

ME-D2N: Multi-Expert Domain Decompositional Network for Cross-Domain Few-Shot Learning

Yuqian Fu*
fuyq20@fudan.edu.cn
Shanghai Key Lab of Intelligent
Information Processing, School of
Computer Science, Fudan University

Yu Xie*
Yanwei Fu
{yxie18, yanweifu}@fudan.edu.cn
School of Data Science, Fudan
University

Jingjing Chen
Yu-Gang Jiang#
{chenjingjing, ygj}@fudan.edu.cn
Shanghai Key Lab of Intelligent
Information Processing, School of
Computer Science, Fudan University

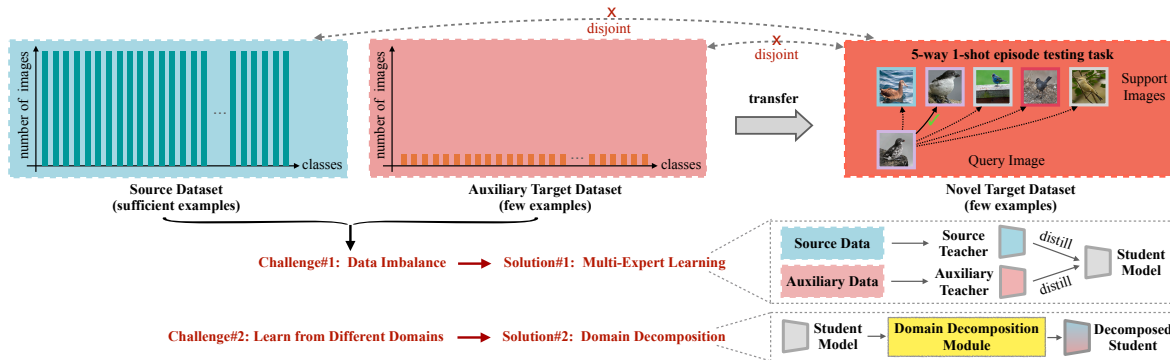


Figure 1: Illustration of our motivation and solutions. We observe two core challenges: 1) serious data imbalance problem between the two training datasets; 2) network is required to learn from different domains simultaneously. Correspondingly, we have the following key solutions: 1) proposing the multi-expert learning that first learns two individual teacher models and then transfers the knowledge to a student model via knowledge distillation; 2) presenting a novel domain decomposition module that learns to decompose the network structure of student model into two domain-related sub-parts.

ABSTRACT

Recently, Cross-Domain Few-Shot Learning (CD-FSL) which aims at addressing the Few-Shot Learning (FSL) problem across different domains has attracted rising attention. The core challenge of CD-FSL lies in the domain gap between the source and novel target datasets. Though many attempts have been made for CD-FSL without any target data during model training, the huge domain gap makes it still hard for existing CD-FSL methods to achieve very satisfactory results. Alternatively, learning CD-FSL models with few labeled target domain data which is more realistic and promising is advocated in previous work [13]. Thus, in this paper, we stick to this setting and technically contribute a novel Multi-Expert Domain Decompositional Network (ME-D2N). Concretely, to solve the data imbalance problem between the source data with sufficient examples and the auxiliary target data with limited examples, we build our model under the umbrella of multi-expert

learning. Two teacher models which can be considered to be experts in their corresponding domain are first trained on the source and the auxiliary target sets, respectively. Then, the knowledge distillation technique is introduced to transfer the knowledge from two teachers to a unified student model. Taking a step further, to help our student model learn knowledge from different domain teachers simultaneously, we further present a novel domain decomposition module that learns to decompose the student model into two domain-related sub-parts. This is achieved by a novel domain-specific gate that learns to assign each filter to only one specific domain in a learnable way. Extensive experiments demonstrate the effectiveness of our method. Codes and models are available at https://github.com/lovelyqian/ME-D2N_for_CDFSL.

CCS CONCEPTS

• **Computing methodologies** → **Computer vision**; *Learning latent representations*.

KEYWORDS

cross-domain few-shot learning, classification for unbalanced data, multi-expert learning, network decomposition.

ACM Reference Format:

Yuqian Fu*, Yu Xie*, Yanwei Fu, Jingjing Chen, and Yu-Gang Jiang#. 2022. ME-D2N: Multi-Expert Domain Decompositional Network for Cross-Domain Few-Shot Learning. In *Proceedings of the 30th ACM International Conference*

* indicates equal contributions, # indicates corresponding author.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

MM '22, October 10–14, 2022, Lisboa, Portugal

© 2022 Association for Computing Machinery.

ACM ISBN 978-1-4503-9203-7/22/10...\$15.00

<https://doi.org/10.1145/3503161.3547995>

on *Multimedia (MM '22)*, October 10–14, 2022, Lisboa, Portugal. ACM, New York, NY, USA, 9 pages. <https://doi.org/10.1145/3503161.3547995>

1 INTRODUCTION

FSL mainly aims at transferring knowledge from a source dataset to a novel target dataset with only one or few labeled examples. Generally, FSL assumes that the images of the source and target datasets belong to the same domain. However, such an ideal assumption may not be easy to be met in real-world multimedia applications. For example, as revealed in [9], a model trained on the Imagenet [11] which is mainly composed of massive and diverse natural images still fails to recognize the novel fine-grained birds. To this end, CD-FSL which is dedicated to addressing the domain gap problem of FSL has invoked rising attention.

Recently, various settings of CD-FSL have been extensively studied in many previous methods [13, 14, 33, 40, 44]. Most of them [14, 40, 44] use only the source domain images for training and pay efforts on improving the generalization ability of the FSL models. Though some achievements have been made, it is still hard to achieve very impressive performance due to the huge domain gap between the source and target datasets. Thus, some works [13, 33] relax the most basic yet strict setting, and allow target data to be used during the training phase. More specifically, STARTUP [33] proposes to make use of relative massive unlabeled target data, whilst Meta-FDMixup [13] advocates utilizing few limited labeled target data. Unfortunately, the massive unlabeled examples in the former one may still be not easy to be obtained in many real-world applications, such as the recognition of endangered wild animals and specific buildings. By contrast, learning CD-FSL with few limited labeled target domain data, *e.g.*, 5 images per class, is more realistic. Thus, in this paper, we stick to the setting proposed in Meta-FDMixup [13] to promote the learning process of models.

Formally, given a source domain dataset with enough examples and an auxiliary target domain dataset with only a few labeled examples, our goal is to learn a good FSL model taking these two sets as training data and achieve good results on the novel target data. Notably, as in Meta-FDMixup, our setting doesn't violate the basic FSL setting, as the class sets of the auxiliary target training data and the novel target testing data are disjoint from each other strictly. This ensures that none of the novel target categories will appear during the training stage. Critically, as shown in Figure 1, we highlight that there are two key challenges: 1) The number of labeled examples for the source dataset and auxiliary target dataset are extremely unbalanced. Models learned on such unbalanced training data will be biased towards the source dataset while performing much worse on the target dataset. 2) Since the source dataset and the auxiliary target belong to two distinct domains, it may be too difficult for a single model to learn knowledge from datasets with different domains simultaneously. Such challenges unfortunately have less been touched in previous works [13].

To address these challenges, this paper presents a novel Multi-Expert Domain Decompositional Network (**ME-D2N**) for CD-FSL. Our key solutions are also illustrated in Figure 1. Specifically, taking unbalanced datasets as training data will lead to the model biased problem [2, 49]. That is, the learned model tends to perform well on the classes with more examples but has a performance degradation

on the categories with fewer examples. To tackle the data imbalance issue, we propose to build our model upon the multi-expert learning paradigm. Concretely, rather than learning a model on the merged data of source and auxiliary target datasets directly, we train two teacher models on the source and the auxiliary dataset, respectively. Models trained in this way can be considered experts in their specialized domain avoiding being affected by training data of another domain. Then, we transfer the knowledge from these two teachers to our student model. This is done by using the knowledge distillation technique which constrains the student model to produce consistent predictions with the teachers. By distilling the *individual knowledge* from both source and target teacher models, our student model picks up the ability to recognize both the source and auxiliary target images, avoiding learning from the *unbalanced datasets*. We take one step further: considering that forcing a unified model to learn from teachers of different domains may be nontrivial. Concretely, since each filter in the network needs to be responsible for extracting all domain features simultaneously, this vanilla learning method may limit the performance of the network. A natural question is whether it is possible to decompose the student model into two parts – one for learning from the source teacher and the another for the auxiliary target teacher? Based on the above insights, a novel domain decomposition module which is also termed as D2N is proposed. Specifically, our D2N aims at building a one-to-one correspondence between the network filters and the domains. That is, each filter is only assigned to be activated by one specific domain. Technically, we achieve this by proposing a novel domain-specific gate that learns the activation state of filters for a specific domain dynamically. We insert the D2N into the feature extractor of the student model and make it learnable together with the model parameters.

We conduct extensive experiments on four different target datasets. Results well indicate that our multi-expert learning strategy helps address the data imbalance problem. Besides, our D2N further improves the performance of the student model showing the advantages of decomposing the student model into two domains.

Contributions. We summarize our contributions as below: 1) For the first time, we introduce the multi-expert learning paradigm into the task of CD-FSL with few labeled target data to prevent the model from learning on unbalanced datasets directly. By learning from two teachers, we avoid our model being biased towards the source dataset with significantly more samples. 2) A novel domain decomposition module (D2N) is proposed to learn to decompose the model's filters into the source and target domain-specific parts. The concept of domain decomposition has less been explored in previous work, especially for the task of CD-FSL. 3) Extensive experiments conducted show the effectiveness of our modules and our proposed full model ME-D2N builds a new state of the art.

2 RELATED WORK

Cross-Domain Few-Shot Learning. Recent study [9] finds that most of the existing FSL methods [12, 15, 17, 28, 37, 37, 39, 41–43, 46, 52–54] that assume the source and target datasets belong to the same distribution fail to generalize to novel datasets with a domain gap. Thus, CD-FSL which aims at addressing FSL across different domains has risen increasing attentions [4, 13, 14, 18, 24,

33, 40, 44, 48]. In this paper, these CD-FSL methods are categorized according to which kind of data are being used for training: 1) CD-FSL with only source data [14, 40, 44, 48]; 2) CD-FSL with unlabeled target data [24, 33]; 3) CD-FSL with labeled target data [13]. Typically, CD-FSL with only source data is the most strict setting that demands model to recognize totally unseen target dataset without any target information. Flagship works including FWT [44], BSCD-FSL [18], LRP [40], ATA [48], wave-SAN [14], RDC [29], NSAE [32], and ConFT [10]. Though many well-designed techniques e.g. readjusting the batch normalization [44], augmenting the difficult of meta tasks [48], spanning style distributions [14], and even fine-tuning models using few target images during the testing stage [10, 18, 29, 32], the performances of them are still greatly limited due to the huge domain gap. By contrast, STARTUP [33] relaxes this strict setting and uses unlabeled target data for training. Another choice that performs CD-FSL with few labeled target data is advocated by Meta-FDMixup [13], since obtaining extremely few labeled target data per class is relatively more realistic and can boost the model performance to a large extent. Thus, in this paper, we mainly stick to the setting of CD-FSL with few labeled target data.

Long-Tailed Recognition. This paper is also related to the long-tailed recognition from the perspective of tackling data imbalance problem. In the literature, many attempts have been made to address the task of long-tailed recognition [16, 59]. Main-stream methods include: 1) re-sampling based methods [2, 3, 19] which under-sample the head classes or over-sample the tail classes; 2) re-weighting based methods [5, 21, 22, 50] which assign different weights for classes or instances; 3) two-stage fine-tuning based methods [7, 26, 57] which train the representation and classifier separately. Another line of work [6, 49, 51] trains multiple experts for the head, medium, and tail classes respectively which is the most related work to us. However, to the best of our knowledge, it is the first time that multi-expert learning is used for the CD-FSL task. Our method trains two teacher models of different domains and transfers the knowledge into the student model. In addition, besides the unbalanced training sets, what intrinsically distinguishes our work from these methods is that we tackle datasets of different domains. The idea of decomposing the student model into different domains has never been explored in these works.

Decomposition of Network Filters. Generally, as studied in [1, 55], the filters of a normal CNN tend to extract mixed features of the input data. Such entangled filters inevitably lead to some unexpected problems, including limiting the representational capability of the network and increasing the uninterpretability. Subsequently, some methods [23, 30, 34, 36] explore decomposing the filters for making a more efficient and compressed network. Other works including [8, 31, 38, 55] decompose the network filters for more interpretable networks via assigning filters dynamically. Generally, we are similar but fundamentally different from methods of this type. They learn the correspondence between filters and the “classes” or “objects” to explain the activation of the model, while our motivation is to decompose the student model into two “domains” so that we can learn from teachers of different domains. Another work that may also be related to us is domain-aware dynamic network [56] which learns different weights for different

domains. However, using soft weights for readjusting the activation of the network essentially can not be seen as a decomposition.

3 METHOD

Problem Definition. For the CD-FSL with few labeled target data, we have two training sets: source training dataset $D_{src} = \{x_{src}, y_{src}\}$ and the auxiliary target dataset $D_{tgt} = \{x_{tgt}, y_{tgt}\}$. The model trained on D_{src} and D_{tgt} is evaluated on the novel target testing dataset $D_{test} = \{x_{test}, y_{test}\}$. The x represents the image examples and the y denotes the corresponding labels of images. Note that all the classes contained in D_{src} , D_{tgt} , and D_{test} are **dis-joint** from each other and there is a domain gap between the source dataset D_{src} and the target datasets D_{tgt} , D_{test} .

We construct meta-learning tasks which also known as the N -way K -shot episodes to train and test our model. Typically, an episode contains a support set $S = \{x_i, y_i\}_i^{N \times K}$ and a query set $Q = \{x_i, y_i\}_i^{N \times M}$. N -way K -shot means that N categories are sampled. Each category of S and Q contains K labeled examples and M testing images, respectively. The images in Q are classified according to the given S . We use the $\{S_{src}, Q_{src}\}$, $\{S_{tgt}, Q_{tgt}\}$, and $\{S_{test}, Q_{test}\}$ to denote episodes sampled from D_{src} , D_{tgt} , and D_{test} , respectively.

Method Overview. The overall illustration of our method is given in Figure 2. We mainly have two training stages: a) optimizing the source and target domain teacher networks (**St-Net** & **Tt-Net**) separately; b) optimizing the multi-expert domain decompositional student network (**ME-D2N**) by distilling the knowledge from St-Net and Tt-Net. The St-Net and Tt-Net are composed of an embedding module E and an FSL classifier G . Besides these two basic modules E, G , ME-D2N also contains a novel domain decomposition module $D2N$ and two global classifiers f_{src}, f_{tgt} which classify the input images into the global source and target categories. Note that f_{src}, f_{tgt} are only used during the training phase. As for the object functions, the St-Net and Tt-Net are optimized using the FSL classification loss \mathcal{L}_{fsl} alone. The ME-D2N is optimized by the \mathcal{L}_{fsl} , the knowledge distillation loss \mathcal{L}_{kd} , and the global classification loss \mathcal{L}_{cls} simultaneously. Note that we use the “STD path” and the “DSG path” in the figure to denote the standard forward path and the domain-specific gate forward path which is guided by the D2N, respectively. During testing, the ME-D2N is utilized to obtain the predictions for the novel target episodes.

3.1 Learning the Teacher Networks

As shown in Figure 2.a, we first train our two teacher networks (St-Net & Tt-Net) using episodes sampled from the source training dataset D_{src} and the auxiliary target dataset D_{tgt} , respectively. These trained teachers are considered experts in the corresponding domain to guide the subsequent training of the ME-D2N student network. Both the network structure and training process of St-Net and Tt-Net are exactly the same. Here we take the St-Net as an example to introduce the learning details of the teacher network. For each training iteration, we randomly sample a source episode $\{S_{src}, Q_{src}\}$ from D_{src} as input, and feed it into the embedding module E of St-Net to obtain the feature representations of the S_{src} and Q_{src} . After that, the FSL classifier module G of St-Net is used to predict the class categories of Q_{src} according to the S_{src} resulting in the FSL prediction scores \bar{P}^{src} . Note that to prevent the model

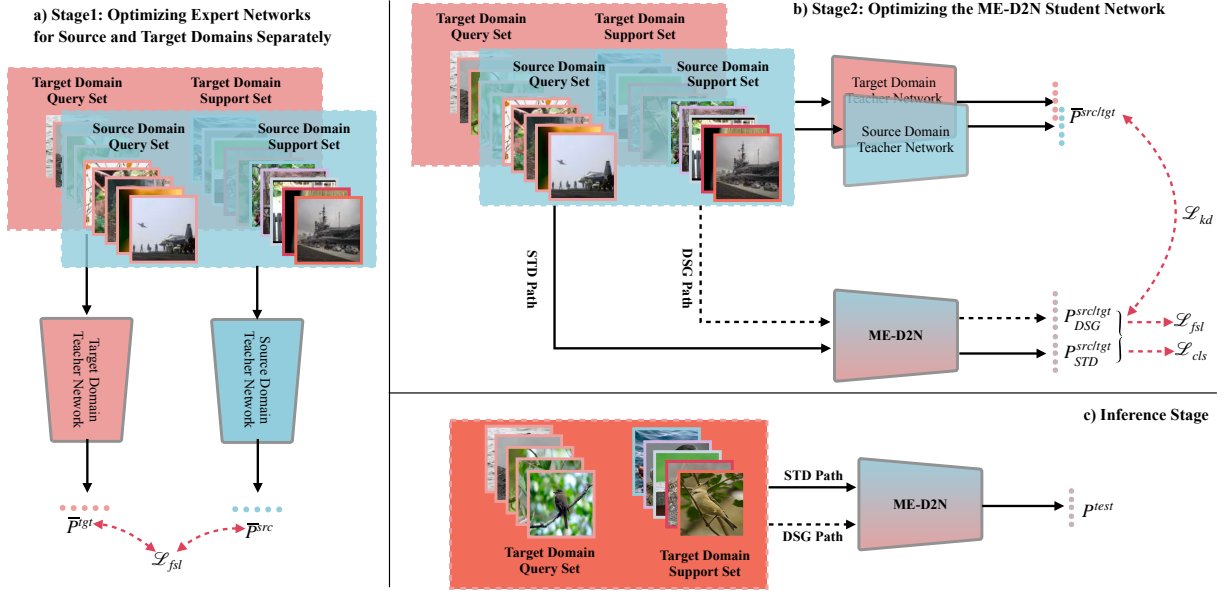


Figure 2: Our method contains two training stages: a) optimizing the experts network for the source and target domains; b) optimizing the ME-D2N student network by distilling knowledge from both source and target domain experts. In the inference stage, only the ME-D2N network is used for prediction.

from learning the correspondence between the inputs images to its global labels y^{src} , meta-learning refines the labels of these N classes as $y_{fsl}^{src} \in (0, 1, \dots, N - 1)$. By calculating the cross entropy loss between the predictions and its corresponding FSL ground truth y_{fsl}^{src} as follows, we obtain its FSL classification loss \mathcal{L}_{fsl}^{src} .

$$\mathcal{L}_{fsl}^{src} = CE(\bar{p}_{fsl}^{src}, y_{fsl}^{src}) \quad (1)$$

where CE indicates the cross entropy loss. In the same way, for the Tt-Net, we get the FSL prediction scores \bar{p}_{fsl}^{tgt} and the FSL classification loss \mathcal{L}_{fsl}^{tgt} . The \mathcal{L}_{fsl}^{src} and the \mathcal{L}_{fsl}^{tgt} are finally used for optimizing the Tt-Net and St-Net, respectively.

3.2 Domain Decomposition Module

Given two experts St-Net and Tt-Net, a direct and commonly used solution for training the student model is distilling knowledge from these two teachers at the same time. However, as we stated in Sec. 1, one key challenge of this task lies in the available training sets D_{src} and D_{tgt} belong to different domains. To that end, the learned teacher models will be biased towards their training domains. It may be difficult for a unified student model to learn knowledge from two teachers of different domains. Thus, our D2N learns to decompose the student model into the source-specific part and target-specific part. Overall, the decomposition is achieved by a novel domain-specific gate (DSG) that learns to assign each filter to only one specific domain dynamically.

As shown in Figure 3, we first randomly initialize the domain-specific gate matrix \tilde{M} . The number of elements \tilde{M} is consistent with the number of filters that need to be decomposed in the network. The element in \tilde{M} can be seen as the probability that the corresponding filter belongs to one specific domain. Typically, we use the \tilde{M} to denote the gate matrix for the source domain, and the

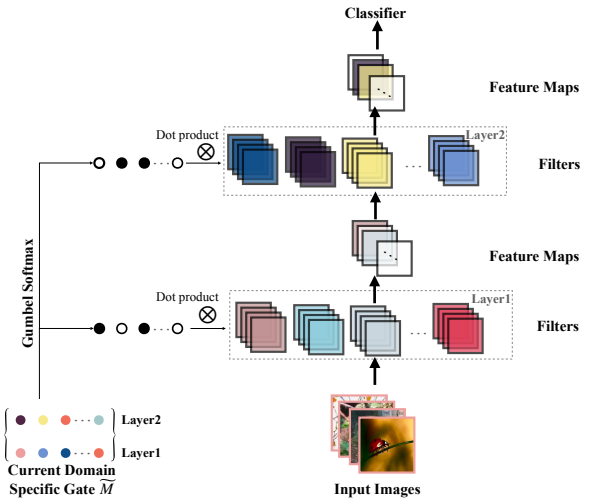


Figure 3: The illustration of our domain decomposition module (D2N). D2N learns a domain-specific gate matrix \tilde{M} to control the activation states of the filters. The Gumbel softmax is utilized to binarize the soft matrix.

gate matrix for the target domain can be easily obtained by $1 - \tilde{M}$. While the soft gate is not enough to achieve the real decomposition, which does not meet our ideal expectation of assigning a filter to only a domain. Thus, the Gumbel softmax [25] which generates the discrete data from a soft categorical distribution is introduced to transform the soft \tilde{M} into the hard one M (denoted as the black and white dots). With the M , the DSG forward path only activates the filters when the gates for them equal 1 thus establishing a one-to-one correspondence between filters and domains.

3.3 Learning the ME-D2N Student Network

As shown in Figure 2.b, to learn knowledge from both the source and target domain teachers, two episodes $\{S_{src}, Q_{src}\}, \{S_{tgt}, Q_{tgt}\}$ are randomly sampled as the input for each training iteration. These source and target episodes are fed into the ME-D2N through two different forward paths. One is the standard (STD) path and the other is the DSG path as introduced in Sec.3.2. Regardless of whether the forward is domain decomposed or not, all the other details e.g. input data and loss functions are the same for these two paths. Thus, for convenience, we first introduce the learning and optimization process of the DSG path as an example. For each input episode e.g. the source episode $\{S_{src}, Q_{src}\}$, we feed it into the embedding module and the FSL classifier module consequently obtaining its FSL predictions P_{DSG}^{src} . Based on the P_{DSG}^{src} , a total of three sub-tasks are performed resulting in three different losses. Firstly and most importantly, the knowledge distillation is performed to keep the ability of ME-D2N with that of the teacher model. Specifically, the same $\{S_{src}, Q_{src}\}$ is fed into the trained St-Net obtaining the FSL predictions \bar{P}^{src} . After that, the Kullback-Leibler divergence loss is used to constrain the consistency between the P_{DSG}^{src} and \bar{P}^{src} . Thus, the knowledge distillation loss $\mathcal{L}_{DSG_kd}^{src}$ is expressed as:

$$\mathcal{L}_{DSG_kd}^{src} = KL(P_{DSG}^{src}, \bar{P}^{src}) \quad (2)$$

where the KL means the Kullback-Leibler divergence loss. Secondly, by calculating the cross entropy loss between the predictions P_{DSG}^{src} and its FSL ground truth as defined in Equa. 1, we obtain its FSL classification loss $\mathcal{L}_{DSG_fsl}^{src}$. Thirdly, we also use the global classifier f_{src} to classify the input images into the its global class categories. This generates the global classification loss $\mathcal{L}_{DSG_cls}^{src}$. In the same way, we obtain the target knowledge distillation loss $\mathcal{L}_{DSG_kd}^{tgt}$, the target FSL classification loss $\mathcal{L}_{DSG_fsl}^{tgt}$, and the target global classification loss $\mathcal{L}_{DSG_cls}^{tgt}$. Formally, we have:

$$\mathcal{L}_{DSG_kd} = \lambda_1 \mathcal{L}_{DSG_kd}^{src} + (1 - \lambda_1) \mathcal{L}_{DSG_kd}^{tgt} \quad (3)$$

$$\mathcal{L}_{DSG_fsl} = \lambda_1 \mathcal{L}_{DSG_fsl}^{src} + (1 - \lambda_1) \mathcal{L}_{DSG_fsl}^{tgt} \quad (4)$$

$$\mathcal{L}_{DSG_cls} = \lambda_1 \mathcal{L}_{DSG_cls}^{src} + (1 - \lambda_1) \mathcal{L}_{DSG_cls}^{tgt} \quad (5)$$

$$\mathcal{L}_{DSG} = \mathcal{L}_{DSG_kd} + \lambda_2 \mathcal{L}_{DSG_fsl} + \lambda_3 \mathcal{L}_{DSG_cls} \quad (6)$$

Similarly, we obtain the loss for the STD path as \mathcal{L}_{STD} . The final loss for our ME-D2N is defined as follows:

$$\mathcal{L} = \mathcal{L}_{DSG} + \lambda_4 \mathcal{L}_{STD} \quad (7)$$

The $\lambda_1, \lambda_2, \lambda_3$, and λ_4 are four hyper-parameters.

Inference Stage. During the inference stage, as shown in Figure 2.c, only the ME-D2N is used for generating the predictions. Given a novel target testing episode $\{S_{test}, Q_{test}\}$, we forward it into the ME-D2N using both the DSG and STD paths resulting in two predictions P_{DSG}^{test} and P_{STD}^{test} . Note that since the Gumbel softmax will cause inconsistencies in each forward process, unlike the training process, we directly binarize the gate matrix \tilde{M} by choosing the domain with a higher value. The final prediction P^{test} takes the mean of these two paths:

$$P^{test} = \frac{1}{2}(P_{DSG}^{test} + P_{STD}^{test}) \quad (8)$$

4 EXPERIMENTS

4.1 Setup

Datasets. Totally five datasets are used to validate the effectiveness of our method. Concretely, mini-Imagenet [35] works as the source dataset. CUB [47], Cars [27], Places [58], and Plantae [45] serve as the target datasets, respectively. As for the splits of D_{src} , D_{tgt} , and D_{test} , we strictly follow the meta-FDMixup [13]. Specifically, each class category in D_{tgt} has only 5 labeled examples.

Network Modules. Resnet-10 [20] is used as the embedding module E . GNN [17] is selected as the FSL classifier G . Totally same basic modules E and G with previous CD-FSL methods ensures the fair comparisons. Besides, same to wave-SAN [14], we divide the embedding module E into four blocks and decompose the filters of the last two blocks. The global classifiers f_{src} and f_{tgt} are represented by a fully connected layer, respectively.

Implementation Details. We conduct our experiments in the form of 5-way 1-shot and 5-way 5-shot meta tasks. For training, we train the Tt-Net 100 and 400 epochs for 5-way 1-shot and 5-way 5-shot, respectively. The training of both the St-Net and ME-D2N takes 400 epochs. The Adam with the initial learning rate of 0.001 is uniformly used as the optimizer. Besides, as a common practice in CD-FSL [13, 44], the E pre-trained on the source dataset D_{src} with the standard classification tasks is used to warm start the training of St-Net, Tt-Net, and ME-D2N. As for the testing stage, the average accuracy of 1000 episodes randomly sampled from the novel target set D_{test} is reported. Without searching the best hyper-parameters for different target dataset, we uniformly set the $\lambda_1, \lambda_2, \lambda_3$, and λ_4 as 0.2, 0.05, 0.05, and 0.2, respectively.

4.2 Baselines & Competitors

We introduce our baselines and competitors as below. **1) Typical FSL Methods.** Totally three flagship methods in the FSL community including MatchingNet [46], RelationNet [41], and GNN [17] are introduced. These methods demonstrate how the typical FSL methods perform under the CD-FSL setting. **2) Several Baselines.** Three baselines namely “St-Net”, “Tt-Net”, and “M-base” are included for comparisons. Concretely, as stated in Sec. 3, “St-Net” and “Tt-Net” represent our two teacher models. “M-base” is obtained by training model on the merged data of source and auxiliary datasets. The network architecture of “M-base” is exactly the same as that of “St-Net” and “Tt-Net”. These three baselines play an important role in analyzing the effectiveness of our method. **3) CD-FSL Methods.** Since the setting of CD-FSL with few labeled target data is proposed very recently, meta-FDMixup [13] is the only competitor that can be compared directly. To better illustrate the competitiveness of our method, we further adapt several other CD-FSL methods including FWT [44] which improves the generalization ability of the model by feature-wise transformation, LRP [40] which utilizes the results of the explanation model to guide the learning process, ATA [48] which augments the meta tasks via adversarial attack, and wave-SAN [14] which tackles the domain gap from the perspective of augmenting the style distributions of the source dataset. These methods are initially designed for the most strict CD-FSL setting. The adaption is achieved by training the models using the merged training data as used for “M-base”. To that end, we obtain the competitors “M-FWT”, “M-LRP”, “M-ATA”, and “M-waveSAN”.

5-way	1-shot	D_{tgt}	CUB	Cars	Places	Plantae	Avg.
FSL	MatchingNet [46]	-	35.89 \pm 0.51	30.77 \pm 0.47	49.86 \pm 0.79	32.70 \pm 0.60	37.31
	RelationNet [41]	-	42.44 \pm 0.77	29.11 \pm 0.60	48.64 \pm 0.85	33.17 \pm 0.64	38.34
	GNN [17]	-	45.69 \pm 0.68	31.79 \pm 0.51	53.10 \pm 0.80	35.60 \pm 0.56	41.54
Baselines	St-Net	-	46.10 \pm 0.68	31.05 \pm 0.54	54.22 \pm 0.81	37.11 \pm 0.60	42.12
	Tt-Net	✓	52.35 \pm 0.79	39.16 \pm 0.65	49.49 \pm 0.78	44.54 \pm 0.75	46.39
	M-base	✓	57.65 \pm 0.80	46.03 \pm 0.72	55.70 \pm 0.79	48.25 \pm 0.74	51.91
CD-FSL	FWT [44]	-	47.47 \pm 0.75	31.61 \pm 0.53	55.77 \pm 0.79	35.95 \pm 0.58	42.70
	LRP [40]	-	48.29 \pm 0.51	32.78 \pm 0.39	54.83 \pm 0.56	37.49 \pm 0.43	43.35
	ATA [48]	-	45.00 \pm 0.50	33.61 \pm 0.40	53.57 \pm 0.50	34.42 \pm 0.40	41.65
	wave-SAN [14]	-	50.25 \pm 0.74	33.55 \pm 0.61	57.75 \pm 0.82	40.71 \pm 0.66	45.57
	M-FWT [44]	✓	61.16 \pm 0.81	49.01 \pm 0.76	57.89 \pm 0.82	50.49 \pm 0.81	54.64
	M-LRP [40]†	✓	59.23 \pm 0.58	46.88 \pm 0.53	57.92 \pm 0.58	49.11 \pm 0.54	53.29
	M-ATA [48]†	✓	57.73 \pm 0.57	45.19 \pm 0.49	55.39 \pm 0.55	48.07 \pm 0.52	51.60
	M-waveSAN [14]†	✓	63.59 \pm 0.85	50.06 \pm 0.76	59.89 \pm 0.86	51.99 \pm 0.81	56.38
	meta-FDMixup [13]	✓	63.24 \pm 0.82	51.31 \pm 0.83	58.22 \pm 0.82	51.03 \pm 0.81	55.95
Ours	ME-D2N	✓	65.05 \pm 0.83	49.53 \pm 0.79	60.36 \pm 0.86	52.89 \pm 0.83	56.96
5-way	5-shot	D_{tgt}	CUB	Cars	Places	Plantae	Avg.
FSL	MatchingNet [46]	-	51.37 \pm 0.77	38.99 \pm 0.64	63.16 \pm 0.77	46.53 \pm 0.68	50.01
	RelationNet [41]	-	57.77 \pm 0.69	37.33 \pm 0.68	63.32 \pm 0.76	44.00 \pm 0.60	50.61
	GNN [17]	-	62.25 \pm 0.65	44.28 \pm 0.63	70.84 \pm 0.65	52.53 \pm 0.59	57.48
Baselines	St-Net	-	66.89 \pm 0.66	46.26 \pm 0.67	72.87 \pm 0.67	55.13 \pm 0.66	60.29
	Tt-Net	✓	64.72 \pm 0.69	52.32 \pm 0.69	69.37 \pm 0.68	59.23 \pm 0.70	61.41
	M-base	✓	78.08 \pm 0.60	63.27 \pm 0.70	75.90 \pm 0.67	66.69 \pm 0.68	70.99
CD-FSL	FWT [44]	-	66.98 \pm 0.68	44.90 \pm 0.64	73.94 \pm 0.67	53.85 \pm 0.62	59.92
	LRP [40]	-	64.44 \pm 0.48	46.20 \pm 0.46	74.45 \pm 0.47	54.46 \pm 0.46	59.89
	ATA [48]	-	66.22 \pm 0.50	49.14 \pm 0.40	75.48 \pm 0.40	52.69 \pm 0.40	60.88
	wave-SAN [14]	-	70.31 \pm 0.67	46.11 \pm 0.66	76.88 \pm 0.63	57.72 \pm 0.64	62.76
	M-FWT [44]	✓	79.14 \pm 0.62	65.42 \pm 0.70	78.59 \pm 0.60	68.26 \pm 0.68	72.85
	M-LRP [40]†	✓	77.07 \pm 0.44	64.38 \pm 0.48	77.73 \pm 0.45	67.90 \pm 0.47	71.77
	M-ATA [48]†	✓	73.96 \pm 0.46	68.58 \pm 0.45	76.73 \pm 0.42	66.45 \pm 0.46	71.43
	M-waveSAN [14]†	✓	82.29 \pm 0.58	66.93 \pm 0.71	80.01 \pm 0.60	71.27 \pm 0.70	75.13
	meta-FDMixup [13]	✓	79.46 \pm 0.63	66.52 \pm 0.70	78.92 \pm 0.63	69.22 \pm 0.65	73.53
Ours	ME-D2N	✓	83.17 \pm 0.56	69.17 \pm 0.68	80.45 \pm 0.62	72.87 \pm 0.67	76.42

Table 1: The 5-way 1(5)-shot classification results (%) on four novel target datasets. “Avg.” is short for “Average Accuracy”. The checkmark indicates whether the auxiliary target data D_{tgt} is used for training. Notation † denotes that we adapt the methods into our setting. In most cases, our ME-D2N outperforms all the FSL, baselines, and CD-FSL competitors.

4.3 Main Results

Main Results on Target Datasets. The comparison results of our ME-D2N against all the typical FSL methods, baselines, CD-FSL with or without auxiliary target training data D_{tgt} are given in Table 1. What can be apparently seen from the results is that our ME-D2N achieves the best results beating all the baselines and competitors in most cases. Specifically, under the 5-way 5-shot setting, we achieve 83.17%, 69.17%, 80.45%, and 72.87% on the cub, cars, places, and plantae, respectively. Compared with the GNN, our ME-D2N has an average improvement of 15.42% and 18.94% on 5-way 1-shot and 5-shot tasks. Such a performance growth is contributed by both the use of auxiliary target data and the effectiveness of our technical solutions. Besides, there are some other points worth mentioning. 1) Firstly, by comparing the results of “M-base” with that of FWT, LRP, ATA, and wave-SAN, we observe that by merging the source and target datasets together directly as

the training data, the “M-base” shows advantages over these methods that are carefully designed for CD-FSL. Another observation is that after adapting these methods to our setting e.g. adapting the FWT to M-FWT, an obvious improvement can be found. These two phenomena together show the superiority of introducing the few auxiliary target data and well explain why we stick to this setting; 2) Though the M-base has achieved relatively good performance, our ME-D2N still outperforms it by a large margin. This basically shows that we further address the data imbalance issue thus the knowledge of the training data is utilized to a larger extent; 3) We also notice that the performances of our teacher models St-Net and Tt-Net are not so good. Take the Tt-Net as an example, it has only an average accuracy of 46.39% and 61.41% on 1-shot and 5-shot settings. However, based on these “ordinary” teachers, our final ME-D2N still achieves very high results. This indicates that the knowledge learning process of our student model is effective and

Methos	mini-Img CUB	mini-Img Cars	mini-Img Places	mini-Img Plantae	Avg.
St-Net	81.36 \pm 0.57	81.36 \pm 0.57	81.36 \pm 0.57	81.36 \pm 0.57	81.36
Tt-Net	52.02 \pm 0.66	53.91 \pm 0.70	64.77 \pm 0.71	51.63 \pm 0.69	55.58
M-base	78.94 \pm 0.58	80.75 \pm 0.55	79.99 \pm 0.58	80.51 \pm 0.55	80.05
M-FWT	81.88 \pm 0.57	80.89 \pm 0.58	81.32 \pm 0.56	82.28 \pm 0.55	81.59
M-LRP [40] †	80.84 \pm 0.40	81.15 \pm 0.41	81.07 \pm 0.40	81.51 \pm 0.38	81.14
M-ATA [48] †	77.84 \pm 0.39	78.49 \pm 0.40	77.57 \pm 0.39	78.39 \pm 0.40	78.07
M-waveSAN [14] †	83.08 \pm 0.56	82.96 \pm 0.57	82.79 \pm 0.56	83.20 \pm 0.58	83.01
meta-FDMixup [13]	82.29 \pm 0.57	81.00 \pm 0.58	81.37 \pm 0.56	79.64 \pm 0.59	81.08
ME-D2N	84.10 \pm 0.53	83.89 \pm 0.53	85.45 \pm 0.53	83.74 \pm 0.56	84.30

Table 2: The 5-way 5-shot results (%) on the testing set of mini-Imagenet (abbreviated as mini-Img). “mini-Img | target set” indicates the model is trained on which target dataset. Note that “St-Net” doesn’t need any target data thus its results on four target datasets are the same. “Avg.” is short for “Average Accuracy”. Our ME-D2N shows clear advantages over other competitors.

Method	ME	D2N	CUB	Cars	Places	Plantae	Avg.
M-base	-	-	78.08 \pm 0.60	63.27 \pm 0.70	75.90 \pm 0.67	66.69 \pm 0.68	70.99
ME	✓	-	82.22 \pm 0.56	66.59 \pm 0.73	79.63 \pm 0.64	71.30 \pm 0.68	74.94
ME + D2N (ours)	✓	✓	83.17 \pm 0.56	69.17 \pm 0.68	80.45 \pm 0.62	72.87 \pm 0.67	76.42

Table 3: The effectiveness of our main technical contributions – multi-expert learning (abbreviated as ME) and domain decomposition module (D2N) are shown. Experiments were conducted under the 5-way 5-shot setting.

partially shows that our domain decomposition module makes the student model not limited to the performance of teachers.

As for the CD-FSL competitors with auxiliary target data, generally, the M-ATA performs worst, then follows the M-LRP and M-FWT. Meta-FDMixup and M-waveSAN are the most competitive methods. The relatively good performance of meta-FDMixup is not easy to understand since it is purely proposed for this setting. Among these competitors, the M-waveSAN performs best with very competitive results. This shows that the idea of augmenting styles is also a helpful solution to narrow the domain gap. But, generally, M-waveSAN is still inferior to our method.

Main Results on Source Dataset. To further test the performance of our method on the original source domain, we compare it against the baselines and CD-FSL competitors. The 5-way 5-shot results on the testing set of the mini-Imagenet (disjoint from D_{src}), abbreviated as mini-Img, are given in Table 2. Since our setting is target dataset specific, the results are reported in the form of “mini-Img | target set”. Notably, the St-Net is trained using only source data. Thus, the results of St-Net on four target datasets are the same.

We mainly have the following observations. 1) Our ME-D2N outperforms all the three baseline methods and five CD-FSL competitors achieving an average accuracy of 84.30%. This demonstrates that our model keeps the best ability of recognizing the novel source images; 2) By comparing the results of St-Net, Tt-Net, and M-base, we find that St-Net performs best since it is totally trained using source data. The Tt-Net performs worst and the performance of the M-base also has a degradation compared to that of St-Net. This indicates that merging the target data into the source data is harmful to the source domain. However, this negative effect is addressed by our ME-D2N. We even improve the St-Net by 2.94% on average. This shows that decomposing the student network makes our model has sufficient capacity to learn both the knowledge of the source and target domains.

4.4 Ablation Studies.

Ablation Study on Network Modules. The effectiveness of our main technical contributions – multi-expert learning and domain decomposition module are studied. Results of the 5-way 5-shot setting are provided in Table 3. We use the “ME” to refer to our model learned under the umbrella of multi-expert learning without applying domain decomposition to the student model. Correspondingly, the “ME + D2N” denotes the model equipped with both the multi-expert learning and domain decomposition module. Thus, “ME + D2N” also equals our full ME-D2N network. Comparing “ME” against the M-base, we notice that the multi-expert learning strategy improves the M-base by up to 3.95% on average. This illustrates that we do alleviate the data imbalance problem of the M-base. Similarly, the effectiveness of our domain decomposition module can be drawn through the advantages of “ME + D2N” over “ME”.

Ablation Study on the Number of Decomposed Blocks. As stated in Sec 4.1, we divide our embedding module into four blocks, thus decomposing which blocks is an important question. To that end, we conduct experiments of decomposing different number of blocks and give the 5-way 5-shot results in Table 4. As indicated in previous work [1], the low-level filters convey more generic information thus naturally entangled, while the high-level filters are more semantic-related and much easier to be decomposed. Thus, the upper blocks are decomposed with higher priorities. Take “3 decomposed blocks” as an example, it means the last three blocks are decomposed. Results show that our choice of “2 decomposed blocks” performs best, then follows the “1 decomposed block”, “3 decomposed blocks”, and “4 decomposed blocks” consecutively. This phenomenon basically keeps the consistency of the conclusions as in [1]. Different from those works strive for a more interpretable network that only decomposes the last semantic layer [31, 38, 55], in this paper, decomposing the last two blocks is the best choice for us. This reveals that the domain information is also conveyed in the relatively low-level filters.

Choices	CUB	Cars	Places	Plantae	Avg.
4 decomposed blocks	78.05 ± 0.62	63.27 ± 0.67	77.53 ± 0.64	67.85 ± 0.72	71.68
3 decomposed blocks	81.76 ± 0.57	64.90 ± 0.67	78.55 ± 0.62	70.05 ± 0.70	73.82
2 decomposed blocks (ours)	83.17 ± 0.56	69.17 ± 0.68	80.45 ± 0.62	72.87 ± 0.67	76.42
1 decomposed block	81.39 ± 0.60	68.27 ± 0.70	80.39 ± 0.62	72.18 ± 0.67	75.56
STD path	83.32 ± 0.58	67.85 ± 0.68	80.60 ± 0.61	72.52 ± 0.70	76.07
DSG path	78.15 ± 0.62	65.37 ± 0.67	75.15 ± 0.67	68.85 ± 0.70	71.88
STD + DSG paths (ours)	83.17 ± 0.56	69.17 ± 0.68	80.45 ± 0.62	72.87 ± 0.67	76.42

Table 4: Ablation studies of our method. We conduct experiments of decomposing filters on different numbers of blocks and using different testing strategies for model inference. Results of 5-way 5-shot tasks are reported.

Ablation Study on Testing Strategies. Recall that ME-D2N has two forward paths – STD path and DSG path, thus we report the results of different testing strategies in Table 4. The “STD path” and “DSG path” mean only a single forward path is used while the “STD + DSG paths” denotes both of them are utilized. From the results, it can be observed that the STD path performs better than the DSG path. This is not difficult to understand since the STD path receives knowledge from both domains. However, the contribution of our D2N still hold as the STD path is hosted under the D2N module. On average, utilizing two paths generally improves the final results.

4.5 More Analysis

To provide more analysis of our domain decomposition module, we show the number of filters assigned for different domains in Figure 4. Typically, block3 and block4 have a total of 256 and 512 filters, respectively. Firstly, we observe that the models trained on different target sets share similar distribution of the decomposed filters. Secondly, we notice that the number of target-specific filters is comparable with that of source-specific ones with slight advantages at block3. While the filters of block4 have an obvious bias towards the target domain with a rough ratio of 3:2 for target: source. This illustrates that our D2N module learns to assign more capacity to the target domain.

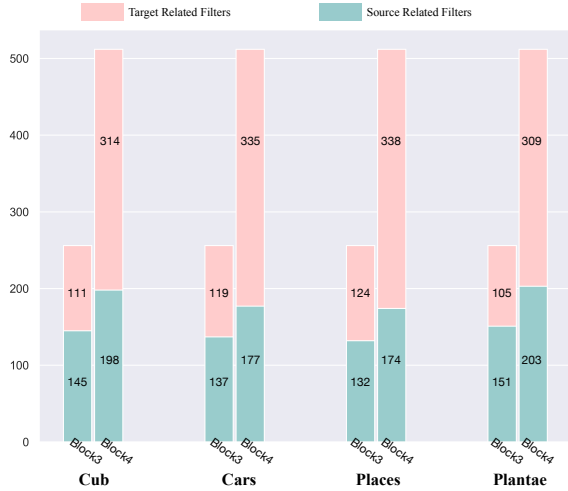


Figure 4: The number of the filters assigned for each domain.

In addition, to have an intuitive understanding of our D2N, we visualize the activation maps of two different domain-specific filters on two different domain images. The demonstrated “source filter” and “target filter” are sampled from the last block of the embedding

module. As shown in Figure 5, the activation results of these two filters towards the same input image are significantly different. The domain-specific filters can accurately focus on the effective features for the input image of the same domain. This further verifies the effectiveness of our D2N module.

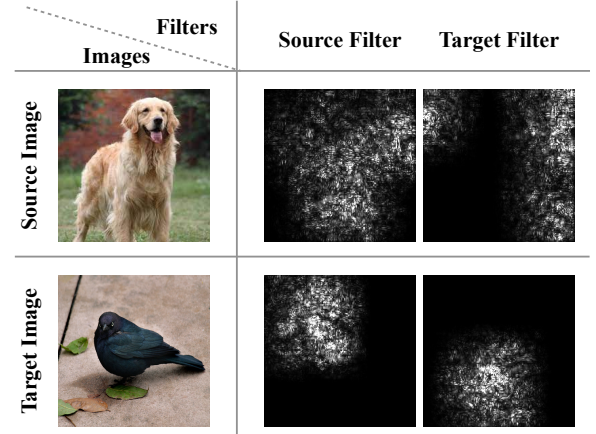


Figure 5: Visualization of activation maps for input images on the source and target-related filters.

5 CONCLUSION

To conclude, we mainly aim at promoting the CD-FSL methods with few labeled target data. To achieve this, we first observe two key challenges lay in this task – data imbalance issue and learning model from two different domains. To address these problems, we thus contribute two novel modules. One is the multi-expert learning mechanism together with the knowledge distillation technique, which enables us to learn “individual knowledge” from two teacher models of different domains rather than learning from the “unbalanced data”. Another is the domain decomposition module which learns to decompose the filters of our student model into the source-specific and target-specific sub-parts. In this way, we prevent our model from learning knowledge of the source domain and target domain at the same time. Based on these two modules, we build our multi-expert domain decomposition network. Experimental results show that our network alleviates the above-mentioned challenges well and achieves state-of-the-art results.

6 ACKNOWLEDGEMENT

This work was supported in part by National Natural Science Foundation of China Project (No. 62072116) and Shanghai Pujiang Program (No. 20PJ1401900).

REFERENCES

- [1] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2017. Network dissection: Quantifying interpretability of deep visual representations. In *CVPR*. 6541–6549.
- [2] Mateusz Buda, Atsuto Maki, and Maciej A Mazurowski. 2018. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks* 106 (2018), 249–259.
- [3] Jonathon Byrd and Zachary Lipton. 2019. What is the effect of importance weighting in deep learning?. In *ICML*. PMLR, 872–881.
- [4] John Cai, Bill Cai, and Shen Sheng Mei. 2021. DAMSL: Domain Agnostic Meta Score-based Learning. In *CVPR*. 2591–2595.
- [5] Jiarui Cai, Yizhou Wang, Hung-Min Hsu, Jenq-Neng Hwang, Kelsey Magrane, and Craig S Rose. 2022. LUNA: Localizing Unfamiliarity Near Acquaintance for Open-Set Long-Tailed Recognition. In *AAAI*, Vol. 36. 131–139.
- [6] Jiarui Cai, Yizhou Wang, and Jenq-Neng Hwang. 2021. Ace: Ally complementary experts for solving long-tailed recognition in one-shot. In *ICCV*. 112–121.
- [7] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arachis, and Tengyu Ma. 2019. Learning imbalanced datasets with label-distribution-aware margin loss. *NeurIPS* 32 (2019).
- [8] Jie Chen, Shouzheng Chen, Mingyuan Bai, Jian Pu, Junping Zhang, and Junbin Gao. 2022. Graph Decoupling Attention Markov Networks for Semisupervised Graph Node Classification. *IEEE TNNLS* (2022).
- [9] Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Frank Wang, and Jia-Bin Huang. 2019. A closer look at few-shot classification. *arXiv preprint* (2019).
- [10] Rajshekhar Das, Yu-Xiong Wang, and José MF Moura. 2021. On the importance of distractors for few-shot classification. In *ICCV*. 9030–9040.
- [11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*.
- [12] Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. *arXiv preprint* (2017).
- [13] Yuqian Fu, Yanwei Fu, and Yu-Gang Jiang. 2021. Meta-FDMixup: Cross-Domain Few-Shot Learning Guided by Labeled Target Data. In *ACM Multimedia*. 5326–5334.
- [14] Yuqian Fu, Yu Xie, Yanwei Fu, Jingjing Chen, and Yu-Gang Jiang. 2022. Wave-SAN: Wavelet based Style Augmentation Network for Cross-Domain Few-Shot Learning. *arXiv preprint* (2022).
- [15] Yuqian Fu, Li Zhang, Junke Wang, Yanwei Fu, and Yu-Gang Jiang. 2020. Depth guided adaptive meta-fusion network for few-shot video recognition. In *ACM Multimedia*. 1142–1151.
- [16] Jixiang Gao, Jingjing Chen, Huazhu Fu, and Yu-Gang Jiang. 2022. Dynamic Mixup for Multi-Label Long-Tailed Food Ingredient Recognition. *IEEE Transactions on Multimedia* (2022).
- [17] Victor Garcia and Joan Bruna. 2017. Few-shot learning with graph neural networks. *arXiv preprint* (2017).
- [18] Yunhui Guo, Noel C Codella, Leonid Karlinsky, James V Codella, John R Smith, Kate Saenko, Tajana Rosing, and Rogerio Feris. 2020. A broader study of cross-domain few-shot learning. In *ECCV*.
- [19] Haibo He and Edwardo A Garcia. 2009. Learning from imbalanced data. *IEEE TKDE* 21, 9 (2009), 1263–1284.
- [20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *CVPR*.
- [21] Chen Huang, Yining Li, Chen Change Loy, and Xiaoou Tang. 2016. Learning deep representation for imbalanced classification. In *CVPR*. 5375–5384.
- [22] Chen Huang, Yining Li, Chen Change Loy, and Xiaoou Tang. 2019. Deep imbalanced learning for face recognition and attribute prediction. *IEEE TPAMI* 42, 11 (2019), 2781–2794.
- [23] Yani Ioannou, Duncan Robertson, Jamie Shotton, Roberto Cipolla, and Antonio Criminisi. 2015. Training cnns with low-rank filters for efficient image classification. *arXiv preprint* (2015).
- [24] Ashraf Islam, Chun-Fu Richard Chen, Rameswar Panda, Leonid Karlinsky, Rogerio Feris, and Richard Radke. 2021. Dynamic distillation network for cross-domain few-shot recognition with unlabeled data. *NeurIPS* 34 (2021).
- [25] Eric Jang, Shixiang Gu, and Ben Poole. 2016. Categorical reparameterization with gumbel-softmax. *arXiv preprint* (2016).
- [26] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. 2019. Decoupling representation and classifier for long-tailed recognition. *arXiv preprint* (2019).
- [27] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 2013. 3d object representations for fine-grained categorization. In *ICCVW*.
- [28] Kai Li, Yulun Zhang, Kunpeng Li, and Yun Fu. 2020. Adversarial Feature Hallucination Networks for Few-Shot Learning. In *CVPR*.
- [29] Pan Li, Shaogang Gong, Yanwei Fu, and Chengjie Wang. 2021. Ranking Distance Calibration for Cross-Domain Few-Shot Learning. *arXiv preprint* (2021).
- [30] Yawei Li, Shuhang Gu, Christoph Mayer, Luc Van Gool, and Radu Timofte. 2020. Group sparsity: The hinge between filter pruning and decomposition for network compression. In *CVPR*. 8018–8027.
- [31] Haoyu Liang, Zhihao Ouyang, Yuyuan Zeng, Hang Su, Zihao He, Shu-Tao Xia, Jun Zhu, and Bo Zhang. 2020. Training interpretable convolutional neural networks by differentiating class-specific filters. In *ECCV*. 622–638.
- [32] Hanwen Liang, Qiong Zhang, Peng Dai, and Juwei Lu. 2021. Boosting the Generalization Capability in Cross-Domain Few-shot Learning via Noise-enhanced Supervised Autoencoder. In *ICCV*. 9424–9434.
- [33] Cheng Perng Phoo and Bharath Hariharan. 2020. Self-training for Few-shot Transfer Across Extreme Task Differences. *arXiv preprint* (2020).
- [34] Qiang Qiu, Xiuyuan Cheng, Guillermo Sapiro, et al. 2018. DCFNet: Deep neural network with decomposed convolutional filters. In *ICML*. PMLR, 4198–4207.
- [35] Sachin Ravi and Hugo Larochelle. 2017. Optimization as a model for few-shot learning. In *ICLR*.
- [36] Xiaofeng Ruan, Yufan Liu, Chunfeng Yuan, Bing Li, Weiming Hu, Yangxi Li, and Stephen Maybank. 2020. Edp: An efficient decomposition and pruning scheme for convolutional neural network compression. *IEEE TNNLS* 32, 10 (2020), 4499–4513.
- [37] Andrei A Rusu, Dushyant Rao, Jakub Sygnowski, Oriol Vinyals, Razvan Pascanu, Simon Osindero, and Raia Hadsell. 2018. Meta-learning with latent embedding optimization. *arXiv preprint* (2018).
- [38] Wen Shen, Zhihua Wei, Shikun Huang, Binbin Zhang, Jiaqi Fan, Ping Zhao, and Quanshi Zhang. 2021. Interpretable Compositional Convolutional Neural Networks. *arXiv preprint* (2021).
- [39] Jake Snell, Kevin Swersky, and Richard Zemel. 2017. Prototypical networks for few-shot learning. In *NeurIPS*.
- [40] Jiamei Sun, Sebastian Lapuschkin, Wojciech Samek, Yunqing Zhao, Ngai-Man Cheung, and Alexander Binder. 2020. Explanation-guided training for cross-domain few-shot classification. *arXiv preprint* (2020).
- [41] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. 2018. Learning to compare: Relation network for few-shot learning. In *CVPR*.
- [42] Hao Tang, Zechao Li, Zhimao Peng, and Jinhui Tang. 2020. BlockMix: Meta Regularization and Self-Calibrated Inference for Metric-Based Meta-Learning. In *ACM Multimedia*. 610–618.
- [43] Hao Tang, Chengcheng Yuan, Zechao Li, and Jinhui Tang. 2022. Learning attention-guided pyramidal features for few-shot fine-grained recognition. *PR* 130 (2022), 108792.
- [44] Hung-Yu Tseng, Hsin-Ying Lee, Jia-Bin Huang, and Ming-Hsuan Yang. 2020. Cross-domain few-shot classification via learned feature-wise transformation. In *ICLR*.
- [45] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. 2018. The inaturalist species classification and detection dataset. In *CVPR*.
- [46] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. 2016. Matching networks for one shot learning. In *NeurIPS*.
- [47] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. 2011. The caltech-ucsd birds-200-2011 dataset. (2011).
- [48] Haoqing Wang and Zhi-Hong Deng. 2021. Cross-domain few-shot classification via adversarial task augmentation. *arXiv preprint* (2021).
- [49] Xudong Wang, Long Lian, Zhongqi Miao, Ziwei Liu, and Stella X Yu. 2020. Long-tailed recognition by routing diverse distribution-aware experts. *arXiv preprint* (2020).
- [50] Yu-Xiong Wang, Deva Ramanan, and Martial Hebert. 2017. Learning to model the tail. *NeurIPS* 30 (2017).
- [51] Liuyu Xiang, Guiguang Ding, and Jungong Han. 2020. Learning from multiple experts: Self-paced knowledge distillation for long-tailed classification. In *ECCV*. 247–263.
- [52] Yu Xie, Yanwei Fu, Ying Tai, Yun Cao, Junwei Zhu, and Chengjie Wang. 2022. Learning To Memorize Feature Hallucination for One-Shot Image Generation. In *CVPR*. 9130–9139.
- [53] Ji Zhang, Jingkuan Song, Lianli Gao, Ye Liu, and Heng Tao Shen. 2022. Progressive Meta-learning with Curriculum. *TCSVT* (2022).
- [54] Ji Zhang, Jingkuan Song, Yazhou Yao, and Lianli Gao. 2021. Curriculum-based meta-learning. In *ACM Multimedia*. 1838–1846.
- [55] Quanshi Zhang, Ying Nian Wu, and Song-Chun Zhu. 2018. Interpretable convolutional neural networks. In *CVPR*. 8827–8836.
- [56] Tianyuan Zhang, Bichen Wu, Xin Wang, Joseph Gonzalez, and Kurt Keutzer. 2019. Domain-aware dynamic networks. *arXiv preprint* (2019).
- [57] Boyan Zhou, Quan Cui, Xiu-Shen Wei, and Zhao-Min Chen. 2020. Bbn: Bilateral-branch network with cumulative learning for long-tailed visual recognition. In *CVPR*. 9719–9728.
- [58] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2017. Places: A 10 million image database for scene recognition. *TPAMI* (2017).
- [59] Jiangang Zhu, Zheng Wang, Jingjing Chen, Yi-Ping Phoebe Chen, and Yu-Gang Jiang. 2022. Balanced Contrastive Learning for Long-Tailed Visual Recognition. In *CVPR*. 6908–6917.