Revisiting Prototypical Network for Cross Domain Few-Shot Learning

Fei Zhou[1] Peng Wang[2*] Lei Zhang[1†] Wei Wei[1] Yanning Zhang[1]
[1] Northwestern Polytechnical University [2] University of Wollonggong
zhoufei@mail.nwpu.edu.cn pengw@uow.edu.au
{nwpuzhanglei,weiweinwpu,ynzhang}@nwpu.edu.cn

Abstract

*Prototypical Network is a popular few-shot solver that aims at establishing a feature metric generalizable to novel few-shot classification (FSC) tasks using deep neural networks. However, its performance drops dramatically when generalizing to the FSC tasks in new domains. In this study, we revisit this problem and argue that the devil lies in the simplicity bias pitfall in neural networks. In specific, the network tends to focus on some biased shortcut features (*e.g.*, color, shape,* etc.*) that are exclusively sufficient to distinguish very few classes in the meta-training tasks within a pre-defined domain, but fail to generalize across domains as some desirable semantic features. To mitigate this problem, we propose a* ***L****ocal-global* ***D****istillation* ***P****rototypical Network (LDP-net). Different from the standard Prototypical Network, we establish a two-branch network to classify the query image and its random local crops, respectively. Then, knowledge distillation is conducted among these two branches to enforce their class affiliation consistency. The rationale behind is that since such global-local semantic relationship is expected to hold regardless of data domains, the local-global distillation is beneficial to exploit some cross-domain transferable semantic features for feature metric establishment. Moreover, such local-global semantic consistency is further enforced among different images of the same class to reduce the intra-class semantic variation of the resultant feature. In addition, we propose to update the local branch as Exponential Moving Average (EMA) over training episodes, which makes it possible to better distill cross-episode knowledge and further enhance the generalization performance. Experiments on eight cross-domain FSC benchmarks empirically clarify our argument and show the state-of-the-art results of LDP-net. Code is available in https://github.com/NWPUZhoufei/LDP-Net*

1. Introduction

Prototypical Network (ProtoNet) [1] is a popular few-shot classification (FSC) method, which works by establishing a feature metric generalizable to novel few-shot tasks using deep neural networks. It adopts an episode-based learning strategy, where each episode, *e.g.*, N-way K-shot, is formulated as a contrastive learning task to identify the correct class for each query sample from a set of limited classes represented by prototypes derived from few support samples. Thanks to the simplicity of the framework and appealing few-shot learning performance, ProtoNet has gained great research attention [2–5].

However, the performance of typical ProtoNet declines greatly when generalizing to FSC tasks in new domains, *e.g.*, apply the ProtoNet trained on natural images in *mini*-ImageNet [6] to the fine-grained bird images in *CUB* [7]. This severely restricts the practicality of ProtoNet in real applications. In this work, we propose to re-inspect the intrinsic reason for the limited cross-domain generalization capability of ProtoNet and revive it in the cross-domain setting with right medicine. Specifically, the key for cross-domain generalization, especially in few-shot setting with ProtoNet, lies on exploiting some semantic information of each class that is invariant across different domains. To this end, typical ProtoNet resorts to taking advantages of the great expressive capacity of deep neural networks for feature learning. Obviously, it fails to exploit the desirable cross-domain transferable semantic features. In that case, what feature representation are obtained by the deep neural network? Some recent works [8–10] may have found the possible answer, viz., simplicity bias. It has shown that neural networks exclusively latch on to the simplest feature (*e.g.*, color, shape, *etc.*) and tends to ignore the complex predictive features (*e.g.*, semantics of the object). Inspired by this, we argue that the limited cross-domain generalization capacity of ProtoNet is incurred by the simplicity bias, viz., it tends to exploit some biased shortcut features that are exclusively sufficient to distinguish very few classes in the meta-training tasks within a pre-defined domain, but prone to be variant across different domains.

*F. Zhou and P. Wang contributed equally in this work.
†Corresponding author.

To mitigate this problem, we propose a **L**ocal-global **D**istillation **P**rototypical Network (LDP-net) to identify image features and metric that can generalize better to FSC tasks in new domains. The network employs a two-branch structure. A global branch predicts the class affiliation for each query image, which is akin to standard ProtoNet. A local branch works with image patches randomly cropped from the query image and makes classification predictions for such local crops. We then perform knowledge distillation between these two branches to enforce a global image and its local patches to have consistent class affiliation predictions. The rationale behind are twofold. Firstly, comparing to biased visual patterns, the semantic relationship between global image and local patches can hold more generally regardless of data domains. Secondly, the local-global distillation enables embedding richer semantic information from local features into the final global feature representation, which are proven to be more domain-invariant [11]. Take a step further, we apply such affiliation consistency constraint across images belonging to the same class. By doing this, we can reduce the intra-class semantic variation and further improve the robustness of the image feature representations. In addition, the local branch is updated as Exponential Moving Average (EMA) of the global branch to produce robust classification predictions, which enables our model to distill cross-episode knowledge and enhance the generalization performance. Once the model is trained, only the global branch is retained as a feature extractor for cross-domain FSC evaluation. Notably, by simply freezing the feature extractor in a new domain, the proposed method achieves state-of-the-art results on eight cross-domain FSC benchmark datasets.

The major contributions of this study can be summarized as follows:

- We inspect the limited cross-domain generalization capability of typical ProtoNet from the perspective of simplicity bias and propose a local-global knowledge distillation framework to effectively mitigate this problem.
- The proposed LDP-Net has insightful and innovative designs and can learn a robust feature metric that generalizes better to FSC tasks in new domains.
- The proposed LDP-Net achieves state-of-the-art performance on a set of cross-domain FSC benchmarks.

2. Related Work

Few-shot learning. Few-shot learning (FSL) aims to generalize knowledge learned in some auxiliary base classes to novel classes with very few labeled samples. Popular works solve FSL mainly from prototype-based metric learning [1–3, 6], meta-learning [12–16] and transfer learning [17–19]. Prototype-based metric learning methods, *e.g.*, ProtoNet [1], MatchingNet [6], *etc.*, focus on learning an embedding space that push samples of the same class together and separate samples of different classes apart. In meta-learning based methods, *e.g.*, MAML [12], MetaOptNet [13], *etc.*, focus on fast adaptation through the two-stage optimization. LEO [14] and HT [15] utilize the hypernetwork [20] to generate task-aware parameters to dynamically handle each few-shot task. Transfer learning based methods, *e.g.*, S2M2 [18] and Neg-Cosine [21] focus on learning good feature initialization, and then performing task-level fine-tuning to improve performance.

Cross-domain few-shot learning. Unlike FSL, cross-domain FSL (CD-FSL) focuses on learning a model on the source domain that can effectively generalize to the target domain. According to the training data used, CD-FSL can be divided into three types, *e.g.*, training with only a single source domain [5, 22–25], training with multiple source domains [26], and training with both source and target domain data [27]. Among them, single-source CD-FSL is more challenging and practical, and thus we focus on it in this work.

Some recent works have made progress on single-source CD-FSL. Doersch et al. [4] customize a spatially-aware prototype for each query image based on cross-attention, and unify self-supervised learning into a meta-learning framework to effectively alleviate domain shift. Since complex Transformer [28] are used, this method needs to use large-scale source domain data for training. Li et al. [5] achieve state-of-the-art performance by calibrating the relative distance between support samples and query samples in feature space. Das et al. [25] utilize a feature masker to filter features suitable for the target domain few-shot task. Tseng et al. [29] adopt task-specific affine transformation on features to achieve domain adaptation. Wang et al. [22] perform gradient updates on input samples to improve robustness to domain changes. Guo et al. [23] utilize pre-training combined with fine-tuning to achieve good performance, even better than the sophisticated meta-learning algorithms. Liang et al. [24] devise a feature reconstruction-based loss to fine-tune each few-shot task and achieved significant performance gains. Although these works [5, 22–25] have made progress, they require fine-tuning the model (*i.e.*, feature extractor) to alleviate the domain gap when dealing with few-shot tasks on each target domain. In contrast, the proposed method focuses on learning a model with strong generalization ability, which is able to generalize to wide range of target domains without fine-tuning.

3. Methodology

Problem formulation. In CD-FSC, the model is trained on the source domain dataset $\mathcal{D}_s$, and then tested on a series

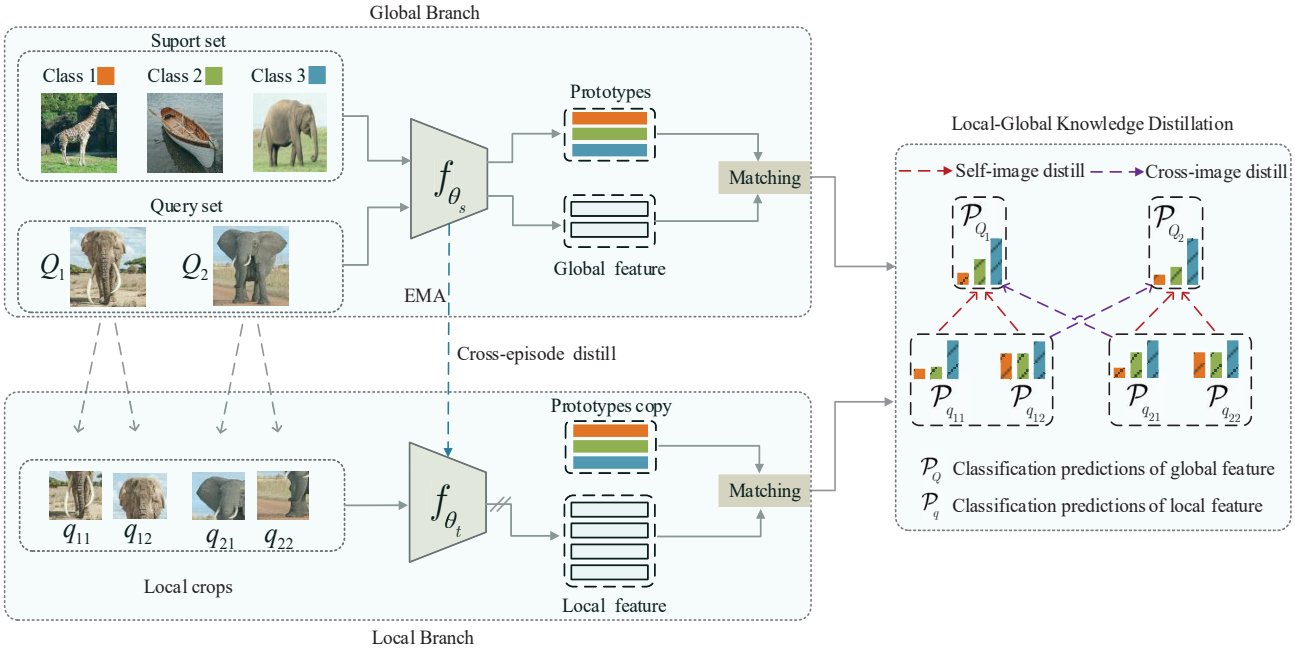


Figure 1. Framework of the proposed LDP-net. LDP-net consists of a two-branch network. The global branch extracts global features while the local branch takes random crops of raw query image as input to extract local features. Then, knowledge distillation is conducted among these two branches to enforce the local-global semantic consistency. In addition, the local branch is updated as the Exponential Moving Average (EMA) of the global branch during knowledge distillation, which makes it possible to better distill cross-episode knowledge.

of N-way K-shot episodes randomly sampled in the target domain dataset $\mathcal{D}_t$. Note that the classes between $\mathcal{D}_s$ and $\mathcal{D}_t$ do not overlap. In each episode $\mathcal{T}$ (*i.e.*, task) , N represents the number of classes, and K represents the number of labeled samples for each class. The $N{\times}K$ labeled samples are called the support set $\mathcal{T}_S$. Besides, each episode contains the query set $\mathcal{T}_Q$ for evaluation, which consists of different samples of the same class as $\mathcal{T}_s$. Usually, in order to mimic few-shot evaluation scenario, the model training is also performed in an episode-based way.

3.1. Preliminary knowledge about the ProtoNet

The ProtoNet is a popular few-shot learner. It constructs a prototype for each class based on its support samples, and then matches the query sample against all prototypes. Formally, given a few-shot episode $\mathcal{T}$, the prototype corresponding to each class is calculated as:

$$\mathcal{C}_n = \frac{1}{K}\sum_{k=1}^{K} f_\theta\left(X_{S_{n,k}}\right), \tag{1}$$

where f_θ represents the feature extractor, $\mathcal{C}_n$ represents the prototype of class n, and $X_{S_{n,k}}$ represents the k-th support sample of class n.

Then, the classification predictions $\mathcal{P}_{Q_i}$ for query sample X_{Q_i} is obtained by matching against all prototypes:

$$\mathcal{P}_{Q_i} = \texttt{matching}\left(f_\theta(X_{Q_i}), \mathcal{C}_n\right), n \in [1, N], \tag{2}$$

where $\texttt{matching}(\cdot)$ represents the similarity matching between two features. The label corresponding to the maximum prediction score is used as the predicted label $\hat{y}_{Q_i}$ for the query sample X_{Q_i}.

Finally, the cross entropy loss $H(\cdot)$ can be calculated as:

$$\mathcal{L}^{ce}_{X_{Q_i}} = H(\hat{y}_{Q_i}, y_{Q_i}), \tag{3}$$

where y_{Q_i} is the ground truth of the query image X_{Q_i}.

3.2. The proposed LDP-net

Overview. As shown in Fig. 1, the proposed LDP-net consists of a two-branch network. Among them, the global branch is utilized to extract global features from the input image, and its structure is the same as the feature extractor in standard ProtoNet. The local branch takes random crops of raw query image as input to extract local features. On top of these two branches, we propose a local-global knowledge distillation module to enforce a consistency constraint between the affiliation predictions made from local and global features, which proves to be invariant across domains. In addition, we propose to distill cross-episode knowledge by updating the local branch as the Exponential Moving Average (EMA) of the global branch over meta-training episodes.

Global branch. In the global branch, we first extract global features for each image through a feature extractor

f_{θ_s}. In this work, we denote the features corresponding to the raw image extracted by the global network as global features. Then, as with the ProtoNet, we customize the prototypes according to Eq. 1. Finally, the classification predictions $\mathcal{P}_{Q_i}$ of the query image X_{Q_i} can be calculated by Eq. 2.

Local branch. In the local branch, we extract local features for the local crops of the query image, and utilize the prototypes defined before to calculate the class affiliation predictions for each local feature. Some recent works [30–32] have shown that local information of the image can be obtained efficiently by multi-crop augmentation. This augmentation first randomly crops the raw image and then resizes it to a lower resolution to obtain local crop. Following this, in this work, we obtain local crops corresponding to each query image by multi-crop augmentation.

Specifically, for a given query image X_{Q_i} in few-shot episode $\mathcal{T}$, we first obtain local image crops $X_{q_{i,r}}$ by multi-crop augmentation, where $r \in [1, R]$, R is the number of crops. Then, we extract local features $f_{\theta_t}\left(X_{q_{i,r}}\right)$ through the local network f_{θ_t}. Similarly, we calculate the classification predictions $\mathcal{P}_{q_{i,r}}$ corresponding to each local feature $f_{\theta_t}\left(X_{q_{i,r}}\right)$ based on the prototypes.

Local-global knowledge distillation. We encourage the global image representations to distill richer semantic information from local crops. By doing so, we can enforce the semantic consistency between the local and global branches. To achieve this goal, we impose consistency constraints on the global and local classification predictions of the query image.

Specifically, for a given query image X_{Q_i}, we utilize the global branch and the local branch to obtain the classification predictions, respectively. Then, the self-image local-global distillation loss can be calculated as:

$$\mathcal{L}_{X_{Q_i}}^{self} = \frac{1}{R}\sum_{r=1}^{R} H\left(\mathcal{P}_{q_{i,r}}, \mathcal{P}_{Q_i}\right), \tag{4}$$

where $\mathcal{P}_{Q_i}$ and $\mathcal{P}_{q_{i,r}}$ are the global classification predictions and the r-th local classification predictions for the given query image X_{Q_i}, respectively. $H(\cdot)$ represents cross entropy loss function.

Moreover, in order to reduce the intra-class semantic variation, we further enforce the local-global semantic consistency between different images from the same category. In implementation, we found that keeping the local and global predictions consistent across all query samples of the same class would lead to the model learning trivial solution, resulting in model collapse [33]. To avoid this problem, for a query image X_{Q_i}, we randomly select one query image X_{Q_j} from same category in the query set to enforce the cross-image semantic consistency. The cross-image local-global distillation loss can be calculated as:

$$\mathcal{L}_{X_{Q_i}}^{cross} = \frac{1}{R}\sum_{r=1}^{R} H\left(\mathcal{P}_{q_{j,r}}, \mathcal{P}_{Q_i}\right), j \neq i, \tag{5}$$

where $\mathcal{P}_{q_{j,r}}$ are the local classification predictions for query image X_{Q_j}.

Cross-episode knowledge distillation. In the proposed method, the feature extraction network in the local branch has the same structure as that in the global branch. A simple approach is to update both branch networks simultaneously according to the loss function. However, this introduces additional learnable parameters and also leads to inefficiencies in the training process. On the other hand, the episode-based training paradigm updates the parameters of the model according to the gradient of the current episode. However, the learning episodes in meta-learning are normally sampled randomly from an auxiliary dataset, which means each episode has different combinations of classes. This is in stark contrast to batch-based training where all the batches share the same set of classes. Independently solving such tasks with dramatic semantic space variation enforces the network to keep switching to different combinations of visual patterns, which is an inefficient way to accumulate knowledge across the learning episodes. To mitigate these problems, we propose to distill cross-episode knowledge by updating the local branch as the Exponential Moving Average (EMA) of the global branch during meta-training, which makes it possible to better learn cross-episode knowledge and further enhance the generalization performance. Specifically, we update the parameters of the local network as:

$$\theta_t \leftarrow m\theta_t + (1-m)\,\theta_s, \tag{6}$$

where θ_t is the parameter of the local branch f_{θ_t}, θ_s is the parameter of the global branch f_{θ_s}, and m is the momentum.

Meta-training. For each query sample, we also compute a cross-entropy loss $\mathcal{L}_{x_{q_i}}^{ce}$ based on its global predictions to facilitate prototype learning.

In summary, for a few-shot episode $\mathcal{T}$, the total loss of the proposed is:

$$\mathcal{L}_{TS} = \sum_{i=1}^{I}\mathcal{L}_{X_{Q_i}}^{ce} + \lambda_1 \cdot \sum_{i=1}^{I}\mathcal{L}_{X_{Q_i}}^{self} + \lambda_2 \cdot \sum_{i=1}^{I}\mathcal{L}_{X_{Q_i}}^{cross}, \tag{7}$$

where I represents the total number of query samples in $\mathcal{T}$, λ_1 and λ_2 are the weight coefficients of each loss function. We utilize the total loss to meta-train the global branch. For the local branch, we update it according to Eq. 6. The detailed meta-training process is summarized in the algorithm

1. Once the entire meta-training is done, we discard the local branch, leaving the global branch as feature extractor for cross-domain FSC evaluation.

Algorithm 1: Meta-training algorithm of the proposed method.

Input: Source domain $\mathcal{D}_s$, feature extractor of global branch f_{θ_s} with parameters θ_s, feature extractor of local branch f_{θ_t} with parameters θ_t.

while *not converged* **do**

1. Sample a few-shot episode $\mathcal{T}$ from $\mathcal{D}_s$;
2. Calculate prototypes according to $\mathcal{T}_S$ based on Eq. 1;
   for *each query image* X_{Q_i} *in* $\mathcal{T}_Q$ **do**
   3. Obtain local image crops $X_{q_{i,r}}$ based on multi-crop augmentation;
   4. Utilize global branch to calculate the global predictions $\mathcal{P}_{Q_i}$;
   5. Utilize local branch to calculate the local predictions $\mathcal{P}_{q_{i,r}}$;
   6. Calculate the self-image distillation loss and cross-image distillation loss according to Eq. 5 and Eq. 6, respectively;
   7. Calculate cross-entropy loss based on Eq. 3;
8. Calculate the total loss according to Eq. 7, and update θ_s based on the total loss;
9. Update θ_t according to Eq. 6.

Output: Feature extractor of global branch f_{θ_s}.

Cross-domain FSC evaluation. In cross-domain FSC evaluation phase, for each few-shot task, we first utilize the global network to extract features for each image. And then, we use the support set to train a Logistic Regression Classifier. Finally, the query samples are classified according to the trained classifier. Notably, the proposed method does not require fine-tuning the feature extractor during testing on the target domain.

4. Experimental Analysis

4.1. Experimental details

Datasets. In this work, we focus on the single source domain CD-FSL problem. Following the standard benchmarks [5, 22, 23], we utilize the meta-training set with 64 classes in *mini*-ImageNet [6] dataset as the source domain for training. Then, we validate the generalization performance on eight target domain datasets, i.e., *CUB*, *Cars*, *Places*, *Plantae*, *ChestX*, *ISIC*, *EuroSAT* and *CropDisease*. Among them, *CUB*, *Cars*, *Places* and *Plantae* proposed in [29] contain natural images of different properties. *ChestX*, *ISIC*, *EuroSAT* and *CropDisease* proposed in [23] are cross-domain datasets from the domain of medicine, agriculture and remote sensing, which observe significant domain shift. All the images are resized to 224×224 pixels following common practice.

Implementation details. As shown in Fig. 1, the proposed method includes a global branch and a local branch. For the global branch, following [5, 22, 23], we use ResNet-10 as feature extraction network, and pre-train it by traditional batch-based supervised classification on the source domain. The feature extraction network in local branch has the same structure.

We meta-train the network for 100 epochs using Adam optimizer with learning rate set to be 0.001. In each epoch, we randomly sample 100 episodes from the source domain. In each episode, without otherwise stated, we set the number of classes to 5, the number of support samples of each class to 5, and the query sample size of each class to 15. For hyper-parameters, we set λ_1=1.0, λ_2=0.15, m=0.998, and R=6. Since we do not have a validation set for model selection, we use the checkpoint after the last epoch as the final model. It is worth noting that the proposed method only needs to meta-train the model once, which can be directly deployed to target domains without fine-tuning.

Evaluation protocol. We validate the proposed method following standard CD-FSC evaluation protocols [5, 23]. In each target domain, we randomly sample 600 N-way K-shot 15-query tasks, and calculate the average accuracy and 95% confidence intervals over these sampled tasks. In all validation experiments, we set N=5 and K=1 or 5.

4.2. Experimental results

4.2.1 Comparison with the ProtoNet baseline

We first conduct some analytical experiments to compare the proposed method with the ProtoNet. We follow the experimental setting in [23] to implement the ProtoNet with the ResNet-10 as backbone on CD-FSC benchmark. Besides, for a fair comparison, we pre-train the backbone of the ProtoNet in the same way as the proposed method. We mark the pre-trained ProtoNet as ProtoNet++.

It is worth noting that, the ProtoNet utilizes the Euclidean distance metric for classification. Therefore, to maintain a fair comparison, we utilize the same distance metric for classification in all ablation experiments. We conduct experiments on two natural image cross-domain datasets, *i.e.*, *CUB* and *Cars*, and two extreme cross-domain datasets, *i.e.*, *EuroSAT* and *ISIC*.

Quantitative comparison. The comparison results between the proposed method and the ProtoNet are shown in Table 1. As can be seen, the proposed method outperforms

Table 1. Ablation study. Average classification accuracies (%) are provided. $^{\diamond}$ indicates that the Euclidean distance metric is used as the classifier. ✔ indicates that this component is used, vice versa. The best results are in bold.

			CUB		Cars		EuroSAT		ISIC	
Method			1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
ProtoNet			41.77	58.98	29.79	41.16	57.50	74.44	30.65	40.42
ProtoNet++			40.34	61.94	31.63	46.56	59.11	**81.44**	31.73	44.01
Ours$^{\diamond}$			**47.70**	**68.94**	**34.65**	**51.61**	**63.70**	80.26	**33.51**	**46.42**
Self-image	**Cross-image**	**Cross-episode**	**1-shot**	**5-shot**	**1-shot**	**5-shot**	**1-shot**	**5-shot**	**1-shot**	**5-shot**
✔	✗	✗	43.99	63.80	32.87	49.58	62.31	80.22	32.88	45.00
✔	✔	✗	44.04	64.01	33.33	49.69	63.03	80.07	33.28	45.35
✔	✔	✔	47.70	68.94	34.65	51.61	63.70	80.26	33.51	46.42

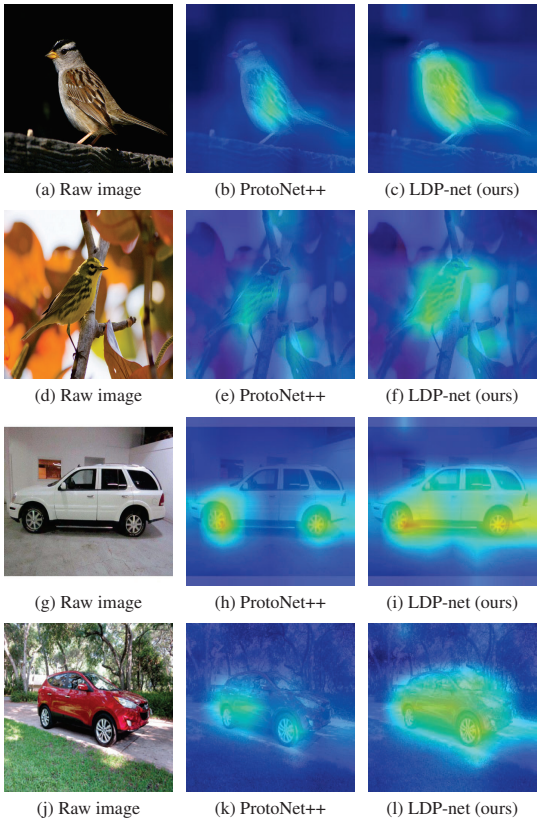
Table 2. Comparison with state-of-the-art methods in 5-way 1-shot setting. Average classification accuracies (%) are provided. † stands for exploiting the full data of FSL task. * stands for fine-tuning on target domain. The best results are in bold.

Methods	Mark	Ft	CUB	Cars	Places	Plantae	Chest	ISIC	EuroSAT	CropDisease	Ave.
MatchingNet [6]	NeurIPS-16	✗	35.89	30.77	49.86	32.70	-	-	-	-	-
RelationNet [34]	CVPR-18	✗	41.27	30.09	48.16	31.23	21.95	30.53	49.08	53.58	38.24
GNN [35]	ICLR-18	✗	44.40	31.72	52.42	33.60	21.94	30.14	54.61	59.19	41.00
RelationNet+FT [29]	ICLR-20	✗	43.33	30.45	49.92	32.57	21.79	30.38	53.53	57.57	39.94
RelationNet+ATA [22]	IJCAI-21	✗	43.02	31.79	51.16	33.72	22.14	31.13	55.69	61.17	41.23
GNN+FT [29]	ICLR-20	✗	45.50	32.25	53.44	32.56	22.00	30.22	55.53	60.74	41.53
GNN+ATA [22]	IJCAI-21	✗	45.00	33.61	53.57	34.42	22.10	33.21	61.35	67.47	43.84
MatchingNet+AFA [36]	ECCV-22	✗	41.02	33.52	**54.66**	37.60	22.11	32.32	61.28	60.71	42.90
GNN+AFA [36]	ECCV-22	✗	46.86	34.25	54.04	36.76	22.92	33.21	63.12	67.61	44.85
LDP-net (ours)	-	✗	**49.82**	**35.51**	53.82	**39.84**	**23.01**	**33.97**	**65.11**	**69.64**	**46.34**
TPN+ATA† [22]	IJCAI-21	✗	50.26	34.18	57.03	39.83	21.67	**34.70**	65.94	77.82	47.68
TPN+AFA† [36]	ECCV-22	✗	50.85	38.43	60.29	40.27	21.69	34.25	66.17	72.44	48.05
RDC† [5]	CVPR-22	✗	47.77	**38.74**	58.82	**41.88**	**22.66**	32.29	67.58	80.88	48.83
LDP-net†(ours)	-	✗	**55.94**	37.44	**62.21**	41.04	22.21	33.44	**73.25**	**81.24**	**50.85**
Fine-tuning* [23]	ECCV-20	✔	43.53	35.12	50.57	38.77	22.13	34.60	66.17	73.43	45.54
TPN+ATA*† [22]	IJCAI-21	✔	51.89	38.07	57.26	40.75	22.45	35.55	70.84	82.47	49.91
RDC*† [5]	CVPR-22	✔	50.09	39.04	61.17	41.30	22.32	36.28	70.51	85.79	50.81

ProtoNe by 3% to 10% on all datasets. Compared with ProtoNet++, the proposed method also achieves significant performance gains in most cases. For example, on the *CUB* dataset, the proposed method outperforms ProtoNet++ by 7.36% and 7.00% in 1-shot and 5-shot settings, respectively. On the *EuroSAT* dataset, although the ProtoNet++ achieves better results in 5-shot setting, the proposed method outperforms ProtoNet++ by 4.6% in 1-shot setting. In short, the proposed method achieves significant performance gains compared to ProtoNet and ProtoNet++. This shows that the proposed method has better cross-domain generalization ability.

Qualitative analysis. In order to verify that the proposed method can learn rich semantic information instead of only focusing on the simplest features, we adopt CAM [37] to visualize the features. The visualization results are shown in Fig. 2. It can be seen that ProtoNet++ only pays attention to some local regions of the object, *e.g.*, Fig. 2 (b), (e), (h), (k). In contrast, the proposed method can focus on a wider range of the object, *e.g.*, Fig. 2 (c), (f), (i), (l), which means that the proposed method can capture more comprehensive semantic information and thus generalize better.

To further illustrate the generalization advantage of the proposed method, we visualize the loss landscape of the model. The loss landscape is a visualization tool proposed by Li et al. [38] for model generalization verification. In implementation, we randomly perturb the model trained in the source domain in 2000 different directions. Then, we perform inference on the target domain against each perturbed model and record the loss value. Finally, we visualize the loss landscape according to the recorded loss values and orientations. In loss landscape, the contour near center describes the optimal solution of the model. The smoother contour and larger the space spanned by the contour corresponding to the optimal solution of the model, indicating that the model has better generalization [38]. We use the *CUB* dataset as the target domain to visualize the loss landscape of the model. The comparison between the proposed LDP-net and ProtoNet++ is shown in Fig. 3. As can be seen, compared with ProtoNet++, the contour corresponding to



(a) Raw image (b) ProtoNet++ (c) LDP-net (ours)

(d) Raw image (e) ProtoNet++ (f) LDP-net (ours)

(g) Raw image (h) ProtoNet++ (i) LDP-net (ours)

(j) Raw image (k) ProtoNet++ (l) LDP-net (ours)

Figure 2. Feature visualization for ProtoNet++ and the proposed LDP-net.

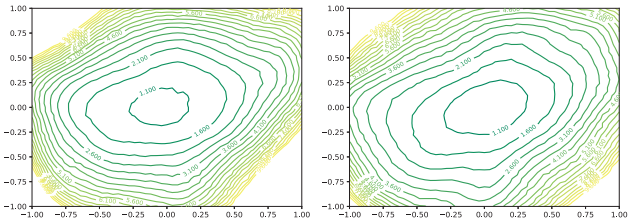
the optimal solution of the proposed method is smoother, and the space spanned by the contour is larger. This reveals that the proposed method has stronger generalization ability. This finding resonates with quantitative experimental results.

In conclusion, the above quantitative and qualitative experiments show that the proposed method can effectively alleviate the simplicity bias pitfall in ProtoNet, and learn transferable semantic knowledge, resulting in better cross-domain generalization.

4.2.2 Comparison with state-of-the-art methods

State-of-the-art methods usually employ fine-tuning or exploit the full data in the few-shot task to improve performance. Among them, fine-tuning refers to updating the feature extractor trained on the source domain in each few-shot task on target domain. Exploiting the full data means that the samples in the query set are also used but in an unsupervised fashion.

We divide the comparative experiments into three cases according to whether fine-tuning is required and whether the full data is used. Case 1: neither fine-tuning is required nor the full data is used, such as RelationNet+ATA [22], GNN+AFA [36] and so on. Case 2: fine-tuning is not required, but full data is used, such as TPN+ATA† [22] and RDC† [5]. Case 3: both fine-tuning and full data are required, such as TPN+ATA*† [22] and RDC*† [5]. In order to maintain a fair comparison with the methods in Case 2, the proposed method (LDP-net†) also exploit the full data in the few-shot task. Specifically, we use a classifier trained on the support set to make predictions on query samples. Then, we select some query samples with high confidence as the augmentation of the support set according to the predictions. Finally, the classifier is re-trained on the augmented support set. We repeat this process several times, and utilize the last classifier to test the query samples as the final result.



(a) ProtoNet++ (b) LDP-net (ours)

Figure 3. Loss landscape.

The experiments are conducted on eight target domains under 5-way 1-shot setting and 5-way 5-shot setting, respectively. For each setting, we calculate the average results for the eight target domains as the overall evaluation. The results of 1-shot and 5-shot are shown in Table 2 and Table 3, respectively.

For Case 1, the proposed method (LDP-net) achieves the best performance on most datasets. Overall, in 1-shot setting, the average result of the proposed method reaches 46.34%. Compared with the second-best method (GNN+AFA), the proposed method achieves 1.49% average improvement. In the 5-shot setting, the proposed method achieves 62.60% average result, outperforming the second-best method (GNN+AFA) by 1.02%. For Case 2, in 1-shot setting, the average result of the proposed method (LDP-net†) reaches 50.85%. Compared with the second-best method (RDC†), the proposed method observes an improvement of 2.02%. In 5-shot setting, the proposed method achieves the best performance on all datasets. In terms of average results, the proposed method (LDP-net†) outperforms the second-best comparison method (RDC†) by 4.27%. For Case 3, despite freezing the feature extractor on the target domain, the proposed method (LDP-net†) still achieves the best average results under both 1-shot and 5-shot settings.

In summary, the proposed method achieves the best average performance in all cases. The performance gains indicate that the proposed method has stronger cross-domain generalization ability. The reason behind this is that the proposed method is able to learn more knowledge in the source domain to promote generalization on the target do-

Table 3. Comparison with state-of-the-art methods in 5-way 5-shot setting. Average classification accuracies (%) are provided.† stands for exploiting the full data of FSL task. * stands for fine-tuning on target domain. The best results are in bold.

Methods	Mark	Ft	CUB	Cars	Places	Plantae	Chest	ISIC	EuroSAT	CropDisease	Ave.
MatchingNet [6]	NeurIPS-16	✗	51.37	38.99	63.16	46.53	22.40	36.74	64.45	66.39	48.75
MAML [12]	ICML-17	✗	-	-	-	-	23.48	40.13	71.70	78.05	-
RelationNet [34]	CVPR-18	✗	56.77	40.46	64.25	42.71	24.07	38.60	65.56	72.86	50.66
MetaOptNet [13]	CVPR-19	✗	-	-	-	-	22.53	36.28	64.44	68.41	-
GNN [35]	ICLR-18	✗	62.87	43.70	70.91	48.51	23.87	42.54	78.69	83.12	56.77
RelationNet+FT [29]	ICLR-20	✗	59.77	40.18	65.55	44.29	23.95	38.68	69.13	75.78	52.17
RelationNet+ATA [22]	IJCAI-21	✗	59.36	42.95	66.90	45.32	24.43	40.38	71.02	78.20	53.57
GNN+FT [29]	ICLR-20	✗	64.97	46.19	70.70	49.66	24.28	40.87	78.02	87.07	57.72
GNN+ATA [22]	IJCAI-21	✗	66.22	49.14	**75.48**	52.69	24.32	44.91	83.75	**90.59**	60.89
MatchingNet+AFA [36]	ECCV-22	✗	59.46	46.13	68.87	52.43	23.18	39.88	69.63	80.07	54.96
GNN+AFA [36]	ECCV-22	✗	68.25	49.28	76.21	54.26	25.02	46.01	**85.58**	88.06	61.58
LDP-net (ours)	-	✗	**70.39**	**52.84**	72.90	**58.49**	**26.67**	**48.06**	82.01	89.40	**62.60**
TPN+ATA† [22]	IJCAI-21	✗	65.31	46.95	72.12	55.08	23.60	45.83	79.47	88.15	59.56
TPN+AFA† [36]	ECCV-22	✗	65.86	47.89	72.81	55.67	23.47	46.29	80.12	85.69	59.73
RDC† [5]	CVPR-22	✗	63.39	52.75	72.83	55.30	25.10	42.10	79.12	88.03	59.83
LDP-net† (ours)	-	✗	**73.34**	**53.06**	**75.47**	**59.64**	**26.88**	**48.44**	**84.05**	**91.89**	**64.10**
Fine-tuning* [23]	ECCV-20	✓	63.76	51.21	70.68	56.45	25.37	49.51	81.59	89.84	61.05
NSAE(CE+CE)* [24]	ICCV-21	✓	68.51	54.91	71.02	59.55	27.10	54.05	83.96	93.14	64.03
ConFeSS* [25]	ICLR-22	✓	-	-	-	-	27.09	48.85	84.65	88.88	-
TPN+ATA*† [22]	IJCAI-21	✓	70.14	55.23	73.87	59.02	24.74	49.83	85.47	93.56	63.98
RDC*† [5]	CVPR-22	✓	67.23	53.49	74.91	57.47	25.07	49.91	84.29	93.30	63.21

main. Another advantage of the proposed method is that the feature extractor can be readily used without fine-tuning, which shows the practicality of the proposed method.

4.2.3 Ablation study

The proposed method mainly consists of three components, self-image distillation ("Self-image"), cross-image distillation ("Cross-image"), and cross-episode distillation ("Cross-episode"). We perform the ablation study for each component. It is worth noting that, we utilize the Euclidean distance metric for classification in all ablation experiments.

The ablation results are shown in Table 1. Firstly, compared with ProtoNet++ baseline, the proposed method performs better in most cases when only using the self-image distillation. For example, on the *CUB* dataset, the self-image distillation improves ProtoNet++ by 3.65% and 1.86% under 1-shot and 5-shot settings, respectively. After incorporating the cross-image distillation, the proposed method can observe performance rise in most cases. In addition, when the cross-episode distillation is added, the performance is further boosted. In particular, on the *CUB* dataset, the performance can be improved by 3.66% in 1-shot and 4.93% in 5-shot, respectively. In summary, the ablation study shows that each component plays an important role in the proposed method and all of them contribute positively to cross-domain generalization.

5. Conclusions

In this study, we inspected the poor cross-domain generalization of standard ProtoNet from the perspective of simplicity bias and proposed a local-global knowledge distillation framework to alleviate this problem in the ProtoNet. By simultaneously enforcing the class affiliation predictions between a global image and local patches from both the same image and other images of the same class, our model is expected to be able to capture more robust semantic information desirable for cross-domain generalization. In addition, we propose a cross-episode knowledge distillation strategy to further improve the generalization performance of the learned feature and metric. The proposed method achieves state-of-the-art results on eight CD-FSC datasets.

Although promising improvement has been achieved for CD-FSC tasks, the performance is still far from satisfactory when generalizing to domains with significant domain shift such as *chest* and *ISIC*. Possible remedies may include increasing the diversity of training data to extract more universal meta-knowledge or proposing smarter model adaptation strategy to integrate the extracted knowledge to the target task in a more data-efficient way. We will leave this as future work.

6. Acknowledgment

This work was supported in part by the National Natural Science Foundation of China under Grand 62101454, Grant 62071387, and Grant U19B2037; in part by the Fundamental Research Funds for the Central Universities; in part by the Shenzhen Fundamental Research Program under Grant JCYJ20190806160210899. P. Wang's participation was in part supported by Australian Research Council Discovery Projects (DP220101784).

References

[1] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, pages 4077–4087, 2017. 1, 2

[2] Yifei Huang, Lijin Yang, and Yoichi Sato. Compound prototype matching for few-shot action recognition. In *European Conference on Computer Vision*, pages 351–368. Springer, 2022. 1, 2

[3] Tao Zhang and Wu Huang. Kernel relative-prototype spectral filtering for few-shot learning. In *European Conference on Computer Vision*, pages 541–557. Springer, 2022. 1, 2

[4] Carl Doersch, Ankush Gupta, and Andrew Zisserman. Crosstransformers: spatially-aware few-shot transfer. *Advances in Neural Information Processing Systems*, 33:21981–21993, 2020. 1, 2

[5] Pan Li, Shaogang Gong, Chengjie Wang, and Yanwei Fu. Ranking distance calibration for cross-domain few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9099–9108, 2022. 1, 2, 5, 6, 7, 8

[6] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. In *Advances in neural information processing systems*, pages 3630–3638, 2016. 1, 2, 5, 6, 8

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

[8] Harshay Shah, Kaustav Tamuly, Aditi Raghunathan, Prateek Jain, and Praneeth Netrapalli. The pitfalls of simplicity bias in neural networks. *Advances in Neural Information Processing Systems*, 33:9573–9585, 2020. 1

[9] Damien Teney, Ehsan Abbasnejad, Simon Lucey, and Anton van den Hengel. Evading the simplicity bias: Training a diverse set of models discovers solutions with superior ood generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16761–16772, 2022. 1

[10] Kaifeng Lyu, Zhiyuan Li, Runzhe Wang, and Sanjeev Arora. Gradient descent on two-layer nets: Margin maximization and simplicity bias. *Advances in Neural Information Processing Systems*, 34:12978–12991, 2021. 1

[11] Adrien Bardes, Jean Ponce, and Yann LeCun. Vicregl: Self-supervised learning of local visual features. In *NeurIPS*, 2022. 2

[12] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1126–1135. JMLR. org, 2017. 2, 8

[13] Kwonjoon Lee, Subhransu Maji, Avinash Ravichandran, and Stefano Soatto. Meta-learning with differentiable convex optimization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10657–10665, 2019. 2, 8

[14] Andrei A Rusu, Dushyant Rao, Jakub Sygnowski, Oriol Vinyals, Razvan Pascanu, Simon Osindero, and Raia Hadsell. Meta-learning with latent embedding optimization. In *International Conference on Learning Representations*, 2019. 2

[15] Andrey Zhmoginov, Mark Sandler, and Maksym Vladymyrov. Hypertransformer: Model generation for supervised and semi-supervised few-shot learning. In *International Conference on Machine Learning*, pages 27075–27098. PMLR, 2022. 2

[16] Sungyong Baik, Myungsub Choi, Janghoon Choi, Heewon Kim, and Kyoung Mu Lee. Meta-learning with adaptive hyperparameters. *Advances in Neural Information Processing Systems*, 33:20755–20765, 2020. 2

[17] Yonglong Tian, Yue Wang, Dilip Krishnan, Joshua B Tenenbaum, and Phillip Isola. Rethinking few-shot image classification: a good embedding is all you need? In *European Conference on Computer Vision*, pages 266–282. Springer, 2020. 2

[18] Puneet Mangla, Nupur Kumari, Abhishek Sinha, Mayank Singh, Balaji Krishnamurthy, and Vineeth N Balasubramanian. Charting the right manifold: Manifold mixup for few-shot learning. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2218–2227, 2020. 2

[19] Chen Liu, Yanwei Fu, Chengming Xu, Siqian Yang, Jilin Li, Chengjie Wang, and Li Zhang. Learning a few-shot embedding model with contrastive learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 8635–8643, 2021. 2

[20] Dominic Zhao, Johannes von Oswald, Seijin Kobayashi, João Sacramento, and Benjamin F Grewe. Meta-learning via hypernetworks. In *4th Workshop on Meta-Learning at NeurIPS 2020 (MetaLearn 2020)*, 2020. 2

[21] Bin Liu, Yue Cao, Yutong Lin, Qi Li, Zheng Zhang, Mingsheng Long, and Han Hu. Negative margin matters: Understanding margin in few-shot classification. In *European conference on computer vision*, pages 438–455. Springer, 2020. 2

[22] Haoqing Wang and Zhi-Hong Deng. Cross-domain few-shot classification via adversarial task augmentation. In Zhi-Hua Zhou, editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 1075–1081. International Joint Conferences on Artificial Intelligence Organization, 8 2021. Main Track. 2, 5, 6, 7, 8

[23] Yunhui Guo, Noel C Codella, Leonid Karlinsky, James V Codella, John R Smith, Kate Saenko, Tajana Rosing, and Rogerio Feris. A broader study of cross-domain few-shot learning. In *European conference on computer vision*, pages 124–141. Springer, 2020. 2, 5, 6, 8

[24] Hanwen Liang, Qiong Zhang, Peng Dai, and Juwei Lu. Boosting the generalization capability in cross-domain few-shot learning via noise-enhanced supervised autoencoder. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9424–9434, 2021. 2, 8

[25] Debasmit Das, Sungrack Yun, and Fatih Porikli. Confess: A framework for single source cross-domain few-shot learning. In *International Conference on Learning Representations*, 2021. 2, 8

[26] Wei-Hong Li, Xialei Liu, and Hakan Bilen. Universal representation learning from multiple domains for few-shot classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9526–9535, 2021. 2

[27] Yuqian Fu, Yanwei Fu, and Yu-Gang Jiang. Meta-fdmixup: Cross-domain few-shot learning guided by labeled target data. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 5326–5334, 2021. 2

[28] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2

[29] Hung-Yu Tseng, Hsin-Ying Lee, Jia-Bin Huang, and Ming-Hsuan Yang. Cross-domain few-shot classification via learned feature-wise transformation. In *International Conference on Learning Representations*, 2019. 2, 5, 6, 8

[30] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in Neural Information Processing Systems*, 33:9912–9924, 2020. 4

[31] Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, and Luc V Gool. Revisiting contrastive methods for unsupervised learning of visual representations. *Advances in Neural Information Processing Systems*, 34:16238–16250, 2021. 4

[32] Tri Huynh, Simon Kornblith, Matthew R Walter, Michael Maire, and Maryam Khademi. Boosting contrastive self-supervised learning with false negative cancellation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2785–2795, 2022. 4

[33] Akash Srivastava, Lazar Valkov, Chris Russell, Michael U Gutmann, and Charles Sutton. Veegan: Reducing mode collapse in gans using implicit variational learning. *Advances in neural information processing systems*, 30, 2017. 4

[34] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1199–1208, 2018. 6, 8

[35] Victor Garcia and Joan Bruna. Few-shot learning with graph neural networks. In *6th International Conference on Learning Representations, ICLR 2018*, 2018. 6, 8

[36] Yanxu Hu and Andy J Ma. Adversarial feature augmentation for cross-domain few-shot classification. In *European Conference on Computer Vision*, pages 20–37. Springer, 2022. 6, 7, 8

[37] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016. 6

[38] Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. Visualizing the loss landscape of neural nets. *Advances in neural information processing systems*, 31, 2018. 6