Explanation-Guided Training for Cross-Domain Few-Shot Classification

Jiamei Sun*, Sebastian Lapuschkin†, Wojciech Samek†, Yunqing Zhao*,
Ngai-Man Cheung*, and Alexander Binder*
*Information System of Technology and Design, Singapore University of Technology and Design
†Department of Video Coding & Analytics, Fraunhofer Heinrich Hertz Institute, Berlin, Germany

Abstract—**Cross-domain few-shot classification task (CD-FSC) combines few-shot classification with the requirement to generalize across domains represented by datasets. This setup faces challenges originating from the limited labeled data in each class and, additionally, from the domain shift between training and test sets. In this paper, we introduce a novel training approach for existing FSC models. It leverages on the explanation scores, obtained from existing explanation methods when applied to the predictions of FSC models, computed for intermediate feature maps of the models. Firstly, we tailor the layer-wise relevance propagation (LRP) method to explain the prediction outcomes of FSC models. Secondly, we develop a model-agnostic explanation-guided training strategy that dynamically finds and emphasizes the features which are important for the predictions. Our contribution does not target a novel explanation method but lies in a novel application of explanations for the training phase. We show that explanation-guided training effectively improves the model generalization. We observe improved accuracy for three different FSC models: RelationNet, cross attention network, and a graph neural network-based formulation, on five few-shot learning datasets: miniImagenet, CUB, Cars, Places, and Plantae.**

I. INTRODUCTION

Few-shot classification(FSC) aims at generalization to new categories with a small sample size [1]–[8]. Few-shot learning has two notable properties. Firstly, it aims at generalization across sets of labeled tasks. Secondly, it relies on small sample sizes.

In other settings, the first choice for better generalization would be obviously to label more training data, which by problem definition, is out of the question for few-shot learning. This restriction motivates research for alternatives that can be employed to improve generalization in few-shot learning in lieu of unavailable training data.

FSC models are commonly evaluated using a test dataset originating from the same domain as the training dataset. Lately, [9] states that FSC methods will meet difficulties in cases exhibiting domain shift between the training data (source domain) and the test data (target domain). Tackling the domain shift problem requires additional efforts to avoid overfitting to the source domain. A recent work addresses the domain shift issue by learning a noise distribution for intermediate layers in the feature encoder [10]. Other approaches rely on adding batch spectral regularization over the encoded image features [11] and employing novel losses [9], [12]. This paper proposes a novel approach for improving CD-FSC models

from a different perspective: we leverage on explanations computed for intermediate feature maps of FSC methods to guide the model to learn features with better performance. For explanations, we refer to methods such as gradient- or Shapley-type methods, LRP [13] or LIME [14] that compute a score for every dimension of a feature map, denoting its importance to the final prediction.

Although a large number of explanation methods have contributed substantial progress to the field of explaining model predictions [13]–[21], they are usually for models in the testing phase, and frequently, do not consider the question of use cases for these methods. Known use cases are the audit of predictions [22], explanation-weighed document representations that are more comprehensive [23], and identification of biases in datasets [18]. We will add a new use case for explanations during the training phase, and consider whether the explanations are suitable to improve model performance in small sample size regimes such as few-shot classification.

Many explanation methods [13], [15], [18] explain predictions on a per-sample basis. With a target label and an input sample, these explanation methods assign scores to each element of each feature map within the model. These scores are related to the importance of a neuron to the target label. Explanations are generated usually with a modified backward-pass. Apart from that, they require no additional trainable parameters inside the model. In this paper, we study whether the explanation scores of intermediate feature maps can be employed to improve model generality in the few-shot classification, which is still a novel question.

Concretely, we adapt LRP-type explanations [13] to FSC models. LRP has been used to explain convolutional neural networks (CNN) [13], recurrent neural networks (RNN) [24], graph neural networks(GNN) [25], and clustering models [26]. It backpropagates the relevance score for a target label through the neural network and assigns the relevance scores to the neurons within the network. The sign and the amplitude of LRP relevance scores reflect the contribution of a neuron to the prediction, as shown in Figure 1. Relying on this property, we propose "explanation-guided training" for FSC models. The LRP relevance scores of intermediate feature maps are employed as weights and used to construct LRP-weighted feature maps. This step emphasizes the feature dimensions, which are more relevant to the model prediction, and downscales less relevant features. The LRP-weighted
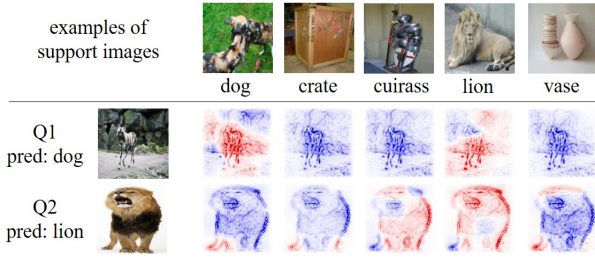
Fig. 1: LRP explanation heatmaps of the input image with 5 target labels. The experiment model is a RelationNet trained on miniImagenet under the 5-way 5-shot setting. The first row illustrates some examples of the support images. The other two rows show the explanation heatmaps of two query images, Q1: *African hunting dog* (denoted as *dog*) and Q2: *lion*. Both images are correctly predicted and the heatmaps are generated using different target labels. Red pixels indicate positive LRP relevance score and blue indicates negative. The strength of the color corresponds to the value of the LRP relevance scores.

features are then fed into the network to guide the training. Since LRP explanations are calculated for each sample-label pair separately, our explanation-guided training adds a label-dependent feature weighting mechanism during training. We will show that this mechanism can reduce overfitting to the source domain.

We remark that the principles used for explanation-guided training strategy are model-agnostic and can be combined with other CD-FSC methods such as the learned feature-wise transformation (LFT) [10] and other explanation methods.

The main contributions of this paper are described as follows.

- We derive explanations for FSC models using LRP.
- We investigate the potential of improving model performance using explanations in the training phase under small sample size settings.
- We propose an explanation-guided training strategy to tackle the domain shift problem in FSC.
- We conduct experiments to show that the explanation-guided training strategy improves the model generalization for a number of FSC models and datasets.
- We combine our explanation-guided training strategy with another recent approach, LFT [10], which shares with our approach the property of being applicable on top of existing models, and observe a synergy of these two by demonstrating that this combination further improves the performance.

II. RELATED WORK

A. Few-shot Classification Methods

Optimization-based and metric-based approaches constitute two prominent directions in few-shot learning. The former one is learning initialization parameters that can be quickly adapted to new categories [2], [6], [7], [27] or are designing a meta-optimizer that learns how to update the model parameters

[28]–[31]. Metric-based methods learn a distance metric to compare the support and query samples and classify the query image to the closest category [1], [3]–[5], [8], [32], [33]. Other approaches are noteworthy. [34], [35] design and add task-conditional layers to the model. [36]–[38] dynamically update the classifier weight for new categories. [39], [40] combine multiple modal information such as the word embedding of the class label. [41] augments the training data by hallucinating new samples. [42], [43] leverage unlabeled training samples and semi-supervising training strategy. [44] equips the model with a self-supervision mechanism. However, recent research discussed that existing FSC methods may meet difficulties with domain shift, a more challenging and practical problem [9].

B. Cross-domain Few-shot Classification Methods

It is common to develop cross-domain few-shot classification methods from existing FSC methods. LFT [10] learns a noise distribution and adds the noise to intermediate feature maps to generate more diverse features during training and improve the model generality. In the recent CVPR Cross-Domain Few-Shot Learning challenges [11], [45] ensembled multiple feature encoders and employed batch spectral regularization over the image features for each encoder. Batch spectral regularization penalizes the singular values of the feature matrix within a batch so that the learned features maintain similar spectra across domains. [46] combined the first-order MAML [2] and the GNN metric-based method [5]. [12] applied a prototypical triplet loss to increase the inter-class distance and a large margin cosine loss to minimize the intra-class distance, which is also studied by [9] that reducing intra-class variation benefits FSC, especially for shallow image feature encoders. In our approach, we do not introduce more parameters like [10]. We are similar to [11] and [12] in adding constraints on the image features. We are different in using LRP-weighted features to guide the model to dynamically correct itself for each instance instead of penalizing feature statistics over a batch. The LRP-weighting idea has been used to generate more comprehensive document representations [23]. We are different from [23] that the re-weighting strategy is embedding into the training phase to improve the model.

C. Explanation for Few-shot Classification Models

There exist explanation methods for deep neural networks(DNN) [13], [17]–[19], [25], [47] that can be adapted to FSC models, since many FSC models adopt CNN to encode image features and many metric-based methods also adopt DNN to learn the distance metric [4], [5], [32]. For FSC models that use non-parametric distance metrics, we refer to [26] that transforms various K-means classifiers into neural net structures and then applies LRP to obtain explanations. In this paper, we have chosen LRP due to its reasonable performance [48], our understanding of its hyperparameters, and its reasonable speed compared to LIME or some theoretically equally well-motivated but exhaustive Shapley-type approaches. While using other explanation methods among the faster ones would be possible, this would not change the

qualitative message of this paper regarding the applicability of explanation methods for few-shot training. The results here are meant as a case for explanation methods in general, even when they are demonstrated for one approach.

III. EXPLANATION-GUIDED TRAINING

Before presenting our explanation-guided training, we first introduce the cross-domain few-shot learning task and the notations. For a K-way N-shot task, denoted as an episode, we are given a support set $\mathcal{S} = \{(x_s, y_s)\}_{s=1}^{K*N}$ containing K classes and N labeled samples per class for training and a query set $\mathcal{Q} = \{(x_q, y_q)\}_{q=1}^{n_q}$ from the same classes as $\mathcal{S}$ for testing. A CD-FSC task is to train an FSC model using episodes $\{\mathcal{S}_i, \mathcal{Q}_i\}$ randomly sampled from a base domain $\mathcal{D}_{seen}$ and test the model with episodes sampled from an unseen domain $\mathcal{D}_{unseen}$. We consider FSC models that can be outlined as Figure 2 in our study. This includes many metric-based FSC models.

The support set $\mathcal{S}$ and query set $\mathcal{Q}$ are encoded by a CNN [4], [8], possibly with augmented layers [10], [34] to obtain the support image features f_s and the query image features f_q. f_s and f_q are further processed before classification, for example, [4] simply averages the f_s over classes and concatenate the averaged class representations pairwise with f_q, [8] designs an attention module and generate the attention-weighted support and query image features, [32] applies GNN on f_s and f_q to obtain graph structured features. The processed features are fed into a classifier for predictions. The classifier can be cosine similarity [8], Euclidean distances [3], Mahalanobis distance [35], or neural nets [4], [5].

Explanation-guided training for FSC models involves the following steps. For each training episode:

Step1: One forward-pass through the model and obtain the prediction p, illustrated as the blue path in Figure 2.

Step2: **Explaining the classifier**. We initialize the LRP relevance for each label and apply LRP to explain the classifier. We can obtain the relevance of the classifier input $R(f_p)$, illustrated as the **Explain** block.

For FSC models that implement a neural network as the classifier, the relevance scores for each label can be initialized with their logits. For the models using non-parametric distance measures such as cosine similarity and Euclidean distance, the predicted scores are positive for all labels, which will result in similar explanations. For such cases, we refer to the logit function in [26] to initialize the relevance scores. Taking the cosine similarity as an example, we first calculate the probability for each class using the exponential function via equation (1) [1].

$$P(y_c|f_p) = \frac{exp(\beta \cdot cs_c(f_p))}{\sum_{k=1}^{K} exp(\beta \cdot cs_k(f_p))} \quad (1)$$

$cs_k(\cdot)$ means the cosine similarity between a query sample and class k. f_p is the processed feature fed to the classifier. β is a

[1]For distance measures such as Euclidean distance, we need to use the negative distance to replace the similarity metric.
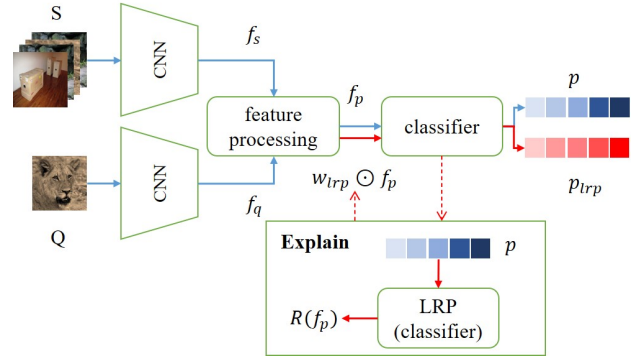

Fig. 2: Explanation-guided training. Blue paths denote the conventional FSC training. The red paths are originating from the explanation method. They are added after one step following the blue paths. The support samples $\mathcal{S}$ and the query sample $\mathcal{Q}$ are fed into an image encoder to obtain features f_s and f_q, which are compared by a *feature processing* module. The output of *feature processing* f_p is fed into a classifier to make predictions. Both the *feature processing* and *classifier* modules vary across different FSC methods. The **Explain** block explains the model prediction p and generate the explanations for f_p, denoted as $R(f_p)$, which are used to calculate the LRP weight w_{lrp}. The LRP-weighted feature $w_{lrp} \odot f_p$ is fed into the classifier resulting in the updated prediction p_{lrp}.

constant scale parameter to strengthen the highest probability. Using the probability defined above, the relevance score of class c is defined as:

$$R_c = log\left(\frac{P(y_c|f_p)}{1 - P(y_c|f_p)}(K-1)\right) \quad (2)$$

$R_c, c = 1 \dots K$ is positive when the $P(y_c|f_p)$ is larger than $1/K$. In other words, the class label whose probability is larger than the random guessing probability receives a positive relevance score. With the relevance score of each target label R_c, standard LRP is applicable to backpropagate R_c through the classifier to generate the explanations.

Consider the forward pass from layer l to layer $l+1$ as:

$$y_j^{l+1} = \sum_i w_{ij} z_i^l + b_j$$
$$z_j^{l+1} = f(y_j^{l+1}) \quad (3)$$

where i and j are the indices of neurons in l^{th} and $l+1^{th}$ layer, $f(\cdot)$ is an activation function. Let $R(\cdot)$ denote the relevance of a neuron and $R_{i \leftarrow j}$ denote the relevance attribution from z_j^{l+1} to z_i^l. We rely on two established LRP backpropagation mechanisms here, the ϵ-rule and the $\alpha\beta$-rule [13].

1) ϵ-**rule**

$$R_{i \leftarrow j} = R(z_j^{l+1}) \frac{z_i^l w_{ij}}{y_j^{l+1} + \epsilon \odot \text{sign}(y_j^{l+1})} \quad (4)$$

ϵ is a small positive number and $\epsilon \text{sign}(y_j^{l+1})$ guarantees safe division.

2) $\alpha\beta$-rule

$$R_{i\leftarrow j} = R(z_j^{l+1}) \left(\alpha \frac{(z_i^l w_{ij})^+}{(y_j^{l+1})^+} - \beta \frac{(z_i^l w_{ij})^-}{(y_j^{l+1})^-} \right) \quad (5)$$

where $\alpha \geqslant 1$ controls the ratio of positive relevance to backpropagate and $\beta = \alpha - 1$. $(y_j^{l+1})^+ = \max(y_j^{l+1}, 0)$ and $(y_j^{l+1})^- = \min(y_j^{l+1}, 0)$.

The relevance of z_i^l is the summation of all the relevance attribution flowing to it.

$$R(z_i^l) = \sum_j R_{i\leftarrow j} \quad (6)$$

We adopt the ϵ-rule for linear layers and the $\alpha\beta$-rule for convolutional layers to obtain $R(f_p)$ [49]. $R(f_p)$ is normalized by its maximal absolute value.

Step3: **LRP-weighted features**. To emphasize the features which are more relevant to the prediction and downscale the less relevant ones, we define the LRP weights and the LRP-weighted features as

$$w_{lrp} = 1 + R(f_p) \quad (7)$$
$$f_{p-lrp} = w_{lrp} \odot f_p \quad (8)$$

where $\odot$ is the element-wise product. Note that $R(f_p) \in [-1, 1]$ after normalization, thus w_{lrp} magnifies the features with positive relevance scores and downscales those with negative relevance scores. The maximal feature scaling after weighting with w_{lrp} is 2.

Step4: Finally, we forward the LRP-weighted features to the classifier to generate the explanation-guided predictions p_{lrp}. The objective function merges both the model prediction p and the explanation-guided prediction p_{lrp}.

$$\mathcal{L} = \xi \mathcal{L}_{ce}(y, p) + \lambda \mathcal{L}_{ce}(y, p_{lrp}) \quad (9)$$

where $\mathcal{L}_{ce}$ is the cross entropy loss. ξ and λ are positive scalars that control how much information from p and p_{lrp} are used. In our experiment, ξ and λ are empirically adjusted for different FSC models.

IV. EXPERIMENTS

We evaluate the proposed explanation-guided training on RelationNet(RN) [4] and two of the state-of-the-art models, cross attention network(CAN) [8] and GNN network [5]. The correspondence of the three FSC models to the framework in Figure 2 is summarized in Table I. We will demonstrate that explanation-guided training improves the performance of the three models on 4 cross-domain test sets.

Moreover, we also combine explanation-guided training with another approach, the learning to learn feature-wise transformation (LFT) [10]. We show that explanation-guided training is compatible with LFT and the combination further improves the performance.

TABLE I: The correspondence of RelationNet(RN), cross attention network(CAN), and graph neural network(GNN) to the framework in Figure 2.

	feature processing	classifier
RN	pairwise concatenation	relation module
CAN	cross attention module	cosine similarity
GNN	*fc* layer and concatenation	graph neural network

A. Dataset and Model Preparation

Five datasets are used in our experiment including miniImagenet [50], CUB [51], Cars [52], Places [53], and Plantae [54], which are introduced in [10]. Each dataset consists of train/val/test splits. We choose miniImagenet as the $\mathcal{D}_{seen}$ and train the FSC models on the training set, validate the models on the validation set of miniImagenet, and adopt the test sets of the other four datasets for testing.

We use Resnet10 [55] as the image encoder for RN and GNN, and Resnet12 for CAN model. The three models are trained under 5-way 5-shot and 5-way 1-shot settings. For RN and GNN models, we set the parameters of explanation-guided training as $\xi = 1, \lambda = 0.5$ for 5-way 5-shot setting and $\xi = 1, \lambda = 1$ for 5-way 1-shot setting. CAN model employs cosine similarity as the classifier, thus we set β in eq(1) as 7, the same as the original model and $\xi = 0, \lambda = 1$ for eq(9). The LRP parameters are $\alpha = 1, \epsilon = 0.001$ for all the experiments.

We follow the same implementation details as [10][2] and [8][3] to train the RelationNet, GNN, and CAN model. At test time, we evaluate the performance over 2000 randomly sampled episodes, with 16 query images per episode.

B. Evaluation for Explanation-Guided Training on Cross-Domain Setting

In this section, we evaluate the performance of RN, GNN, and CAN models trained with and without explanation-guided training on CD-FSC tasks. For more comprehensive analyses, we also implement the **Transductive inference** proposed by [8]. Transductive inference iteratively augments the support set using the confidently classified query images during the test phase. Specifically, we first predict the label of query images with the trained model; second, we choose the query images with higher predicted scores as the candidate images. The candidate images and their predicted label are augmented to the support set. This is an iterative process. In our experiment, we implement the transductive operation for two iterations with 35 candidates for the first iteration and 70 for the second iteration, the same strategy as [8]. GNN requires a fixed number of support images, thus we implement the transductive inference on RN and CAN models.

Table II and Table III summarise the accuracy of the RN, CAN, and GNN models trained with and without explanation-guided training. We can observe a consistent improvement after implementing explanation-guided training on both the seen-domain and the cross-domain test sets. The results are

[2]https://github.com/hytseng0509/CrossDomainFewShot
[3]https://github.com/blue-blue272/fewshot-CAN

TABLE II: Evaluation of explanation-guided training on cross-domain datasets using RN and CAN. We report the average accuracy of over 2000 episodes with 95% confidence intervals. The models are trained on the miniImagenet training set and tested on the test set of various domains. **LRP-** means explanation-guided training using LRP. **T** indicates transductive inference.

miniImagenet	1-shot	1-shot-T	5-shot	5-shot-T
RN	58.31±0.47%	61.52±0.58%	72.72±0.37%	73.64±0.40%
LRP-RN	**60.06±0.47%**	**62.65±0.56%**	**73.63±0.37%**	**74.67±0.39%**
CAN	**64.66±0.48%**	67.74±0.54%	79.61±0.33%	80.34±0.35%
LRP-CAN	64.65±0.46%	**69.10±0.53%**	**80.89±0.32%**	**82.56±0.33%**

mini-CUB	1-shot	1-shot-T	5-shot	5-shot-T
RN	41.98±0.41%	42.52±0.48%	58.75±0.36%	59.10±0.42%
LRP-RN	**42.44±0.41%**	**42.88±0.48%**	**59.30±0.40%**	**59.22±0.42%**
CAN	44.91±0.41%	46.63±0.50%	63.09±0.39%	62.09±0.43%
LRP-CAN	**46.23±0.42%**	**48.35±0.52%**	**66.58±0.39%**	**66.57±0.43%**

mini-Cars	1-shot	1-shot-T	5-shot	5-shot-T
RN	29.32±0.34%	28.56±0.37%	38.91±0.38%	37.45±0.40%
LRP-RN	**29.65±0.33%**	**29.61±0.37%**	**39.19±0.38%**	**38.31±0.39%**
CAN	31.44±0.35%	30.06±0.42%	41.46±0.37%	40.17±0.40%
LRP-CAN	**32.66±0.46%**	**32.35±0.42%**	**43.86±0.38%**	**42.57±0.42%**

mini-Places	1-shot	1-shot-T	5-shot	5-shot-T
RN	**50.87±0.48%**	**53.63±0.58%**	66.47±0.41%	67.43±0.43%
LRP-RN	50.59±0.46%	53.07±0.57%	**66.90±0.40%**	**68.25±0.43%**
CAN	56.90±0.49%	60.70±0.58%	72.94±0.38%	74.44±0.41%
LRP-CAN	**56.96±0.48%**	**61.60±0.58%**	**74.91±0.37%**	**76.90±0.39%**

mini-Plantae	1-shot	1-shot-T	5-shot	5-shot-T
RN	33.53±0.36%	33.69±0.42%	47.40±0.36%	46.51±0.40%
LRP-RN	**34.80±0.37%**	**34.54±0.42%**	**48.09±0.35%**	**47.67±0.39%**
CAN	36.57±0.37%	36.69±0.42%	50.45±0.36%	48.67±0.40%
LRP-CAN	**38.23±0.45%**	**38.48±0.43%**	**53.25±0.36%**	**51.63±0.41%**

TABLE III: Evaluation of explanation-guided training on cross-domain datasets using GNN. We report the average accuracy of over 2000 episodes with 95% confidence intervals. The models are trained on the miniImagenet training set and tested on the test set of various domains. **LRP-** means explanation-guided training using LRP.

5-way 1-shot	miniImagenet	Cars	Places	CUB	Plantae
GNN	64.47±0.55%	30.97±0.37%	54.64±0.56%	46.76±0.50%	37.39±0.43%
LRP-GNN	**65.03±0.54%**	**32.78±0.39%**	**54.83±0.56%**	**48.29±0.51%**	**37.49±0.43%**

5-way 5-shot	miniImagenet	Cars	Places	CUB	Plantae
GNN	80.74±0.41%	42.59±0.42%	72.14±0.45%	63.91±0.47%	**54.52±0.44%**
LRP-GNN	**82.03±0.40%**	**46.20±0.46%**	**74.45±0.47%**	**64.44±0.48%**	54.46±0.46%

also competitive with the recent work on LFT [10] which learns a noise distribution by adding feature-wise transformation layers to the image encoder while explanation-guided training does not introduce more training parameters. To show that our approach exploits a different mechanism to improve the model, we also combine the LFT and our explanation-guided training in the next section.

C. Synergies in Combining Explanation-guided Training with Feature-wise Transformation

To compare and to combine our idea with the LFT method, we apply the explanation-guided training to the multiple domain experiment as [10]. The LFT model is trained using the pseudo-seen domain and pseudo-unseen domains. In our experiment, the miniImagenet is the pseudo-seen domain. Three of the other four datasets are the pseudo-unseen domains and the model is tested on the last domain. The pseudo-unseen domains are used to train the feature-wise transformation layers and the pseudo-seen domain is used to update the other trainable parameters of the model. If the parameters of the feature-wise transformation layers are fixed, we will get the

FT method that adds the noise with a fixed distribution on certain intermediate layers.

The performance of the standard RN, the FT and LFT methods, explanation-guided training, and its combination with LFT are shown in Table IV. These models are trained with the same random seed, learning rate, optimizer, and datasets. The combination of our explanation-guided training and LFT(**LFT-LRP-RN**) achieves the best accuracy. Comparing the results of **FT-RN** and **LRP-RN**, we can see explanation-guided training is even better without introducing more trainable parameters to the model.

We remark that the improvement observed when combining explanation-guided training with LFT shows that both optimize the model from different angles. This demonstrates the independence of both approaches as well as both their strength.

D. Explaining the Effect of Explanation-Guided Training

In this section, we provide an intuition for the improvement of FSC models by explanation-guided training. It is known from the information bottleneck framework that training a discriminative classifier implies learning to filter irrelevant

TABLE IV: The results of multiple domains experiment using RelationNet. We report the average accuracy of over 2000 episodes with 95% confidence intervals. **FT** and **LFT** indicate the feature-wise transformation layer with fixed or trainable parameters. **LRP-** means explanation-guided training using LRP. **LFT-LRP** is the combination of LFT and explanation-guided training.

5-way 1-shot	Cars	Places	CUB	Plantae
RN	29.40±0.33%	48.05±0.46%	44.33±0.43%	34.57±0.38%
FT-RN	30.09±0.36%	48.12±0.45%	44.87±0.44%	35.53±0.39%
LRP-RN	30.00±0.32%	48.74±0.45%	45.64±0.42%	36.04±0.38%
LFT-RN	30.27±0.34%	48.07±0.46%	47.35±0.44%	35.54±0.38%
LFT-LRP-RN	**30.68±0.34%**	**50.19±0.47%**	**47.78±0.43%**	**36.58±0.40%**

5-way 5-shot	Cars	Places	CUB	Plantae
RN	40.01±0.37%	64.56±0.40%	62.50±0.39%	47.58±0.37%
FT-RN	40.52±0.40%	64.92±0.40%	61.87±0.39%	48.54±0.38%
LRP-RN	41.05±0.37%	66.08±0.40%	62.71±0.39%	48.78±0.37%
LFT-RN	41.51±0.39%	65.35±0.40%	64.11±0.39%	49.29±0.38%
LFT-LRP-RN	**42.38±0.40%**	**66.23±0.40%**	**64.62±0.39%**	**50.50±0.39%**

features [56]. This compression of task-irrelevant information is also acknowledged in recent works that shed critical light on the application of the information bottleneck to deep networks [57]. There is a difference between traditional classification and few-shot classification regarding removable information. The removable information means that some intermediate feature channels related to these removable features are not activated. The traditional classification task is to classify a fixed set of classes, therefore removing information irrelevant to these classes will not influence the discriminative capability. For example, a classifier for cat breeds will likely learn rather the features of eyes, tails, and legs than features of sofas or grass. In FSC, the classes vary across episodes. Thus, the irrelevant information of one episode can be discriminative for the next episode. Excessive information removal can be detrimental for FSC that requires generalization across new classes.

From a classifier trained to classify a fixed set of classes, one would expect that in higher-layer feature maps, a few channels are highly activated somewhere in the spatial dimensions, while most channels show only low values overall. Explanation-guided training adopts explanation scores of the predicted class to re-weight intermediate features. If a classifier is overfitting and frequently predicts a wrong class label, then the explanation-guided training will identify the relevant features for the wrongly predicted class(**step2** in SectionIII), upscale them, and the subsequent loss minimization will penalize these upscaled features more (**step3&4** in SectionIII). Thus, it avoids the intermediate features from being too specialized towards a fixed set of classes and achieves better generalization.

To quantify the above intuition, we analyze the CNN encoded image features of the RelationNet, GNN, and CAN models trained with or without explanation-guided training under the 5-way 5-shot setting, the same models as SectionIV-B. We use the test set images of the four cross-domain datasets in this experiment. Each CNN encoded image feature is of a shape $f_{CNN} \in \mathbb{R}^{C \times H \times W}$. We first perform a pooling over the spatial dimensions $[H, W]$, then compute a statistic over channels C, and then average the statistics over the test
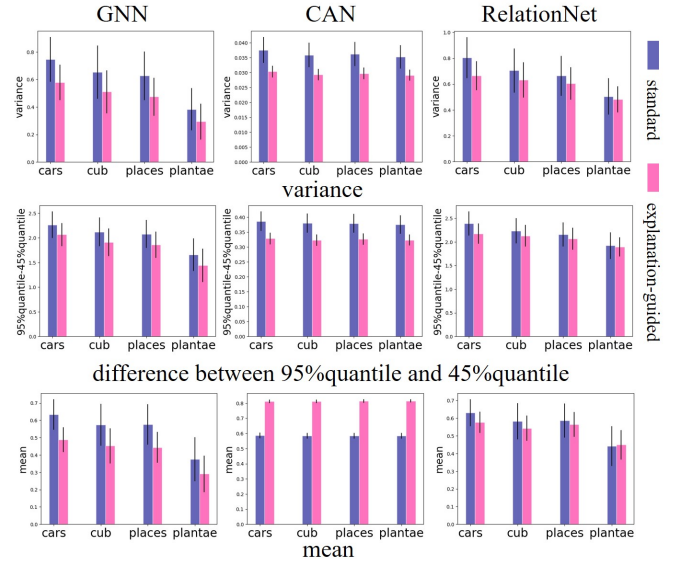

Fig. 3: The *variance* (the first row) and *quantile difference* (the second row) across the channels of the CNN encoded image feature vectors. We report the mean and the standard deviation of the two feature vector statistics over all the test set images of four cross-domain datasets. The experiment models are RelationNet(RN), GNN, and CAN with(dark-pink)/without(dark-blue) explanation-guided training

images. We use the 95% quantile for spatial pooling, resulting in a vector $f \in \mathbb{R}^C$. We do not use spatial average pooling due to the spatial sparsity of features as discriminative parts are usually present only in a small region of an image. For the same reason, median pooling would yield zeros mostly.

To verify that explanation-guided training indeed reduces excessive information removal, we observe the variance and intervals between the quantiles of the image feature vectors f, $S^2 = (\sum_{i=1}^{C} (f^i - \bar{f})^2)/C$ and the $95\% - 45\%$ quantile difference. Each image is with two statistics and we calculate the mean and standard deviation of the statistics over all the test set images of four cross-domain datasets, as illustrated in

TABLE V: The *variance* (S^2) and **quantile difference** across the channels of the CNN encoded image feature vectors. We report the mean and the standard deviation over all the test set images of four cross-domain datasets. The experiment models are RelationNet(RN), GNN, and CAN w/wo explanation-guided training(**LRP-**).

S^2	Cars	Places	CUB	Plantae
RN	0.804±0.1576	0.664±0.1557	0.705±0.1709	0.504±0.1409
LRP-RN	**0.664±0.1118**	**0.604±0.1255**	**0.632±0.1355**	**0.481±0.1007**
CAN	0.037±0.0044	0.036±0.0040	0.036±0.0041	0.035±0.0039
LRP-CAN	**0.030±0.0020**	**0.030±0.0019**	**0.029±0.0019**	**0.029±0.0019**
GNN	0.745±0.1614	0.626±0.1764	0.652±0.1939	0.383±0.1522
LRP-GNN	**0.578±0.1292**	**0.474±0.1376**	**0.510±0.1559**	**0.293±0.1292**
$95\% - 45\%$quantile	Cars	Places	CUB	Plantae
RN	2.385±0.2558	2.156±0.2559	2.234±0.2685	1.919±0.2791
LRP-RN	**2.172±0.2138**	**2.0654±0.2298**	**2.128±0.2273**	**1.895±0.2032**
CAN	0.387±0.0327	0.379±0.0315	0.380±0.0321	0.375±0.0310
LRP-CAN	**0.329±0.0195**	**0.326±0.0190**	**0.323±0.0192**	**0.324±0.0188**
GNN	2.265±0.2690	2.077±0.2844	2.119±0.2915	1.658±0.3298
LRP-GNN	**2.065±0.2355**	**1.859±0.2670**	**1.908±0.2792**	**1.442±0.3351**

Figure3. Detail numbers are listed in TableV

Lower S^2 and quantile difference mean that the features are not focused on a few channels but are more balanced over every channel, which preserves more diverse information and results in better generalization for new classes. The consistent decrease of S^2 and quantile difference over four cross-domain datasets after applying explanation-guided training provides some evidence that the explanation-guided training effectively avoids excessive information removal and avoids overfitting on the source domain.

We note that the lower S^2 and quantile difference are not due to lower first-order statistics such as the mean. For the CAN model, we observe an increased mean of f and a decreased S^2 with explanation-guided training for all the cross-domain datasets.

Furthermore, the standard deviation of some first-order statistics calculated over the test set also decrease with explanation-guided training, which can be observed in Table V. This is comparable to the effect of batch normalization, while batch normalization is naturally less effective for FSC.

E. Qualitative Results of LRP Explanation for FSC Models

The above experiments have demonstrated that, by leveraging the LRP explanation of the intermediate feature map to re-weight the same feature map, explanation-guided training effectively improves the performances of FSC models and successfully reduces the domain gap. In this section, we visualize the LRP explanation of the input images as heatmaps. From the LRP heatmaps, we can easily observe which parts of the image are used by the model to make the predictions, in other words, what features have the model learned to differentiate classes. To our best knowledge, this is the first attempt to explain the FSC models though many existing explanation methods are in principle applicable.

Figure 1 has already presented some heatmaps for the RelationNet. We further illustrate the LRP explanations of the CAN model for the 5-way 1-shot setting in Figure 4. Since there is only one training sample per class, we also show the LRP heatmap and the attention map for the support images. For the correctly classified $Q1$ and $Q3$, LRP heatmaps for the
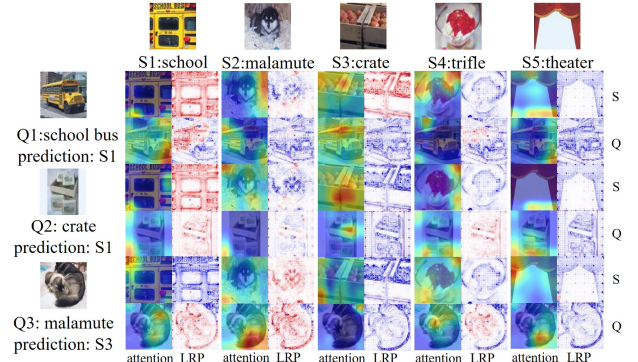

Fig. 4: LRP heatmaps and the attention heatmaps of the CAN model from one episode. The model is trained under the 5-way 1-shot setting. The first row shows the support images of each class. For each query image, we illustrate the attention heatmaps and the LRP heatmaps of both the support images and the query images with 5 target labels.

correct label highlight the relevant features. Specifically, the LRP heatmaps can capture the features of the window frames for the *bus* and the head features for the *malamute*.

On the other hand, the LRP heatmaps of the other wrong labels show more negative evidence, while we can still find some interesting resemblance between the query image and the explained label. For example, in Figure 1, when we explain the label *lion* for *Q1:African hunting dog*, the LRP heatmap highlights the legs of the *African hunting dog* and when we explain the label *cuirass* (a kind of medieval soldiers' armor) for *Q2:lion*, the LRP heatmap emphasizes the round contour that resembles an armor plate. In Figure 4, when we explain the label *trifle* for *Q3:malamute*, the LRP heatmap highlights the texture within a circle structure.

V. CONCLUSION

This paper shows the usefulness of explanation methods for few-shot learning during the training phase, exemplified by, but not limited to LRP. We find two points noteworthy.

Firstly, explanation-guided training successfully addresses the domain shift problem in few-shot learning, as demonstrated in the cross-domain few-shot classification task. Secondly, when combining explanation-guided training with feature-wise transformation, the model performance is further improved, indicating that these two approaches optimize the model in a non-overlapping manner. We conclude that applying explanation methods to the few-shot classification can not only provide intuitive and informative visualizations but can also be used to improve the models.

REFERENCES

[1] O. Vinyals, C. Blundell, T. Lillicrap, D. Wierstra *et al.*, "Matching networks for one shot learning," in *NIPS*, 2016, pp. 3630–3638.

[2] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017, pp. 1126–1135.

[3] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," in *NIPS*, 2017, pp. 4077–4087.

[4] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. Torr, and T. M. Hospedales, "Learning to compare: Relation network for few-shot learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1199–1208.

[5] V. G. Satorras and J. B. Estrach, "Few-shot learning with graph neural networks," in *ICLR*, 2018. [Online]. Available: https://openreview.net/forum?id=BJj6qGbRW

[6] A. A. Rusu, D. Rao, J. Sygnowski, O. Vinyals, R. Pascanu, S. Osindero, and R. Hadsell, "Meta-learning with latent embedding optimization," in *ICLR*, 2019. [Online]. Available: https://openreview.net/forum?id=BJgklhAcK7

[7] Q. Sun, Y. Liu, T.-S. Chua, and B. Schiele, "Meta-transfer learning for few-shot learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 403–412.

[8] R. Hou, H. Chang, M. Bingpeng, S. Shan, and X. Chen, "Cross attention network for few-shot classification," in *NIPS*, 2019, pp. 4005–4016.

[9] W.-Y. Chen, Y.-C. Liu, Z. Kira, Y.-C. F. Wang, and J.-B. Huang, "A closer look at few-shot classification," in *ICLR*, 2019. [Online]. Available: https://openreview.net/forum?id=HkxLXnAcFQ

[10] H.-Y. Tseng, H.-Y. Lee, J.-B. Huang, and M.-H. Yang, "Cross-domain few-shot classification via learned feature-wise transformation," in *ICLR*, 2020.

[11] B. Liu, Z. Zhao, Z. Li, J. Jiang, Y. Guo, H. Shen, and J. Ye, "Feature transformation ensemble model with batch spectral regularization for cross-domain few-shot classification," *arXiv preprint arXiv:2005.08463*, 2020.

[12] J.-F. Yeh, H.-Y. Lee, B.-C. Tsai, Y.-R. Chen, P.-C. Huang, and W. H. Hsu, "Large margin mechanism and pseudo query set on cross-domain few-shot learning," *arXiv preprint arXiv:2005.09218*, 2020.

[13] S. Bach, A. Binder, G. Montavon, F. Klauschen, K. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PloS one*, vol. 10, no. 7, p. e0130140, 2015.

[14] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should i trust you?: Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. ACM, 2016, pp. 1135–1144.

[15] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," in *ICLR (workshop track)*, 2014.

[16] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017, pp. 3319–3328.

[17] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," *CoRR*, vol. abs/1312.6034, 2013.

[18] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[19] G. Montavon, S. Lapuschkin, A. Binder, W. Samek, and K. Müller, "Explaining nonlinear classification decisions with deep taylor decomposition," *Pattern Recognition*, vol. 65, pp. 211–222, 2017.

[20] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," in *International Conference on Machine Learning*, 2017, pp. 3145–3153.

[21] P. Kindermans, K. T. Schütt, M. Alber, K. Müller, D. Erhan, B. Kim, and S. Dähne, "Learning how to explain neural networks: Patternnet and patternattribution," in *ICLR*, 2018.

[22] S. Lapuschkin, S. Wäldchen, A. Binder, G. Montavon, W. Samek, and K. Müller, "Unmasking clever hans predictors and assessing what machines really learn," *Nature Communications*, vol. 10, no. 1, p. 1096, 2019. [Online]. Available: https://doi.org/10.1038/s41467-019-08987-4

[23] L. Arras, F. Horn, G. Montavon, K. Müller, and W. Samek, "" what is relevant in a text document?": An interpretable machine learning approach." *Plos one*, vol. 12, no. 8, pp. e0181142–e0181142, 2017.

[24] L. Arras, G. Montavon, K. Müller, and W. Samek, "Explaining recurrent neural network predictions in sentiment analysis," in *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, 2017, pp. 159–168.

[25] T. Schnake, O. Eberle, J. Lederer, S. Nakajima, K. T. Schütt, K.-R. Müller, and G. Montavon, "Xai for graphs: Explaining graph neural network predictions by identifying relevant walks," *arXiv preprint arXiv:2006.03589*, 2020.

[26] J. Kauffmann, M. Esders, G. Montavon, W. Samek, and K.-R. Müller, "From clustering to cluster explanations via neural networks," *arXiv preprint arXiv:1906.07633*, 2019.

[27] A. Nichol and J. Schulman, "Reptile: a scalable metalearning algorithm," *arXiv preprint arXiv:1803.02999*, vol. 2, no. 3, p. 4, 2018.

[28] S. Ravi and H. Larochelle, "Optimization as a model for few-shot learning," 2017.

[29] A. Santoro, S. Bartunov, M. Botvinick, D. Wierstra, and T. Lillicrap, "Meta-learning with memory-augmented neural networks," in *International conference on machine learning*, 2016, pp. 1842–1850.

[30] N. Mishra, M. Rohaninejad, X. Chen, and P. Abbeel, "A simple neural attentive meta-learner," in *ICLR*, 2018. [Online]. Available: https://openreview.net/forum?id=B1DmUzWAW

[31] T. Munkhdalai and H. Yu, "Meta networks," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, 2017, pp. 2554–2563.

[32] Y. Liu, J. Lee, M. Park, S. Kim, E. Yang, S. Hwang, and Y. Yang, "Learning to propagate labels: transductive propagation network for few-shot learning," in *ICLR*, 2019. [Online]. Available: https://openreview.net/forum?id=SyVuRiC5K7

[33] A. Devos and M. Grossglauser, "Subspace networks for few-shot classification," *arXiv preprint arXiv:1905.13613*, 2019.

[34] B. Oreshkin, P. R. López, and A. Lacoste, "Tadam: Task dependent adaptive metric for improved few-shot learning," in *NIPS*, 2018, pp. 721–731.

[35] P. Bateni, R. Goyal, V. Masrani, F. Wood, and L. Sigal, "Improved few-shot visual classification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 14493–14502.

[36] S. Gidaris and N. Komodakis, "Dynamic few-shot visual learning without forgetting," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4367–4375.

[37] H. Qi, M. Brown, and D. G. Lowe, "Low-shot learning with imprinted weights," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5822–5830.

[38] S. Gidaris and N. Komodakis, "Generating classification weights with gnn denoising autoencoders for few-shot learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 21–30.

[39] A. Li, W. Huang, X. Lan, J. Feng, Z. Li, and L. Wang, "Boosting few-shot learning with adaptive margin loss," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 12576–12584.

[40] C. Xing, N. Rostamzadeh, B. Oreshkin, and P. O. Pinheiro, "Adaptive cross-modal few-shot learning," in *NIPS*, 2019, pp. 4848–4858.

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

[42] X. Li, Q. Sun, Y. Liu, Q. Zhou, S. Zheng, T.-S. Chua, and B. Schiele, "Learning to self-train for semi-supervised few-shot classification," in *NIPS*, 2019, pp. 10276–10286.

[43] M. Ren, E. Triantafillou, S. Ravi, J. Snell, K. Swersky, J. B. Tenenbaum, H. Larochelle, and R. S. Zemel, "Meta-learning for semi-supervised few-shot classification," 2018.

[44] S. Gidaris, A. Bursuc, N. Komodakis, P. Perez, and M. Cord, "Boosting few-shot visual learning with self-supervision," in *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[45] Y. Guo, N. C. Codella, L. Karlinsky, J. R. Smith, T. Rosing, and R. Feris, "A new benchmark for evaluation of cross-domain few-shot learning," *arXiv preprint arXiv:1912.07200*, 2019.

[46] J. Cai and S. M. Shen, "Cross-domain few-shot learning with meta fine-tuning," *arXiv preprint arXiv:2005.10544*, 2020.

[47] J. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, "Striving for simplicity: The all convolutional net," in *ICLR (workshop track)*, 2015.

[48] N. Poerner, B. Roth, and H. Schütze, "Evaluating neural network explanation methods using hybrid documents and morphosyntactic agreement." Association for Computational Linguistics (ACL), 2018.

[49] M. Kohlbrenner, A. Bauer, S. Nakajima, A. Binder, W. Samek, and S. Lapuschkin, "Towards best practice in explaining neural network decisions with lrp," in *Proceedings of International Joint Conference on Neural Networks*, 2020.

[50] S. Ravi and H. Larochelle, "Optimization as a model for few-shot learning," 2016.

[51] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The caltech-ucsd birds-200-2011 dataset," 2011.

[52] J. Krause, M. Stark, J. Deng, and L. Fei-Fei, "3d object representations for fine-grained categorization," in *Proceedings of the IEEE international conference on computer vision workshops*, 2013, pp. 554–561.

[53] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 6, pp. 1452–1464, 2017.

[54] G. Van Horn, O. Mac Aodha, Y. Song, Y. Cui, C. Sun, A. Shepard, H. Adam, P. Perona, and S. Belongie, "The inaturalist species classification and detection dataset," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8769–8778.

[55] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[56] N. Tishby and N. Zaslavsky, "Deep learning and the information bottleneck principle," in *2015 IEEE Information Theory Workshop (ITW)*. IEEE, 2015, pp. 1–5.

[57] A. M. Saxe, Y. Bansal, J. Dapello, M. Advani, A. Kolchinsky, B. D. Tracey, and D. D. Cox, "On the information bottleneck theory of deep learning," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2019, no. 12, p. 124020, 2019.