Frequency Guidance Matters in Few-Shot Learning

Hao Cheng[1] Siyuan Yang[1] Joey Tianyi Zhou[2,3] Lanqing Guo[1] Bihan Wen[1*]
[1]Nanyang Technological University [2]Centre for Frontier AI Research (CFAR), A*STAR, Singapore
[3]Institute of High Performance Computing (IHPC), A*STAR, Singapore
{hao006,siyuan005,lanqing001,bihan.wen}@ntu.edu.sg zhouty@cfar.a-star.edu.sg

Abstract

Few-shot classification aims to learn a discriminative feature representation to recognize unseen classes with few labeled support samples. While most few-shot learning methods focus on exploiting the spatial information of image samples, frequency representation has also been proven essential in classification tasks. In this paper, we investigate the effect of different frequency components on the few-shot learning tasks. To enhance the performance and generalizability of few-shot methods, we propose a novel Frequency-Guided Few-shot Learning framework (dubbed FGFL), which leverages the task-specific frequency components to adaptively mask the corresponding image information, with a novel multi-level metric learning strategy including a triplet loss among original, masked and unmasked image as well as a contrastive loss between masked and original support and query sets to exploit more discriminative information. Extensive experiments on four benchmarks under several few-shot scenarios, i.e.*, standard, cross-dataset, cross-domain, and coarse-to-fine annotated classification, are conducted. Both qualitative and quantitative results show that our proposed FGFL scheme can attend to the class-discriminative frequency components, thus integrating those information towards more effective and generalizable few-shot learning.*

1. Introduction

Deep learning typically relies on large-scale annotated datasets to achieve superior results on image classification tasks. However, the learned deep models may fail to generalize under some severe conditions, *e.g.*, few data or annotations. Correspondingly, the task to classify a query sample using only a few labeled data is known as *few-shot classification* [10, 40]. The key challenge is to learn the discriminative feature representation that can be generalized from the given training set to novel testing classes. Recently, several few-shot methods [36, 48, 58] have applied the meta-learning framework [40] to build a discriminative feature space by simulating numerous meta-tasks to obtain the generalized feature representation, achieving superior performance for recognizing novel classes.

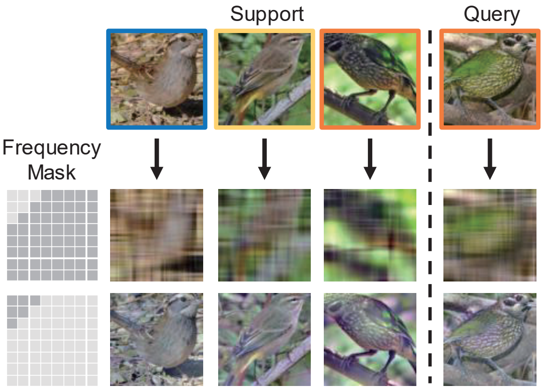


Figure 1: Visualizations of samples with different frequency masks on the CUB dataset [41]. Samples with the same color box belong to the same class. It is easier to distinguish two classes with masked specific frequency components. Each square on the left represents the masked status of corresponding frequency components, *i.e.*, the dark and light color for masked and reserved, respectively. For each row (column), top-to-bottom and left-to-right represent low frequency to high frequency, respectively.

However, very recent studies [3, 17, 26, 39] report that existing methods may fail to generalize well to novel classes due to large domain gaps between training and testing sets, *e.g.*, from general to fine-grained tasks or a specific domain. In these cases, the learned feature embeddings based on spatial images may capture excessive domain-variant information from the training patterns, leading to degraded performance in the test scenario with a large domain gap. Several domain adaptation-based few-shot algorithms [26, 37, 39, 57] have been proposed to alleviate the domain shift issues on few-shot classification. These

[*]Corresponding author.

methods focus on learning a flexible feature embedding or fine-tuning on less novel support samples to transfer knowledge across domains. However, they cannot achieve stable performance improvements for different few-shot scenarios. Moreover, most domain adaption-based methods only focused on exploiting the image spatial features.

Beyond spatial features, the frequency-domain representation plays a vital role for various computer vision tasks, *e.g.*, deep fake recognition [12, 24, 27], domain generalization [19] and image classification [4, 32, 42, 50]. Different frequency components naturally convey discriminative and interpretive features for visual understanding [19, 32], *e.g.*, the selected frequency components (FCs) of input images can be used to interpret the learned CNN features in classification [42, 50]. In few-shot learning, Chen *et al.* [4] is the first attempt to apply image frequency spectrum for classification, in which frequency information is naively concatenated as an extra feature, without exploiting different effects of its FCs. Figure 1 provides an example, clearly showing that certain FCs are more discriminative, thus should be selected and highlighted for more effective classification. However, no work to date has systematically investigated the roles of different FCs and how to adaptively select important FCs for effective few-shot learning.

In this paper, we propose a novel Frequency-Guided Few-shot Learning (FGFL) framework towards highly effective few-shot classification and generalization. For each input image, FGFL adaptively generates a task-specific class-discriminative mask in the frequency domain using Grad-CAM [35], where the activated FCs represent the important frequency information associated with the current task. Then we convert the masked and unmasked frequency components output back to the spatial domain for generating the masked and unmasked images, which contain limited and rich task-specific class-discriminative information, respectively. After that, for each image in the meta-task, we construct its triplet with its corresponding unmasked and masked images as positive and negative samples. We also construct two additional few-shot tasks by regrouping the masked and original support and query images, *i.e.*, $\mathcal{T}_{ms}$ with masked support set and original query set, and $\mathcal{T}_{mq}$ with original support set and masked query set. In addition, we propose two types of ranking losses, including the triplet and contrastive loss functions at single images and tasks (*i.e.*, $\mathcal{T}_{ms}$ and $\mathcal{T}_{mq}$) levels, respectively, to force the model to capture task-specific class-discriminative information that can well generalize to the novel classes. Moreover, the unmasked images are also added to the support set to enhance the discriminability of feature embeddings.

The contributions of this work are fourfold. *First*, we propose a Frequency-Guided Few-shot Learning (FGFL) framework to utilize the frequency information to enhance the performance and generalizability of few-shot methods in the spatial domain. *Second*, we propose two types of ranking loss, including the triplet and contrastive loss functions at both sample and task levels based on original and generated images, to force the network to focus on class-discriminative frequency components for each few-shot meta-task. In addition, the generated unmasked images are also utilized to augment the support set to enhance the discriminability of feature embeddings. *Third*, we investigate the impact of frequency components on current methods under several generalized few-shot settings, namely cross-dataset, cross-domain, and coarse-to-fine annotated generalization. To the best of our knowledge, this is the first work that exploits the effect of different frequency components on few-shot classification and generalization tasks explicitly. *Last*, experimental results and visualizations demonstrate that the proposed FGFL can improve the performance and generalizability of few-shot methods and achieve superior results under several few-shot settings.

2. Related Work

2.1. Few-shot Classification

Few-shot learning aims to recognize novel query samples with only a few available support samples. Recently, existing deep few-shot methods can be divided into three main categories. First, optimization-based meta-learning methods [2, 11, 20, 28, 33] design a good model initialization or optimization strategy that can quickly adapt to novel tasks. Second, metric-based methods [7, 29, 30, 36, 48, 55, 58] focus on constructing an appropriate latent space to extract discriminative feature embeddings and then predict based on the similarity between support and query samples. Third, data generative and augmentation-based methods focus on generating more variant samples to learn a more accurate classifier, *e.g.*, applying data rotation or transformation [21] by self-supervised learning, synthesizing new data with a generative model [14, 23], or using external data with pseudo labels [18, 44, 52].

Most above-mentioned algorithms only employ spatial images as the network input without exploiting the frequency information. In our work, we investigate the effect of the frequency information in the Discrete Cosine Transformation (DCT) domain and leverage it to guide few-shot classification in the spatial domain.

2.2. Learning in the Frequency Domain

Frequency components contain rich patterns for image understanding and have been widely studied in several computer vision tasks such as compressed sensing [43, 51, 56], visual pre-training [47], domain transfer [19, 53], and image classification [4, 8, 32, 42, 50]. Yang *et al.* [53] use the Fast Fourier Transformation (FFT) to achieve domain alignment by simply swapping the low-frequency spectrum between

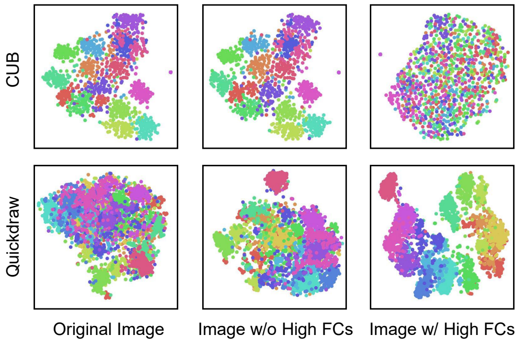


Figure 2: T-SNE visualization of all image features of 20 test classes in the CUB (top) and 15 test classes in the Quickdraw (bottom) dataset. The few-shot model is trained on the ***mini*ImageNet** dataset. A cluster that is clearer and more compact within each class is better.

Method	***mini*ImageNet**	**CUB**	**Quickdraw**
ProtoNet [36]	**79.86**	66.29	64.32
ProtoNet w/ High FCs	56.47	37.96	**66.26**
ProtoNet w/o High FCs	77.75	**67.09**	65.99

Table 1: Few-shot classification results under the 5-way 5-shot setting trained with the *mini*ImageNet dataset. ProtoNet is trained on *mini*ImageNet dataset and evaluated on several datasets with specific frequency components.

source and target domain samples. GFNet [32] considers the attention in the FFT frequency domain to replace the self-attention module in the vision transformer. However, FFT requires double the memory for complex coefficients. Therefore, some recent works [4, 31, 50] consider applying the Discrete Cosine Transformation (DCT) to learn in the frequency domain. For example, Wang *et al.* [43] first applies the DCT to speed up the network. Instead of inputting RGB pixels, Gueguen *et al.* [16] convert the JPEG-compressed image to its 8×8 block DCT coefficients as the input of the neural network. FcaNET [31] adopts DCT to generalize the compression of the channel attention mechanism in the frequency domain. Specifically, Chen *et al.* [4] select pre-defined channels in the DCT frequency domain to improve the performance for few-shot classification.

In this work, we also employ the DCT to transform the images to the frequency domain. Unlike previous work [4], we consider the interaction between spatial and frequency domains and learn to generate the discriminative frequency mask to enhance learning in the spatial domain.

3. Proposed Method

In this section, we first investigate the impact of frequency components in different few-shot classification scenarios. Then the proposed FGFL framework is described and explained in detail, followed by the loss function of our method and empirical validations.

3.1. Frequency Analysis on Few-shot Classification

We first investigate the impact of frequency components on the generalized performance of current few-shot methods under different few-shot settings by removing and preserving partial frequency components of each image in the meta-task for class prediction.

To generate the frequency representation of an input RGB image $X \in \mathbb{R}^{H \times W \times 3}$, we apply the 2D-Discrete Cosine Transform [1] (2D-DCT) function as $D = DCT(X) \in \mathbb{R}^{H \times W \times 3}$. We can then remove or preserve partial frequency components of each input image, followed by an Inverse 2D-DCT function as $\hat{X} = IDCT(\hat{D})$ to transform it back into the spatial domain with the original input size.

With the help of the DCT functions, we can decompose the raw data into two parts, $X = \{X^O, X^H\}$, where X^H and X^O represent high-frequency components (dubbed w/ high FCs), and other components (dubbed w/o high FCs). For a given few-shot task $\mathcal{T}_{\text{test}} = (\mathcal{S}, \mathcal{Q})$ on the novel testing set, we can reconstruct images with partial frequency information to update the testing task as $\tilde{\mathcal{T}}_{\text{test}} = \{\tilde{\mathcal{S}}, \tilde{\mathcal{Q}}\}$. Then we adopt the existing few-shot methods to make predictions for each query sample $\tilde{X}_j$ with $\tilde{\mathcal{S}}$.

Table 1 shows classification accuracy on different test domains and Figure 2 plots the t-SNE visualizations of image features extracted by the trained feature encoder. We can observe that removing specific frequency components can improve the generalized performance of few-shot models but degrade standard few-shot performance in the same testing scenario.

To this end, frequency information plays different roles in different classification scenarios. Through evaluation and observation, it can be conjectured that the generalization behaviors of few-shot methods are relevant to frequency components which may vary across datasets and domains. [1]

3.2. Frequency-Guided Few-shot Learning

Different FCs encode specific image information, we propose the Frequency-Guided Few-shot Learning (FGFL) framework, which utilizes the frequency mask to guide the training in the spatial domain. As shown in Figure 3, FGFL has two branches with two encoders E^{fq} and E^{sp} for feature extraction in the frequency and spatial domains, respectively. FGFL generates the frequency mask based on the task-specific attention map for the given few-shot task, which is then adopted to generate additional images, guiding the network updates in the spatial domain. Moreover, we propose a multi-level metric to force the network to capture specific frequency components for classification.

[1] More implementation details and evaluation results for all frequency components of more few-shot methods are provided in Supplementary.

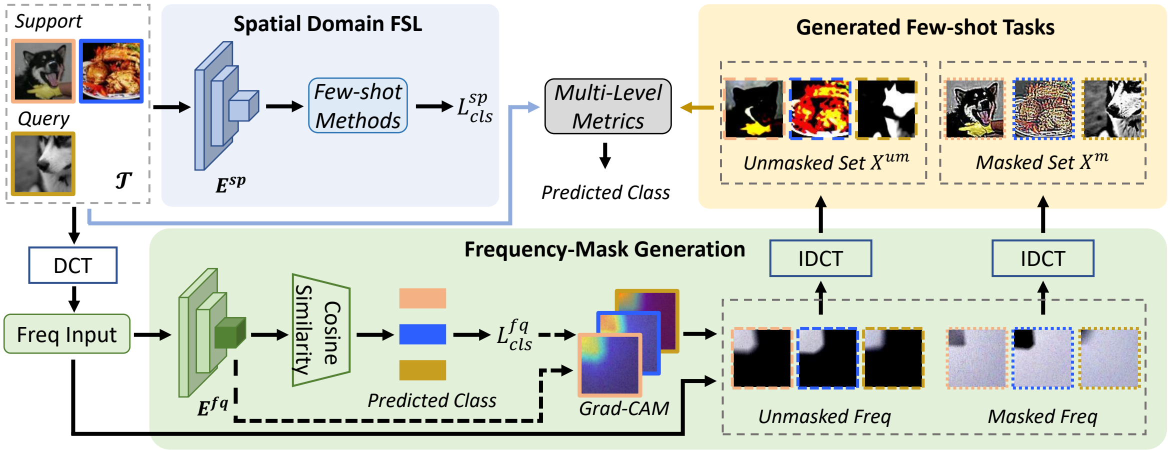


Figure 3: Overview of our proposed Frequency-Guided Few-shot Learning framework (FGFL). We first generate the task-specific class-discriminative frequency mask based on the classification in the frequency domain. Then we convert the masked and unmasked frequency samples back to the spatial domain and generate two image sets (masked set X_m and unmasked set X_{um}) according to the mask and unmask regions, respectively. Furthermore, we adopt novel multi-level metrics (shown in Figure 4) on the original and generated image sets to make class predictions.

3.2.1 Task-Specific Frequency Mask Generation

To obtain a task-specific attention mask of frequency components for each few-shot task, we first transform the input spatial input image X to D in the frequency domain. Then we generate the attention map by Grad-CAM [35] based on the output feature map $M \in \mathbb{R}^{C \times h \times w}$ of the last convolution layer in the frequency encoder (*e.g.*, ResNet-12), where C is the channel number of the feature map, h and w indicate the height and width of the feature map, respectively.

For its ground-truth class c, we compute the gradient of the score s^c with respect to the feature map activation $M_k \in \mathbb{R}^{h \times w}$ as $\frac{\partial s^c}{\partial M_k}$, where k indicates the channel index of the feature map. These gradients flowing back will pass through a global average pooling layer (denoted as $GAP(\cdot)$) to obtain the neuron importance weights w_k^c:

$$w_k^c = GAP(\frac{\partial s^c}{\partial M_k}). \tag{1}$$

The weight w_k^c reflects the importance of activation map M_k to make the prediction of class c for the input D. We then perform a weighted combination with a ReLU function to integrate the positive influence of feature maps across different channels of M for class c as:

$$A^c = ReLU(\sum_k w_k^c M_k), \tag{2}$$

where $A^c \in \mathbb{R}^{h \times w}$ is the heatmap reflecting frequency components that have a positive impact on the ground-truth class c of a given frequency input.

Note that previous works [35, 22] utilize a fully connected (fc) layer to obtain the classification score s^c for the ground-truth class c with the same training and testing class set. However, since the training and testing classes in few-shot scenarios are non-overlap, directly applying fc layers for gradient computation may be prone to overfitting.

To this end, we compute the cosine similarity between each sample and class prototype as the classification score for each given N-way K-shot task in few-shot learning, where N and K indicates the number of selected classes and labeled support samples from each class, respectively. Specifically, given a few-shot task with support set $\mathcal{S}_f = \{D_m, y_m\}_{m=1}^{NK}$, and a query set $\mathcal{Q}_f = \{D_n, y_n\}_{n=1}^{Q}$ in the frequency domain, we compute the cosine similarity score s^c between each query sample $\mathbf{q} \in \mathbb{R}^C$ and each class prototype $\mathbf{v}_c \in \mathbb{R}^C$ as:

$$s^c = p(y = c \mid \mathbf{q}) = \frac{\mathbf{q}^T \mathbf{v}_c}{\|\mathbf{q}\| \, \|\mathbf{v}_c\|}, \tag{3}$$

where $\mathbf{v}_c = \frac{1}{K} \sum_{k=1}^{K} \mathbf{z}_k$ is the mean feature vector of the support features $\mathbf{z}_k$ of the c-th class. Note that, it is easy to prove Eq. (3) is equivalent to an fc layer when neglecting the bias of the fc layer. Moreover, the classification weights $\mathbf{v}_c$ are flexible and task-dependent, *i.e.*, only based on the current few-shot task. Hence, the network can learn to generate a task-specific heatmap A^c of each sample to find specific frequency components that contribute most to the current task, which avoids over-fitting to the training set.

We then up-sample A^c to the original size and generate the frequency mask following previous works [45, 22] as:

$$Mask^c = \mathbf{1} - Sigmoid(Up(A^c)), \tag{4}$$

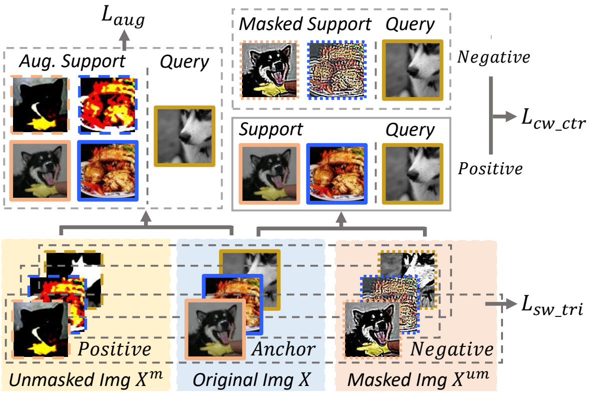


Figure 4: Illustration of the proposed multi-level metrics. Specifically, for each image in the image set X in the meta-task, we construct its triplet from the corresponding unmasked and masked set X^m and X^{um} as the positive and negative pair, respectively, and adopt the sample-wise triplet loss L_{sw_tri} to exploit the correlations them. At the task level, we construct two additional few-shot tasks by regrouping the images in X and X^m. (Here we take the query set as an example.) Then we propose a class-wise contrastive loss L_{cw_ctr} between positive pairs (original support and query samples) and negative pairs (support and query from different sets). Moreover, we use unmasked images to augment the support set for classification with a loss L_{aug}.

where $\mathbf{1}$ is an all-ones matrix, $Up(\cdot)$ is the upsampling operation, and *Sigmoid*$(\cdot)$ is applied to make it derivable and constrains the value to the range $[0, 1]$.

Mask^c is then applied to D and converted back to the spatial domain to generate the masked spatial image X^m and corresponding unmasked spatial image X^{um} as:

$$X^m = IDCT(Mask^c \odot D), X^{um} = IDCT((1-Mask^c)\odot D), \tag{5}$$

where $\odot$ is the dot product function. Note that X^m and X^{um} are considered to contain limited and rich information belonging to the target class since the regions on the attention map beyond the class-aware range should be hard to identify the true class. We can then utilize X^m and X^{um} to guide few-shot learning in the spatial domain.

3.2.2 Mask-Guided Multi-Level Metrics

With the help of task-specific frequency mask generation, we already generate additional samples (*i.e.*, masked images X^m and unmasked images X^{um}). To incorporate frequency information to encourage the spatial network to capture class discriminative information, we propose two ranking losses at the sample and task levels, (*i.e.*, sample-wise triplet loss L_{sw_tri} and class-wise contrastive loss L_{cw_ctr}, respectively.) to fully exploit the mask information in the frequency signal, shown in Figure 4. Meanwhile, we also adopt the unmasked images to enhance the few-shot classification performance with an additional classification loss L_{aug}.

Sample-wise triplet loss. For each original image X_i in the meta-task, this loss aims to enhance the feature similarity between X_i and its corresponding unmasked image X_i^{um}, while reducing that between X_i and its corresponding masked image X_i^m. Formally, we construct the triplet for each image in the meta-task as (X_i, X_i^m, X_i^{um}), where X_i, X_i^{um}, and X_i^m are the anchor, positive, and negative samples, respectively. We define $E^{sp} = f_\phi(\cdot)$ as the feature extractor with parameters ϕ to obtain the feature vector. Then the sample-level triplet loss can be formulated as:

$$\begin{aligned} L_{sw\_tri}(X_i, X_i^{um}, X_i^m) = \max\{&d\left(f_\phi(X_i), f_\phi(X_i^{um})\right) \\ &-d\left(f_\phi(X_i), f_\phi(X_i^m)\right) + \eta, 0\}, \end{aligned} \tag{6}$$

where $d(\cdot,\cdot)$ denotes the distance between two features and η is the margin between positive and negative pairs.

Class-wise contrastive loss. To encourage the network to capture class discriminative information, we construct contrastive pairs for each query instance in the original and masked sets. Specifically, given masked support and query sets $\mathcal{S}_m = \{X_i^m, y_i\}_{i=1}^{NK}$ and $\mathcal{Q}_m = \{X_j^m, y_j\}_{j=1}^{Q}$, we first construct two additional masked few-shot tasks by regrouping the masked and original sets as:

$$\mathcal{T}_{ms} = (\mathcal{S}_m, \mathcal{Q}), \quad \mathcal{T}_{mq} = (\mathcal{S}, \mathcal{Q}_m), \tag{7}$$

where $\mathcal{S} = \{X_i, y_i\}_{i=1}^{NK}$ and $\mathcal{Q} = \{X_j, y_j\}_{j=1}^{Q}$ are original support and query set, respectively.

Each masked sample consists of less class-discriminative information. To avoid being misled by class-irrelevant information (*i.e.*, preserved in $\mathcal{S}_m$ and $\mathcal{Q}_m$), we regard the support and query samples from tasks $\mathcal{T}_{ms}$ and $\mathcal{T}_{mq}$ as negative pairs, and those from the original task with the same class as positive pairs. Then the contrastive loss corresponding to query sample $X_q \in \mathcal{Q}$ can be written as:

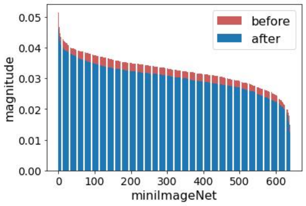
$$L_q = -\log \frac{\sum_{y_j=y_q, X_j\in S} e^{d(z_j,z_q)/\tau}}{\sum_{y_j=y_q,X_j\in S} e^{d(z_j,z_q)/\tau} + \sum_{X_k\in S_m} e^{d(z_k,z_q)/\tau}}, \tag{8}$$

where $z_i = f_\phi(X_i)$ represents the feature vector extracted by the spatial backbone. Similarly, the contrastive loss for support sample $X_s \in \mathcal{S}$ can be written as:

$$L_s = -\log \frac{\sum_{y_j=y_s, X_j\in Q} e^{d(z_j,z_s)/\tau}}{\sum_{y_j=y_s,X_j\in Q} e^{d(z_j,z_s)/\tau} + \sum_{X_k\in Q_m} e^{d(z_k,z_s)/\tau}}. \tag{9}$$

Overall, for one episode, L_{cw_ctr} can be written as:

$$L_{cw\_ctr} = \frac{1}{Q}\sum_{q=1}^{Q} L_q + \frac{1}{NK}\sum_{s=1}^{NK} L_s. \tag{10}$$



(a) Visualization results on Mean Magnitude features of Channels (MMC).

Method	*mini*	CUB
Pretrained	0.5567	1.5584
FEAT [55]	0.5386	1.5367
Ours	**0.1256**	**1.2550**

(b) Maximum Mean Discrepancy (MMD) results on different test datasets.

Figure 5: (a) MMC before and after applying the proposed FGFL method. (b) MMD results of few-shot methods trained over *mini*ImageNet. The lower MMD value indicates a smaller gap and better performance. 'Pretrained' denotes the feature extractor (*i.e.*, ResNet-12) pre-trained on the training set of *mini*ImageNet.

Augmented classification loss. To better classify samples in the query set $\mathcal{Q}$, we use generated support images $\mathcal{S}_{um}$ in X^{um} to enhance the support set for classification. Specifically, we build a new support set by combining the original support and unmasked support set as $\mathcal{S}_{aug} = \mathcal{S} \cup \mathcal{S}_{um}$. Hence, we conduct few-shot classification on query set $\mathcal{Q}$ based on $\mathcal{S}_{aug}$ and minimize the cross-entropy loss:

$$L_{aug} = \sum_{j=1}^{Q} L_{CE}\left(f_\theta\left(X_j; \mathcal{S}_{aug}\right), y_j\right). \tag{11}$$

3.2.3 Objective, Model Training, and Inference

Combining with the standard few-shot classification, the overall objective function is defined as:

$$L_{total} = L_{cls}^{sp} + L_{cls}^{fq} + \lambda_1 \cdot L_{sw\_tri} + \lambda_2 \cdot L_{cw\_ctr} + \lambda_3 \cdot L_{aug}, \tag{12}$$

where L_{cls}^{sp} and L_{cls}^{fq} represent the cross-entropy loss for few-shot classification in the spatial and frequency branch, respectively. λ_1, λ_2, and λ_3 denote the weighting parameters of the corresponding losses, respectively.

During the training phase, we first generate frequency masks based on the frequency branch and then adopt the proposed multi-level metrics for model training. During the inference time, our proposed method strictly follows the inductive setting. We only feed the support set into the frequency branch to generate unmasked support images to refine the support set for classification.

3.3. Empirical Validation

The proposed FGFL framework aims to utilize frequency information to guide the spatial feature extractor towards class-discriminative features and avoid over-fitting to the training patterns, thus generalizing well to testing scenarios. To validate the effectiveness of the proposed method, we conduct experiments about learned features on two aspects: channel distribution shift and domain distance. [2]

Mean Magnitude of Channels (MMC) Luo et al. [26] shows the channel bias problem affects the generalizability of few-shot methods and utilizes MMC to visualize and measure the channel response to features, where effective few-shot methods might have a more uniform MMC curve in the testing set. Inspired by this work, we visualize the MMC before and after applying FGFL over *mini*ImageNet. We can observe a channel magnitude change in Figure 5 (a), which validates that our proposed method can alleviate the channel bias problem, thus improving the performance.

Maximum Mean Discrepancy (MMD) The MMD [15] is a distance measure between two domains based on the embedding of distribution measures in a reproducing kernel Hilbert space, which has been widely applied in transfer learning problems. We adopt class-wise MMD to measure the distribution gap between the training and testing sets with different methods trained over *mini*ImageNet. Figure 5 (b) shows different gaps between one training set and different testing sets. We can also observe that our proposed method helps to reduce the domain shift.

4. Experiments

4.1. Experimental Setup

Datasets. We conduct extensive experiments on four few-shot datasets, *i.e.*, *mini*ImageNet [40], *tiered*ImageNet [34], CUB [41], and FS-DomainNet [6].

Implementation Details. For fair comparisons with previous methods, we adopt the ResNet-12 network with $[64, 160, 320, 640]$ channels as our backbone in each branch, and set the input image size as 84×84 for all datasets. We combine our proposed FGFL with FEAT [55], which adopts a linear projection layer for feature adaptation. We randomly select 15 samples of each class to construct the query set for 5-way 1-shot and 5-way 5-shot settings. The backbone networks are pre-trained on the training set following the typical classification task. The weighting parameters λ_1, λ_2, and λ_3 are all set to 0.1 to balance the scale between the classification loss and the corresponding losses for all datasets. [3]

4.2. Few-shot Classification

We conduct experiments on two standard few-shot settings, *i.e.*, general and fine-grained classification, shown in Table 2. Table 2 (a) shows that our proposed FGFL can

[2] More details are included in Supplementary.

[3] More details about datasets and experimental settings are included in Supplementary.

Method	*mini*ImageNet		*tiered*ImageNet	
	5-way 1-shot	5-way 5-shot	5-way 1-shot	5-way 5-shot
ProtoNet [36] †	61.83 ± 0.20	79.86 ± 0.14	66.84 ± 0.23	84.54 ± 0.16
Meta-Baseline [5]	63.17 ± 0.23	79.26 ± 0.17	68.62 ± 0.27	83.29 ± 0.18
Good-Embed [38]	64.82 ± 0.60	82.14 ± 0.43	71.52 ± 0.69	86.03 ± 0.58
DeepEMD [58]	65.91 ± 0.82	82.41 ± 0.56	71.16 ± 0.87	86.03 ± 0.58
FRN [46]	66.45 ± 0.19	82.83 ± 0.13	71.16 ± 0.22	86.01 ± 0.15
FEAT [55] †	66.52 ± 0.20	81.46 ± 0.14	70.30 ± 0.23	84.55 ± 0.16
BML [60]	67.04 ± 0.63	83.63 ± 0.29	68.99 ± 0.50	85.49 ± 0.34
IEPT [59]	67.05 ± 0.44	82.90 ± 0.30	72.24 ± 0.50	86.73 ± 0.34
MELR [9]	67.40 ± 0.43	83.40 ± 0.28	72.14 ± 0.51	87.01 ± 0.35
MCL-Katz [25]	67.51	83.99	72.01	86.02
CSEI [21]	67.59 ± 0.83	81.93 ± 0.36	72.57 ± 0.95	85.72 ± 0.63
Meta DeepBDC [48]	67.83 ± 0.43	84.46 ± 0.28	72.34 ± 0.49	**87.31 ± 0.32**
DFR [6]	68.12 ± 0.81	82.79 ± 0.56	72.38 ± 0.95	86.00 ± 0.61
Yang et al. [54]	**70.19 ± 0.46**	**84.66 ± 0.29**	**72.62 ± 0.51**	86.62 ± 0.33
Ours	**69.14 ± 0.80**	**86.01 ± 0.62**	**73.21 ± 0.88**	**87.21 ± 0.61**

(a) Average classification accuracy (%) on ***mini*ImageNet** and ***tiered*ImageNet** datasets with the ResNet-12 as backbone.

Method	Backbone	**CUB**	
		5-way 1-shot	5-way 5-shot
MELR [9]	ConvNet-4	70.26 ± 0.50	85.01 ± 0.32
ProtoNet [36] †	ResNet-12	72.25 ± 0.21	87.47 ± 0.13
DeepEMD [58]	ResNet-12	75.65 ± 0.83	88.69 ± 0.50
FEAT [55] †	ResNet-12	75.68 ± 0.20	87.91 ± 0.13
BML [60]	ResNet-12	76.21 ± 0.63	90.45 ± 0.36
Good-Embed [38] ‡	ResNet-18	77.92 ± 0.46	89.94 ± 0.26
DFR [6]	ResNet-12	78.07 ± 0.79	89.74 ± 0.51
VFD [49]	ResNet-12	79.12 ± 0.83	91.48 ± 0.39
FRN [46] †	ResNet-12	**82.90 ± 0.19**	**92.61 ± 0.10**
Ours	ResNet-12	**80.77 ± 0.90**	**92.01 ± 0.71**

(b) Average classification accuracy (%) on **CUB** dataset.

Table 2: Comparison with state-of-the-art methods for general (a) and fine-grained (b) few-shot image classification. The best and second best results under each setting and dataset are highlighted as Red and Blue, respectively. †: reimplemented based on public code with our setting. ‡: reported in [48].

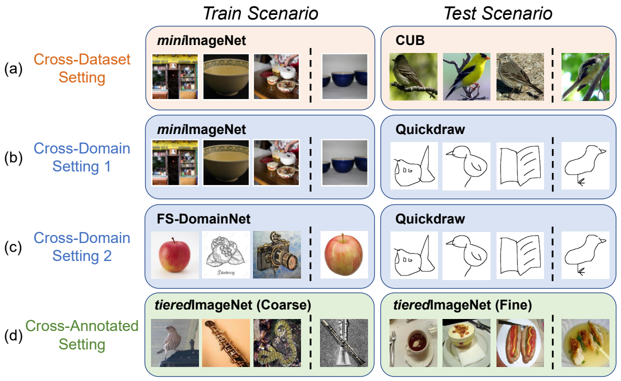


Figure 6: Illustration of three challenging few-shot generalization scenarios under the 3-way 1-shot setting.

significantly improve the performance of few-shot methods and outperform state-of-the-art methods for 5-shot setting on *mini*ImageNet and 1-shot setting on *tiered*ImageNet datasets. Moreover, FGFL is still on par with state-of-the-art methods under other settings. For CUB, as shown in Table 2 (b), we can observe that FGFL can improve the performance of FEAT by a large margin, *i.e.*, 5.09% and 4.1% for 1-shot and 5-shot settings, and is competitive with the state-of-the-art methods.

4.3. Few-Shot Generalization

To evaluate the generalized performance of few-shot methods, we propose three few-shot generalization settings for cross-domain classification based on challenging test scenarios, shown in Figure 6.

4.3.1 Cross-Dataset Generalization

In this setting, we only consider the distribution gaps of classes between training and testing datasets with the same style, *e.g.*, natural images. As shown in Figure 6 (a), we train on the *mini*ImageNet dataset with natural images and evaluate the fine-grained CUB dataset with various bird species, *i.e.*, *mini* → CUB.

Table 3 shows that the proposed FGFL significantly improves the performance of the baseline FEAT method and outperforms the other methods. Figure 7 (a) shows that the proposed FGFL can force the network to capture discriminative task-specific frequency components for classification, *e.g.*, high-frequency edge features for the first class and low-frequency color features for the third class.

4.3.2 Cross-Domain Generalization

Unlike the cross-dataset setting, the visual similarity between the training (source) and testing (target) domains may have a significant gap, as the target domain may contain unnatural images, *e.g.*, quickdraw. As shown in Figure 6 (b) and Figure 6 (c), we split the cross-domain scenario into two specific settings according to the number of the source domains, *i.e.*, single (*e.g.*, *mini*ImageNet) or multiple domains (*e.g.*, FS-DomainNet). For the second setting, each source domain only contains images with the same style such as paintings.

Results in Table 3 show that our proposed FGFL outperforms other few-shot methods by a large margin under two cross-domain generalization settings, *i.e.*, *mini* → Quickdraw and FS-DomainNet → Quickdraw. Figure 7 (b) also verifies that our proposed method can indeed learn to ex-

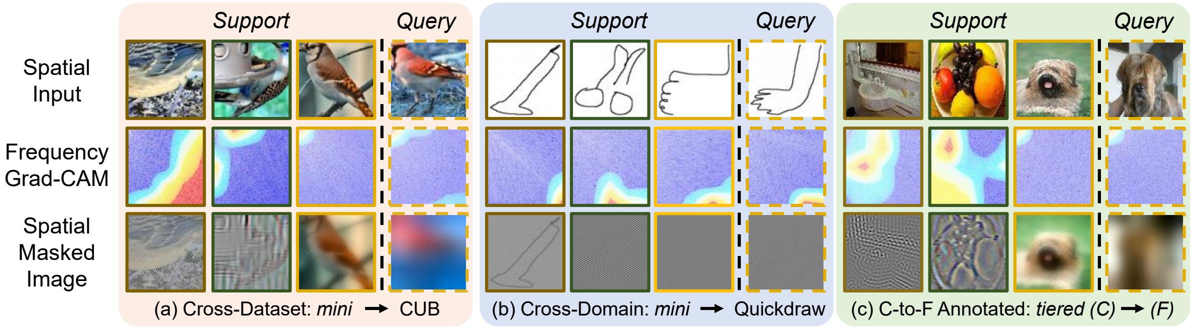

Support
Query
Support
Query
Support
Query
Spatial Input
Frequency Grad-CAM
Spatial Masked Image
(a) Cross-Dataset: mini → CUB
(b) Cross-Domain: mini → Quickdraw
(c) C-to-F Annotated: tiered (C) → (F)

Figure 7: Visualizations for few-shot generalization on three test datasets. (a), (b), and (c) represent different settings, *i.e.*, cross-dataset, cross-domain, and cross-to-fine (C-to-F) annotated settings, respectively.

ploit high-frequency patterns rare in natural images to distinguish different classes on the quickdraw dataset. Another interesting observation is that few-shot methods generally perform better when generalizing from FS-DomainNet to quickdraw than from *mini*ImageNet to quickdraw. A reasonable explanation is that compared to the *mini*ImageNet dataset with small domain variations, the FS-DomainNet dataset includes multiple domains with obvious differences, which enables the network to capture domain-specific information from each meta-task for classification, resulting in better generalization ability to the unseen domain. Furthermore, our FGFL can still boost performance in this setting, illustrating the necessity of frequency domain information.

4.3.3 Coarse-to-Fine Annotated Generalization

Different from existing settings, we consider the annotation difference between training and testing sets for few-shot generalization. In practice, an image may have different levels of class annotations due to the requirements of labeling tasks. For example, we only need a training set with two labels "Furniture" and "Person" to train a powerful classifier for a binary classification task. However, when we want to use the above classifier to identify the species of furniture, we need additional fine-grained labels such as "Desk" or "Chair". This is a practical scenario since fine-grained labeling requires a high cost of human collaboration. We conduct experiments on *tiered*ImageNet with hierarchical annotations to evaluate the generalization performance under this setting, shown in Figure 6 (d). All competing methods are trained on the coarse-annotated training set with 20 super-classes and evaluated on the fine-annotated testing set with 160 sub-classes same as the standard few-shot setting, denoted as *tiered*(C) → (F).

As shown in Table 3, all methods perform worse than general classification results in Table 2. The reason is that the model tends to overfit the general features of high-level class prediction tasks while ignoring within-class differences. We observe that our proposed FGFL can still outperform all methods. Figure 7 (c) shows that our method can focus on specific frequency components (*e.g.*, low and middle FCs corresponding to color and shape spatial features), which is discriminative for class prediction.

Method	Cross-Dataset	Cross-Domain		Cross-Annotated
	***mini* → CUB**	***mini* → Q**	**F → Q**	***tiered*(C) → (F)**
MatchingNet [39]	51.37	-	-	-
GNN [13]	62.25	-	-	-
GNN + FT [39]	66.98	-	-	-
Meta DeepBDC [48]	77.87	-	-	-
FEAT [55] †	62.28	55.96	77.23	64.67
ProtoNet [36] †	66.29	64.32	74.43	64.08
BML [60] †	71.22	75.47	65.38	52.63
DeepEMD [58] †	77.34	58.04	78.52	65.87
Ours	**81.35**	**76.34**	**89.34**	**67.84**

Table 3: Few-shot generalization results under the 5-way 5-shot setting. ***mini***, **Q**, **F**, and ***tiered*(C), (F)** denote *mini*ImageNet, QuickDraw, FS-DomainNet, *tiered*ImageNet with coarse-grained or fine-grained annotations. †: reimplement based on public code with our setting. *tiered*ImageNet with coarse-grained (super-classes) or fine-grained annotations (sub-classes). †: reimplemented based on public code with our setting.

Model	***mini*ImageNet**	
	5-way 1-shot	5-way 5-shot
Baseline	66.52 ± 0.20	81.46 ± 0.14
$+L_{sw\_tri}$	67.38 ± 0.80	83.61 ± 0.61
$+L_{cw\_ctr}$	68.44 ± 0.81	85.60 ± 0.60
$+L_{aug}$	**69.14 ± 0.80**	**86.01 ± 0.62**

Table 4: Ablation studies for our proposed FGFL framework over *mini*ImageNet.

4.4. Ablation Study

We perform ablation studies to investigate the effectiveness of the key components of FGFL on *mini*ImageNet under the 5-way settings with ResNet-12 as the backbone. As shown in Table 4, each proposed loss contributes to performance improvement. Moreover, we find that L_{sw_tri} and L_{cw_ctr} significantly improve the performance (1.92% and 4.14% for 1-shot and 5-shot, resp.), indicating that the proposed multi-level metrics can better utilize the frequency information to improve the generalizability of the model and avoid over-fitting to spatial features, leading to further improvements over spatial masks. Results also show that the two losses play different roles and complement each other, which further proves that our proposed method can utilize frequency information from different levels to help with classification. Noted that the performance gain under the 5-shot setting is much higher than the 1-shot setting. One plausible reason is that more supervised information introduced by support samples helps to generate a more precise frequency mask, which helps reduce the intra-class variation to learn a strong classifier, aligning with our idea.

5. Conclusion

In this paper, we proposed a novel Frequency-Guided Few-shot Learning (FGFL) framework for few-shot classification and generalization. Specifically, the frequency masks generated by the classification gradients in the DCT domain can emphasize the class-discriminative frequency components in the current meta-task. Our proposed multi-level metrics, including sample-wise triplet loss and class-wise contrastive loss, can better utilize the frequency information at sample and task levels and encourage the model to capture more task-relevant frequency information to help the classification in the spatial domain. Besides, we systematically investigated the effects of frequency information on few-shot generalization behaviors. Extensive experiments demonstrated that FGFL achieves competitive results with state-of-the-art methods on the standard few-shot settings, and significantly improves the generalized performance on unseen testing domains.

Acknowledgement

This research is supported in part by the SUG and MOE AcRF Tier1 (RG61/22) Grant. This research work is also partially supported by Joey Tianyi Zhou's A*STAR SERC Central Research Fund (Use-inspired Basic Research), and the Singapore Government's Research, Innovation and Enterprise 2020 Plan (Advanced Manufacturing and Engineering domain) under Grant A18A1b0045.

References

[1] Nasir Ahmed, T_ Natarajan, and Kamisetty R Rao. Discrete cosine transform. *IEEE transactions on Computers*, 100(1):90–93, 1974. 3

[2] Sungyong Baik, Janghoon Choi, Heewon Kim, Dohee Cho, Jaesik Min, and Kyoung Mu Lee. Meta-learning with task-adaptive loss function for few-shot learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9465–9474, 2021. 2

[3] Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Frank Wang, and Jia-Bin Huang. A closer look at few-shot classification. In *International Conference on Learning Representations*, 2019. 1

[4] Xiangyu Chen and Guanghui Wang. Few-shot learning by integrating spatial and frequency representation. In *2021 18th Conference on Robots and Vision (CRV)*, pages 49–56. IEEE, 2021. 2, 3

[5] Yinbo Chen, Zhuang Liu, Huijuan Xu, Trevor Darrell, and Xiaolong Wang. Meta-baseline: Exploring simple meta-learning for few-shot learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9062–9071, 2021. 7

[6] Hao Cheng, Yufei Wang, Haoliang Li, Alex C Kot, and Bihan Wen. Disentangled feature representation for few-shot image classification. *IEEE Transactions on Neural Networks and Learning Systems*, 2023. 6, 7

[7] Hao Cheng, Joey Tianyi Zhou, Wee Peng Tay, and Bihan Wen. Graph neural networks with triple attention for few-shot learning. *IEEE Transactions on Multimedia*, 2023. 2

[8] Max Ehrlich and Larry S Davis. Deep residual learning in the jpeg transform domain. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3484–3493, 2019. 2

[9] Nanyi Fei, Zhiwu Lu, Tao Xiang, and Songfang Huang. Melr: Meta-learning via modeling episode-level relationships for few-shot learning. In *International Conference on Learning Representations*, 2021. 7

[10] 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. 1

[11] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, pages 1126–1135. PMLR, 2017. 2

[12] Joel Frank, Thorsten Eisenhofer, Lea Schönherr, Asja Fischer, Dorothea Kolossa, and Thorsten Holz. Leveraging frequency analysis for deep fake image recognition. In *International conference on machine learning*, pages 3247–3258. PMLR, 2020. 2

[13] Victor Garcia and Joan Bruna. Few-shot learning with graph neural networks. In *International Conference on Learning Representations*, 2018. 8

[14] Spyros Gidaris and Nikos Komodakis. Generating classification weights with gnn denoising autoencoders for few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21–30, 2019. 2

[15] Arthur Gretton, Karsten Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex Smola. A kernel method for the two-sample-problem. *Advances in neural information processing systems*, 19, 2006. 6

[16] Lionel Gueguen, Alex Sergeev, Ben Kadlec, Rosanne Liu, and Jason Yosinski. Faster neural networks straight from jpeg. In *International Conference on Learning Representations*, 2018. 3

[17] Yunhui Guo, Noel C Codella, Leonid Karlinsky, James V Codella, John R Smith, Kate Saenko, Tajana Rosing, and Rogerio Feris. A broader study of cross-domain few-shot learning. In *European Conference on Computer Vision*, pages 124–141. Springer, 2020. 1

[18] Bharath Hariharan and Ross Girshick. Low-shot visual recognition by shrinking and hallucinating features. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3018–3027, 2017. 2

[19] Jiaxing Huang, Dayan Guan, Aoran Xiao, and Shijian Lu. Fsdr: Frequency space domain randomization for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6891–6902, 2021. 2

[20] Eugene Lee, Cheng-Han Huang, and Chen-Yi Lee. Few-shot and continual learning with attentive independent mechanisms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9455–9464, 2021. 2

[21] Junjie Li, Zilei Wang, and Xiaoming Hu. Learning intact features by erasing-inpainting for few-shot classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8401–8409, 2021. 2, 7

[22] Kunpeng Li, Ziyan Wu, Kuan-Chuan Peng, Jan Ernst, and Yun Fu. Tell me where to look: Guided attention inference network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9215–9223, 2018. 4

[23] Kai Li, Yulun Zhang, Kunpeng Li, and Yun Fu. Adversarial feature hallucination networks for few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13470–13479, 2020. 2

[24] Honggu Liu, Xiaodan Li, Wenbo Zhou, Yuefeng Chen, Yuan He, Hui Xue, Weiming Zhang, and Nenghai Yu. Spatial-phase shallow learning: rethinking face forgery detection in frequency domain. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 772–781, 2021. 2

[25] Yang Liu, Weifeng Zhang, Chao Xiang, Tu Zheng, Deng Cai, and Xiaofei He. Learning to affiliate: Mutual centralized learning for few-shot classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14411–14420, 2022. 7

[26] Xu Luo, Jing Xu, and Zenglin Xu. Channel importance matters in few-shot image classification. In *International conference on machine learning*, pages 14542–14559. PMLR, 2022. 1, 6

[27] Yuchen Luo, Yong Zhang, Junchi Yan, and Wei Liu. Generalizing face forgery detection with high-frequency features. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16317–16326, 2021. 2

[28] Alex Nichol, Joshua Achiam, and John Schulman. On first-order meta-learning algorithms. *arXiv preprint arXiv:1803.02999*, 2018. 2

[29] Jaehoon Oh, Hyungjun Yoo, ChangHwan Kim, and Se-Young Yun. {BOIL}: Towards representation change for few-shot learning. In *International Conference on Learning Representations*, 2021. 2

[30] Mihir Prabhudesai, Shamit Lal, Darshan Patil, Hsiao-Yu Tung, Adam W Harley, and Katerina Fragkiadaki. Disentangling 3d prototypical networks for few-shot concept learning. In *International Conference on Learning Representations*, 2021. 2

[31] Zequn Qin, Pengyi Zhang, Fei Wu, and Xi Li. Fcanet: Frequency channel attention networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 783–792, 2021. 3

[32] Yongming Rao, Wenliang Zhao, Zheng Zhu, Jiwen Lu, and Jie Zhou. Global filter networks for image classification. *Advances in Neural Information Processing Systems*, 34:980–993, 2021. 2, 3

[33] Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. In *International Conference on Learning Representations*, 2017. 2

[34] 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 *International Conference on Learning Representations*, 2018. 6

[35] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 2, 4

[36] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 4080–4090, 2017. 1, 2, 3, 7, 8

[37] Takeshi Teshima, Issei Sato, and Masashi Sugiyama. Few-shot domain adaptation by causal mechanism transfer. In *International Conference on Machine Learning*, pages 9458–9469. PMLR, 2020. 1

[38] 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 *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, pages 266–282. Springer, 2020. 7

[39] Hung-Yu Tseng, Hsin-Ying Lee, Jia-Bin Huang, and Ming-Hsuan Yang. Cross-domain few-shot classification via learned feature-wise transformation. *arXiv preprint arXiv:2001.08735*, 2020. 1, 8

[40] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pages 3637–3645, 2016. 1, 6

[41] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 1, 6

[42] Haohan Wang, Xindi Wu, Zeyi Huang, and Eric P Xing. High-frequency component helps explain the generalization of convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8684–8694, 2020. 2

[43] Yunhe Wang, Chang Xu, Chao Xu, and Dacheng Tao. Packing convolutional neural networks in the frequency domain. *IEEE transactions on pattern analysis and machine intelligence*, 41(10):2495–2510, 2018. 2, 3

[44] Yu-Xiong Wang, Ross Girshick, Martial Hebert, and Bharath Hariharan. Low-shot learning from imaginary data. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7278–7286, 2018. 2

[45] Yunchao Wei, Jiashi Feng, Xiaodan Liang, Ming-Ming Cheng, Yao Zhao, and Shuicheng Yan. Object region mining with adversarial erasing: A simple classification to semantic segmentation approach. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1568–1576, 2017. 4

[46] Davis Wertheimer, Luming Tang, and Bharath Hariharan. Few-shot classification with feature map reconstruction networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8012–8021, 2021. 7

[47] Jiahao Xie, Wei Li, Xiaohang Zhan, Ziwei Liu, Yew Soon Ong, and Chen Change Loy. Masked frequency modeling for self-supervised visual pre-training. *arXiv preprint arXiv:2206.07706*, 2022. 2

[48] Jiangtao Xie, Fei Long, Jiaming Lv, Qilong Wang, and Peihua Li. Joint distribution matters: Deep brownian distance covariance for few-shot classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7972–7981, 2022. 1, 2, 7, 8

[49] Jingyi Xu, Hieu Le, Mingzhen Huang, ShahRukh Athar, and Dimitris Samaras. Variational feature disentangling for fine-grained few-shot classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8812–8821, 2021. 7

[50] Kai Xu, Minghai Qin, Fei Sun, Yuhao Wang, Yen-Kuang Chen, and Fengbo Ren. Learning in the frequency domain. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1740–1749, 2020. 2, 3

[51] Guang Yang, Simiao Yu, Hao Dong, Greg Slabaugh, Pier Luigi Dragotti, Xujiong Ye, Fangde Liu, Simon Arridge, Jennifer Keegan, Yike Guo, et al. Dagan: deep de-aliasing generative adversarial networks for fast compressed sensing mri reconstruction. *IEEE transactions on medical imaging*, 37(6):1310–1321, 2017. 2

[52] Shuo Yang, Lu Liu, and Min Xu. Free lunch for few-shot learning: Distribution calibration. In *International Conference on Learning Representations*, 2021. 2

[53] Yanchao Yang and Stefano Soatto. Fda: Fourier domain adaptation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4085–4095, 2020. 2

[54] Zhanyuan Yang, Jinghua Wang, and Yingying Zhu. Few-shot classification with contrastive learning. In *European conference on computer vision*, pages 293–309. Springer, 2022. 7

[55] Han-Jia Ye, Hexiang Hu, De-Chuan Zhan, and Fei Sha. Few-shot learning via embedding adaptation with set-to-set functions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8808–8817, 2020. 2, 6, 7, 8

[56] Yi Yu, Yufei Wang, Wenhan Yang, Shijian Lu, Yap-Peng Tan, and Alex C Kot. Backdoor attacks against deep image compression via adaptive frequency trigger. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12250–12259, 2023. 2

[57] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13834–13844, 2021. 1

[58] Chi Zhang, Yujun Cai, Guosheng Lin, and Chunhua Shen. Deepemd: Few-shot image classification with differentiable earth mover's distance and structured classifiers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2020. 1, 2, 7, 8

[59] Manli Zhang, Jianhong Zhang, Zhiwu Lu, Tao Xiang, Mingyu Ding, and Songfang Huang. Iept: Instance-level and episode-level pretext tasks for few-shot learning. In *International Conference on Learning Representations*, 2020. 7

[60] Ziqi Zhou, Xi Qiu, Jiangtao Xie, Jianan Wu, and Chi Zhang. Binocular mutual learning for improving few-shot classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8402–8411, 2021. 7, 8