DG-FSC, we propose FS-BAN (Sec. V) that builds upon the baseline BAN method (Sec. IV). FS-BAN consists of novel multi-task learning objectives: (i) Mutual Regularization (MR): We extend BAN with additional feedback *from the student to the teacher*, encouraging the teacher to continue to improve using soft predictions from the student. The student's soft prediction provides additional regularization to alleviate overfitting in the teacher model. This technique achieves significant improvements in all experiments. (ii) Mismatched teachers (MM): To address domain shift, we propose a technique of *mismatched teacher*: a teacher model which is trained on a domain different from that of the current training task. Our proposed mismatch teacher is an imitation procedure so that an FSC model has exposure to domain shift during the training stage. We show in experiments that this imitation in training leads to a better generalization of unseen domains and achieves better domain robustness. (iii) Meta-control temperature (MCT): Temperature is an important parameter to control the distillation of knowledge in BAN training [22]. It is usually regarded as a hyperparameter and manually pre-set to a fixed value for the entire training (regardless of different domains and tasks). In contrast, we propose to *meta-learn the temperature* during training to improve adaptation to diverse domains.

The proposed FS-BAN can be readily applied to existing FSC models without modification of the structure. Experiment results show that FS-BAN achieves new state-of-the-art results for DG-FSC on six benchmark datasets, with three popular FSC baseline models. We further show in comprehensive ablation studies that the different learning objectives in FS-BAN indeed address these challenges proposed above.

Our contributions in this paper are summarized as:

1) As a pioneer work, we propose BAN episodic training as our first contribution (Sec.IV). We carefully study its effectiveness for DG-FSC and compare it to related work. We empirically validate its improved generalization.
2) As our second contribution, we propose FS-BAN for DG-FSC (Sec.V). FS-BAN consists of multi-task learning objectives that can better address the unique challenges posed by DG-FSC: few labeled support data in an episode and domain shift in the testing phase. FS-BAN overcomes the challenges, and such efforts have not been done before.
3) We conduct extensive experiments and show that FS-BAN consistently improves three baseline FSC models on six public datasets. Our approach outperforms the state-of-the-art in both the conventional FSC and DG-FSC setups. We also perform detailed ablation studies to demonstrate the effectiveness of FS-BAN.

II. Related Works

In this section, we perform a literature review from different perspectives, as our work involves FSC, domain generalization, and effective knowledge transfer. We highlight the different and challenging problem setups compared to closely related traditional FSC and domain generalization tasks.

A. Metric Learning for Few-Shot Classification

FSC [10], [49] models aim to recognize novel classes given few labeled data. Among them, metric learning based methods [49], [60], [53], [41] learn to compare the relation between the unlabeled query data and the labeled support data. The prediction result of each query image is a confidence (probability) distribution assigned to each category that belongs to a training task. Metric learning based ideas have attracted a

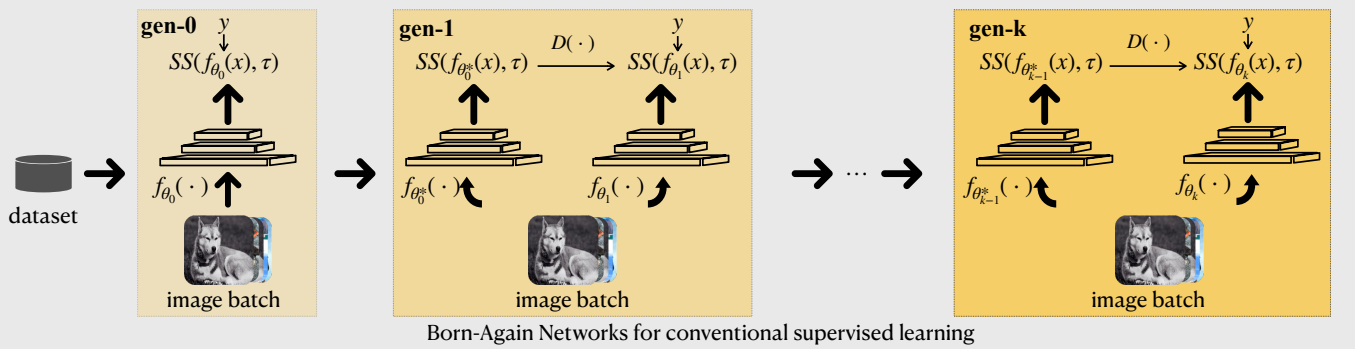

Fig. 2: In conventional supervised learning, BAN samples a batch of images $\{(x, y) \in (\mathcal{X}, \mathcal{Y})\}$ of all categories in the dataset and distills the knowledge from the teacher model to the student in each generation. In related work, Tian *et al.*[55] conducted the born-again process in generations to obtain a powerful backbone network and transferred it to the downstream FSC task.

fair amount of attention on FSC tasks. Meanwhile, there is no need to further fine-tune the model parameters or select the hyperparameters in test time [55], [6].

In this paper, we set our experiments to focus on three popular metric-based FSC models as baseline methods, similar to a recent work [57]: MatchingNet [60], RelationNet [53] and Graph Neural Network (GNN) [13] due to their simplicity and easy implementation. However, these models often fail to make predictions on novel tasks from unseen domains, due to the domain shift [6], [67], [56] and overfitting on the base classes data from the source domains seen in training. Therefore, our proposed FS-BAN builds up on their models and aims to obtain further improvement and generalization.

B. Domain Generalization FSC

Traditional domain adaptation (DA) problem often enables the model to learn with sufficient unlabeled data from the target domain [36], [34], [66], [1], [19] in the training stage. Therefore, the domain discrepancy between the source and the target domains could be explicitly reduced. Different from DA, domain generalization (DG) [30], [42] aims to learn good feature representations that generalize well on unseen domains in test time [70], [69], [71]. For the traditional supervised classification tasks, [31], [14] propose to add regularization objectives in the training stage to improve the generalization performance. However, the label space for training and testing is shared therefore there is still prior knowledge of the target domain.

In DG-FSC, models are needed to recognize samples of novel categories from unseen domains, given only few (*e.g.*, 5-shot) labeled support data. Very recently, [57] applies the learned feature-wise transformation layer (LFT) [44] to modulate the channel-wise scale and shift parameters, trying to produce diverse and entangled feature representations of different domains. [51] applies the explanation-guided layer-wise relevance propagation (LRP) to enhance the discriminative features during training with multiple seen domains. [68] address a similar problem but they use the unlabelled data from target domains in the training phase.

Our proposed FS-BAN, differently, aims to improve the generalization of FSC models for episodic training in DG-FSC setup, by less overfitting to hard targets and it is more robust to arbitrary unseen domains, with disjoint label space (*e.g.*, train on Cars domain [27] but test on Birds species [21]) during evaluation. Our setup is more challenging compared to conventional supervised learning but closer to the real-world applications and model deployment environment.

C. Knowledge Distillation and Born-Again Network

Knowledge distillation (KD) [22], [4] often aims to transfer the "knowledge" of a larger and stronger machine learning model (*the teacher*) learned on a large-scale dataset, to another compact model (*the student*) with a small training dataset [33], [72]. KD has shown empirical benefits in some applications, *e.g.*, model compression [62] and transfer learning [65]. Usually, KD method can train the student network that benefits from the teacher's knowledge and obtains a good performance.

Born-Again Network (BAN) [12] is a special case of KD that transfers knowledge from well-trained teacher(s) to the student *with an identical network structure and training data*. Taking advantage of this, BAN can generate multiple generations by repeating the knowledge-transfer process (we discuss this Sec. III). Surprisingly, previous works [12], [7] discover that the student can outperform the teacher consistently in terms of prediction accuracy on conventional supervised learning tasks, which suggests improved generalization to the test data. Recently, [55] applies BAN in conventional classification task (*i.e.*, Figure 2) to obtain a backbone network. Then, they apply the standard transfer learning pipeline on the student model to handle the downstream single-domain FSC tasks. In this work, we design FS-BAN for DG-FSC that takes the advantage of BAN the improved generalization without the need for modification to the network structure and additional training data. Compared to the similar work [55], our designs are clearly different, as shown in Figures 3 and Figure 5. Comparison results with [55] show the superiority of our designs (see Table V).

III. Preliminary

In this section, we discuss the concepts of BAN and DG-FSC. Concretely, in Sec. III-A, we review the mechanism of BAN in conventional supervised image classification; in Sec. III-B, we formulate the DG-FSC problem setup and the episodic training process of existing FSC models.

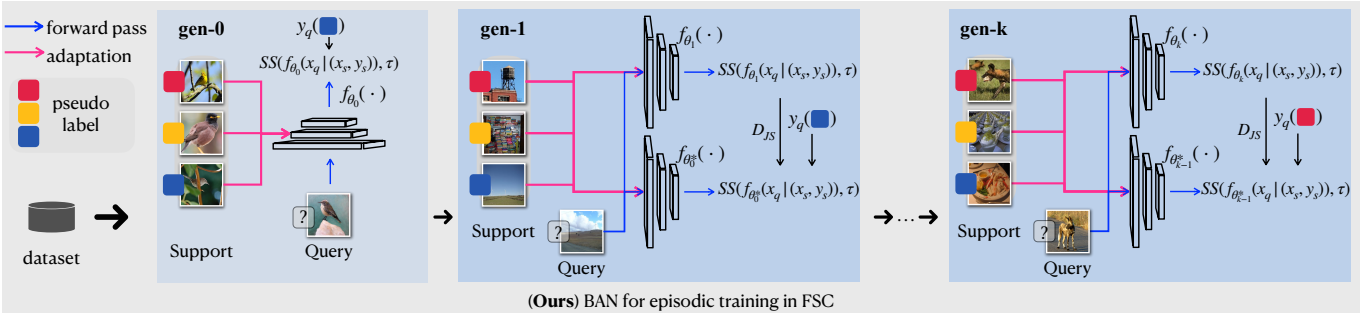

Fig. 3: Our proposed BAN episodic training. A task $\mathcal{T}$ with N_w categories is sampled (here $N_w = 3$). The support set of $\mathcal{T}$ is applied to adapt the teacher and student models. Then, the teacher conditioning on the support set predicts the query samples of $\mathcal{T}$ and transfers the knowledge to the student. Compared to [55] (see Sec. IV) that adopts the transfer learning approach, we directly apply BAN in episodic training that simulates the realistic setting in the evaluation phase for FSC.

A. BANs for Conventional Supervised Image Classification

We follow the definition of BAN in conventional classification problems [12]. Consider a dataset containing image samples $\mathcal{X}$ and true labels $\mathcal{Y}$. Generally, the prediction of the input samples $\mathcal{X}$ is parameterized by a network $f_{\theta_0}(\mathcal{X})$. $f_{\theta_0^*}(\cdot)$ is called *the teacher* network and it can be obtained by minimizing the cross-entropy loss to the ground truth labels:

$$\theta_0^* = \arg\min_{\theta_0} \mathcal{L}_{ce}(\mathcal{Y}, \hat{\mathcal{Y}}^{\theta_0}), \tag{1}$$

where $\hat{\mathcal{Y}}^{\theta_0} = SS(f_{\theta_0}(\mathcal{X}), \tau)$. $SS(\cdot, \tau)$ is the SoftMax function with a temperature τ over N training classes:

$$SS(z, \tau) = \frac{e^{z/\tau}}{\sum_{c=1}^{N} e^{z_c/\tau}}, \tag{2}$$

where we assume z is input to the SoftMax layer. Normally, Eqn. 2 is considered to soften or harden the soft predictions when $\tau > 1$ or $\tau < 1$. As Figure 2, BAN enables another model ($f_{\theta_1}(\cdot)$, *the student*) to exploit the rich information contained in the predicted probability distribution of the teacher, by minimizing the distance (D) between the output distribution of the teacher $f_{\theta_0^*}(\cdot)$ and that of the student $f_{\theta_1}(\cdot)$:

$$\mathcal{L}_{ce}(\mathcal{Y}, \hat{\mathcal{Y}}^{\theta_1}) + D(SS(f_{\theta_1}(\mathcal{X}), \tau), SS(f_{\theta_0^*}(\mathcal{X}), \tau)), \tag{3}$$

where the first term is the classification loss to the one-hot ground truth, and the second term employs the soft prediction of the fixed teacher model for knowledge transfer.

Since the student has the identical structure and training data of the teacher, this *born-again* process can be applied sequentially with multiple generations: In k-th generation (gen-k, $k > 1$), the student $f_{\theta_k}(\cdot)$ is trained to optimize a sum of cross-entropy loss and the distance between its prediction and the soft targets from the student obtained in gen-(k-1):

$$\mathcal{L}_{ce}(\mathcal{Y}, \hat{\mathcal{Y}}^{\theta_k}) + D(SS(f_{\theta_k}(\mathcal{X}), \tau), SS(f_{\theta_{k-1}^*}(\mathcal{X}), \tau)). \tag{4}$$

The student $f_{\theta_{k-1}^*}(\cdot)$ obtained in (k-1)-th generation now becomes the new teacher. In particular, $f_{\theta_0^*}(\cdot)$ indicates the first teacher that is trained with only cross-entropy loss to one-hot labels in gen-0. Interestingly, previous work [12], [55] reported *improved generalization* of student network with BAN training in this conventional supervised learning setup, which motivates us to investigate BAN for DG-FSC. We discuss it in Sec. IV.

B. Metric-based Models for DG-FSC

Here, we discuss the problem setup in this work. We follow the popular meta-learning algorithms [10], [49], [53], [64], [60], [13], [48] to define episodic training for DG-FSC.

Episodic training. We denote the input images as $\mathcal{X}$ and the corresponding labels as $\mathcal{Y}$. In each training iteration, instead of sampling a batch of images with their true labels directly (as in conventional supervised learning), we sample an N_w-Way (number of classes) N_s-Shot (number of labeled samples per class) task $\mathcal{T}$ of a source domain $\mathcal{D}$ from several seen domains $\{\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_n\}$ [57]. Each $\mathcal{T}$ consists of a support set $\mathcal{S} = \{(\mathcal{X}_s, \mathcal{Y}_s)\}$, and a query set $\mathcal{Q} = \{(\mathcal{X}_q, \mathcal{Y}_q)\}$. The support set $\mathcal{S}$ and the query set $\mathcal{Q}$ are formed by randomly selecting N_s and N_q samples of each of N_w categories (usually, $N_w = 5$), respectively. In this context, the batch size is one task (or an episode), and the samples in $\mathcal{S}$ and $\mathcal{Q}$ are pseudo-labeled which will change in different episodes.

Metric learning based FSC. Suppose a metric-based FSC model f is parameterized by θ. For each sampled task $\mathcal{T}$, $f_\theta(\cdot)$ firstly extracts the feature embeddings of both support $\mathcal{S}$ and query samples $\mathcal{Q}$, then it predicts the label of each query sample by comparing its relation to support sample features (*i.e.*, conditioned on the labeled support set):

$$\hat{\mathcal{Y}}_q^\theta = SS(f_\theta(\mathcal{X}_q | (\mathcal{X}_s, \mathcal{Y}_s)), \tau), \tag{5}$$

where $\hat{\mathcal{Y}}_q^\theta$ is the prediction results of query samples over N_w classes. Generally, we aim to minimize the prediction error on the query set with cross-entropy loss w.r.t. one-hot labels:

$$\mathcal{L}_{ce}(\mathcal{Y}_q, \hat{\mathcal{Y}}_q^\theta). \tag{6}$$

In the testing phase, we evaluate the accuracy of the query set of tasks sampled from *novel* classes of *unseen* domains. We follow the DG setup [57], [51], [31] that we do not approach any samples from unseen domains in the training phase. Therefore, our FSC models are expected to learn robust and discriminative knowledge that can be well transferred to other domains. Note that in this DG-FSC setup, the label spaces of source domains and target unseen domains are disjoint, different from some recent DG literature [35], [32], [31].

IV. Born-Again Episodic Training for DG-FSC

BAN episodic training. Motivated by the theoretical analysis in [2], and improved generalization observed in conven-

TABLE I: Accuracy (%) of the proposed BAN episodic training (Figure 3) for DG-FSC. Model is trained with 5-Way 1-Shot tasks of base classes of miniImageNet. RelationNet [53] is the baseline model and ResNet-10 [20] is the backbone network. **Top**: Model tested on novel classes of different unseen domains. **Bottom**: Performance of BAN episodic training in different generations on novel classes of miniImageNet (seen) and CUB (unseen). See Sec. IV. for more details.

Method	**Source**	CUB	Cars	Places	Plantae
RelationNet (gen-0)	miniImageNet	42.44	29.11	48.64	33.17
+BAN (gen-1)	miniImageNet	**43.35**	**29.71**	**51.30**	**33.81**

Dataset	gen-0	gen-1	gen-2	gen-3	gen-4
miniImageNet (base ↦ novel)	57.80	60.45	60.79	**61.47**	61.39
miniImageNet ↦ CUB	42.44	43.35	43.64	**43.92**	43.87

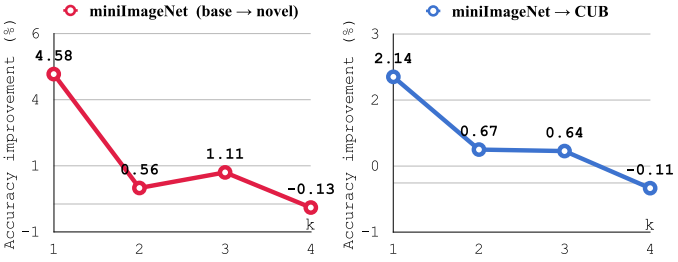

Fig. 4: Accuracy improvement (%) of gen-(k-1) ↦ gen-k of Table I. It is clear that the major gain is obtained at gen-0 ↦ gen-1. The deeper generations come with expensive training costs and lead to diminishing increment, and the negative impact is observed after the empirical optimal generation (gen-3 in Table I). We note that this observation is consistent with that of BAN in conventional supervised learning [12], [55].

tional supervised learning [12], in this section, we propose BAN episodic training for DG-FSC. We conduct a rigorous study and show the effectiveness of BAN for the existing FSC model under domain shift, which motivates us to propose FS-BAN (discussed in the next section).

As Figure 3 and description in Sec. III-B, in each training iteration of DG-FSC during the k-th generation of BAN, rather than sampling a batch of images of all classes, we instead sample a task $\mathcal{T}$ with N_w categories. We apply the support set of $\mathcal{T}$ to adapt both the teacher and student models. Then, the models conditioning on the support set are used to predict query samples of $\mathcal{T}$. After that, similar to Eqn. 4, we optimize the student network $f_{\theta_k}(\cdot)$ by leveraging the one-hot label and the soft targets predicted by the teacher network $f_{\theta^*_{k-1}}(\cdot)$ on the same query set $\mathcal{Q}$:

$$\mathcal{L}_{BAN} = \lambda_1 \mathcal{L}_{ce}(\mathcal{Y}_q, \hat{\mathcal{Y}}_q^{\theta_k}) + \lambda_2 \tau^2 D_{JS}(\hat{\mathcal{Y}}_q^{\theta_k}, \hat{\mathcal{Y}}_q^{\theta^*_{k-1}}), \quad (7)$$

where λ_1 and λ_2 are coefficients of the weighted sum, and $\hat{\mathcal{Y}}_q^{\theta_k} = SS(f_{\theta_k}(\mathcal{X}_q|(\mathcal{X}_s, \mathcal{Y}_s)), \tau)$. We use *JS* divergence [11] as the distance metric. Meanwhile, since the magnitudes of the gradients produced by the soft targets are scaled by $\frac{1}{\tau^2}$, we multiply the second term of Eqn. 7 by τ^2 to maintain the balance [22]. In the meta-testing phase, the temperature is set to τ=1 to evaluate the accuracy of novel tasks. The teacher is discarded, hence the outcome of BAN episodic training is the student model without any additional parameters.

Experiment setups. To validate the effectiveness of the proposed BAN episodic training for DG-FSC, we design

TABLE II: Accuracy (%) of BAN via transfer learning [55] and the proposed BAN episodic training for DG-FSC. For both methods, ResNet-10 [20] is the backbone network and ProtoNet [49] is used as the classifier head for a fair comparison. **Top:** miniImageNet (base ↦ novel), **Bottom:** miniImageNet ↦ CUB (unseen). Other experiment setups are the same as Table I. See detailed analysis in Sec. IV.

Method	gen-0	gen-1	gen-2	gen-3	gen-4
BAN transfer learning [55]	42.62	46.61	47.53	**47.92**	47.81
BAN episodic training (Ours)	50.39	53.10	54.61	**55.08**	54.91

Method	gen-0	gen-1	gen-2	gen-3	gen-4
BAN transfer learning [55]	38.19	39.07	**40.78**	40.66	40.32
BAN episodic training (Ours)	38.26	39.66	40.63	**41.10**	40.77

two experiment setups: (a) We train the student with one generation, *i.e.*, k=1, and test its performance on tasks of novel classes from various unseen domains. (b) We evaluate the performance of different born-again generations on novel classes of both seen and unseen domains. We employ a popular metric-based FSC model RelationNet [53] as the baseline method in this experiment. Follow [57], we use ResNet-10 [20] as the backbone network. We train each student network with 800 epochs (100 tasks in each epoch) of 5 generations. To enable the episodic training, in each iteration, we sample a 5-Way 1-Shot task from the base classes of miniImageNet [45]. In the testing stage, we randomly sample 1000 tasks from novel classes of either miniImageNet or different unseen domains to evaluate the performance of BAN in setup (a) and setup (b), with the average accuracy reported. We include the detailed dataset information in Sec. VI.

Results and analysis. The experiment results are shown in Figure 1 (qualitatively), Table I, and Figure 4 (quantitatively). Empirically, our observations can be summarized as follows:

1) Table I **(Top)**: Similar to the observation in conventional supervised learning, BAN episodic training can achieve consistent improvement on various novel unseen classes and unseen domains. This suggests the potential of BAN in boosting the generalization of other FSC models in domain generalization setups.
2) Table I **(Bottom)**: Multiple BAN generations lead to diminishing improvements. Compared to the born-again learning process of gen-0 ↦ gen-1, the improvement becomes small in deeper generations. We even observe the performance drop after the empirical optimal generation. Similar observations are also found in other applications of conventional supervised learning [55], [63]. See detailed analysis in Figure 4.
3) Visualization: We extract and analyze the features by the backbone network of a novel task during evaluation. Compared to the baseline model, we observe that BAN can lead to more discriminative features with better decision boundaries, see details in Figure 1.

Comparison with BAN transfer learning for FSC. Recently, Tian *et al.*[55] proposed to adopt BAN training in conventional supervised learning (as Figure 2) to obtain a powerful backbone network as the feature encoder. Then, in evaluation, they transfer it to the unseen FSC task ($\mathcal{T}$), extract features of the support set of $\mathcal{T}$, fit a new classifier, and predict query samples. In Table II, we compare the proposed BAN

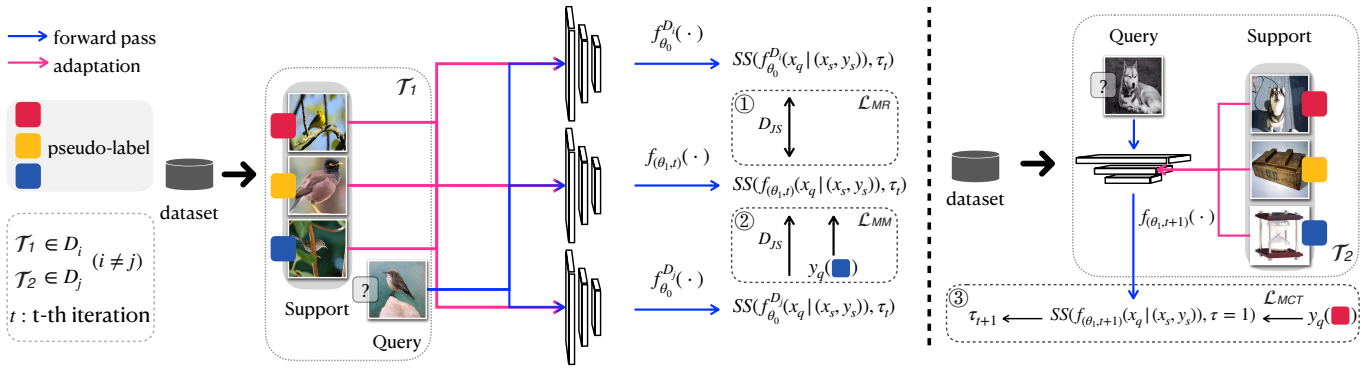

Fig. 5: Overview of our proposed FS-BAN for DG-FSC. ①**Mutual regularization**. To overcome the potential overfitting of teachers due to limited labeled data in an episode, we propose to regularize the teacher to match the soft distribution from the student. ② **Mismatched teacher**. To explicitly consider domain-shift in training, for a task sampled from domain $\mathcal{D}_i$, we propose to select a mismatched teacher trained on $\mathcal{D}_j$ for the knowledge transfer, where $i \neq j$. ③ **Meta-control the temperature**. The temperature τ is meta-updated in different iterations by evaluating the performance of the updated student on task from $\mathcal{D}_j (i \neq j)$.

episodic training with [55]. For a fair comparison, for both methods, ResNet-10 [20] is the feature encoder, and ProtoNet [49] is the classifier, which computes the feature distance between the query and the center of support samples of each class (*i.e.*, the "prototype") for prediction. We show that our proposed BAN episodic training can achieve competitive performance as [55] in different (DG-)FSC setups. On the other hand, episodic training attempts to simulate a realistic setting in evaluation by learning to solve FSC tasks, and it has been shown very useful to tackle novel, unseen classes given limited labeled data [49], [53], [60], [13]. Therefore, we are motivated to apply BAN directly in episodic training for (DG-)FSC, as Figure 3. **Critically**, in contrast to Tian *et al.*[55], taking advantage of episodic training, we do not modify the network structure or remove/add any layers during the entire training/test phase, and the classifier of our proposed method is compatible with many existing FSC models, which potentially can achieve better performance (see experiments in Sec. VI).

Next, we propose our improved method of BAN episodic training to tackle unique tasks in DG-FSC. In order to pursue an efficient learning process and prevent the computationally expensive sequential training, we exploit the major gain of BAN episodic training at gen-0 $\mapsto$ gen-1 and only train one generation student (*i.e.*, k=1 in Eqn. 7) in the rest of the paper.

V. Few-Shot BAN

We show in Sec. IV the promising results of BAN episodic training for DG-FSC, which indicate better generalization on novel class tasks from unseen domains during evaluation. However, the improvement of baseline BAN could have been inhibited due to several unique challenges of DG-FSC:

1) The particularity of BAN lies in that the teacher network is trained with an identical structure and the same training data as the student. In this few-shot scenario, overfitting of the teacher network could degrade the knowledge transferred to the student.
2) DG-FSC requires the FSC model to recognize novel tasks from unseen domains that are not accessible during training. Inspired by a recent DG work [31] for conventional image classification, it is useful to imitate such domain shift during training so that domain robustness can be improved.
3) The key hyper-parameter temperature τ in BAN is often pre-set to be a fixed value for different source domains, which could be sub-optimal. For DG-FSC tasks, we expect to find a proper temperature that is suitable for various seen domains and such that the student model can be better generalized to unseen domains.

To address the issues, we propose few-shot born-again networks (FS-BAN), including novel multi-task learning objectives with different teacher-student interactions, as Figure 5. We show in experiments that, these challenges are greatly mitigated with a marginal increment of the training cost.

A. Mutual Regularization

We show in Table I that BAN improves DG-FSC before the optimal generation. We attribute this to that the teacher at gen-k ($k > 1$) learned the cross-category knowledge [63] in the last generation. However, since we train $f_{\theta_k}(\cdot)$ only if $f_{\theta_{k-1}}(\cdot)$ converges, BAN suffers from the sequential training and it severely reduces the training efficiency.

To make the teacher reap the benefits of the soft knowledge, [63] emphasizes the importance of high-quality secondary information in prediction distribution, and Top Score Difference (TSD) regularization is proposed to make the prediction distribution less peaked to the primary class of the input samples. Differently, to avoid sequential training, we propose an alternative method that makes use of the student prediction distribution of each task. Concretely, we add a feedback path from the student to the teacher and mutually regularize (MR) both the teacher and the student with the soft prediction from each other. Besides the student is learned by Eqn. 7 (k=1), the well-trained teacher network is further fine-tuned by

$$\mathcal{L}_{MR} = \lambda_2 \tau^2 (D_{JS}(\hat{\mathcal{Y}}_q^{\theta_1}, \hat{\mathcal{Y}}_q^{\theta_0})). \quad (8)$$

As improvements on unseen domains and TSD analysis shown in the ablation study, $\mathcal{L}_{MR}$ reliably counteracts the

TABLE III: Meta-test accuracy (%) of DG-FSC with our proposed FS-BAN (Figure 5). We follow the experiment setup as in [57]. We let All={miniImageNet, CUB, Cars, Places, Plantae} be the union of all domains for training and testing. In *training phase*, we sample tasks from multiple seen domains, *e.g.*, All$\setminus$ {CUB}. In *testing phase*, we evaluate the model on tasks sampled from the leave-one-out selected unseen domain, *e.g.*, CUB. miniImageNet is always the source domain. FS-BAN-lite indicates that we do not include $\mathcal{L}_{MCT}$ in FS-BAN since it requires more GPU memory for training.

Method	All$\setminus$ {CUB}$\mapsto$CUB		All$\setminus$ {Cars}$\mapsto$Cars		All$\setminus$ {Places}$\mapsto$Places		All$\setminus$ {Plantae}$\mapsto$Plantae	
	5-Way 1-Shot	5-Way 5-Shot	5-Way 1-Shot	5-Way 5-Shot	5-Way 1-Shot	5-Way 5-Shot	5-Way 1-Shot	5-Way 5-Shot
MatchingNet [60]	37.90 ± 0.55	51.92 ± 0.80	28.96 ± 0.45	39.87 ± 0.51	49.01 ± 0.65	61.82 ± 0.57	33.21 ± 0.51	47.29 ± 0.51
+FT [57]	41.74 ± 0.59	56.29 ± 0.80	28.30 ± 0.44	39.58 ± 0.54	48.77 ± 0.65	62.32 ± 0.58	32.15 ± 0.50	46.48 ± 0.52
+LFT [57]	43.29 ± 0.59	61.41 ± 0.57	30.62 ± 0.48	43.08 ± 0.55	52.51 ± 0.67	64.99 ± 0.59	35.12 ± 0.54	48.32 ± 0.57
+FS-BAN-lite (Our)	45.22 ± 0.65	$\mathbf{62.83 \pm 0.59}$	$\mathbf{31.90 \pm 0.50}$	42.44 ± 0.56	$\mathbf{53.53 \pm 0.68}$	69.84 ± 0.55	39.83 ± 0.62	$\mathbf{54.87 \pm 0.57}$
+FS-BAN (Our)	$\mathbf{45.27 \pm 0.57}$	61.34 ± 0.52	31.71 ± 0.62	$\mathbf{45.01 \pm 0.57}$	53.33 ± 0.67	$\mathbf{70.09 \pm 0.60}$	$\mathbf{40.02 \pm 0.70}$	53.89 ± 0.64
RelationNet [53]	44.33 ± 0.59	62.13 ± 0.74	29.53 ± 0.45	40.64 ± 0.54	47.76 ± 0.63	64.34 ± 0.57	33.76 ± 0.52	46.29 ± 0.56
+FT [57]	44.87 ± 0.44	61.87 ± 0.39	30.09 ± 0.36	40.52 ± 0.40	48.12 ± 0.45	64.92 ± 0.40	35.53 ± 0.39	48.54 ± 0.38
+LFT [57]	48.38 ± 0.63	64.99 ± 0.54	32.21 ± 0.51	43.44 ± 0.59	50.74 ± 0.66	67.35 ± 0.54	35.00 ± 0.52	50.39 ± 0.52
+LRP [51]	45.64 ± 0.42	62.71 ± 0.39	30.00 ± 0.32	41.05 ± 0.37	48.74 ± 0.45	66.08 ± 0.40	36.04 ± 0.38	48.78 ± 0.37
+FS-BAN-lite (Our)	$\mathbf{48.69 \pm 0.65}$	65.37 ± 0.58	33.33 ± 0.57	44.35 ± 0.59	53.43 ± 0.66	$\mathbf{70.64 \pm 0.56}$	38.29 ± 0.62	53.40 ± 0.58
+FS-BAN (Our)	47.67 ± 0.59	$\mathbf{65.55 \pm 0.56}$	$\mathbf{33.43 \pm 0.57}$	$\mathbf{45.78 \pm 0.57}$	$\mathbf{53.50 \pm 0.68}$	69.72 ± 0.58	$\mathbf{38.75 \pm 0.61}$	$\mathbf{53.55 \pm 0.57}$
GNN [13]	49.46 ± 0.73	69.26 ± 0.68	32.95 ± 0.56	48.91 ± 0.67	51.39 ± 0.80	72.59 ± 0.67	37.15 ± 0.60	58.36 ± 0.68
+FT [57]	48.24 ± 0.75	70.37 ± 0.68	33.26 ± 0.56	47.68 ± 0.63	54.81 ± 0.81	74.48 ± 0.70	37.54 ± 0.62	57.85 ± 0.68
+LFT [57]	51.51 ± 0.80	73.11 ± 0.68	34.12 ± 0.63	49.88 ± 0.67	56.31 ± 0.80	77.05 ± 0.65	42.09 ± 0.68	58.84 ± 0.66
+FS-BAN-lite (Our)	$\mathbf{52.33 \pm 0.77}$	72.16 ± 0.67	34.50 ± 0.66	49.29 ± 0.68	58.86 ± 0.85	77.74 ± 0.62	41.28 ± 0.68	61.32 ± 0.67
+FS-BAN (Our)	52.07 ± 0.74	$\mathbf{73.70 \pm 0.66}$	$\mathbf{34.87 \pm 0.66}$	$\mathbf{50.66 \pm 0.65}$	$\mathbf{58.91 \pm 0.87}$	$\mathbf{78.57 \pm 0.67}$	$\mathbf{41.72 \pm 0.76}$	$\mathbf{61.85 \pm 0.66}$

overfitting of the teacher. Since N_w classes are randomly selected for each FSC task, the true categories corresponding to the pseudo-label vary on different tasks. This allows the teacher to learn meaningful cross-category information from samples instead of remembering the pseudo-label. Besides, the student can leverage the one-hot label (in Eqn. 7) to ensure that both teachers and students are on the correct updating directions, resulting in non-degenerate solutions.

Design Choices. There are several potential variants to regularize the teacher network, including using both classification loss and $\mathcal{L}_{MR}$. However, our goal is to make the teacher a regulator to provide soft knowledge and guide the approximate training direction such that the overfitting will not be passed to the students. Updating the teacher with cross-entropy loss may retain overfitting. Moreover, applying $\mathcal{L}_{MR}$ to intermediate network layers is also possible, but it is computationally complex and hard to find the perfect design. Therefore, we choose a simple way to perform $\mathcal{L}_{MR}$ in the output space.

B. The Mismatched Teachers

One core issue in DG-FSC is that we cannot get access to statistics from the target domain during training. Therefore, the reasonable way to improve the performance of an FSC model on *unseen domains* is to improve the robustness and make it produce more stable predictions on *various seen domains*.

To formulate a training scheme that is similar to the test phase on unseen domains, inspired by the recent work [31], we train the student in a way that exposes it to domain shift, making it robust to the mismatched source domain on which the current teacher is trained. Concretely, for each source domain $\mathcal{D}_i$, we train a teacher network using the training data of $\mathcal{D}_i$ via Eqn. 6, and we denote the teacher obtained on $\mathcal{D}_i$ as $f_{\theta_0^*}^{D_i}(\cdot)$. In each iteration, for a task $\mathcal{T}$ sampled from $\mathcal{D}_i$, the student is updated in the same way as Eqn. 7 but the teacher is obtained from a different domain $\mathcal{D}_j$ that has never seen $\mathcal{D}_i$ before. We update the student using the ground truth and the mismatched soft outputs of $f_{\theta_0^*}^{D_j}(\cdot)$, where $i \neq j$, as $\mathcal{L}_{MM}$:

$$\mathcal{L}_{MM} = \lambda_1 \mathcal{L}_{ce}(\mathcal{Y}_q, \hat{\mathcal{Y}}_q^{\theta_1}) + \lambda_3 \tau^2 D_{JS}(\hat{\mathcal{Y}}_q^{\theta_1}, \hat{\mathcal{Y}}_q^{(\theta_0, D_j)}). \quad (9)$$

How can a mismatched teacher (MM) help DG-FSC? Our insight is, if the student can be adapted to predict accurately on tasks from domain $\mathcal{D}_i$ while guided by a mismatched teacher obtained on domain $\mathcal{D}_j$ ($i \neq j$), then its robustness to domain-shift in the testing phase has increased. As minima quality analysis [31] shown in Sec. VI, $\mathcal{L}_{MM}$ improves domain-robustness compared to the baseline model. We further note that $\mathcal{L}_{MM}$ improves the generalization by a large margin on fine-grained unseen domains.

Design Choices. In each task, the teacher is randomly selected which is mismatched to the domain of the current task. Compared to the teacher from the same source domain applied conventionally, the student $f_{\theta_1}(\cdot)$ is penalized for the wrong prediction given the mismatched teacher that performs poorly on the current source domain. To minimize the total loss, the student model must learn to solve the task from the correct labels, but regularized by the teacher that is *under domain-shift*. In the ablation study, we show the separate benefit of $\mathcal{L}_{MM}$ that it outperforms the baseline BAN consistently.

C. Meta-Control the Temperature

A fixed temperature (*e.g.*, $\tau = 4$) is often applied in the BAN and KD training process to soften the prediction probability distribution. Therefore, the student model can learn the inter-class relationships predicted by the well-trained teacher network. However, in the DG-FSC setup, the fixed temperature is applied to various source domains, of which there may be large differences and it leads to sub-optimal performance. For example, in some tasks, a higher temperature may result in less difference between classes. We propose to use meta-learned temperature tuning on different source domains. Our idea is that with the adaptively tuned τ that is proper to various seen domains the student can learn appropriate inter-class knowledge and improve the performance on unseen domains.

Instead of directly updating τ, we propose a meta-learning scheme [10], [18], [38], [1] to efficiently tune the temperature (MCT): In iteration t, we sample two subtasks from two different source domains: $\mathcal{T}_1 \in D_i$, and $\mathcal{T}_2 \in D_j$. Firstly,

TABLE IV: Meta-test accuracy (%) for DG-FSC. Models are trained on miniImageNet and tested on various unseen domains. Note that in the following experiments there is only one domain involved in training, therefore only $\mathcal{L}_{MR}$ in FS-BAN is used. However, we show that FS-BAN still can improve the baseline FSC models consistently.

Method	miniImageNet ↦ CUB		miniImageNet ↦ Cars		miniImageNet ↦ Places		miniImageNet ↦ Plantae	
	5-Way 1-Shot	5-Way 5-Shot	5-Way 1-Shot	5-Way 5-Shot	5-Way 1-Shot	5-Way 5-Shot	5-Way 1-Shot	5-Way 5-Shot
MatchingNet [60]	35.89 ± 0.51	51.37 ± 0.77	$\mathbf{30.77 \pm 0.68}$	38.99 ± 0.64	49.86 ± 0.79	63.16 ± 0.77	32.70 ± 0.60	46.53 ± 0.68
+FT [57]	36.64 ± 0.53	55.23 ± 0.83	29.82 ± 0.44	$\mathbf{41.24 \pm 0.65}$	51.07 ± 0.72	64.55 ± 0.75	34.48 ± 0.50	41.69 ± 0.63
+Baseline BAN (Our)	36.47 ± 0.53	51.07 ± 0.58	28.71 ± 0.43	37.29 ± 0.49	51.05 ± 0.70	64.19 ± 0.61	35.01 ± 0.53	46.78 ± 0.53
+FS-BAN (Our)	$\mathbf{41.03 \pm 0.58}$	$\mathbf{55.54 \pm 0.56}$	30.38 ± 0.49	40.75 ± 0.54	$\mathbf{53.88 \pm 0.66}$	$\mathbf{68.55 \pm 0.55}$	$\mathbf{36.05 \pm 0.53}$	$\mathbf{50.68 \pm 0.51}$
RelationNet [53]	42.44 ± 0.77	57.77 ± 0.69	29.11 ± 0.60	37.33 ± 0.68	48.64 ± 0.85	63.32 ± 0.76	33.17 ± 0.64	44.00 ± 0.60
+FT [57]	44.07 ± 0.77	59.46 ± 0.71	28.63 ± 0.59	39.91 ± 0.69	50.68 ± 0.87	66.28 ± 0.72	33.14 ± 0.62	45.08 ± 0.59
+LRP [51]	42.44 ± 0.41	59.30 ± 0.40	29.65 ± 0.33	39.19 ± 0.38	50.59 ± 0.46	66.90 ± 0.40	34.80 ± 0.37	$\mathbf{48.09 \pm 0.35}$
+Baseline BAN (Our)	43.35 ± 0.60	60.79 ± 0.55	29.71 ± 0.46	39.27 ± 0.53	51.30 ± 0.68	67.62 ± 0.54	33.81 ± 0.51	46.26 ± 0.51
+FS-BAN (Our)	$\mathbf{44.41 \pm 0.60}$	$\mathbf{61.31 \pm 0.55}$	$\mathbf{30.80 \pm 0.49}$	$\mathbf{40.47 \pm 0.54}$	$\mathbf{53.97 \pm 0.72}$	$\mathbf{70.21 \pm 0.56}$	$\mathbf{35.36 \pm 0.54}$	47.95 ± 0.54
GNN [13]	45.69 ± 0.68	62.25 ± 0.65	31.79 ± 0.51	44.28 ± 0.63	53.10 ± 0.80	70.84 ± 0.65	35.60 ± 0.56	52.53 ± 0.59
+FT [57]	47.47 ± 0.75	66.98 ± 0.68	31.61 ± 0.53	44.90 ± 0.64	53.77 ± 0.79	73.94 ± 0.67	35.95 ± 0.58	53.85 ± 0.62
+LRP [51]	48.29 ± 0.51	64.44 ± 0.48	32.78 ± 0.39	46.20 ± 0.46	54.83 ± 0.56	74.45 ± 0.47	37.49 ± 0.43	54.46 ± 0.46
+Baseline BAN (Our)	47.08 ± 0.70	68.30 ± 0.68	32.22 ± 0.56	44.65 ± 0.63	54.82 ± 0.80	74.94 ± 0.66	36.71 ± 0.63	55.34 ± 0.63
+FS-BAN (Our)	$\mathbf{50.04 \pm 0.76}$	$\mathbf{69.83 \pm 0.66}$	$\mathbf{33.21 \pm 0.60}$	$\mathbf{46.48 \pm 0.66}$	$\mathbf{59.70 \pm 0.84}$	$\mathbf{75.91 \pm 0.65}$	$\mathbf{38.87 \pm 0.64}$	$\mathbf{56.09 \pm 0.66}$

TABLE V: Conventional FSC results with model trained and tested solely on miniImageNet or tieredImageNet. Follow [57], we use ResNet-10 as the backbone and we show that our method can surpass the state-of-the-art methods with fewer parameters.

Method	Backbone	miniImageNet (base ↦ novel)		tieredImageNet (base ↦ novel)	
		5-Way 1-Shot	5-Way 5-Shot	5-Way 1-Shot	5-Way 5-Shot
TADAM [41]	ResNet-12	58.50 ± 0.30	76.70 ± 0.30	–	–
MTL [52]	ResNet-12	61.20 ± 1.80	75.50 ± 0.80		
CAN [23]	ResNet-12	63.85 ± 0.48	79.44 ± 0.34	69.89 ± 0.34	84.23 ± 0.37
MetaOptNet [29]	ResNet-12	64.09 ± 0.62	80.00 ± 0.45	65.99 ± 0.72	81.56 ± 0.53
Neg-Cosine [39]	ResNet-12	63.85 ± 0.81	81.57 ± 0.43	–	–
RFS [55]	ResNet-12	64.82 ± 0.60	82.14 ± 0.56	$\mathbf{71.52 \pm 0.69}$	86.03 ± 0.49
LEO [48]	WRN-28-10	61.76 ± 0.08	77.59 ± 0.12	66.33 ± 0.05	81.44 ± 0.09
DAE-GNN [16]	WRN-28-10	62.96 ± 0.15	78.85 ± 0.10	68.18 ± 0.16	83.09 ± 0.12
AWGIM [64]	WRN-28-10	63.12 ± 0.08	78.40 ± 0.11	67.69 ± 0.11	82.82 ± 0.13
GNN [13]	ResNet-10	60.77 ± 0.75	80.87 ± 0.56	66.37 ± 1.09	85.79 ± 0.51
+FT [57]	ResNet-10	66.32 ± 0.80	81.98 ± 0.55	–	–
+FS-BAN (Our)	ResNet-10	$\mathbf{68.82 \pm 0.78}$	$\mathbf{84.89 \pm 0.50}$	70.55 ± 0.78	$\mathbf{88.80 \pm 0.26}$

we update the student network $f_{(\theta_1,t)}(\cdot)$ on $\mathcal{T}_1$, given a predetermined τ_t. Then, for task $\mathcal{T}_2$, we fix the weights of the student $f_{(\theta_1,t+1)}(\cdot)$ and evaluate the effectiveness of the temperature τ_t that is applied to train the student on $\mathcal{T}_1$, by testing the performance of $f_{(\theta_1,t+1)}(\cdot)$. In this step, we use only cross-entropy loss (τ=1), which is the same as the testing phase. The temperature τ_{t+1} is obtained by evaluating $f_{(\theta_1,t+1)}(\cdot)$ on query set $\mathcal{Q}_2 = \{\mathcal{X}_{(q,2)}, \mathcal{Y}_{(q,2)}\}$ of $\mathcal{T}_2$:

$$\mathcal{L}_{MCT} = \mathcal{L}_{ce}(\mathcal{Y}_{(q,2)}, \hat{\mathcal{Y}}_{(q,2)}^{(\theta_1,t+1)}). \tag{10}$$

τ_{t+1} is used in the next iteration *t+1*. With the adaptively fine-tuned temperature, we obtain a meta-learned hyperparameter that is trained to adapt to diverse domains.

Design Choices. Potentially, we have several ways to tune the temperature: (i) The simplest way is that the temperature is updated directly as a normal learnable parameter in episodic training. (ii) We update the student on $\mathcal{T}_1$ and evaluate the effectiveness of the temperature on $\mathcal{T}_2$, but both tasks are from the same domain, *i.e.*, $\mathcal{T}_1, \mathcal{T}_2 \in \mathcal{D}_1$. (iii) (Proposed $\mathcal{L}_{MCT}$ in FS-BAN) We update the student on $\mathcal{T}_1$ and evaluate the effectiveness of the temperature on $\mathcal{T}_2$, and the two tasks are from different source domains *i.e.*, $\mathcal{T}_1 \in \mathcal{D}_i, \mathcal{T}_2 \in \mathcal{D}_j$, as Figure 5. In Sec. VI, the temperature with setup (iii) converges gradually in the training process and gains better performance in the evaluation stage, indicating that we find a temperature suitable to diverse domains and training tasks. Therefore, we choose this setup for FS-BAN.

D. Multi-Task Learning Objectives

The final learning objective of FS-BAN is:

$$\mathcal{L} = \mathcal{L}_{MR} + \mathcal{L}_{MM} + \mathcal{L}_{MCT}. \tag{11}$$

In the next, we conduct comprehensive experiments to evaluate the effectiveness of FS-BAN on public datasets with popular FSC models as baselines. Detailed ablation studies and analyses are performed both qualitatively and quantitatively.

VI. Experiments

In this section, we discuss the experiment settings and evaluate the proposed FS-BAN on six publicly available datasets with three popular metric baseline FSC models. We also conduct detailed ablation studies.

A. Datasets

We evaluate the proposed FS-BAN on six publicly available datasets: miniImageNet [45], tierdImageNet [47], Caltech-UCSD Birds 200 (CUB) [61], Stanford Cars (Cars) [27], Places [73] and Plantae [58]. We follow the dataset split protocol as previous work [57] for a fair comparison, and we summarize it in Table VII. In the meta-training phase, we use the standard data augmentation skills, including image jittering, random crop, random horizontal flip, and normalization for better generalization. In the meta-valid and meta-test stages, we do not use data augmentation.

TABLE VI: Ablation study of FS-BAN with meta-test accuracy (%). Model is trained on several seen source domains and evaluated with 5-Way 5-Shot tasks on the leave-one-out selected unseen domain. We show that each proposed component in FS-BAN improves the baseline models separately and they are complementary to each other.

5-Way 5-Shot	$\mathcal{L}_{MR}$	$\mathcal{L}_{MCT}$	$\mathcal{L}_{MM}$	All\ {CUB}↦CUB	All\ {Cars}↦Cars	All\ {Places}↦Places	All\ {Plantae}↦Plantae
MatchingNet [60]	-	-	-	51.92 ± 0.80	39.87 ± 0.51	61.82 ± 0.57	47.29 ± 0.51
FT[57]	-	-	-	56.29 ± 0.80	39.58 ± 0.54	62.32 ± 0.58	46.48 ± 0.52
LFT [57]	-	-	-	61.41 ± 0.57	43.08 ± 0.55	64.99 ± 0.59	48.32 ± 0.57
Baseline BAN	-	-	-	53.47 ± 0.58	39.60 ± 0.51	62.37 ± 0.60	48.42 ± 0.57
	✓	-	-	59.75 ± 0.56	42.03 ± 0.55	69.34 ± 0.57	54.61 ± 0.58
	-	✓	-	55.97 ± 0.59	41.97 ± 0.55	64.37 ± 0.58	50.61 ± 0.59
	-	-	✓	57.28 ± 0.58	44.83 ± 0.59	68.21 ± 0.57	55.35 ± 0.55
	-	✓	✓	58.25 ± 0.59	42.91 ± 0.56	66.22 ± 0.55	51.99 ± 0.54
	✓	✓	-	61.64 ± 0.59	42.18 ± 0.56	69.80 ± 0.58	$\mathbf{56.38 \pm 0.60}$
FS-BAN-lite	✓	-	✓	$\mathbf{62.83 \pm 0.59}$	42.44 ± 0.56	69.84 ± 0.55	54.87 ± 0.57
FS-BAN	✓	✓	✓	61.34 ± 0.52	$\mathbf{45.01 \pm 0.57}$	$\mathbf{70.09 \pm 0.60}$	53.89 ± 0.64

TABLE VII: Collection of domains and the class split.

Domain	miniImageNet	tieredImageNet	CUB	Cars	Places	Plantae
# Training classes	64	351	100	98	183	100
# Valid classes	16	97	50	49	91	50
# Test classes	20	160	50	49	91	50

B. Baseline Models

Since FS-BAN does not require additional learnable parameters and can be readily used to existing FSC methods, we apply FS-BAN to three popular metric-based FSC models to validate the effectiveness of FS-BAN: MatchingNet [60], RelationNet [53] and Graph Neural Network (GNN) [13]. All these baseline models share the same feature extractor as the backbone network and only differ in the metric-based classifier head for prediction. For other DG-FSC methods, we compare with [57] that applies feature-wise transformation layers (LFT) to improve the generalization. We also compare with layer-wise relevance propagation (LRP, [51]) and more state-of-the-art FSC models in both single domain and DG setups.

C. Experiment Setups

For a fair comparison, we follow [57] to assume there are multiple seen source domains in training. Nevertheless, to comprehensively evaluate different methods, in the main experiments, we perform three experiment setups:

1) Models are trained on tasks of base classes of multiple seen source domains and tested on a target unseen domain. The source domains are selected from All={miniImageNet, CUB, Cars, Places, Plantae}, *e.g.*, All\ {CUB}. The unseen domain for testing is the held-out domain during training, *e.g.*, CUB.
2) Models are trained on a single source domain, *i.e.*, base classes of miniImageNet, and tested on novel classes of various unseen domains, *e.g.*, All\ {miniImageNet}.
3) We further perform conventional FSC [53], [64] experiments, where base classes for training and novel classes for evaluation are from the same domain (*e.g.*, miniImageNet).

Note that in setups 2) and 3) there is only one source domain involved in training, therefore only $\mathcal{L}_{MR}$ in FS-BAN is applicable. However, we show that this partial FS-BAN can still outperform other methods.

For all experiment setups, follow [57], [51], we use ResNet-10 [20] as the backbone network for baseline models and our method. We initialize the temperature with $\tau = 4$, and it is activated by a SoftPlus function to ensure it is non-negative:

$$\tau = \text{SoftPlus}(\tau) = \ln(1 + e^{\tau}), \tag{12}$$

where τ is updated in each iteration, as described in Sec. V-C.

D. Implementation Details

We strictly follow the standard FSC setups [48], [64], [53], [49]: either 5-Way 1-Shot or 5-Way 5-Shot tasks are sampled in training and testing stages. In each task, we sample $N_q = 16$ query images per category to compute the loss and accuracy. We train FS-BAN with 800 epochs (100 tasks are sampled from a random source domain in each epoch). We apply the Adam optimizer [26] to train the models with default hyperparameters, *e.g.*, learning rate 0.001. In the testing phase, we sample 1,000 tasks of novel classes from the unseen target domain in setup 1) and 2) and the same source domain in setup 3) for evaluation, respectively. We select the model checkpoints with the best validation accuracy and report the average accuracy on the test set with 95% confidence interval.

On the other hand, we follow the prior works [57], [48] to pre-train the backbone model (ResNet-10 [20] feature encoder with a linear layer as the classifier) on 64 base classes of mini-ImageNet, by minimizing a standard cross-entropy loss, as Eqn. 1. After that, we remove the classifier head and use the pre-trained backbone weights to initialize the student network for episodic training for DG-FSC. Therefore, at the beginning of the meta-training stage, the student is equipped with a feature encoder that can extract discriminative features. We use this technique in all our experiments as it has been shown very useful in FSC in prior works [57], [48], [15], [37].

E. Experiment Results

The results of experiment setup 1), 2) and 3) are shown in Table III, Table IV, and Table V, respectively. In all setups, our proposed FS-BAN consistently improves the different baseline FSC models to state-of-the-art, presenting desirable performance on unseen domains. Since there is no true label

TABLE VIII: Meta-test accuracy (%) with different implementation techniques for mutual regularization. Model is trained on several seen source domains and evaluated on the leave-one-out selected unseen domain with 5-Way 5-Shot tasks. The feature encoders of all models are pre-trained on mini-ImageNet.

Method	All\ {CUB}↦CUB	All\ {Cars}↦Cars	All\ {Places}↦Places	All\ {Plantae}↦Plantae
MatchingNet [60]	51.92 ± 0.80	39.87 ± 0.51	61.82 ± 0.57	47.29 ± 0.51
+ Baseline BAN	53.47 ± 0.58	39.60 ± 0.51	62.37 ± 0.60	48.42 ± 0.57
+ $\mathcal{L}_{MR}$ (w/o student warmup)	55.98 ± 0.54	40.51 ± 0.57	64.84 ± 0.60	50.37 ± 0.57
+ $\mathcal{L}_{MR}$ (w/ student warmup)	58.46 ± 0.55	41.82 ± 0.55	68.56 ± 0.61	52.77 ± 0.54
+ $\mathcal{L}_{MR}$ (w/ student warmup + reduced teacher lr)	$\mathbf{59.75 \pm 0.56}$	$\mathbf{42.03 \pm 0.55}$	$\mathbf{69.34 \pm 0.57}$	$\mathbf{54.61 \pm 0.58}$

TABLE IX: Ablation study of coefficients for $\mathcal{L}_{MR}$ and $\mathcal{L}_{MM}$ of FS-BAN. We use the same experiment setup as Table III.

5-Way 5-Shot	λ_1	λ_2	All\ {CUB}↦CUB	All\ {Cars}↦Cars	All\ {Places}↦Places	All\ {Plantae}↦Plantae
MatchingNet + $\mathcal{L}_{MR}$	1	0	51.92 ± 0.80	39.87 ± 0.51	61.82 ± 0.57	47.29 ± 0.51
	1	0.2	56.27 ± 0.57	41.64 ± 0.56	66.55 ± 0.58	52.71 ± 0.56
	1	0.5	57.50 ± 0.50	40.31 ± 0.52	67.88 ± 0.56	52.73 ± 0.56
	1	0.8	$\mathbf{59.75 \pm 0.56}$	$\mathbf{42.03 \pm 0.55}$	$\mathbf{69.34 \pm 0.57}$	$\mathbf{54.61 \pm 0.58}$
5-Way 5-Shot	λ_1	λ_3	All\ {CUB}↦CUB	All\ {Cars}↦Cars	All\ {Places}↦Places	All\ {Plantae}↦Plantae
MatchingNet + $\mathcal{L}_{MM}$	1	0	51.92 ± 0.80	39.87 ± 0.51	61.82 ± 0.57	47.29 ± 0.51
	1	0.2	55.42 ± 0.60	42.89 ± 0.57	66.77 ± 0.57	52.87 ± 0.55
	1	0.5	$\mathbf{57.28 \pm 0.58}$	44.83 ± 0.59	$\mathbf{68.21 \pm 0.57}$	$\mathbf{55.35 \pm 0.55}$
	1	0.8	55.32 ± 0.57	$\mathbf{45.45 \pm 0.58}$	67.98 ± 0.56	53.26 ± 0.57

in episodic training for DG-FSC (*i.e.*, samples are pseudo-labeled in different tasks), the results imply that our obtained models indeed learn generalizable knowledge that can help tackle different tasks on novel classes of unseen domains, as analysis in Table XII, where we show the improved accuracy and lower top-score difference in the prediction distributions.

Compared to the prior state-of-the-art method that introduces additional learnable parameters [57], our proposed FS-BAN can address unique issues of DG-FSC, including overfitting and domain shift, benefits the network generalizability on unseen target domains, and improves the performance without additional inference cost in the deployment stage. We further note that in setup 3) where base classes for training and novel classes for evaluation are from the same domain (hence the domain gap is reduced), our method can still achieve considerable improvement consistently, even with fewer parameters of the backbone network, as Table V.

As we show in the ablation studies in the next, our proposed learning objectives in FS-BAN successfully address the unique challenges posed in DG-FSC, and the generalizability is greatly improved on unseen domains.

F. Ablation Study

Ablation study of learning objectives of FS-BAN. To evaluate the effectiveness of each individual component in the multi-task learning objectives of the proposed FS-BAN, we conduct comprehensive ablation studies and observe the empirical performance of FS-BAN in the DG-FSC setup. We use MatchingNet as the baseline model. We sample 5-Way 5-Shot tasks for training and evaluation, and other settings are the same as setup 1). The results are shown in Table VI.

We show that each separate learning objective ($\mathcal{L}_{MR}$, $\mathcal{L}_{\mathcal{MM}}$, $\mathcal{L}_{MCT}$) in FS-BAN improves the baseline models effectively and they are complimentary to each other. On the other hand, the FS-BAN with the full multi-task learning objectives achieves a good balance and performance on different unseen domains (the last row in Table VI).

In practice, to ensure that the feedback from the student prediction for ($\mathcal{L}_{MR}$) is reasonable and will not mislead the fine-tuning of the teacher network, *esp.,* at the beginning of student training, we introduce a "student warmup" process to let the student train 10 epochs with randomly sampled tasks before its feedback to the well-trained teacher network. We note that the backbone of the student model is pre-trained on mini-ImageNet training classes. We also reduce the learning rate (*lr*) of the teacher network by a factor of 5 compared to that of the student network, such that the teacher model is only moderately updated. In Table VIII, we study the impact of "student warmup" and "reduced *lr* for teacher" for $\mathcal{L}_{MR}$, and we show that the adopted techniques can improve the performance for $\mathcal{L}_{MR}$ by a considerable margin.

On the other hand, interestingly, when we only use FS-BAN with the mismatched teacher ($\mathcal{L}_{MM}$), the performance is better than all baselines and the state-of-the-art models with Cars being the unseen domain. This suggests improved generalization on the fine-grained datasets.

Ablation study of coefficients in learning objectives. We perform a grid search to select the coefficients of the loss objectives of λ_2 in $\mathcal{L}_{MR}$ and λ_3 in $\mathcal{L}_{MM}$ in our work. We fix $\lambda_1 = 1$ for cross-entropy loss and tune λ_2 for $\mathcal{L}_{MR}$ and λ_3 for $\mathcal{L}_{MM}$. In Table IX, we show the accuracy with different choices of coefficients. The student is trained using setup 1) with 5-Way 5-Shot tasks. MatchingNet is the baseline model. We observe that the performance of our proposed model is not very sensitive to different coefficients, and our methods can outperform the baseline method (where $\lambda_2 = \lambda_3 = 0$) significantly. Nevertheless, we note that it is possible that the mismatched teacher may harm the accuracy of the student, especially when such *domain-shift* training in $\mathcal{L}_{MM}$ is over-emphasized. Here, we found that $\lambda_3 = 0.5$ is the best weight of $\mathcal{L}_{MM}$, which indicates that it is not over-emphasized. Based

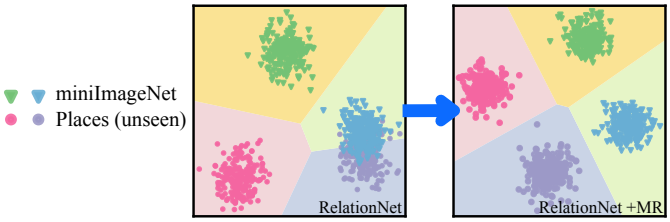

Fig. 6: Qualitative evaluation of the class separation. We show the projection of novel class features of the first and second components of LDA. We sample 200 images from miniImageNet and Places (unseen) separately. It is clear that $\mathcal{L}_{MR}$ brings better decision boundaries for the DG-FSC setup.

TABLE X: Quantitative class separation evaluation for miniImageNet → Places. 'RN' and 'MN' indicate the RelationNet, and MatchingNet, respectively. Following [17], lower values correspond to better feature clustering of novel tasks.

Metric (↓)	RN	RN+$\mathcal{L}_{MR}$	MN	MN+$\mathcal{L}_{MR}$	GNN	GNN+$\mathcal{L}_{MR}$
R_{FC}	7.94	**6.40**	2.24	**2.22**	6.32	**5.85**
R_{HV}	1.81	**1.73**	1.78	**1.59**	1.75	**1.71**

on the empirical results in Table IX, we choose $\lambda_1 = 1, \lambda_2 = 0.8, \lambda_3 = 0.5$ as the coefficients in $\mathcal{L}_{MR}$ and $\mathcal{L}_{MM}$, which performs well in most experiments.

Mutual Regularization leads to better separation boundaries. In this analysis, we validate the effectiveness of $\mathcal{L}_{MR}$. For simplicity, models are trained on base classes of miniImageNet with 5-Way 1-Shot tasks.

In Figure 7, we visualize the performance of the teacher and the student in the meta-valid phase on tasks sampled from novel classes of miniImageNet. Compared to the baseline model and the original BAN (*i.e.*, teacher without $\mathcal{L}_{MR}$), we observe that both the teacher and the student gain better generalization performance on novel classes. Meanwhile, the student can consistently outperform the improved teacher network, which suggests that $\mathcal{L}_{MR}$ maintains the advantage of the baseline BAN and brings non-degenerate solutions.

Where does this improved performance come from? Qualitatively, in Figure 6, we follow [17] to sample tasks from novel classes (miniImageNet) and unseen domain (Places) and project the extracted features (via backbone network) of query samples onto the first two components of LDA [40], on directions that minimize the intra-class to inter-class variance ratio. In the plots, we observe that $\mathcal{L}_{MR}$ obtains better class separability, which leads to better generalization ability on novel classes of unseen domains.

Quantitatively, we further follow [17] to analyze the quality of the learned features for few-shot tasks, via feature clustering (R_{FC}) and hyperplane variation (R_{HV}). For measurement of R_{FC}, we explicitly compute the intra-class to inter-class variance ratio. Denote the data in class i and index j by $\{x_{i,j}\}$, feature extractor by E. μ_i is the centroid feature of class i and μ is the centroid feature of all classes, we have:

$$R_{FC}(E, \{x_{i,j}\}) = \frac{N_w}{N_q} \frac{\sum_{ij} \|E(x_{i,j}) - \mu_i\|_2^2}{\sum_i \|\mu_i - \mu\|_2^2}, \qquad (13)$$

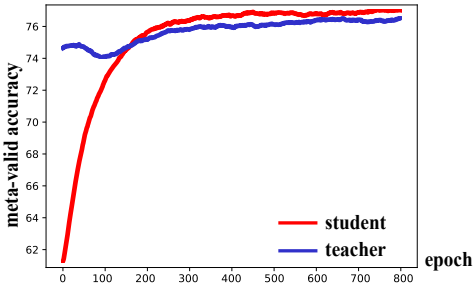

Fig. 7: Valid accuracy (%) on novel classes of miniImageNet with only $\mathcal{L}_{MR}$ in FS-BAN. We show that, regularized by $\mathcal{L}_{MR}$, the well-trained teacher network can be continually improved and gain better performance on unseen novel classes. It in turn leads to a better student with improved generalizability that even outperforms the teacher network consistently.

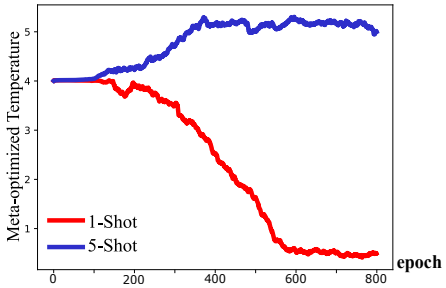

Fig. 8: Visualization of the temperature (τ) value during training with $\mathcal{L}_{MCT}$. Compared to a fixed τ, our proposed $\mathcal{L}_{MCT}$ can help find a proper temperature that is suitable for diverse domains. In Table XI we demonstrate the performance with different design choices of updating τ and show the effectiveness of the proposed $\mathcal{L}_{MCT}$.

where N_w and N_q are number of classes and query samples per class. When R_{FC}=0, samples of the same category are mapped to a single point, and there is no uncertainty of hyperplane when separating arbitrary samples from two classes. Similarly, Hyperplane Variation (R_{HV}) measures the sensitivity of separating hyperplanes to data sampling. For both R_{FC} and R_{HV}, the lower value corresponds to better class separation. We compute R_{FC} and R_{HV} by sampling 200 query images per category, averaging over 1000 novel 5-Way 1-Shot tasks of the unseen domain. These numerical results are shown in Table X. Furthermore, as TSD analysis in Sec. VI.G, it is clear that the improvement comes from the awareness of cross-category information of the teacher network, thus a simple $\mathcal{L}_{MR}$ brings better class separation and feature clustering performance on unseen domains.

Temperature convergence and analysis with $\mathcal{L}_{MCT}$. We visualize the meta-learned temperature (initialized by τ=4) trained with experiment setup i), as Figure 8. In both 1-Shot and 5-Shot training processes, the meta-controlled temperature gradually converges, finding its own equilibrium. Therefore, the adaptively tuned hyperparameter on diverse domains is the reason that we obtain the improvements, as numerical results in Table VI. Meanwhile, in Table XI, we further note that directly updating the temperature as a learnable parameter does not bring improvement and introduces overfitting.

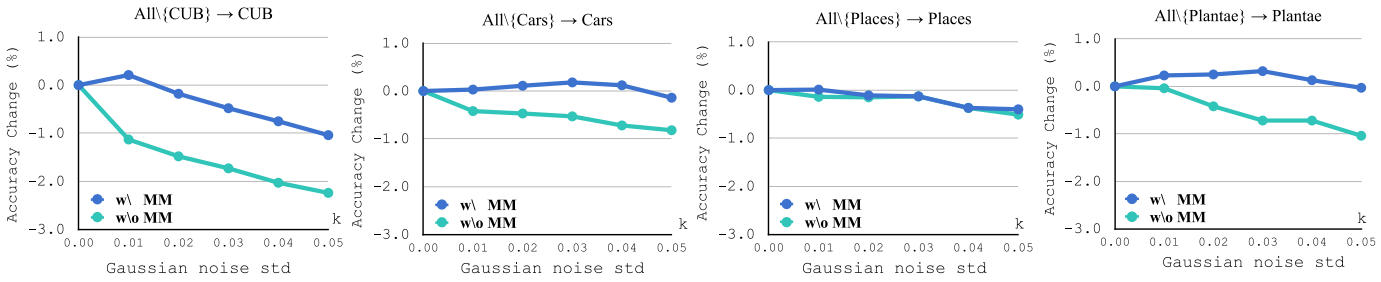

Fig. 9: Minima quality analysis of $\mathcal{L}_{MM}$: Baseline FSC model vs. FS-BAN (only with $\mathcal{L}_{MM}$). We apply experiment setup i) with 5-Way 1-Shot tasks for training and testing, and observe the performance change by adding Gaussian noise with different std to all model parameters. We show that, compared to the baseline model, the domain robustness has been clearly improved.

TABLE XI: Meta-test accuracy (%) of DG-FSC with experiment setup 1) by updating the temperature with different design choices on 5-Way 5-Shot tasks.

Models	setting	All\{CUB}↦CUB	All\{Cars}↦Cars
MatchingNet	no update τ	51.92 ± 0.80	39.87 ± 0.51
	directly update τ	50.58 ± 0.59	40.57 ± 0.52
	$\mathcal{L}_{MCT}$	$\mathbf{55.97 \pm 0.59}$	$\mathbf{41.97 \pm 0.55}$
		All\{Places}↦Places	All\{Plantae}↦Plantae
MatchingNet	no update τ	61.82 ± 0.57	47.29 ± 0.51
	directly update τ	61.89 ± 0.59	48.95 ± 0.57
	$\mathcal{L}_{MCT}$	$\mathbf{64.37 \pm 0.58}$	$\mathbf{50.61 \pm 0.59}$

TABLE XII: We train models on base classes of miniImageNet using 5-Way 5-Shot tasks. We present the average Top-score difference (TSD) of the prediction distribution of the teacher in training, and the performance of the student on novel classes from the unseen domain (Places) in testing.

(a) TSD of the teacher with M=3 in the training process.

	Baseline BAN	Baseline BAN + $\mathcal{L}_{MR}$
TSD	0.78	**0.64**

(b) Meta-test accuracy (%) on novel classes on different domains.

Dataset	Baseline BAN	Baseline BAN + $\mathcal{L}_{MR}$
miniImageNet (base ↦ novel)	73.79	**75.31**
miniImageNet↦Places	64.19	**68.55**

Mismatched Teachers improve solution robustness. How to understand that a randomly selected and mismatched teacher of FS-BAN improves robustness to DG-FSC (see Table III)? One ideal case is that converging to a 'wide' minima leads to a more robust solution of the model. Recently, some DG literature on conventional supervised learning [5], [31], [25] analyze the model robustness in terms of evaluating the solution minima quality.

Following [31], [25], we compare the model robustness, by adding Gaussian noise to model parameters and observe the accuracy change in the testing phase, as Figure 9. In most cases, we can observe that FS-BAN (with only $\mathcal{L}_{MM}$) brings higher robustness facing perturbation, which suggests better minima quality and generalization on held-out unseen domains. Another interesting observation is that in some cases we obtain an incremental improvement by introducing noise to model weights, which is a by-product that is related to a recent work [57].

TABLE XIII: The teacher is trained on miniImageNet and we evaluate its characteristics on different domains.

Teacher Model	miniImageNet	mini↦CUB	mini↦Cars	mini↦Places	mini↦Plantae
Accuracy (%)	**70.96**	51.37	38.99	63.16	46.53
TSD	**0.64**	0.39	0.23	0.48	0.35

G. Top-score Difference Analysis

In their previous work [63], they show that a better student model can be obtained with a more tolerant teacher, which is less focused on the primary class when making predictions. That is, the teacher passes the reasonable inter-class knowledge to the student (*i.e.*, probability prediction to all categories). Following their findings, in episodic training, we measure the Top-score difference (TSD) of the probability predictions produced by the teacher network:

$$\text{TSD} = f_{\theta_0,a1}(\cdot) - \frac{1}{M-1}\sum_{m=2}^{M} f_{\theta_0,am}(\cdot), \tag{14}$$

where $f_{\theta_0,am}(\cdot)$ is short for m-th largest value in the probability distribution $f_{\theta_0}(\cdot)$. We set a fixed $M = 3$ which represents the number of potential semantically similar classes for each image in the episode, including the primary class (the class assigned the highest probability). Then, we calculate the gap between the prediction probabilities of the primary class and the average of other $M-1$ classes with the highest scores.

TSD for $\mathcal{L}_{MR}$. In FS-BAN, $\mathcal{L}_{MR}$ requires the teacher network, *i.e.* $f_{\theta_0}(\cdot)$, to match the soft distribution produced from the student, *i.e.* $f_{\theta_1}(\cdot)$. Therefore, the teacher can learn the cross-category similarity information from the student. Here, we quantify these benefits via statistical measurements during training. As shown in Table XII, in the training process, $\mathcal{L}_{MR}$ indirectly reduces TSD of the teacher network, which suggests that the produced soft predictions are less picked and the similarity knowledge is well preserved. In the testing phase, $\mathcal{L}_{MR}$ for FS-BAN has higher accuracy. Therefore, the teacher network is less overfitting and preserves the meaningful soft knowledge transferred from the student.

TSD for $\mathcal{L}_{MM}$. What does the student learn from the mismatched teacher? To understand the working mechanism of FS-BAN mismatched teachers, a potentially ideal way is to observe the behavior of the mismatched teacher. We select the different source domains, then we observe the performance of the teacher training on the miniImageNet (hence, when

TABLE XIV: Meta-test accuracy (%) with different backbone networks of teacher models. Model is trained on several seen source domains and evaluated on the leave-one-out selected unseen domain with 5-Way 5-Shot tasks.

Method	**Backbone**	**Backbone of Teacher**	All\ {CUB}↦CUB	All\ {Cars}↦Cars	All\ {Places}↦Places	All\ {Plantae}↦Plantae
MatchingNet	Conv-4	-	50.27 ± 0.54	37.75 ± 0.52	56.72 ± 0.55	43.22 ± 0.59
+FS-BAN	Conv-4	Conv-4	51.95 ± 0.61	43.20 ± 0.56	62.45 ± 0.49	44.28 ± 0.61
+FS-BAN	Conv-4	Conv-6	$\mathbf{53.20 \pm 0.58}$	$\mathbf{44.95 \pm 0.52}$	$\mathbf{64.38 \pm 0.54}$	$\mathbf{48.14 \pm 0.60}$
MatchingNet	ResNet-10	-	51.92 ± 0.80	39.87 ± 0.51	61.82 ± 0.57	47.29 ± 0.51
+FS-BAN	ResNet-10	ResNet-10	61.34 ± 0.52	45.01 ± 0.57	70.09 ± 0.60	53.89 ± 0.64
+FS-BAN	ResNet-10	ResNet-18	$\mathbf{61.58 \pm 0.55}$	$\mathbf{46.73 \pm 0.58}$	$\mathbf{70.31 \pm 0.61}$	$\mathbf{54.44 \pm 0.57}$

TABLE XV: We compare the meta-test accuracy (%) on unseen domains to the method proposed by Tian *et al.*[55]. Model is trained on several seen source domains and evaluated on the leave-one-out selected unseen domain with 5-Way 5-Shot tasks.

Method	All\ {CUB}↦CUB	All\ {Cars}↦Cars	All\ {Places}↦Places	All\ {Plantae}↦Plantae
[43] (tian *et al.*[55])	56.07 ± 0.77	41.22 ± 0.59	67.73 ± 0.61	52.97 ± 0.50
MatchingNet [60]	51.92 ± 0.80	39.87 ± 0.51	61.82 ± 0.57	47.29 ± 0.51
+ FS-BAN (Ours)	$\mathbf{61.34 \pm 0.52}$	$\mathbf{45.01 \pm 0.57}$	$\mathbf{70.09 \pm 0.60}$	$\mathbf{53.89 \pm 0.64}$
RelationNet [53]	62.13 ± 0.74	40.64 ± 0.54	64.34 ± 0.57	46.29 ± 0.56
+ FS-BAN (Ours)	$\mathbf{65.55 \pm 0.56}$	$\mathbf{45.78 \pm 0.57}$	$\mathbf{69.72 \pm 0.58}$	$\mathbf{53.55 \pm 0.57}$
GNN [13]	69.26 ± 0.68	48.91 ± 0.67	72.59 ± 0.67	58.36 ± 0.68
+ FS-BAN (Ours)	$\mathbf{73.70 \pm 0.66}$	$\mathbf{50.66 \pm 0.65}$	$\mathbf{78.57 \pm 0.67}$	$\mathbf{61.85 \pm 0.66}$

miniImageNet is not the source domain, the teacher becomes a mismatched teacher). We sample 5-Way 5-Shot tasks from the novel classes of each domain, and we measure the TSD and the accuracy of the teacher. As Table XIII, when we evaluate the teacher network on a mismatched source domain, we find that the accuracy is far beyond the random prediction. Therefore, the mismatched teacher is at least meaningful since it is better than randomly guessing. On the other hand, it has apparently lower TSD compared to that of miniImageNet in meta-testing. In this DG-FSC scenario, the attention of the mismatched teacher has transited to predicting the inter-class similarity, and the student is trained to adapt unseen domain by adapting to the "unseen" (mismatched) teacher. At the same time, the student model can be optimized by cross-entropy loss to the ground truth, which guarantees its correct updating directions.

In literature, to improve the model generalizability of unseen samples for classification tasks, several regulators such as Label Smoothing [54] or Confidence Penalty [43] have been proposed to penalize the overconfidence prediction of the classifier, such that the overfitting for the training data is mitigated. However, we note that these regulators have a common drawback: they encourage the probabilities to be uniformly distributed over all training classes, regardless if these classes are really similar to each other. In contrast to this, in our proposed method, the student in $\mathcal{L}_{MR}$ and $\mathcal{L}_{MM}$ is regularized to match a soft and better confidence prediction, which is designed specifically for overfitting and domain-shift for DG-FSC, and they achieve considerable improvement.

H. Training Student with a Stronger Teacher

In Sec. V, to find a good balance between the performance and the training cost, the proposed FS-BAN does not involve sequential training in generations, and we only train one generation of the student. Therefore, the architecture and size of the student are not limited to being the same as the teacher.

Ideally, one possible way to further improve the performance of the student network is to introduce a stronger teacher network, *i.e.*, more parameters with higher capacity. In Table XIV, we conduct a study to empirically validate this assumption: We set the scale of the teacher backbone equal (*born-again networks setup*) or larger (*common knowledge distillation setup*) than that of the student. We consider different types of backbone networks that are popular in FSC [55], [57], [10], [49], [53] as the feature encoder: Conv-4/6 (4/6-layer convolutional networks), and ResNet-10/18 [20]. We use the same setting as experiment setup 1): the student is trained on multiple seen source domains and tested on the leave-one-out selected target domain with 5-Way 5-Shot tasks. We use MatchingNet [60] as the baseline model.

As can be observed in Table XIV, when the backbone networks of the teacher and the student are the same, our proposed FS-BAN improves the performance of the student by a considerable margin. On the other hand, if we choose a teacher network with a larger backbone, the performance of the student network can be further improved.

I. Comparison to BAN with Transfer Learning

In this section, we compare our proposed method with the simple baseline [55] that leverages BAN with transfer learning approach for FSC, using the experiment setup 1), where we have multiple source domains: for each seen source domain, we follow [55] to initialize a linear layer as the classifier head and they share the feature encoder (ResNet10 [20]). In each training epoch, we randomly select the source domain and the corresponding classifier head, and the model is optimized by minimizing a standard cross-entropy loss as Eqn. 1. In DG-FSC evaluation, we follow [55] to transfer the obtained feature encoder on novel tasks and fit a new linear classifier for the prediction of query samples. For a fair comparison, we apply the same backbone and data augmentation skills as our method. The results are in Table XV. We show that our proposed method can achieve competitive performance on different DG-FSC setups. Moreover, we note that we follow [55] to conduct BAN training for two generations, therefore their training cost is higher than that of our method.

VII. Discussion

Conclusion. In this work, we first propose Born-Again Network (BAN) episodic training for domain generalization few-shot classification (DG-FSC) and reveal that BAN leads to more discriminative features and generates better decision boundaries on novel tasks from unseen domains. This suggests that similar to the observation in conventional supervised learning, BAN is also promising for DG-FSC tasks. To the best of our knowledge, this is the first study of BAN for episodic training. Motivated by this, we propose Few-Shot BAN (FS-BAN) as our main contribution. FS-BAN consists of multi-task learning objectives: Mutual Regularization, Mismatched Teacher, and Meta-Control of the Temperature. They aim to address the unique challenges posted specifically in DG-FSC: overfitting and domain shift. The effectiveness of FS-BAN is demonstrated by competitive accuracy on six benchmark datasets, three baseline FSC models, and qualitative and quantitative ablation studies.

Limitation. We follow exactly previous work (*e.g.*, [57]) in the choice of domains and datasets for a fair comparison. However, given the extremely wide range of domains to which DG-FSC can be applied, it is not feasible for us to validate our findings for all possible domains. On the other hand, our comprehensive qualitative and quantitative experiment results supported by our analysis provide supportive evidence that our method could be generalized to other domains. Meanwhile, FS-BAN does not impact the inference stage since we do not modify the model structure. Therefore, the effectiveness of FS-BAN on other domains in the open world can be easily validated with existing FSC models.

Future Work. While the performance of the state-of-the-art FSC algorithms has been largely improved within the single domain and unseen domains that include diverse classes, the accuracy on fine-grained domains remains poor, and an example can be observed in the results of the Cars domain in Table III. Future work will consider the different types of unseen domains, including this fine-grained setup that is challenging for all current FSC models.

References

[1] Milad Abdollahzadeh, Touba Malekzadeh, and Ngai-Man Man Cheung. Revisit multimodal meta-learning through the lens of multi-task learning. *Advances in Neural Information Processing Systems*, 34:14632–14644, 2021.

[2] Zeyuan Allen-Zhu and Yuanzhi Li. Towards understanding ensemble, knowledge distillation and self-distillation in deep learning. *arXiv preprint arXiv:2012.09816*, 2020.

[3] Leo Breiman and Nong Shang. Born again trees, 1996. https://www.stat.berkeley.edu/ breiman/BAtrees.pdf.

[4] Keshigeyan Chandrasegaran, Ngoc-Trung Tran, Yunqing Zhao, and Ngai-Man Cheung. Revisiting label smoothing and knowledge distillation compatibility: What was missing? In *International Conference on Machine Learning*, pages 2890–2916. PMLR, 2022.

[5] Pratik Chaudhari, Anna Choromanska, Stefano Soatto, Yann LeCun, Carlo Baldassi, Christian Borgs, Jennifer Chayes, Levent Sagun, and Riccardo Zecchina. Entropy-sgd: Biasing gradient descent into wide valleys. *Journal of Statistical Mechanics: Theory and Experiment*, 2019(12):124018, 2019.

[6] Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Frank Wang, and Jia-Binn Huang. A closer look at few-shot classification. In *ICLR*, 2019.

[7] Kevin Clark, Minh-Thang Luong, Urvashi Khandelwal, Christopher D. Manning, and Quoc V. Le. BAM! born-again multi-task networks for natural language understanding. In *ACL*, pages 5931–5937, July 2019.

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009.

[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 network. In *ICML*, 2017.

[11] Bent Fuglede and Flemming Topsoe. Jensen-shannon divergence and hilbert space embedding. In *International Symposium onInformation Theory, 2004. ISIT 2004. Proceedings.*, page 31. IEEE, 2004.

[12] Tommaso Furlanello, Zachary Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. Born again neural networks. In *International Conference on Machine Learning*, pages 1607–1616. PMLR, 2018.

[13] Victor Garcia and Joan Bruna. Few-shot learning with graph neural networks. In *ICLR*, 2018.

[14] Golnaz Ghiasi, Tsung-Yi Lin, and Quoc V Le. Dropblock: A regularization method for convolutional networks. In *NeurIPS*, pages 10727–10737, 2018.

[15] Spyros Gidaris and Nikos Komodakis. Dynamic few-shot visual learning without forgetting. In *CVPR*, pages 4367–4375, 2018.

[16] Spyros Gidaris and Nikos Komodakis. Generating classification weights with gnn denoising autoencoders for few-shot learning. In *Proc. CVPR*, pages 21–30, 2019.

[17] Micah Goldblum, Steven Reich, Liam Fowl, Renkun Ni, Valeriia Cherepanova, and Tom Goldstein. Unraveling meta-learning: Understanding feature representations for few-shot tasks. In *ICML*, pages 8857–8866, 2020.

[18] Jia Gong, Zhipeng Fan, Qiuhong Ke, Hossein Rahmani, and Jun Liu. Meta agent teaming active learning for pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11079–11089, 2022.

[19] Jia Gong, Foo Lin Geng, Zhipeng Fan, and Jun Liu. Diffpose: Toward more reliable 3d pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[21] Nathan Hilliard, Lawrence Phillips, Scott Howland, Artëm Yankov, Courtney D Corley, and Nathan O Hodas. Few-shot learning with metric-agnostic conditional embeddings. *arXiv preprint arXiv:1802.04376*, 2018.

[22] Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. In *NeurIPS Deep Learning and Representation Learning Workshop*, 2015.

[23] Ruibing Hou, Hong Chang, MA Bingpeng, Shiguang Shan, and Xilin Chen. Cross attention network for few-shot classification. In *NeurIPS*, pages 4003–4014, 2019.

[24] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[25] Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836*, 2016.

[26] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

[27] J. Krause, M. Stark, J. Deng, and L. Fei-Fei. 3d object representations for fine-grained categorization. In *2013 IEEE International Conference on Computer Vision Workshops*, pages 554–561, 2013.

[28] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[29] Kwonjoon Lee, Subhransu Maji, Avinash Ravichandran, and Stefano Soatto. Meta-learning with differentiable convex optimization. In *CVPR*, pages 10657–10665, 2019.

[30] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *Proc. ICCV*, pages 5542–5550, 2017.

[31] Da Li, Jianshu Zhang, Yongxin Yang, Cong Liu, Yi-Zhe Song, and Timothy M Hospedales. Episodic training for domain generalization. In *Proc. CVPR*, pages 1446–1455, 2019.

[32] Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. Domain generalization with adversarial feature learning. In *Proc. CVPR*, pages 5400–5409, 2018.

[33] Honglin Li, Chenglu Zhu, Yunlong Zhang, Yuxuan Sun, Zhongyi Shui, Wenwei Kuang, Sunyi Zheng, and Lin Yang. Task-specific fine-tuning

via variational information bottleneck for weakly-supervised pathology whole slide image classification. *arXiv preprint arXiv:2303.08446*, 2023.

[34] Shuang Li, Mixue Xie, Kaixiong Gong, Chi Harold Liu, Yulin Wang, and Wei Li. Transferable semantic augmentation for domain adaptation. In *Proc. CVPR*, pages 11516–11525, 2021.

[35] Ya Li, Xinmei Tian, Mingming Gong, Yajing Liu, Tongliang Liu, Kun Zhang, and Dacheng Tao. Deep domain generalization via conditional invariant adversarial networks. In *ECCV*, pages 624–639, 2018.

[36] Yanghao Li, Naiyan Wang, Jianping Shi, Jiaying Liu, and Xiaodi Hou. Revisiting batch normalization for practical domain adaptation. In *ICLR Workshop*. OpenReview.net, 2017.

[37] Yann Lifchitz, Yannis Avrithis, Sylvaine Picard, and Andrei Bursuc. Dense classification and implanting for few-shot learning. In *CVPR*, pages 9258–9267, 2019.

[38] Foo Lin Geng, Jia Gong, Zhipeng Fan, and Jun Liu. System-status-aware adaptive network for online streaming video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[39] 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*, 2020.

[40] Sebastian Mika, Gunnar Ratsch, Jason Weston, Bernhard Scholkopf, and Klaus-Robert Mullers. Fisher discriminant analysis with kernels. In *Neural networks for signal processing IX: Proceedings of the 1999 IEEE signal processing society workshop (cat. no. 98th8468)*, pages 41–48. Ieee, 1999.

[41] Boris Oreshkin, Pau Rodríguez López, and Alexandre Lacoste. Tadam: Task dependent adaptive metric for improved few-shot learning. In *NeurIPS*, pages 721–731, 2018.

[42] Prashant Pandey, Mrigank Raman, Sumanth Varambally, and Prathosh AP. Domain generalization via inference-time label-preserving target projections. *arXiv preprint arXiv:2103.01134*, 2021.

[43] Gabriel Pereyra, George Tucker, Jan Chorowski, Łukasz Kaiser, and Geoffrey Hinton. Regularizing neural networks by penalizing confident output distributions. *arXiv preprint arXiv:1701.06548*, 2017.

[44] Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron C. Courville. Film: Visual reasoning with a general conditioning layer. In *AAAI*, 2018.

[45] Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017.

[46] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proc. CVPR*, pages 779–788, 2016.

[47] 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. In *Proc. ICLR*, 2018.

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

[49] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *NeurIPS*, pages 4077–4087, 2017.

[50] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *Proc. ICCV*, pages 843–852, 2017.

[51] Jiamei Sun, Sebastian Lapuschkin, Wojciech Samek, Yunqing Zhao, Ngai-Man Cheung, and Alexander Binder. Explanation-guided training for cross-domain few-shot classification. *ICPR*, 2020.

[52] Qianru Sun, Yaoyao Liu, Tat-Seng Chua, and Bernt Schiele. Meta-transfer learning for few-shot learning. In *CVPR*, pages 403–412, 2019.

[53] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. Learning to compare: Relation network for few-shot learning. In *CVPR*, pages 1199–1208, 2018.

[54] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proc. CVPR*, pages 2818–2826, 2016.

[55] 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*, 2020.

[56] Eleni Triantafillou, Tyler Zhu, Vincent Dumoulin, Pascal Lamblin, Utku Evci, Kelvin Xu, Ross Goroshin, Carles Gelada, Kevin Jordan Swersky, Pierre-Antoine Manzagol, and Hugo Larochelle. Meta-dataset: A dataset of datasets for learning to learn from few examples. In *ICLR*, 2020.

[57] Hung-Yu Tseng, Hsin-Ying Lee, Jia-Bin Huang, and Ming-Hsuan Yang. Cross-domain few-shot classification via learned feature-wise transformation. In *ICLR*, 2020.

[58] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *CVPR*, 2018.

[59] Thibaut Vidal and Maximilian Schiffer. Born-again tree ensembles. In *Proc. ICML*, pages 9743–9753. PMLR, 2020.

[60] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. In *NeurIPS*, pages 3630–3638, 2016.

[61] 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.

[62] Ji Wang, Weidong Bao, Lichao Sun, Xiaomin Zhu, Bokai Cao, and S Yu Philip. Private model compression via knowledge distillation. In *Proc. AAAI*, volume 33, pages 1190–1197, 2019.

[63] Chenglin Yang, Lingxi Xie, Siyuan Qiao, and Alan L Yuille. Training deep neural networks in generations: A more tolerant teacher educates better students. In *Proc. AAAI*, pages 5628–5635, 2019.

[64] Ngai-Man Cheung Yiluan Guo. Attentive weights generation for few shot learning via information maximization. In *CVPR*, 2020.

[65] Hongxu Yin, Pavlo Molchanov, Jose M Alvarez, Zhizhong Li, Arun Mallya, Derek Hoiem, Niraj K Jha, and Jan Kautz. Dreaming to distill: Data-free knowledge transfer via deepinversion. In *CVPR*, pages 8715–8724, 2020.

[66] Xiangyu Yue, Zangwei Zheng, Shanghang Zhang, Yang Gao, Trevor Darrell, Kurt Keutzer, and Alberto Sangiovanni Vincentelli. Prototypical cross-domain self-supervised learning for few-shot unsupervised domain adaptation. In *Proc. CVPR*, pages 13834–13844, 2021.

[67] 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*, pages 12203–12213, 2020.

[68] An Zhao, Mingyu Ding, Zhiwu Lu, Tao Xiang, Yulei Niu, Jiechao Guan, and Ji-Rong Wen. Domain-adaptive few-shot learning. In *WACV*, pages 1390–1399, 2021.

[69] Yunqing Zhao, Keshigeyan Chandrasegaran, Milad Abdollahzadeh, and Ngai-man Cheung. Few-shot image generation via adaptation-aware kernel modulation. In *Advances in Neural Information Processing Systems*, 2022.

[70] Yunqing Zhao, Henghui Ding, Houjing Huang, and Ngai-Man Cheung. A closer look at few-shot image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022.

[71] Yunqing Zhao, Chao Du, Milad Abdollahzadeh, Tianyu Pang, Min Lin, Shuicheng YAN, and Ngai-Man Cheung. Exploring incompatible knowledge transfer in few-shot image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023.

[72] Yunqing Zhao, Tianyu Pang, Chao Du, Xiao Yang, Ngai-Man Cheung, and Min Lin. A recipe for watermarking diffusion models. *arXiv preprint arXiv: 2303.10137*, 2023.

[73] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017.

ACKNOWLEDGMENT

This work was supported in part by the National Research Foundation, Singapore, under its AI Singapore Programmes (AISG) under Award AISG2-RP-2021-021 and Award AISG2-TC-2022-007; and in part by the Singapore University of Technology and Design under Project PIE-SGP-AI-2018-01. This project was also based on the research/work support in part by the Changi General Hospital and Singapore University of Technology and Design, under the HealthTech Innovation Fund (HTIF Award No. CGH-SUTD-2021-004). The authors would like to thank anonymous reviewers for their helpful comments to improve the paper and also would like to thank Yiluan Guo, Jiamei Sun and Milad Abdollahzadeh for their valuable comments, feedback and discussion.