CROSS-LEVEL DISTILLATION AND FEATURE DENOISING FOR CROSS-DOMAIN FEW-SHOT CLASSIFICATION

Hao ZHENG*
Tokyo Institute of Technology
zheng.h.ad@m.titech.ac.jp

Runqi Wang*
Huawei Noah's Ark Lab
runqiwangstu@hotmail.com

Jianzhuang Liu
Huawei Noah's Ark Lab
liu.jianzhuang@huawei.com

Asako Kanezaki†
Tokyo Institute of Technology
kanezaki@c.titech.ac.jp

ABSTRACT

The conventional few-shot classification aims at learning a model on a large labeled base dataset and rapidly adapting to a target dataset that is from the same distribution as the base dataset. However, in practice, the base and the target datasets of few-shot classification are usually from different domains, which is the problem of cross-domain few-shot classification. We tackle this problem by making a small proportion of unlabeled images in the target domain accessible in the training stage. In this setup, even though the base data are sufficient and labeled, the large domain shift still makes transferring the knowledge from the base dataset difficult. We meticulously design a cross-level knowledge distillation method, which can strengthen the ability of the model to extract more discriminative features in the target dataset by guiding the network's shallow layers to learn higher-level information. Furthermore, in order to alleviate the overfitting in the evaluation stage, we propose a feature denoising operation which can reduce the feature redundancy and mitigate overfitting. Our approach can surpass the previous state-of-the-art method, Dynamic-Distillation, by 5.44% on 1-shot and 1.37% on 5-shot classification tasks on average in the BSCD-FSL benchmark. The implementation code will be available at https://gitee.com/mindspore/models/tree/master/research/cv/CLDFD.

1 INTRODUCTION

Deep learning has achieved great success on image recognition tasks with the help of a large number of labeled images. However, it is the exact opposite of the human perception mechanism which can recognize a new category by learning only a few samples. Besides, a large amount of annotations is costly and unavailable for some scenarios. It is more valuable to study few-shot classification which trains a classification model on a base dataset and rapidly adapts it to the target dataset. However, due to the constraint that the base data and the target data need to be consistent in their distributions, the conventional few-shot classification may not cater to the demands in some practical scenarios. For example, it may fail in scenarios where the training domain is natural images, but the evaluation domain is satellite images. Considering this domain shift in practical applications, we focus on cross-domain few-shot classification (CD-FSC) in this paper. Previous methods, such as (Mangla et al., 2020; Adler et al., 2020; Tseng et al., 2020), can handle this problem with small domain gaps. However, the CD-FSC problem with a large domain gap is still a challenge.

BSCD-FSC (Guo et al., 2020) is a suitable benchmark for studying this problem, where the base dataset has natural images and the target datasets contain satellite images, crop disease images, skin disease images and X-ray images of sundry lung diseases. On this benchmark, previous methods following the traditional CD-FSC protocol train their models on the base dataset and evaluate them

*Co-first author. Part of this work was done during an internship in Huawei Noah's Ark Lab.
†Corresponding author.

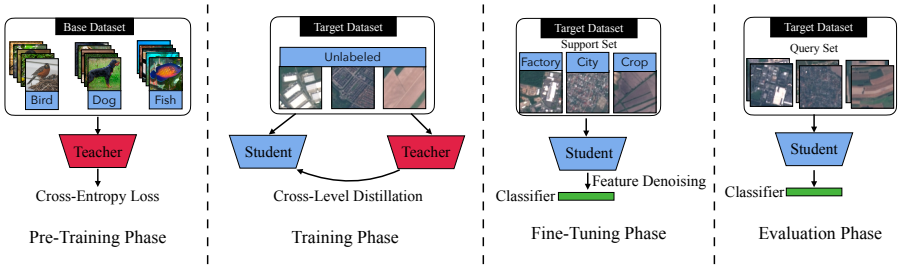


Figure 1: Our CD-FSC framework. The first phase is pre-training, which trains the teacher network on the labeled base dataset by optimizing the cross-entropy loss. The second phase trains the student network using our proposed cross-level distillation (CLD). The third phase fine-tunes a linear classifier on a few labeled images in the target domain, and feature denoising (FD) is conducted to remove the noise in the final feature vectors. The final phase classifies images in the target domain.

on the target dataset, but their performances are far from satisfactory. STARTUP (Phoo & Hariharan, 2021) and Dynamic-Distillation (Islam et al., 2021a) introduce a more realistic setup that makes a small portion of the unlabeled target images accessible during the training phase. These target images bring a prior to the model and dramatically promote the model's performance on the target datasets. Inspired by that, we follow their setup to explore the CD-FSC problem with a large domain shift. In this work, we propose a cross-level distillation (CLD), which can effectively transfer the knowledge from the base dataset and improve the performance of the student network on the target domain. Besides, we propose feature denoising (FD) to remove the noise in the features during the fine-tuning stage. Our CD-FSC framework is given in Figure 1.

The detail of CLD is shown in Figure 2, which distills a teacher's deeper layers to a student's shallower layers, where the student and the teacher share the same structure. Unlike the distillation methods in STARTUP and Dynamic-Distillation, which only distill the teacher's last layer to the student's last layer, our CLD leads the shallow layers of the student to mimic the features generated from the deeper levels of the teacher so that the student can learn more deeper semantic information and extract more discriminative features on the target dataset. Additionally, since the teacher networks in STARTUP and Dynamic-Distillation are pre-trained on the base dataset only, the teacher's observation of the target data is biased. In order to calibrate the bias, we design an iterative process by building another network which shares the same structure and parameters with the historical student network, named old student network. In each training iteration, the features from the teacher and the old student in the same layers are dynamically fused to guide the corresponding layers of the student. The latter the training iteration, the fewer fusion features from the teacher network, and the more from the old student network. Due to the target data used in training is unlabeled, the self-supervised loss is introduced to excavate the target domain information further.

The self-supervised loss not only supports the network in mining valuable information on the target domain, but also brings a phenomenon where the final feature vector for classification has a small number of dominant (strongly activated) elements with the others are close to zero (Hua et al., 2021; Kalibhat et al., 2022). We find that during the fine-tuning phase in Figure 1, these small activated elements are redundant and considered as noise. Our FD operation keeps the top h largest elements and sets the others to zero. It is experimentally verified that FD can greatly improve the model's performance.

Above all, our main contributions are summarized below:

- We propose a cross-level distillation (CLD) framework, which can well transfer the knowledge of the teacher trained on the base dataset to the student. We also use an old student network mechanism is also necessary to calibrate the teacher's bias learned from the base data.
- Considering the noisy feature activations, we design a feature denoising (FD) operation that can significantly improve the performance of our model.
- Extensive experiments are conducted to verify that our proposed CLD and FD can achieve state-of-the-art results on the BSCD-FSL benchmark with large domain gaps.

2 RELATED WORK

Cross-domain few-shot classification. The cross-domain few-shot classification is firstly defined by (Chen et al., 2018), which trains a model on the base dataset and evaluates it on the target dataset in a different domain from the base dataset. LFT (Tseng et al., 2020) simulates the domain shift from the base dataset to the target dataset by meta-learning and inserts linear layers into the network to align the features from the different domains. Meta-FDMixup (Fu et al., 2021) uses several labeled images from the target dataset for domain-shared feature disentanglement and feeds the domain-shared features to the classifier. FLUTE (Triantafillou et al., 2021) learns the universal templates of features across multi-source domains to improve the transferability of the model. However, all these methods concentrate on the CD-FSC problem with small domain shifts. Some methods handle CD-FSC with large domain gaps, in which the target datasets have obvious dissimilarity from the base dataset on perspective distortion, semantics, and/or color depth. For example, ConfeSS (Das et al., 2021) extracts useful feature components from labeled target images. ATA (Wang & Deng, 2021) does not require any prior of the target dataset and proposes a plug-and-play inductive bias-adaptive task augmentation module. CI (Luo et al., 2022) trains an encoder on the base dataset and converts the features of the target data with a transformation function in the evaluation stage. UniSiam (Lu et al., 2022) adopts a self-supervised approach to address the CD-FSC problem. Among the methods dealing with large domain shifts, STARTUP (Phoo & Hariharan, 2021) is a strong baseline, which uses a few unlabeled target images in the training stage. It firstly trains a teacher network on the base dataset in a supervised fashion and transfers the teacher's knowledge to the student by knowledge distillation (KD). It jointly optimizes the cross-entropy loss with the base dataset, contrastive loss of the unlabeled target images and the KD loss to upgrade the student network. Dynamic-Distillation (Islam et al., 2021a) also uses a small number of unlabeled images and KD. The main difference between Dynamic-Distillation and STARTUP is that the former upgrades the pre-trained teacher dynamically by exponential moving averages, while the latter fixes the teacher. In our work, we follow their data setup allowing a small proportion of the unlabeled target images to be seen during the training phase. Different from the two methods that perform KD at the last layers of the teacher and the student, our KD is carried out at cross levels. Besides, our denoising operation further improves the performance.

Self-supervised learning. Self-supervised learning is widely used in the scenarios where labels are not available for training. It defines a "pretext task" to pre-train a network. For example, (Gidaris et al., 2018) pre-trains the model by predicting the rotation angle of the image. One popular method is contrastive learning, such as SimCLR (Chen et al., 2020) which pulls different augmented versions of the same image closer and pushes the versions from different images away. Beyond contrastive learning, BYOL (Grill et al., 2020) and SimSiam (Chen & He, 2021) rely only positive pairs.

Knowledge distillation. (Hinton et al., 2015) firstly propose knowledge distillation by guiding a compact network (student) to mimic the output of a large network (teacher). Since features in the intermediate layers are informative, some previous methods distill the teacher's intermediate features (Romero et al., 2014) or attention maps of the features (Zagoruyko & Komodakis, 2017). Besides, self-distillation methods, such as BYOT (Zhang et al., 2019), distill the last layer of the network to its own shallower layers. BYOT's teacher and student are the same network.

In our framework (Figure 2), in addition to the KD in the intermediate layers, we design an old student that shares the same structure as the student but with different parameters. The introduction of this old student not only alleviates the teacher's bias learned from the base dataset, but also has the effect of assembling multiple historic students during training.

3 METHODOLOGY

3.1 PRELIMINARY

During the training period, we follow the setting in STARTUP (Phoo & Hariharan, 2021) and Dynamic-Distillation (Islam et al., 2021a), where a labeled base dataset $\mathcal{D}_{\mathrm{B}}$ and a few unlabeled target images sampled from the target dataset $\mathcal{D}_{\mathrm{T}}$ are available. In the testing stage, the support set $\mathcal{D}_{\mathrm{S}}$ comprises N classes, and K samples are randomly selected from each class in $\mathcal{D}_{\mathrm{T}}$, which is the so-called N-way K-shot task. The support set $\mathcal{D}_{\mathrm{S}}$ is for fine-tuning a new classifier with the frozen encoder (the student network in this work). The images in the query set $\mathcal{D}_{\mathrm{Q}}$ are randomly picked

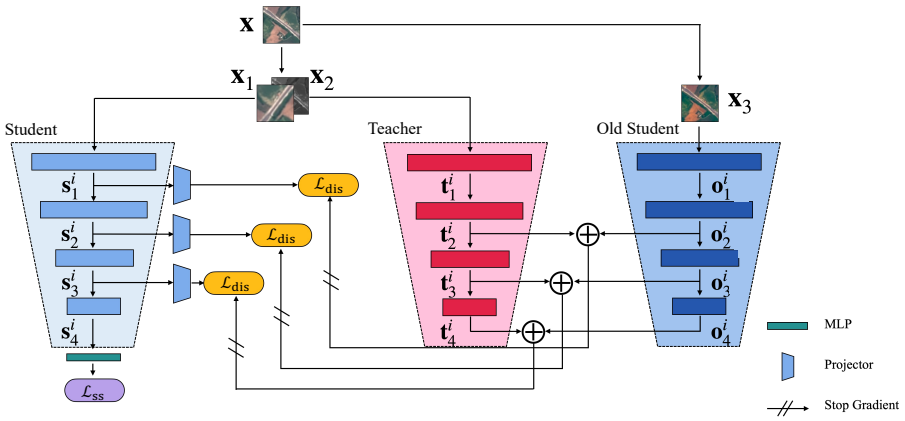


Figure 2: Framework of CLD for knowledge distillation (KD). The teacher network is pre-trained on the labeled base dataset with the cross-entropy loss and is fixed during KD. When training the student network, the target image $\mathbf{x}$ is augmented into $\mathbf{x}_1$, $\mathbf{x}_2$ and $\mathbf{x}_3$. Then $\mathbf{x}_1, \mathbf{x}_2$ is fed to both the student and the teacher, and $\mathbf{x}_3$ is fed to the old student. At the i-th iteration, the parameters of the old student are a copy of those of the student at $(i-\tau)$-th iteration. The feature $\mathbf{s}_l^i$ of the student is firstly projected by ω_l for dimensionality alignment, where l is the block index. Then we fuse the features $\mathbf{t}_{l+1}^i$ and $\mathbf{o}_{l+1}^i$ obtaining $\mathbf{u}_{l+1}^i$, which are from the $(l+1)$-th block of the teacher and the old student, respectively. The KD is conducted by forcing $\omega_l(\mathbf{s}_l^i)$ to mimic $\mathbf{u}_{l+1}^i$. Additionally, the self-supervised loss $\mathcal{L}_{ss}$ is introduced on the student network.

from the selected N classes for evaluating the classification accuracy. The support set $\mathcal{D}_{\mathrm{S}}$ and the query set $\mathcal{D}_{\mathrm{Q}}$ have no overlap.

3.2 Cross-level distillation

The proposed cross-level distillation (CLD) framework is shown in Figure 2. The teacher network f_{t} is pre-trained on $\mathcal{D}_{\mathrm{B}}$ with the cross-entropy loss. The student network f_{s} is expected to inherit the knowledge of the teacher and extract discriminative features of $\mathcal{D}_{\mathrm{T}}$. However, the teacher's observation of the target data is biased since it is pre-trained on the base dataset only. In the i-th training iteration, if the features extracted by the student f_{s}^i directly mimic the features of the teacher, the teacher's bias will be transferred to the student. To reduce the bias, we introduce an *old student network* f_{o}^i, which is a copy of $f_{\mathrm{s}}^{i-\tau}$, where the hyper-parameter τ denotes the training iteration interval between $f_{\mathrm{s}}^{i-\tau}$ and f_{s}^i.

To simplify the KD complexity, we divide each backbone of f_{s}, f_{o}, and f_{t} into L residual blocks. Let $\mathbf{s}_l^i$, $\mathbf{o}_l^i$, and $\mathbf{t}_l^i$ be the features obtained by the student, the old student, and the teacher in the l-th block at the i-th iteration. The fusion between $\mathbf{t}_l^i$ and $\mathbf{o}_l^i$ is defined as:

$$\mathbf{u}_l^i = \alpha^i \mathbf{o}_l^i + (1-\alpha^i)\mathbf{t}_l^i, \tag{1}$$

where $\alpha^i = \frac{i}{T}$ is a dynamic weight with T being the total number of training iterations. The KD loss $\mathcal{L}_{\mathrm{dis}}^i$ in the i-th iteration is defined as:

$$\mathcal{L}_{\mathrm{dis}}^i = \begin{cases} \sum_{l=1}^{L} \|\omega_l(\mathbf{s}_l^i) - \mathbf{t}_{l+1}^i\|_2^2 & \text{if } i \leq \tau \\ \sum_{l=1}^{L} \|\omega_l(\mathbf{s}_l^i) - \mathbf{u}_{l+1}^i\|_2^2 & \text{otherwise,} \end{cases} \tag{2}$$

where $\omega_l(\cdot)$ is a projector of the l-th student's block for feature dimensionality alignment. It is comprised of a convolutional layer, a batch normalization operator and ReLU activation. Note that the KD in Equation 2 is from the $(l+1)$-th block of the teacher to the l-th block of the student. In our experiments, we find that this style is better than others (see Section 4.3).

The total loss $\mathcal{L}$ for training the student network is:

$$\mathcal{L} = \mathcal{L}_{\mathrm{ss}} + \lambda \mathcal{L}_{\mathrm{dis}}, \tag{3}$$

where $\mathcal{L}_{\text{ss}}$ is a self-supervised loss, and λ is a weight coefficient of loss function for balancing the two losses.

For $\mathcal{L}_{\text{ss}}$, off-the-shelf self-supervised losses like SimCLR (Chen et al., 2020) or BYOL (Grill et al., 2020) can be used. The contrastive loss in SimCLR is:

$$\mathcal{L}_{\text{simclr}} = -\frac{1}{|\mathbf{B}|} \sum_{m,n \in \mathbf{B}} \log \frac{\exp\left(\text{sim}\left(\mathbf{z}_m, \mathbf{z}_n\right)/\gamma\right)}{\sum_{q=1, q \neq m}^{2|\mathbf{B}|} \exp\left(\text{sim}\left(\mathbf{z}_m, \mathbf{z}_q\right)/\gamma\right)}, \tag{4}$$

where $\mathbf{z}_m$, $\mathbf{z}_n$ and $\mathbf{z}_q$ are the projected embeddings of different augmentations, (m, n) is a positive pair from the same image, $\text{sim}(\cdot)$ is a similarity function, $\mathbf{B}$ is a mini-batch of unlabeled target images, and γ is a temperature coefficient. The self-supervised loss in BYOL is:

$$\mathcal{L}_{\text{byol}} = \sum_{m \in \mathbf{B}} 2 - 2 \cdot \frac{\text{sim}(p(\mathbf{z}_m), \mathbf{z}'_m)}{\|p(\mathbf{z}_m)\|_2 \cdot \|\mathbf{z}_{m'}\|_2}, \tag{5}$$

where $\mathbf{z}_m$ and $\mathbf{z}'_m$ are embeddings of the online network and the target network, respectively, and $p(\cdot)$ is a linear predictor.

Note that the last convolution block of the student network is not involved in KD and is trained by minimizing $\mathcal{L}_{\text{ss}}$ only. The reason is that the last block mainly discovers semantic information that is highly domain-specific. Therefore, we constrain it on the target data rather than letting it learn from the teacher that is pre-trained on the base data.

3.3 Feature Denoising

The self-supervised loss brings a phenomenon where the final feature vector for classification has a small number of strongly activated elements while the others are close to zero (Hua et al., 2021; Kalibhat et al., 2022). These elements of small magnitudes are regarded as noise, which may cause overfitting. We propose a feature denoising (FD) operation to remove their contribution during the fine-tuning phase (see Figure 1). FD keeps the largest h elements of the feature from the student network and zeros the other elements. The FD operation is only performed on the feature $\mathbf{s}_L$ in the last layer of the student network. Specifically, let $\mathbf{s}_L = [s_{(L,1)}, s_{(L,2)}, \ldots, s_{(L,D_L)}]$ and $\widetilde{\mathbf{s}}_L = [\widetilde{s}_{(L,1)}, \widetilde{s}_{(L,2)}, \ldots, \widetilde{s}_{(L,D_L)}]$ be the features before and after the FD operation, respectively, where D_L is the feature's dimensionality. Then the FD operation is defined as:

$$\widetilde{s}_{L,d} = \begin{cases} (s_{L,d})^{\beta} & \text{if } s_{L,d} \in top_h(\mathbf{s}_L),\ d = 1, 2, ..., D_L \\ 0 & \text{otherwise,} \end{cases} \tag{6}$$

where β is a hyper-parameter which makes the non-zero elements more distinguishable and $top_h(\cdot)$ is the operator selecting the largest h elements of the feature. Finally, $\widetilde{\mathbf{s}}_M$ is fed to the classifier for fine-tuning.

4 Experiments

4.1 Datasets and Implementation

Datasets. We evaluate the proposed CLD and FD for the CD-FSC problem on the BSCD-FSL benchmark (Guo et al., 2020) with large domain gaps. The miniImageNet dataset (Vinyals et al., 2016) serves as the base dataset $\mathcal{D}_B$ which has sufficient labeled images. EuroSAT (Helber et al., 2019), CropDisease (Mohanty et al., 2016), ISIC (Codella et al., 2019) and ChestX (Wang et al., 2017) in BSCD-FSL are the unlabeled target datasets $\mathcal{D}_T$. We follow the training protocol in STARTUP (Phoo & Hariharan, 2021) and Dynamic-Distillation (Islam et al., 2021a), which allows the whole labeled training set of miniImageNet and a small proportion (20%) of unlabeled target images available during the training period. The remaining 80% of target images are utilized for fine-tuning and evaluating by building 5-way K-shot tasks, $K \in \{1, 5\}$.

Implementation details. We implement our model using the MindSpore Lite tool (Mindspore). For a fair comparison, all the methods use ResNet-10 (Guo et al., 2020) as the backbone. Our

Table 1: The averaged 5-way 1-shot and 5-shot averaged accuracy and 95% confidence interval among 600 episodes are given. The reported results of SimCLR (Base) and previous state-of-the-art Dynamic-Distillation are from (Islam et al., 2021a). The results of CI are from (Luo et al., 2022). The results of ATA are from (Wang & Deng, 2021). The performance of ConFeSS refers to (Das et al., 2021) and they do not give the confidence interval. The champion results are marked in bold.

	EuroSAT		CropDisease		ISIC		ChestX	
	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
Transfer	58.42±0.94	75.31±0.71	68.45±0.87	89.12±0.52	32.82±0.60	47.13±0.58	22.48±0.41	26.65±0.43
SimCLR (Base)	58.28±0.90	80.83±0.64	68.26±0.86	83.44±0.61	32.15±0.59	45.90±0.58	22.37±0.42	26.63±0.46
CI	58.82±0.92	76.26±0.70	61.58±0.88	89.25±0.51	32.43±0.56	44.04±0.55	23.23±0.41	27.20±0.44
Transfer+SimCLR	65.92±0.88	81.83±0.59	81.39±0.80	94.85±0.38	33.86±0.61	47.25±0.59	22.91±0.45	27.03±0.43
STARTUP	65.24±0.88	81.60±0.59	78.18±0.83	92.86±0.44	34.15±0.62	46.42±0.58	22.86±0.43	26.98±0.43
ATA	65.94±0.50	79.47±0.30	77.82±0.50	88.15±0.50	34.70±0.40	45.83±0.30	21.67±0.20	23.60±0.20
BYOL	64.38±0.93	82.44±0.56	67.15±1.17	89.40±0.63	29.21±0.51	41.44±0.51	21.66±0.31	25.99±0.43
ConFeSS	–	84.65	–	88.88	–	48.85	–	27.09
Dynamic-Distillation	73.14±0.84	89.07±0.47	82.14±0.78	95.54±0.38	34.66±0.58	49.36±0.59	23.38±0.43	28.31±0.46
SimCLR	75.10±0.85	90.17±0.43	88.54±0.80	96.09±0.39	35.71±0.66	48.84±0.61	22.00±0.42	24.84±0.41
BYOL+CLD+FD (ours)	72.78±0.84	88.50±0.45	86.94±0.73	95.85±0.38	37.70±0.67	51.19±0.63	**24.34±0.44**	**30.15±0.44**
SimCLR+CLD+FD (ours)	**82.52±0.76**	**92.89±0.34**	**90.48±0.72**	**96.58±0.39**	**39.70±0.69**	**52.29±0.62**	22.39±0.44	25.98±0.43

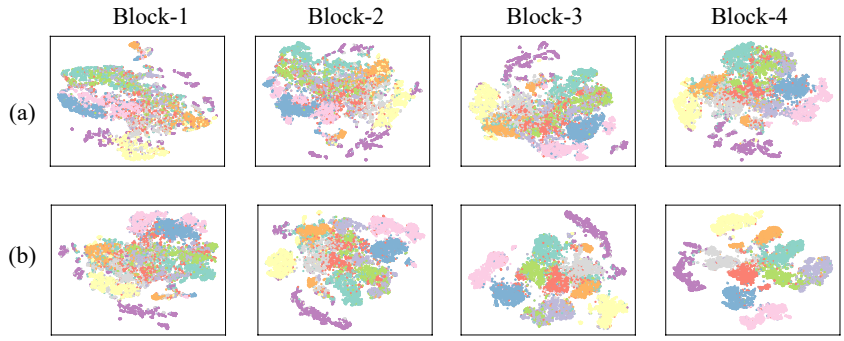
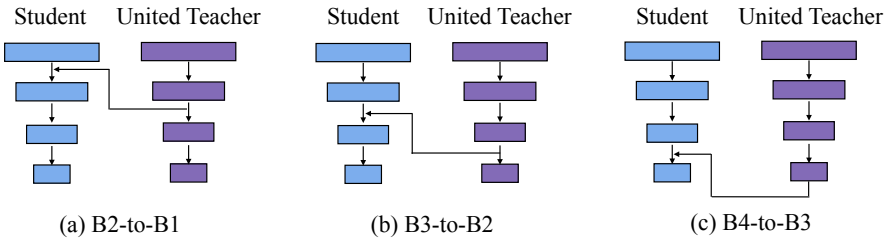


Figure 3: t-SNE results of (a) STARTUP and (b) SimCLR+CLD+FD at different blocks on the EuroSAT dataset. It is obvious that our method can extract more discriminative features.

model is optimized by SGD with the momentum 0.9, weight decay 1e-4, and batch size 32 for 600 epochs. The learning rate is 0.1 at the beginning and decays by 0.1 after the 300th epoch and 500th epoch. The hyper-parameters λ in Equation 3, and h and β in Equation 6 are set to 2, 64, and 0.4, respectively. For fine-tuning and evaluating the student network, we randomly sample 600 episodes of 5-way K-shot tasks. The performance is represented by the average classification accuracy over the 600 episodes within the 95% confidence interval. In each episode, the parameters of the student are frozen. An additional linear classifier is fine-tuned by minimizing the cross-entropy loss. The above mentioned hyperparameters are determined based on the EuroSAT dataset only, and then they are used for evaluation on all the datasets, showing the generalization ability of our method.

4.2 Main Results

We select several basic and competitive methods for comparison in Table 1. The basic model is Transfer that trains the encoder on the labeled base dataset with the cross-entropy loss. SimCLR (Base) (Chen et al., 2020) trains the model on the base dataset (miniImageNet) with the contrastive loss of SimCLR. CI (Luo et al., 2022) trains the encoder on the base dataset and converts the features of the target dataset with a transformation function in the evaluation stage. Transfer+SimCLR is proposed in (Islam et al., 2021b) and exhibits good transferability, which simultaneously optimizes the cross-entropy loss on the base dataset and the contrastive loss of SimCLR on the target dataset. STARTUP (Phoo & Hariharan, 2021) trains the model with three loss functions: cross-



(a) B2-to-B1 (b) B3-to-B2 (c) B4-to-B3

Figure 4: Different single-block distillation structures. B2-to-B1 means that the second block of the teacher is distilled to the first block of the student. B3-to-B2 and B4-to-B3 are similar.

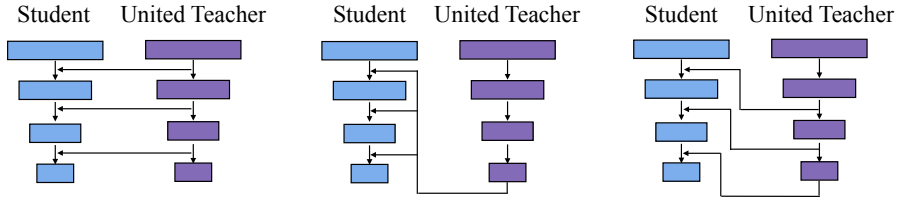
Table 2: Results of different single-block distillation stuctures on EuroSAT (5-way 1-shot). B2-to-B1, B3-to-B2, and B4-to-B3 are defined in Figure 4.

SimCLR	B2-to-B1	B3-to-B2	B4-to-B3	CLD
75.10±0.85	78.54±0.80	78.38±0.77	77.92±0.83	79.38±0.82

entropy loss on the labeled base dataset, and contrastive loss and KD loss on the unlabeled target domain. ATA (Wang & Deng, 2021) designs a plug-and-play inductive bias-adaptive task augmentation module. BYOL (Grill et al., 2020) is trained on the target dataset. ConfeSS (Das et al., 2021) uses labeled target images to find out the useful components of the target image features. Dynamic-Distillation (Islam et al., 2021a) designs a distillation loss that draws support from the dynamically updated teacher network. SimCLR trains the encoder on unlabeled target data only and shows good generalization ability. For our method, we build two models BYOL+CLD+FD and SimCLR+CLD+FD, which use Equations 5 and 4 for self-supervised losses, respectively.

Table 1 gives the comparisons among our models and the baselines. All the best results on the four datasets are obtained by either of our two models. In particular on EuroSAT, our SimCLR+CLD+FD can bring 9.38% and 7.42% gains over Dynamic-Distillation and SimCLR on the 5-way 1-shot task, respectively. On average, on the four datasets, SimCLD+CLD+FD outperforms Dynamic-Distillation significantly (58.77% vs. 53.33%) for 1-shot; 66.94% vs. 65.57% for 5-shot. Besides, although SimCLR+CLD+FD and BYOL+CLD+FD are based on SimCLR and BYOL, respectively, the performances of SimCLR and BYOL are improved greatly. Finally, we visualize the feature distributions of STARTUP and our SimCLR+CLD+FD in Figure 3. It can be seen that SimCLR+CLD+FD exhibits better clustering results at all the blocks, especially at the last one.

4.3 Ablation Study on Cross-level Distillation



(a) same-level distillation (b) one-to-all distillation (c) CLD

Figure 5: Different multi-blocks distillation structures. (a) and (b) are two common KD methods. (c) is our cross-level KD.

For simplicity, the combination (according to Equation 1) of the teacher and the old student in Figure 2 is denoted as the united teacher in Figure 4 and Figure 5. In these experiments, $\mathcal{L}_{\text{simclr}}$ is used as $\mathcal{L}_{\text{ss}}$ and FD is not employed. Table 2 gives the results of SimCLR, different single-block distillation structures, and CLD on EuroSAT (5-way 1-shot). We can notice that all the single-block distillation structures perform better than SimCLR, and CLD can outperform all the single-block distillation

Table 3: Results of different multi-blocks KD structures.

	EuroSAT		CropDisease		ISIC		ChestX	
	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
SimCLR	75.10±0.85	90.17±0.43	88.54±0.80	**96.09±0.39**	35.71±0.66	48.84±0.61	22.00±0.42	24.84±0.41
same-level distillation	75.48±0.85	90.71±0.40	88.14±0.80	95.76±0.44	36.56±0.65	50.51±0.58	**22.12±0.42**	**25.20±0.42**
one-to-all distillation	77.70±0.83	91.80±0.38	89.11±0.79	95.83±0.44	36.69±0.64	50.64±0.62	22.10±0.43	24.77±0.43
CLD	**79.38±0.82**	**92.78±0.35**	**89.43±0.78**	96.08±0.43	**37.82±0.68**	**52.28±0.61**	22.08±0.43	25.08±0.42

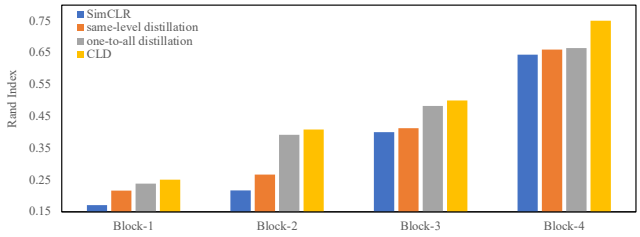


Figure 6: Rand Index values in the residual blocks of the methods in Table 3.

structures. Table 3 gives the results of SimCLR, different multi-blocks distillation structures and CLD. Compared with SimCLR without KD, all the multi-blocks KD structures improve the model's overall performance. Second, the one-to-all KD outperforms the same-level KD in most cases. Finally, our CLD performs best on average among the three multi-blocks KD structures.

We further explore the reason that CLD can exceed other methods. Rand Index (Rand, 1971) is a metric to reflect the quality of feature clustering. The bigger Rand Index is, the better clustering the method provides. Figure 6 is comprised of the Rand Index values of the features in each residual block on the EuroSAT test dataset. Our CLD increases more than all the other methods in each block. It means that our CLD can pick up more discriminative information at each level so that the model gradually hunts the useful features as the network deepens.

Next, we examine the usefulness of the old student and the effect of the training iteration interval τ (Section 3.2). Table 4 shows the experimental results of different settings. First, CLD without the old student outperforms SimCLR. Second, using the old student is better than not using it. Considering the performance and the memory requirement, we choose $\tau = 1$ in all the other experiments on the four datasets.

Table 4: Effects of the old student and the training iteration interval τ. All the methods in the table use $\mathcal{L}_{\text{simclr}}$ as $\mathcal{L}_{\text{ss}}$.

	EuroSAT		CropDisease		ISIC		ChestX	
	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
SimCLR	75.10±0.85	90.17±0.43	88.54±0.80	**96.09±0.39**	35.71±0.66	48.84±0.61	22.00±0.42	24.84±0.41
CLD (w/o old student)	76.85±0.83	91.58±0.40	88.72±0.78	95.85±0.43	37.04±0.65	50.90±0.60	21.40±0.32	23.91±0.39
CLD ($\tau = 0$)	78.56±0.82	92.32±0.35	89.11±0.81	95.69±0.45	37.82±0.69	51.17±0.62	22.03±0.43	25.12±0.44
CLD ($\tau = 1$)	**79.38±0.82**	**92.78±0.35**	89.43±0.78	96.08±0.43	37.82±0.68	**52.28±0.61**	**22.08±0.43**	25.08±0.42
CLD ($\tau = 10$)	78.18±0.81	92.17±0.36	89.24±0.80	95.83±0.44	**38.02±0.68**	51.77±0.62	22.02±0.43	25.04±0.43
CLD ($\tau = 100$)	78.79±0.80	92.44±0.37	**89.56±0.81**	96.06±0.42	36.81±0.65	50.60±0.62	22.29±0.43	**25.17±0.43**

4.4 Ablation Study on Feature Denoising

On EuroSAT, the model SimCLR+CLD+FD is used to find optimal h and β on the 5-way 1-shot tasks. As shown in Figure 7, $h = 64$ and $\beta = 0.4$ are chosen for the best accuracy.

With $h = 64$ and $\beta = 0.4$ in FD, we compare using FD or not in Table 5. We can see that the first five models are improved with FD, while the last two are not. The different positions, which the

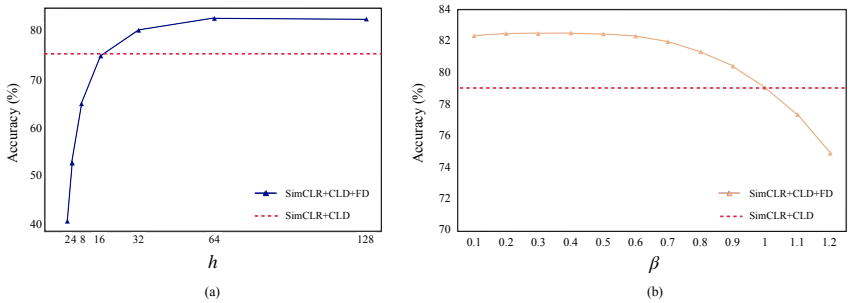


Figure 7: (a) Accuracy *vs.* h when $\beta = 0.4$. (b) Accuracy *vs.* β when $h = 64$.

Table 5: Comparison of using FD or not on EuroSAT (5-way 1-shot).

	SimCLR	BYOL	SimCLR+CLD	BYOL+CLD	one-to-all distillation	Transfer+SimCLR	STARTUP
w/o FD	75.10±0.85	64.38±0.93	79.38±0.82	69.47±0.84	77.70±0.83	65.92±0.88	65.24±0.88
w/ FD	78.02±0.77	69.80±0.85	82.52±0.76	72.78±0.84	80.62±0.77	63.97±0.85	60.87±0.88
Δ	+2.92	+5.42	+3.20	+3.31	+2.92	-1.95	-4.37

auxiliary loss (cross-entropy and/or KD) is applied to in the two groups, are shown in Figures 8(a) and 8(b), respectively. The reason why FD does not help in Figure 8(b) is that the loss $\mathcal{L}_{\text{aux}}$ used in the final layer reduces the noisy elements of the final feature vector for classification.

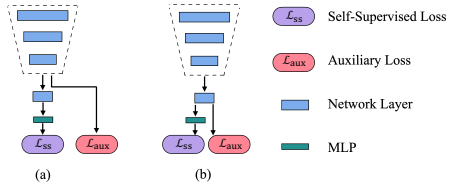


Figure 8: $\mathcal{L}_{\text{ss}}$ is the self-supervised loss and $\mathcal{L}_{\text{aux}}$ includes other losses such as the cross-entropy and KD losses. (a) $\mathcal{L}_{\text{aux}}$ is applied to intermediate layers. (b) $\mathcal{L}_{\text{aux}}$ is applied to the final layer.

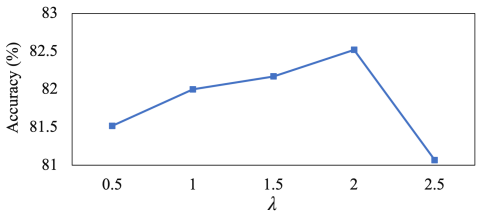


Figure 9: Results of SimCLR+CLD+FD with the different λ values on the EuroSAT dataset (5-way 1-shot).

4.5 Ablation Study on Weight Coefficient of Loss Function

We give the results of SimCLR+CLD+FD with the different λ values in Equation 3 on the EuroSAT dataset (5-way 1-shot), as shown in Figure 9. We can see that the best result is obtained with $\lambda = 2$. So we set $\lambda = 2$ in all the related experiments.

5 Conclusion

In this work, we handle the CD-FSC problems with large domain gaps between the base dataset and the target datasets. We propose the cross-level distillation KD for better transferring the knowledge of the base dataset to the student. We also present the feature denoising operation for mitigating the overfitting. Our method improves the performance of a strong baseline Dynamic-Distillation by 5.44% on 1-shot and 1.37% on 5-shot classification tasks on average in the BSCD-FSL benchmark, establishing new state-of-the-art results.

ACKNOWLEDGEMENTS

We gratefully acknowledge the support of MindSpore (Mindspore), CANN (Compute Architecture for Neural Networks) and Ascend AI Processor used for this research. We would also like to express our sincere gratitude to Mr. Haichen ZHENG, Mrs. Jing LIN, and Ms. Xiao WANG for their unwavering support and confidence in our work.

REFERENCES

Thomas Adler, Johannes Brandstetter, Michael Widrich, Andreas Mayr, David Kreil, Michael Kopp, Günter Klambauer, and Sepp Hochreiter. Cross-domain few-shot learning by representation fusion. *arXiv:2010.06498*, 2020.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020.

Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Frank Wang, and Jia-Bin Huang. A closer look at few-shot classification. In *ICLR*, 2018.

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *CVPR*, 2021.

Noel Codella, Veronica Rotemberg, Philipp Tschandl, M Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, et al. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). *arXiv:1902.03368*, 2019.

Debasmit Das, Sungrack Yun, and Fatih Porikli. Confess: A framework for single source cross-domain few-shot learning. In *ICLR*, 2021.

Yuqian Fu, Yanwei Fu, and Yu-Gang Jiang. Meta-fdmixup: Cross-domain few-shot learning guided by labeled target data. In *ACM MM*, 2021.

Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *ICLR*, 2018.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. In *NeurIPS*, 2020.

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 *ECCV*, 2020.

Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, pp. 2217–2226, 2019.

Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. Distilling the knowledge in a neural network. *NeurIPS Workshop*, 2015.

Tianyu Hua, Wenxiao Wang, Zihui Xue, Sucheng Ren, Yue Wang, and Hang Zhao. On feature decorrelation in self-supervised learning. In *ICCV*, 2021.

Ashraful Islam, Chun-Fu Richard Chen, Rameswar Panda, Leonid Karlinsky, Rogerio Feris, and Richard J Radke. Dynamic distillation network for cross-domain few-shot recognition with unlabeled data. In *NeurIPS*, 2021a.

Ashraful Islam, Chun-Fu Richard Chen, Rameswar Panda, Leonid Karlinsky, Richard Radke, and Rogerio Feris. A broad study on the transferability of visual representations with contrastive learning. In *ICCV*, 2021b.

Neha Mukund Kalibhat, Kanika Narang, Hamed Firooz, Maziar Sanjabi, and Soheil Feizi. Towards better understanding of self-supervised representations. In *ICML Workshop*, 2022.

Yuning Lu, Liangjian Wen, Jianzhuang Liu, Yajing Liu, and Xinmei Tian. Self-supervision can be a good few-shot learner. In *ECCV*, 2022.

Xu Luo, Jing Xu, and Zenglin Xu. Channel importance matters in few-shot image classification. In *ICML*, 2022.

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 *WACV*, 2020.

Mindspore. https://www.mindspore.cn/.

Sharada P Mohanty, David P Hughes, and Marcel Salathé. Using deep learning for image-based plant disease detection. *Frontiers in plant science*, pp. 1419–1432, 2016.

Cheng Perng Phoo and Bharath Hariharan. Self-training for few-shot transfer across extreme task differences. In *ICLR*, 2021.

William M Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, pp. 846–850, 1971.

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *arXiv:1412.6550*, 2014.

Eleni Triantafillou, Hugo Larochelle, Richard Zemel, and Vincent Dumoulin. Learning a universal template for few-shot dataset generalization. In *ICML*, 2021.

Hung-Yu Tseng, Hsin-Ying Lee, Jia-Bin Huang, and Ming-Hsuan Yang. Cross-domain few-shot classification via learned feature-wise transformation. In *ICLR*, 2020.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. In *NeurIPS*, 2016.

Haoqing Wang and Zhi-Hong Deng. Cross-domain few-shot classification via adversarial task augmentation. In *IJCAI*, 2021.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *CVPR*, 2017.

Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In *ICLR*, 2017.

Linfeng Zhang, Jiebo Song, Anni Gao, Jingwei Chen, Chenglong Bao, and Kaisheng Ma. Be your own teacher: Improve the performance of convolutional neural networks via self distillation. In *ICCV*, 2019.

A APPENDIX

A.1 RESULTS OF THE DIFFERENT BATCH SIZES

We present the results of the SimCLR+CLD+FD trained with different batch sizes on the EuroSAT dataset (5-way 1-shot) in Figure 10. We can see that the optimal batch size is 32, so we utilize this value in all the experiments of our method.

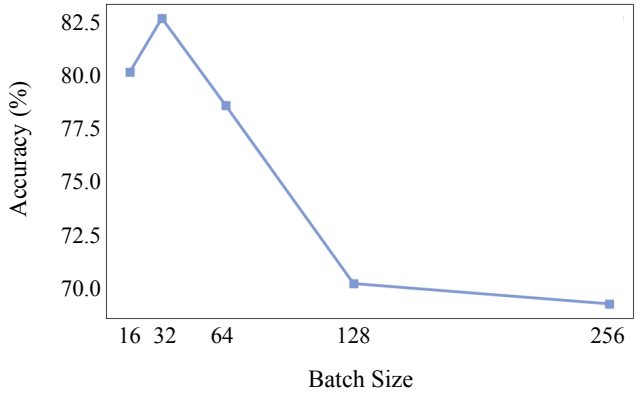


Figure 10: Results of the SimCLR+CLD+FD with the different batch sizes on the EuroSAT dataset (5-way 1-shot).

A.2 RESULTS OF THE DIFFERENT PERCENTAGES OF UNLABELED IMAGES

Figure 11 shows that as more unlabeled target images become available, the accuracy of the model will gradually converge to a plateau.

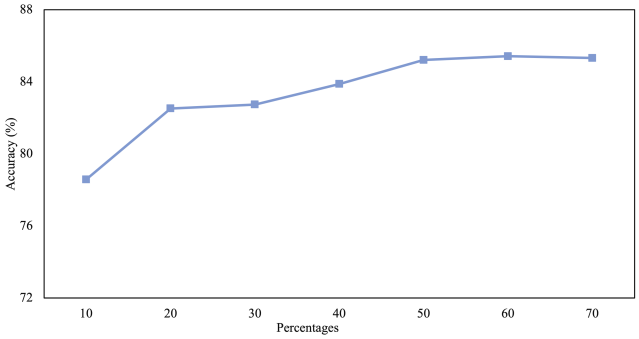


Figure 11: Results of the SimCLR+CLD+FD with the different percentages on the EuroSAT dataset (5-way 1-shot).

A.3 RESULTS WITH SELF-SUPERVISED PRE-TRAINED TEACHER

Table 6: Results with self-supervised & supervised pre-trained teacher.

	EuroSAT		CropDisease		ISIC		ChestX	
	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
SimCLR+CLR+FD (self-supervised)	80.30±0.72	91.63±0.36	89.94±0.72	96.50±0.35	37.42±0.46	49.36±0.64	22.58±0.43	25.96±0.43
SimCLR+CLR+FD (supervised)	82.52±0.76	92.89±0.34	90.48±0.72	96.58±0.39	39.70±0.69	52.29±0.62	22.39±0.44	25.98±0.43

On the EuroSAT dataset (5-way 1-shot), we give the results of the SimCLR+FD+CLD where the teacher is pre-trained in a self-supervised scheme. The performances of the SimCLR+CLD+FD are shown in Table 6.