Disentangled Feature Representation for Few-shot Image Classification

Hao Cheng[1], **Yufei Wang**[1], **Haoliang Li**[2], **Alex C. Kot**[1], **Bihan Wen**[1]*

[1] School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore
[2] Department of Electrical Engineering, City University of Hong Kong, China
{hao006, yufei001, eackot, bihan.wen}@ntu.edu.sg, haoliang.li@cityu.edu.hk

Abstract

Learning the generalizable feature representation is critical for few-shot image classification. While recent works exploited task-specific feature embedding using meta-tasks for few-shot learning, they are limited in many challenging tasks as being distracted by the excursive features such as the background, domain and style of the image samples. In this work, we propose a novel Disentangled Feature Representation framework, dubbed DFR, for few-shot learning applications. DFR can adaptively decouple the discriminative features that are modeled by the classification branch, from the class-irrelevant component of the variation branch. In general, most of the popular deep few-shot learning methods can be plugged in as the classification branch, thus DFR can boost their performance on various few-shot tasks. Furthermore, we propose a novel FS-DomainNet dataset based on DomainNet, for benchmarking the few-shot domain generalization tasks. We conducted extensive experiments to evaluate the proposed DFR on general and fine-grained few-shot classification, as well as few-shot domain generalization, using the corresponding four benchmarks, i.e., mini-ImageNet, tiered-ImageNet, CUB, as well as the proposed FS-DomainNet. Thanks to the effective feature disentangling, the DFR-based few-shot classifiers achieved the state-of-the-art results on all datasets.

Introduction

While deep neural networks achieved superior results on image classification via supervised learning from large-scale datasets, it is challenging to classify a query sample using only few labelled data, which is known as *few-shot classification* (Fei-Fei, Fergus, and Perona 2006). How to learn the discriminative feature representation that can be generalized from the training set to new classes in testing is critical for few-shot tasks. Popular few-shot methods applied ***meta-learning*** (Vinyals et al. 2016) by episodic training from a large amount of simulated meta-tasks, to obtain a task-specific feature embedding associated with a distance metric (e.g., cosine or Euclidean distance) for classification.

In practice, many excursive features of image data, e.g., style, domain and background, are typically class-irrelevant.

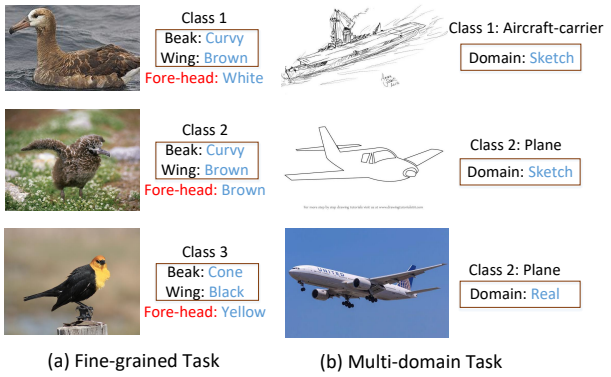


Figure 1: Excursive features (highlighted in boxes) that distract few-shot classification for fine-grained and multi-domain tasks.

Figure 1 shows two such examples in fine-grained and multi-domain classification tasks, respectively, which are challenging for few-shot learning: (1) Only the subtle traits are critical to characterize and differentiate the objects of fine-grained classes; (2) The style and domain information dominates the image visual presence, but they are in fact the excursive and class-irrelevant features. As the subtle traits vary in different simulated meta-tasks, they can hardly be preserved by the learned embedding. On the contrary, the excursive features usually distract the feature embedding (Tokmakov, Wang, and Hebert 2019; Zhang et al. 2020), leading to the degraded few-shot classification results. To rectify such limitations, most recent few-shot methods attempted to suppress excursive features or propose proper metrics, e.g., LCR (Tokmakov, Wang, and Hebert 2019), DeepEMD (Zhang et al. 2020), FEAT (Ye et al. 2020) and CNL (Zhao et al. 2021). However, none of the existing methods explicitly extract the class-specific representation from the excursive image features.

In this paper, we present a novel approach to incorporate deep disentangling for few-shot image classification. Such approach can selectively extract the subtle traits for each task, while maintaining the model generalization. First, we propose a novel Disentangled Feature Representation (DFR)

*Bihan Wen is the corresponding author.

framework which can be applied to most few-shot learning methods. DFR contains two branches: the *classification branch* extracts the discriminative features of the image sample, while the *variation branch* encodes the class-irrelevant information that complements the image representation. A RelationNet (Sung et al. 2018) is applied in the variation branch to measure the feature similarity of each sample pair. A hybrid loss is applied for training DFR, including a reconstruction loss to ensure image information preservation, as well as the translation, discriminative and cross-entropy losses for class-specific feature disentangling. At the inference stage, only the disentangled features from the classification branch are used for class prediction. Second, we integrate the proposed DFR framework into representative baselines for few-shot classification, including the popular ProtoNet (Snell, Swersky, and Zemel 2017) and the state-of-the-art DeepEMD (Zhang et al. 2020) and FEAT (Ye et al. 2020), to carefully investigate the behaviour of DFR with feature visualization and analysis. Extensive experiments are conducted on a set of few-shot tasks, i.e., general image classification, fine-grained image classification, and domain generalization over four benchmarks to demonstrate the effectiveness of our DFR framework.

Our main contributions are summarized as follows:

- We propose a novel disentangled feature representation (DFR) framework, which can be easily applied to most of the few-shot learning methods to extract class-specific features from excessive information.
- We propose a novel benchmark named FS-DomainNet based on DomainNet (Peng et al. 2019) and fully study the few-shot domain generalization task with two evaluation settings.
- We evaluate the DFR framework over four few-shot benchmarks, i.e., mini-ImageNet, tiered-ImageNet, CUB-200-2011, and the proposed FS-DomainNet dataset. Results show that incorporating DFR into existing few-shot algorithms, including both baseline and state-of-the-art methods, can generate consistent gains for multi few-shot classification tasks under both 5-way 1-shot and 5-way 5-shot settings.

Related Work

Few-shot Learning

According to the meta-learning framework (Vinyals et al. 2016), there are mainly three types of few-shot learning methods. Firstly, the gradient-based methods utilize a good model initialization (Finn, Abbeel, and Levine 2017; Nichol, Achiam, and Schulman 2018) or optimization strategy (Ravi and Larochelle 2017; Rusu et al. 2019; Lee et al. 2019; Liu, Schiele, and Sun 2020) to quickly adapt to novel tasks. Secondly, the data augmentation-based methods focus on generating (Gidaris and Komodakis 2019; Li et al. 2020a) or gathering augmented data (Hariharan and Girshick 2017; Wang et al. 2018; Yang, Liu, and Xu 2021) to enable classification from limited samples. In this work, we focus on the third type, namely the metric learning-based methods, i.e., to learn the discriminative feature embedding for distinguishing different image classes. For example, ProtoNet (Snell, Swersky, and Zemel 2017) considered the class-mean representation as the prototype of each class and applied the Euclidean distance metric for classification. LCR (Tokmakov, Wang, and Hebert 2019) applied the subspace-based embedding for each class and DeepEMD (Zhang et al. 2020) adopted the earth mover's distance as the metric function to compare the similarity between two feature maps in a structured way. FEAT (Ye et al. 2020) defined four kinds of set-to-set transformation including self-attention transformer (Jaderberg et al. 2015) to learn a task-specific feature embedding for few-shot learning. Based on prior knowledge, COMET (Cao, Brbic, and Leskovec 2021) mapped some high-level visual concepts into a semi-structured metric space and then learned an ensemble classifier by combining the outputs of independent concept learners. Tang et al. (Tang, Wertheimer, and Hariharan 2020) also uses a semi-structured feature space based on independent prior knowledge concepts to do pose normalization for fine-grained tasks. Our work does not intend to propose new metrics, but focuses on extracting the class-specific features from the variations distracting the metric learning, thus to improve few-shot classification.

Disentangled Feature Representations

Disentangled feature representation aims to learn an interpretable representation for image variants, which has been widely studied in tasks such as face generation (Chen et al. 2016), style translation (Lee et al. 2020; Liu et al. 2019), image restoration (Li et al. 2020b), video prediction (Hsieh et al. 2018) and image classification (Prabhudesai et al. 2021; Li et al. 2021). InfoGAN (Chen et al. 2016) applied an unsupervised method to learn interpretable and disentangled representations by maximized mutual information. DRIT (Lee et al. 2020) embedded images into a content space and a domain-specific attribute space and applied a cycle consistency loss for style translation. FDR (Li et al. 2020b) applied channel-wise feature disentanglement to reduce the interference between hybrid distortions for hybrid-distorted image restoration. Li et al. (Li et al. 2021) proposed a disentangled-VAE to excavate category-distilling information from visual and semantic features for generalized zero-shot learning.

It is noteworthy that the very recent D3DP (Prabhudesai et al. 2021) also adopted a feature disentangling scheme for few-shot detection and VQA, by dividing high-dimensional data (e.g., RGB-D) into individual objects and other attributes. However, our DFR significantly differs from D3DP in the following aspects: DFR is a general-purpose feature extractor for image classification, while D3DP only disentangles individual object to tackle more specific object detection and VQA in 3D scenes. Besides, our DFR works on real images from few-shot benchmarks while D3DP only works with synthetic scenes. Moreover, the DFR framework is much simpler comparing to D3DP with fewer parameters and more efficient algorithms. Thus DFR can serve an enhanced feature extractor for classic backbones that are widely used in most of the existing few-shot methods.

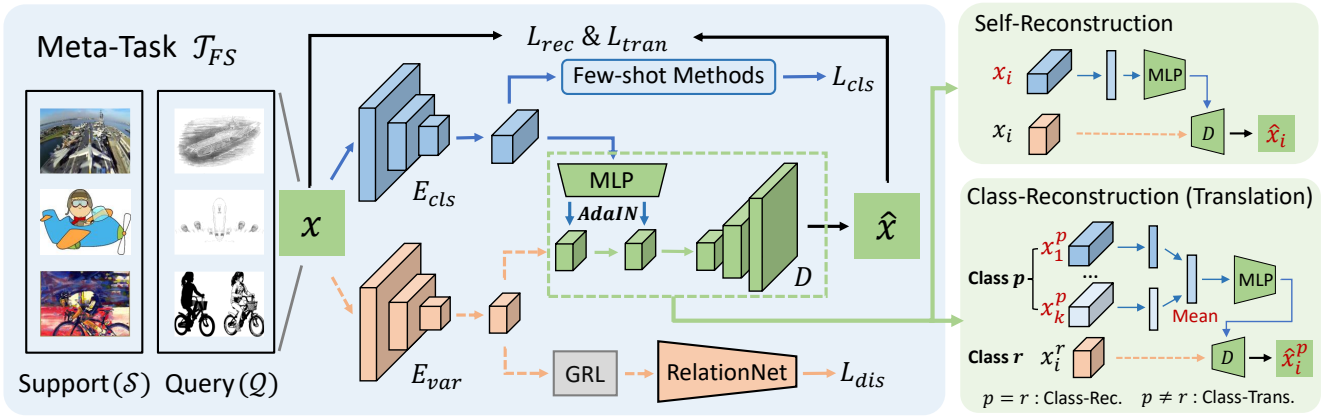


Figure 2: DFR for few-shot image classification: Given a few-shot meta-task $\mathcal{T}_{FS}$ with support ($\mathcal{S}$) and query ($\mathcal{Q}$) image sets, the encoders (E_{cls} and E_{var}) of the classification and variation branches extract the class-specific and class-irrelevant features, respectively. E_{cls} is a classic backbone used in the few-shot methods (e.g., ResNet-12 in this work) which follows the blue stream. The output of E_{cls} is used for few-shot classification, and the output of E_{var} is fed to the RelationNet with a gradient reverse layer (GRL) to remove any class-related information. An MLP block extracts the class information from the classification branch to guide the image reconstruction and translation. Specifically, the Decoder D can achieve self-reconstruction, class-reconstruction, or class-translation according to the different inputs of MLP.

Proposed Method

In this section, we start with a brief introduction to few-shot learning. Then the proposed DFR framework is explained in detail, followed by the loss function of our model and why our DFR works well.

Problem Definition

Given a training image set with the base classes $\mathcal{C}_{train}$, few-shot image classification task aims to predict the novel classes $\mathcal{C}_{test}$ from the testing set, i.e., $\mathcal{C}_{train} \cap \mathcal{C}_{test} = \emptyset$. Thus, the trained classifier from $\mathcal{C}_{train}$ needs to be generalized to $\mathcal{C}_{test}$ in the testing stage with only few labeled samples. In this paper, we follow the meta-learning strategy (i.e., the N-way K-shot setting) (Vinyals et al. 2016) to simulate meta-tasks in the training set that are similar to the few-shot setting at the testing stage, i.e., each meta-task $\mathcal{T}_{FS}$ contains a support set $\mathcal{S}$, and a query set $\mathcal{Q}$. The support set $\mathcal{S}$ contains N classes with K labeled samples (N and K are both very small) and a query set $\mathcal{Q}$ with unlabeled query samples from N classes is used to evaluate the performance.

Disentangled Feature Representation

Figure 2 is an overview of the proposed DFR framework. With a few-shot task $\mathcal{T}_{FS}$ of support set $\mathcal{S}$ and query set $\mathcal{Q}$, the objective is to extract discriminative features for classification from the excursive information of each image x_i. The proposed DFR consists of two branches with two encoders, i.e., E_{cls} and E_{var} for the classification and variation branches, respectively, and one decoder D, as well as a discriminator with a gradient reverse layer and a relation module.

Classification Branch. In principle, any classic metric-based backbone for few-shot learning can be applied as E_{cls} in this branch, to extract the class-specific features of each x_i. In this work, the commonly-used ResNet-12 backbone is adopted as E_{cls}, and the classifier $f(\cdot)$ varies for different few-shot learning baselines being used (e.g., ProtoNet (Snell, Swersky, and Zemel 2017), DeepEMD (Zhang et al. 2020) and FEAT (Ye et al. 2020) are applied in this work, with the corresponding models denoted as +DFR). Therefore, the query sample $x_i^{\mathcal{Q}}$ can be classified based on the support samples $x^{\mathcal{S}}$ as

$$\hat{y}_i = f(E_{cls}(x_i^{\mathcal{Q}}); \{E_{cls}(x^{\mathcal{S}}), y^{\mathcal{S}}\}). \quad (1)$$

Variation Branch. The role of the variation branch is to encode the class-irrelevant information of image samples, which consists of an encoder E_{var} followed by a discriminator. The feature map dimension (i.e., $h \times w$) of $E_{var}(x_i)$ is set to be higher than that of $E_{cls}(x_i)$, to contain more excursive image features. Moreover, we apply instance normalization (IN) and adaptive instance normalization (AdaIN) instead of batch normalization to achieve information transfer for variation encoder E_{var} and decoder D, respectively. The discriminator is formed by a gradient reversal layer (GRL) and a RelationNet r_φ (Sung et al. 2018) to measure the variation feature similarity between any two samples. To be specific, the GRL acts as an identity transform in forward pass, and it multiples the gradient from the subsequent level by a constant -λ during back-propagation. In training, we construct positive and negative pairs from the meta-task by their real labels. The relation module r_φ outputs a score $s_i \in [0, 1]$ indicating the probability that the pair x_{i1} and x_{i2} are from the same class as

$$s_i = r_\varphi \left(E_{var}(x_{i1}), E_{var}(x_{i2})\right). \quad (2)$$

Decoder Module. To preserve the image information and achieve feature disentanglement, a decoder module with a

MLP module and a decoder network D combines the classification and variation branches for image reconstruction and translation.

To be specific, the output feature of the classification branch is fed to the MLP module g to extract class-specific information (μ, σ) of each sample for scaling the feature of the variation branch in the follow-up decoder. The decoder can reconstruct or translate the source image based on different sources of (μ, σ), shown in Figure 2, as

$$\hat{x}_i = D(E_{var}, g(X)), \tag{3}$$

where X can be the feature of the i-th sample itself for self-reconstruction, or the mean feature of class y_i for class-reconstruction or another class y_j with $j \neq i$ for class-translation.

Loss Function

The objective function consists of the discriminative loss L_{dis}, cross-entropy loss L_{cls}, reconstruction loss L_{rec} and translation loss L_{tran}.

Discriminative Loss. To remove the class-specific information in the variation branch, we incorporate the binary cross-entropy loss to optimize the variation feature maps based on the score of RelationNet as

$$L_{dis} = -\sum_{i=1}^{P} \left(l_i \cdot log(s_i) + (1 - l_i) \cdot log(1 - s_i)\right), \tag{4}$$

where P denotes the number of training pairs, s_i is the relation score of the i-th pair which is calculated by (2), and $l_i = 0$ or 1 indicates the ground truth whether the i-th training pair is positive. We minimize L_{dis} in training, and apply GRL to reverse the gradient during back-propagation to achieve feature disentangling, i.e., to minimize the class-specific information captured by the variation feature.

Cross-Entropy Loss. To preserve class-related features for few-shot classification, we minimize the cross-entropy loss L_{cls} for the classification branch for query samples of all classes as

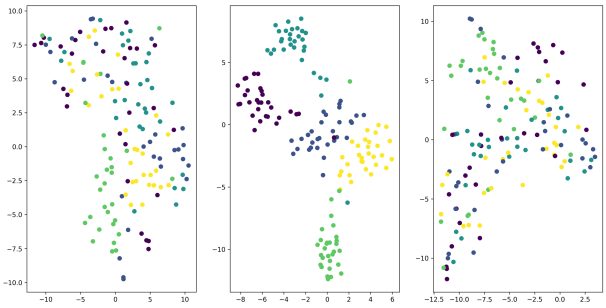
$$L_{cls} = -\sum_{i=1}^{Q} y_i \log P\left(\hat{y}_i = y_i \mid \mathcal{T}_{FS}\right), \tag{5}$$

where Q is the number of query samples in a meta-task $\mathcal{T}$, y_i and $\hat{y}_i$ denote the true and predicted class label of each query sample x_i, respectively.

Reconstruction and Translation Loss. To ensure that the disentangled classification and variation features can jointly restore the input image, an ℓ_1-norm penalty for image reconstruction and a perceptual loss (Johnson, Alahi, and Fei-Fei 2016) are applied after decoding for self-reconstruction and class-reconstruction as

$$L_{rec} = \frac{1}{M}\sum_{i=1}^{M} \|x_i - \hat{x}_i\|_1 + \frac{1}{M}\sum_{i=1}^{M} \|\phi(x_i) - \phi(\hat{x}_i^{c_i})\|_1, \tag{6}$$

where M denotes the number of samples in a meta-task $\mathcal{T}_{FS}$. $\hat{x}_i$ and $\hat{x}_i^{c_i}$ are the reconstructed images of x_i based on the feature of the i-th sample itself and mean feature c_i of class y_i using (3), respectively.



(a) w/o DFR (b) Classification Branch (c) Variation Branch

Figure 3: The t-SNE visualization of the feature representations: (a) the learned features of ResNet-12 backbone for methods w/o DFR, (b) the output features of the classification branch, and (c) the output features of the variation branch.

Moreover, the perceptual loss is also adapted to measure perceptual differences between the output image $\hat{x}_i^{c_j}$ and the support set of the j-th class for class-translation to achieve feature disentanglement as

$$L_{tran} = \frac{1}{N}\sum_{i=1}^{M}\sum_{l=1}^{K} \|\phi(x_l^{\mathcal{S}_j}) - \phi(\hat{x}_i^{c_j})\|_1. \tag{7}$$

The total loss for training DFR can be formulated as

$$L_{total} = \lambda_1 \cdot L_{dis} + \lambda_2 \cdot L_{rec} + \lambda_3 \cdot L_{tran} + L_{cls}, \tag{8}$$

where λ_1, λ_2 and λ_3 denote the weights parameters of L_{dis}, L_{rec} and L_{tran} relative to L_{cls}, respectively.

Why It Works

DFR framework aims to extract only class-related information for classification. Different from other attempts towards more adaptive embedding using attention mechanism (Hou et al. 2019; Li et al. 2019; Ye et al. 2020), our classification and variation branches play as the adversaries by minimizing L_{cls} and L_{dis} simultaneously. In practice, the classification and variation features of an image are always complementary, thus the image reconstruction quality is enforced after fusion by minimizing L_{rec}. It is essential to preserve the image representation in DFR for few-shot classification: as the class-specific features can be task-varying thus hard to be generalized, any information loss throughout the inter-state flow may potentially limit the model performance. Such design is contrast to the classic feature embedding for few-shot learning, in which image features are always projected onto the lower-dimensional manifolds (Simon et al. 2020). Our E_{cls} feature has much lower dimension comparing to the E_{var} feature, as the class-irrelevant information (e.g., image style and background) are typically excessive. To this end, a more restrictive classification feature will significantly reduce the model bias, thus to enhance its generalizability in few-shot tasks.

We visualize the feature distributions w/o and w/ DFR using t-SNE (Van der Maaten and Hinton 2008) to verify our

Method	mini-ImageNet		tiered-ImageNet	
	5-way 1-shot	5-way 5-shot	5-way 1-shot	5-way 5-shot
TADAM (Oreshkin, López, and Lacoste 2018)	58.50 ± 0.30	76.70 ± 0.30	-	-
AFHN (Li et al. 2020a)	62.38 ± 0.72	78.16 ± 0.56	-	-
MetaOptNet (Lee et al. 2019)	62.64 ± 0.82	78.63 ± 0.46	65.99 ± 0.72	81.56 ± 0.53
DSN (Simon et al. 2020)	62.64 ± 0.66	78.83 ± 0.45	66.22 ± 0.75	82.79 ± 0.48
MatchNet (Vinyals et al. 2016)	63.08 ± 0.80	75.99 ± 0.60	68.50 ± 0.92	80.60 ± 0.71
E^{3}BM (Liu, Schiele, and Sun 2020)	63.80 ± 0.40	80.10 ± 0.30	71.20 ± 0.40	85.30 ± 0.30
CAN (Hou et al. 2019)	63.85 ± 0.48	79.44 ± 0.34	69.89 ± 0.51	84.23 ± 0.37
CTM (Li et al. 2019)	64.12 ± 0.82	80.51 ± 0.13	68.41 ± 0.39	84.28 ± 1.73
P-Transfer (Shen et al. 2021)	64.21 ± 0.77	80.38 ± 0.59	-	-
RFS (Tian et al. 2020)	64.82 ± 0.60	82.14 ± 0.43	71.52 ± 0.69	86.03 ± 0.49
ConstellationNet (Xu et al. 2020)	64.89 ± 0.23	79.95 ± 0.17	-	-
FRN (Wertheimer, Tang, and Hariharan 2021)	66.45 ± 0.19	82.83 ± 0.13	71.16 ± 0.22	86.01 ± 0.15
infoPatch (Liu et al. 2021)	67.67 ± 0.45	82.44 ± 0.31	71.51 ± 0.52	85.44 ± 0.35
ProtoNet (Snell, Swersky, and Zemel 2017)	61.83 ± 0.20	79.86 ± 0.14	66.84 ± 0.23	84.54 ± 0.16
ProtoNet + DFR	**64.84 ± 0.20**	**81.10 ± 0.14**	**70.22 ± 0.23**	**84.74 ± 0.16**
DeepEMD (Zhang et al. 2020)	64.93 ± 0.29	81.73 ± 0.57	70.47 ± 0.33	84.76 ± 0.61
DeepEMD + DFR	**65.41 ± 0.28**	**82.18 ± 0.55**	**71.56 ± 0.31**	**86.23 ± 0.58**
FEAT (Ye et al. 2020)	66.52 ± 0.20	81.46 ± 0.14	70.30 ± 0.23	84.55 ± 0.16
FEAT + DFR	**67.74 ± 0.86**	**82.49 ± 0.57**	**71.31 ± 0.93**	**85.12 ± 0.64**

Table 1: Few-shot classification accuracy (%) averaged over mini-ImageNet and tiered-ImageNet with the ResNet backbone.

intuition. Figure 3 (a) shows that the learned features extracted from the ResNet-12 backbone are less discriminative without using the DFR framework. While when applying the DFR framework, the classification branch clusters in Figure 3 (b) are more separable from each other, and the output features of the variation branch in Figure 3 (c) contains more class-irrelevant information that meets our expectations.

Experiment

We conduct extensive experiments on two few-shot benchmarks, i.e., Mini-ImageNet and Tiered-ImageNet on general few-shot classification tasks to evaluate the performance of our proposed DFR framework. After that, we introduce a novel FS-DomainNet dataset with the proposed two evaluation settings for benchmarking few-shot domain generalization task (FS-DG). Moreover, we evaluate the performance of DFR on CUB-200-2011 benchmark on fine-grained few-shot classification task. [1].

Implementation Details

We use ResNet-12 network (Lee et al. 2019) as our backbone E_{cls} for classification branch and set the number of channels as $[64, 160, 320, 640]$, which are similar to the competing methods. The encoder E_{var} consists of four convolutional blocks and two residual blocks. The decoder contains a two-layer MLP block and a decoder network D with residual blocks and upscale convolutional blocks. The I/O channel numbers of variation encoder and decoder are all set to 128. The level ratio λ of GRL layer is set to 1.

[1]The code of the proposed DFR model and FS-DomainNet dataset will be available on https://github.com/chengcv/DFRFS.

Data augmentation including resizing, random cropping, color jitter and random flipping following (Ye et al. 2020) are applied for all methods in training. Our models are all trained using SGD optimizer, with the weight decay as $5e{-}4$, and the momentum as 0.9.

We conduct experiments under both 5-way 1-shot and 5-way 5-shot settings with 15 query images each class, i.e., $5 \times (1\,and\,5) + 5 \times 15$ samples for 1-shot and 5-shot tasks, respectively. We report the mean accuracy of randomly sampled $10k$ tasks as well as the 95% confidence intervals on the testing set as mentioned in (Ye et al. 2020; Zhang et al. 2020). To verify the effectiveness of our proposed DFR framework, we combined DFR with three few-shot algorithms: a commonly used baseline ProtoNet (Snell, Swersky, and Zemel 2017), two state-of-the-art methods DeepEMD (Zhang et al. 2020) and FEAT (Ye et al. 2020).[2] Note that we only adopt the FCN version of DeepEMD for comparison over all datasets.

General Few-shot Classification

We first conduct experiments on two general few-shot benchmarks: mini-ImageNet and tiered-ImageNet.

Mini-ImageNet. Mini-ImageNet (Vinyals et al. 2016) is a subset of the ILSVRC-12 challenge (Krizhevsky, Sutskever, and Hinton 2012) proposed for few-shot classification. It contains 100 diverse classes with 600 images of size $84 \times 84 \times 3$ in each category. Following the class split setting (Ravi and Larochelle 2017) used in previous works, all

[2]We utilized the official codes released by the authors, for implementations of ProtoNet, DeepEMD and FEAT and the corresponding DFR models. The results are all obtained by following the unified setting for fair comparison, which may not exactly match with the results reported in their original papers.

Data Split	Class		Domain	
	Train	Test	Source	Target
Classic DG Setting	–	–	$\triangle$	$\diamondsuit$
Classic FS Setting	$\triangle$	$\mathcal{S}, \mathcal{Q}$	–	–
FS-DG Setting A	$\triangle$	$\mathcal{S}, \mathcal{Q}$	$\triangle$	$\mathcal{S}, \mathcal{Q}$
FS-DG Setting B	$\triangle$	$\mathcal{S}, \mathcal{Q}$	$\triangle$ $\mathcal{S}$	$\mathcal{Q}$

Table 2: Comparison of different settings for DG and FS tasks. $\triangle$: training data selection for DG and FS tasks. $\diamondsuit$: testing data selection for DG tasks. $\mathcal{S}, \mathcal{Q}$: FS support and query data selection. The general FS and DG tasks do not split the domain and class sets, respectively.

100 classes are divided into 64, 16 and 20 classes for training, validation and testing, respectively.

Tiered-ImageNet. Similar to mini-ImageNet, Tiered-ImageNet (Ren et al. 2018) is also a subset of ILSVRC-12, which contains more classes that are organized in a hierarchical structure, i.e., 608 classes from 34 top categories. We follow the setups proposed by (Ren et al. 2018), and split 608 categories into 351, 97 and 160 for training, validation and testing, respectively.

The classification results are shown in Table 1. It is clear that FEAT+DFR achieves state-of-the-art results on mini-ImageNet benchmark, while DeepEMD+DFR achieves state-of-the-art results on tiered-ImageNet benchmark. Moreover, we observe that the improvements by DFR remain inconsistent for all baselines. By adopting the DFR framework, the 5-way 1-shot accuracies by ProtoNet are increased by 3.0% and 3.4% on mini-ImageNet and tiered-ImageNet, respectively, which are even comparable to more sophisticated methods. For the other two methods DeepEMD and FEAT, which are the current state-of-the-art FS methods, their FS classification results can still be further boosted by 1% in average, after applying the DFR framework.

Few-shot Domain Generalization

Domain generalization (DG) aims to learn a domain-agnostic model from multiple sources that can classify data from any target domain. DG tasks become more challenging when there exists a class gap (besides domain gap) between the training and testing sets, i.e., DG under the few-shot setting. General few-shot learning does not consider the influence caused by the domain gap, thus the DG models can hardly be generalized to unseen domains. In this work, we consider a more challenging **Few-Shot Domain Generalization (FS-DG)** problem, i.e., both domain and class gaps exist between the training (source) and testing (target) sets. In our FS-DG experiments (under both the Setting A and B), only the training samples from the source domains are selected in training. Specifically, an N-way K-shot FS-DG task contains support and query samples from N classes on the source domain in the meta-training step, and then the trained model is to predict the query data label out of the testing classes on the target domain. Here we propose two

Method	Setting A		Setting B	
	1-shot	5-shot	1-shot	5-shot
MatchNet	45.23	54.92	40.61	49.09
ProtoNet	47.96	66.64	48.70	67.96
ProtoNet+DFR	**49.29**	**68.73**	**49.76**	**70.34**
DeepEMD	53.20	70.59	51.97	70.62
DeepEMD+DFR	**54.47**	**71.60**	**54.06**	**72.33**
FEAT	51.83	69.26	52.46	71.54
FEAT+DFR	**52.58**	**69.93**	**54.75**	**71.91**

Table 3: FS-DG Classification accuracy ($\%$) averaged on FS-DomainNet with two evaluation settings under the 5-way setting.

FS-DG evaluation settings based on **different domains of support set** $\mathcal{S}$ as: (1) Setting A: Support set is only from the target domain and (2) Setting B: Support set is only from the source domain. Both settings can evaluate the generalizability of the model, i.e., ability to extract domain-invariant and class-specific features. Recent works (Ye et al. 2020; Du et al. 2021) also attempted simple FS-DG tasks to evaluate their proposed FS models. However, only preliminary results are reported following the simple setting (i.e., Setting B in Table 2) without comprehensive investigation of the effect of domain gap on novel classes (test class set). We further conduct experiments with full evaluation settings to validate the proposed DFR for FS-DG tasks using a novel FS-DomainNet Benchmark.

FS-DomainNet Benchmark. We propose FS-DomainNet for benchmarking few-shot domain generalization. Different from the few-shot domainnet (Du et al. 2021) which only contains 200 classes with 1000 images each class, FS-DomainNet captures a much larger subset of DomainNet (Peng et al. 2019), i.e., 569010 images from six distinct domains (i.e., Sketch, Quickdraw, Real, Painting, Clipart and Infograph) with 345 different categories of objects from 24 divisions. We reorganize it for few-shot learning and select all categories (i.e., 527156 images of 299 classes) that include at least the number of samples (i.e., 20) required by the 5-shot setting on each domain. Then we split 299 categories into 191, 47 and 61 for training, validation and testing, respectively, while maintaining the consistency of class split on each domain. More detailed descriptions and data examples of FS-DomainNet are included in our Supplementary Materials. Different from existing few-shot benchmarks, FS-DomainNet additionally includes objects that are collected from multiple domains considering both the domain and class gaps, and the sample size varies greatly between different categories to enable more challenging FS-DG task settings. Additionally, FS-DomainNet can also be utilized for the few-shot domain adaptation and general few-shot classification tasks.

Experimental Setups. Following the classic DG setting, we choose five out of six domains from FS-DomainNet as the source domains and the remaining one as the target domain. We report the average FS-DG accuracies over the

Method	CUB	
	5-way 1-shot	5-way 5-shot
RelationNet (Sung et al. 2018)	66.20 ± 0.99	82.30 ± 0.58
MAML (Finn, Abbeel, and Levine 2017)	67.28 ± 1.08	83.47 ± 0.59
MatchNet (Vinyals et al. 2016)	71.87 ± 0.85	85.08 ± 0.57
COMET (Cao, Brbic, and Leskovec 2021)	72.20 ± 0.90	87.60 ± 0.50
P-Transfer (Shen et al. 2021)	73.88 ± 0.92	87.81 ± 0.48
ProtoNet (Snell, Swersky, and Zemel 2017)	72.25 ± 0.21	87.47 ± 0.13
ProtoNet+DFR	**73.52 ± 0.21**	**87.90 ± 0.13**
DeepEMD (Zhang et al. 2020)	74.88 ± 0.30	88.52 ± 0.52
DeepEMD+DFR	**76.78 ± 0.29**	**89.19 ± 0.52**
FEAT (Ye et al. 2020)	75.68 ± 0.20	87.91 ± 0.13
FEAT+DFR	**77.14 ± 0.21**	**88.97 ± 0.13**

Table 4: Fine-grained few-shot classification accuracy (%) averaged on CUB with the ResNet backbone.

splits with each of the six domains as the target domain.

For 1-shot tasks, we randomly select one support sample only from one random domain for each class; For 5-shot tasks, we select one labeled sample of each source domain for each class, i.e., each meta-task contains the same support samples of each domain. For query samples of each class with multi-domains under both 1- and 5-shot settings, we select the same number of query samples from each domain, i.e., $\|Q\| = 3 \times 5 = 15$.

Results. Table 3 shows the average accuracy of six target domains on the FS-DomainNet benchmark for two evaluation settings. It is clear that DFR can provide consistent improvement on classification accuracies for all FS baselines under both settings. Besides, DFR provides more significant boosting for FS-DG performance under the Setting B, thanks to its effective disentanglement of class-specific features. Comparing to 5-shot tests, DFR provides less help for 1-shot tests, as learning from only one support sample from a random source domain for each category in meta-task is more challenging.

It is worth noting that both ProtoNet and FEAT perform better under the Setting B, while DeepEMD generates better results under the Setting A, comparing to the other setting. It is due to the unique design of DeepEMD by adapting the channel-wise EMD metric based on the feature maps, which inexplicitly incorporates the similarity of domain information. In the FS-DG setting A, the support and query data are from the same domain, which is, in fact, advantageous for DeepEMD, while the domain gap between support and query sets, on the contrary, degrades the DeepEMD performance under the Setting B. After applying the proposed DFR, the feature map in the classification branch removes the interference information, which can always improve DeepEMD under both settings. More experiment results and analysis on the FS-DomainNet dataset can be found in our Supplementary Materials.

Fine-grained Few-shot Classification

We further evaluate DFR on a fine-grained benchmark, i.e., Caltech-UCSD Birds 200-2011 (CUB) (Wah et al. 2011) which was initially proposed for fine-grained image classification, which contains 200 different birds with 11788 images. Following the split in (Chen et al. 2019; Hilliard et al. 2018), 200 classes are divided into 100, 50 and 50 for training, validation and testing, respectively. We also pre-process the data by cropping each image with the provided bounding box according to the prior work (Ye et al. 2020; Wertheimer, Tang, and Hariharan 2021).

Table 4 reports the fine-grained few-shot classification results with both 5-way 1-shot and 5-way 5-shot tests. Comparing to the general and multi-domain few-shot benchmarks that contain significant differences between the categories, fine-grained classification only includes minor intra-class differences. The domain information in the fine-grained dataset may contribute to the category, making it a challenging task. It is clear that the proposed DFR can also significantly and consistently boost all FS baselines, with 0.5% to 1.9% additional improvements on CUB dataset. It demonstrates that DFR can effectively remove the excursive features, and thus highlight the subtle traits which are critical for fine-grained FS classification.

DFR	λ_1	λ_2	λ_3	1-shot	5-shot
✗	-	-	-	66.52	81.46
✓	1.0	-	1.0	66.75	81.98
✓	1.0	1.0	-	66.99	82.16
✓	1.0	1.0	1.0	**67.74**	**82.49**

Table 5: Ablation study on mini-ImageNet dataset of FEAT with the proposed DFR framework.

Ablation Study

We investigate the weights in our formulation and incorporate FEAT as the baseline method. Table 5 shows that FEAT+DFR achieves the best performance when weighting parameters are all set to 1.0. Compared with L_{rec} and L_{tran}, the discriminative loss L_{dis} has a more significant impact on performance as it affects the class-specific information removed from the variation branch, which is directly related to the classification ability of the classification branch. Overall, we find that the performance is minimally affected by loss weight which also shows the robustness of our framework.

Conclusion

We propose a novel and effective Disentangled Feature Representation (DFR) framework for few-shot image classification. Unlike the feature embeddings which may encode the excursive image information, such as background and domain, the proposed DFR aims to extract the class-specific features which is essential in most few-shot learning pipelines. Furthermore, to tackle the challenges of the domain gap in few-shot learning, we propose a novel benchmarking dataset (FS-DomainNet) for the few-shot domain generalization task. We have studied the importance of applying DFR in few-shot tasks by visualizing the t-SNE of the extracted features w/o DFR and disentangled features from the classification and variation branches. Experimental results on four datasets, including three tasks (general image

classification, fine-grained classification, and domain generalization) under the few-shot settings, evaluate the effectiveness of the proposed DFR framework.

References

Cao, K.; Brbic, M.; and Leskovec, J. 2021. Concept Learners for Few-Shot Learning. In *International Conference on Learning Representations*.

Chen, W.-Y.; Liu, Y.-C.; Kira, Z.; Wang, Y.-C. F.; and Huang, J.-B. 2019. A Closer Look at Few-shot Classification. In *International Conference on Learning Representations*.

Chen, X.; Duan, Y.; Houthooft, R.; Schulman, J.; Sutskever, I.; and Abbeel, P. 2016. InfoGAN: interpretable representation learning by information maximizing generative adversarial nets. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, 2180–2188.

Du, Y.; Zhen, X.; Shao, L.; and Snoek, C. G. M. 2021. MetaNorm: Learning to Normalize Few-Shot Batches Across Domains. In *International Conference on Learning Representations*.

Fei-Fei, L.; Fergus, R.; and Perona, P. 2006. One-shot learning of object categories. *IEEE transactions on pattern analysis and machine intelligence*, 28(4): 594–611.

Finn, C.; Abbeel, P.; and Levine, S. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, 1126–1135. PMLR.

Gidaris, S.; and Komodakis, N. 2019. Generating classification weights with gnn denoising autoencoders for few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21–30.

Hariharan, B.; and Girshick, R. 2017. Low-shot visual recognition by shrinking and hallucinating features. In *Proceedings of the IEEE International Conference on Computer Vision*, 3018–3027.

Hilliard, N.; Phillips, L.; Howland, S.; Yankov, A.; Corley, C. D.; and Hodas, N. O. 2018. Few-shot learning with metric-agnostic conditional embeddings. *arXiv preprint arXiv:1802.04376*.

Hou, R.; Chang, H.; Ma, B.; Shan, S.; and Chen, X. 2019. Cross attention network for few-shot classification. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 4003–4014.

Hsieh, J.-T.; Liu, B.; Huang, D.-A.; Fei-Fei, L.; and Niebles, J. C. 2018. Learning to decompose and disentangle representations for video prediction. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 515–524.

Jaderberg, M.; Simonyan, K.; Zisserman, A.; and Kavukcuoglu, K. 2015. Spatial transformer networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems-Volume 2*, 2017–2025.

Johnson, J.; Alahi, A.; and Fei-Fei, L. 2016. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, 694–711. Springer.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, 1097–1105.

Lee, H.-Y.; Tseng, H.-Y.; Mao, Q.; Huang, J.-B.; Lu, Y.-D.; Singh, M.; and Yang, M.-H. 2020. Drit++: Diverse image-to-image translation via disentangled representations. *International Journal of Computer Vision*, 128(10): 2402–2417.

Lee, K.; Maji, S.; Ravichandran, A.; and Soatto, S. 2019. Meta-learning with differentiable convex optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10657–10665.

Li, H.; Eigen, D.; Dodge, S.; Zeiler, M.; and Wang, X. 2019. Finding task-relevant features for few-shot learning by category traversal. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1–10.

Li, K.; Zhang, Y.; Li, K.; and Fu, Y. 2020a. Adversarial feature hallucination networks for few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13470–13479.

Li, X.; Jin, X.; Lin, J.; Liu, S.; Wu, Y.; Yu, T.; Zhou, W.; and Chen, Z. 2020b. Learning Disentangled Feature Representation for Hybrid-distorted Image Restoration. In *European Conference on Computer Vision*, 313–329. Springer.

Li, X.; Xu, Z.; Wei, K.; and Deng, C. 2021. Generalized Zero-Shot Learning via Disentangled Representation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 1966–1974.

Liu, C.; Fu, Y.; Xu, C.; Yang, S.; Li, J.; Wang, C.; and Zhang, L. 2021. Learning a Few-shot Embedding Model with Contrastive Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 8635–8643.

Liu, M.-Y.; Huang, X.; Mallya, A.; Karras, T.; Aila, T.; Lehtinen, J.; and Kautz, J. 2019. Few-shot unsupervised image-to-image translation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10551–10560.

Liu, Y.; Schiele, B.; and Sun, Q. 2020. An Ensemble of Epoch-wise Empirical Bayes for Few-shot Learning. In *European Conference on Computer Vision (ECCV)*.

Nichol, A.; Achiam, J.; and Schulman, J. 2018. On first-order meta-learning algorithms. *arXiv preprint arXiv:1803.02999*.

Oreshkin, B. N.; López, P. R.; and Lacoste, A. 2018. TADAM: Task dependent adaptive metric for improved few-shot learning. In *NeurIPS*.

Peng, X.; Bai, Q.; Xia, X.; Huang, Z.; Saenko, K.; and Wang, B. 2019. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1406–1415.

Prabhudesai, M.; Lal, S.; Patil, D.; Tung, H.-Y.; Harley, A. W.; and Fragkiadaki, K. 2021. Disentangling 3D Prototypical Networks for Few-Shot Concept Learning. In *International Conference on Learning Representations*.

Ravi, S.; and Larochelle, H. 2017. Optimization as a model for few-shot learning. In *International Conference on Learning Representations*.

Ren, M.; Triantafillou, E.; Ravi, S.; Snell, J.; Swersky, K.; Tenenbaum, J. B.; Larochelle, H.; and Zemel, R. S. 2018. Meta-learning for semi-supervised few-shot classification. In *ICLR*.

Rusu, A. A.; Rao, D.; Sygnowski, J.; Vinyals, O.; Pascanu, R.; Osindero, S.; and Hadsell, R. 2019. Meta-Learning with Latent Embedding Optimization. In *International Conference on Learning Representations*.

Shen, Z.; Liu, Z.; Qin, J.; Savvides, M.; and Cheng, K.-T. 2021. Partial Is Better Than All: Revisiting Fine-tuning Strategy for Few-shot Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 9594–9602.

Simon, C.; Koniusz, P.; Nock, R.; and Harandi, M. 2020. Adaptive subspaces for few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4136–4145.

Snell, J.; Swersky, K.; and Zemel, R. 2017. Prototypical networks for few-shot learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 4080–4090.

Sung, F.; Yang, Y.; Zhang, L.; Xiang, T.; Torr, P. H.; and Hospedales, T. M. 2018. Learning to Compare: Relation Network for Few-Shot Learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Tang, L.; Wertheimer, D.; and Hariharan, B. 2020. Revisiting pose-normalization for fine-grained few-shot recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14352–14361.

Tian, Y.; Wang, Y.; Krishnan, D.; Tenenbaum, J. B.; and Isola, P. 2020. 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*, 266–282. Springer.

Tokmakov, P.; Wang, Y.-X.; and Hebert, M. 2019. Learning compositional representations for few-shot recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6372–6381.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11).

Vinyals, O.; Blundell, C.; Lillicrap, T.; Kavukcuoglu, K.; and Wierstra, D. 2016. Matching networks for one shot learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, 3637–3645.

Wah, C.; Branson, S.; Welinder, P.; Perona, P.; and Belongie, S. 2011. The caltech-ucsd birds-200-2011 dataset.

Wang, Y.-X.; Girshick, R.; Hebert, M.; and Hariharan, B. 2018. Low-shot learning from imaginary data. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7278–7286.

Wertheimer, D.; Tang, L.; and Hariharan, B. 2021. Few-Shot Classification With Feature Map Reconstruction Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8012–8021.

Xu, W.; Wang, H.; Tu, Z.; et al. 2020. Attentional Constellation Nets for Few-Shot Learning. In *International Conference on Learning Representations*.

Yang, S.; Liu, L.; and Xu, M. 2021. Free Lunch for Few-shot Learning: Distribution Calibration. In *International Conference on Learning Representations*.

Ye, H.-J.; Hu, H.; Zhan, D.-C.; and Sha, F. 2020. Few-Shot Learning via Embedding Adaptation with Set-to-Set Functions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 8808–8817.

Zhang, C.; Cai, Y.; Lin, G.; and Shen, C. 2020. DeepEMD: Few-Shot Image Classification With Differentiable Earth Mover's Distance and Structured Classifiers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhao, J.; Yang, Y.; Lin, X.; Yang, J.; and He, L. 2021. Looking Wider for Better Adaptive Representation in Few-Shot Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 10981–10989.