Few-Shot Learning with Part Discovery and Augmentation from Unlabeled Images

Wentao Chen[1,2] , **Chenyang Si**[2] , **Wei Wang**[2] , **Liang Wang**[2] , **Zilei Wang**[1] , **Tieniu Tan**[1,2*]
[1]University of Science and Technology of China
[2]Center for Research on Intelligent Perception and Computing, NLPR, CASIA
{wentao.chen, chenyang.si}@cripac.ia.ac.cn, {wangwei, wangliang, tnt}@nlpr.ia.ac.cn,
zlwang@ustc.edu.cn

Abstract

Few-shot learning is a challenging task since only few instances are given for recognizing an unseen class. One way to alleviate this problem is to acquire a strong inductive bias via meta-learning on similar tasks. In this paper, we show that such inductive bias can be learned from a flat collection of unlabeled images, and instantiated as transferable representations among seen and unseen classes. Specifically, we propose a novel part-based self-supervised representation learning scheme to learn transferable representations by maximizing the similarity of an image to its discriminative part. To mitigate the overfitting in few-shot classification caused by data scarcity, we further propose a part augmentation strategy by retrieving extra images from a base dataset. We conduct systematic studies on *mini*ImageNet and *tiered*ImageNet benchmarks. Remarkably, our method yields impressive results, outperforming the previous best unsupervised methods by 7.74% and 9.24% under 5-way 1-shot and 5-way 5-shot settings, which are comparable with state-of-the-art supervised methods.

1 Introduction

Recently, great progress in the computer vision community has been achieved with deep learning, which often needs numerous training data. Unfortunately, there are many practical applications where collecting data is very difficult. To learn a novel concept with only few examples, few-shot learning has been recently proposed and gains extensive attention.[Doersch *et al.*, 2020; Afrasiyabi *et al.*, 2020]

A possible solution to few-shot learning is meta-learning [Finn *et al.*, 2017; Vinyals *et al.*, 2016; Ravi and Larochelle, 2017; Snell *et al.*, 2017; Peng *et al.*, 2019]. It first extracts shared prior knowledge from many similar tasks. After that, this knowledge will be adapted to the target few-shot learning task to restrict the hypothesis space, which makes it possible to learn a novel concept with few examples. Equipped with neural networks, the prior knowledge can be parameterized as an embedding function, or a set of initial parameters of a network, which often needs an extra labeled base dataset for training. Another line of work is based on transfer learning [Tian *et al.*, 2020; Dhillon *et al.*, 2020], which extracts prior knowledge as a pre-trained feature extractor. Typically, a standard cross entropy loss is employed to pre-train the feature extractor, which also needs a labeled base dataset. However, collecting a large labeled dataset is time-consuming and laborious. Besides, these labels are sadly discarded when performing a target few-shot learning task, because they belong to different class spaces. Inspired by recent progress of unsupervised learning, a question is naturally asked: can we can learn prior knowledge only from unlabeled images? If yes, it will be a promising approach for the scenario where many unlabeled images are available but the target task is data-scarce.

Some remarkable works have made an effort for this purpose, e.g., unsupervised meta-learning [Hsu *et al.*, 2019; Khodadadeh *et al.*, 2019]. However, these unsupervised methods are hindered by learning effective class-related representations from images, compared to the supervised counterparts. This is because much unrelated or interfering information, e.g., background clutters, may impose adverse impacts on representation learning under label-free unsupervised setting. Selecting discriminative image regions or target parts is an effective way to reduce this interference during representation learning, which has a consistent motivation with traditional part-based models [Felzenszwalb *et al.*, 2009].

In this paper, we propose a part-based self-supervised learning model, namely Part Discovery Network (PDN), to learn more effective representations from unlabeled images. The key point of this model is mining discriminative target parts from images. Due to the lack of part labels, multiple image regions are first extracted via random crop, which inevitably contain interfering backgrounds or less informative regions. To eliminate these regions, we choose the image region as the most discriminative target part, which keeps the largest average distance to other images (negative samples). The rationale of this selection is that the discriminative part should be able to distinguish the original image from others [Singh *et al.*, 2012], so its average distance to other images should be large enough. With the selected discriminative part, we maximize its similarity to the original image in a similar way to [He *et al.*, 2020].

Another challenge in the few-shot scenario is that the target

[*] Contact Author

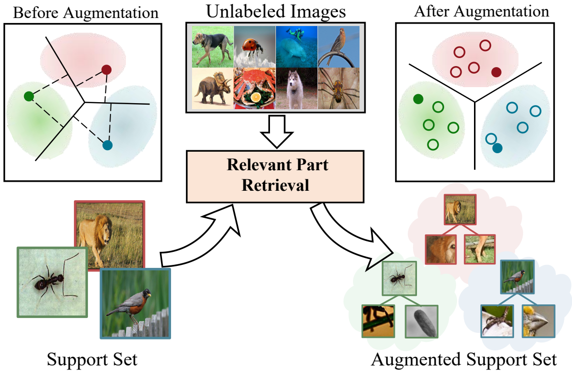

Figure 1: The main idea of the proposed Part Augmentation Network. The classifier is easy to overfit due to the scarcity of support samples in few-shot scenario. We retrieve relevant parts from unlabeled images to augment support set. The augmented support set can help learn a robust classifier with clear decision boundary.

classifier is easy to overfit due to the scarcity of support training samples. An effective way to prevent overfitting is data augmentation, which has been explored in the few-shot learning literature [Schwartz *et al.*, 2018; Wang *et al.*, 2018]. However, most of these methods assume that the variance of seen classes can be directly transferred to unseen classes, which is too strict in most situations. In this paper, we resort to retrieve extra images from a base dataset and extract their part features to augment support set, based on the fact that similar objects generally share some common parts. The core of our method is selecting these part features which match well with image features in the support set. Specifically, we propose a novel Class-Competitive Attention Map (C^{2}AM) to guide the relevant part selection from the retrieved images, and then refine target classifier with these selected parts. Our method is also called Part Augmentation Network (PAN), and Figure 1 illustrates the main idea of PAN.

Our Part Discovery and Augmentation Network (PDA-Net) consisting of both PDN and PAN largely outperforms the stat-of-the-art unsupervised few-shot learning methods, and achieves the comparable results with most of the supervised methods. The contributions of this work can be summarized as:

- We propose a novel self-supervised Part Discovery Network, which can learn more effective representations from unlabeled images for few-shot learning.
- We propose a Part Augmentation Network to augment few support examples with relevant part features, which mitigates overfitting and leads to more accurate classification boundaries.
- Our method outperforms previous unsupervised methods on two standard few-shot learning benchmarks. Remarkably, our unsupervised PDA-Net is also comparable with supervised methods.

2 Related Work

Few-Shot Learning. Few-shot learning aims at learning a novel concept from few examples. A possible solution is meta-learning, which extracts prior knowledge from many similar tasks (episodes). For example, MAML [Finn *et al.*, 2017] learns the optimal initial parameters of a network, which can be quickly adapted to a new task via gradient descent. Matching Networks [Vinyals *et al.*, 2016] classifies an instance based on its nearest labeled neighbor in the learned embedding space. Different from meta-learning, [Dhillon *et al.*, 2020] demonstrate that a strong transfer learning baseline can achieve competitive performance. In order to train models, all of the above methods need a large labeled base dataset. Unsupervised meta-learning methods loose this constraint by constructing training episodes from unlabeled images. [Hsu *et al.*, 2019] propose to acquire pseudo labels by clustering in an unsupervised feature space. [Khodadadeh *et al.*, 2019] use random sampling and augmentation to create synthetic episodes. Our method aims at both learning more effective representation from unlabeled images and increasing support set by data augmentation for unsupervised few-shot learning.

Self-Supervised Learning. Self-supervised learning aims at leveraging unlabeled images to learn good representations for down-stream tasks. Previous work mainly focuses on mining supervision signals from unlabeled data [Gidaris *et al.*, 2018; Doersch *et al.*, 2015; Zhang *et al.*, 2016]. Recently, contrastive learning shows superior performance improvement, which maximizes the similarity of two different views of the same image on global level [Chen *et al.*, 2020] or local level [Ouali *et al.*, 2020], or enforces consistency of cluster assignments between two views [Caron *et al.*, 2020]. In this paper, we explicitly mine discriminative parts for contrastive learning, which learns more effective representations from unlabeled images.

Data Augmentation for Few-Shot Learning. A straightforward way to deal with data deficiency is to synthesize more data. [Schwartz *et al.*, 2018] propose to extract intra-class variance (deltas) from base classes and use them to synthesize samples for novel classes. [Wang *et al.*, 2018] encode the intra-class variance in a hallucinator, which can synthesize new features taking as input a reference feature and noise vectors. To leverage base class samples, [Chen *et al.*, 2019] propose to generate a deformed image by fusing a probe image and a gallery image. Our method augments support set by retrieving extra images from a base dataset and extracting matched part features with the support samples.

3 Notation

In this section, we briefly illustrate the formulation of the few-shot image classification problem. Given a labeled support set $\mathcal{S} = \{(\boldsymbol{x_i}, y_i)\}_{i=1}^{N_s}$ where $\boldsymbol{x_i} \in I$ is an image and $y_i \in \mathcal{C}_{novel}$ is its label, we are supposed to predict the labels of a query set $\mathcal{Q} = \{(\boldsymbol{x_i}, y_i)\}_{i=1}^{N_q}$, which also belongs to $\mathcal{C}_{novel}$. The number of classes $|\mathcal{C}_{novel}|$ is called *way* and the number of samples in each class is called *shot*. For few-shot learning, the shot is very small, like 1-shot or 5-shot. Due to the scarcity of support samples, it is very hard to train a classification model from scratch. Therefore, we are given an extra large base dataset $\mathcal{D}^{base}$ to mine prior knowledge. Here, $\mathcal{D}^{base}$ is from base classes $\mathcal{C}_{base}$, and $\mathcal{C}_{base} \cap \mathcal{C}_{novel} = \phi$. Previous

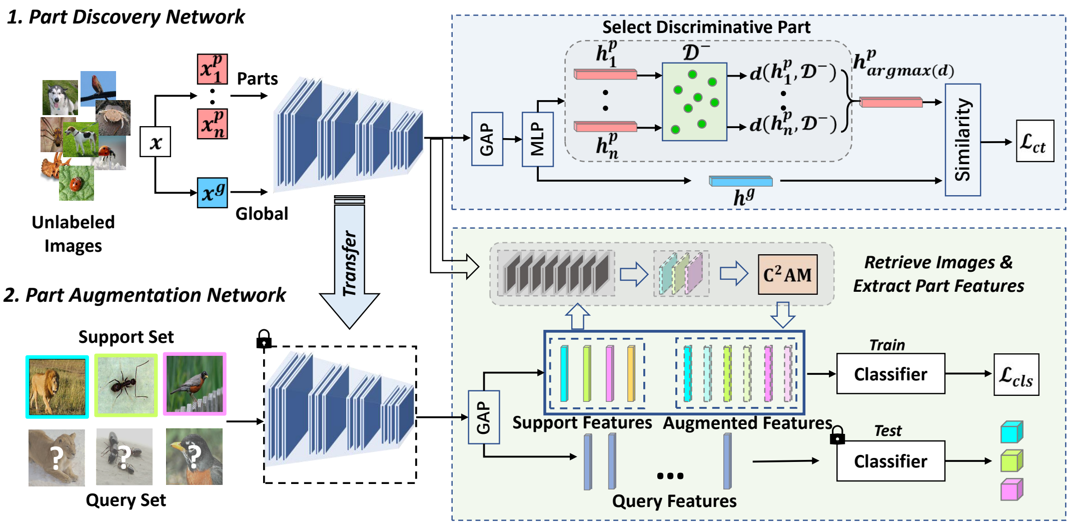

Figure 2: The framework of the proposed Part Discovery and Augmentation Network. In Part Discovery Network (PDN), we first extract multiple parts from an image, and then select the most discriminative one. Through maximizing the similarity of the global view to the selected part, we learn effective representations from unlabeled images. In Part Augmentation Network (PAN), we retrieve extra images that are similar to the support set, and then create a Class-Competitive Attention Map (C^{2}AM) to select relevant parts as augmented features. Finally, a classifier is trained on both support features and augmented features.

works usually need base labels to construct training episodes or pre-train the classification model. In this paper, we only use unlabeled images in $\mathcal{D}^{base}$, i.e., $\mathcal{D}^{base} = \{\boldsymbol{x}_i\}_{i=1}^{N_b}$.

4 Method

Our method follows a transfer learning protocol which includes two stages, namely the representation learning stage and the few-shot learning stage. On the first stage, we aim at learning a feature extractor f from $\mathcal{D}^{base}$. On the second stage, the learned feature extractor f is transferred to the target few-shot learning task, followed by training a linear classifier using support set $\mathcal{S}$. The final classification results of query set $\mathcal{Q}$ are predicted by the learned classifier. From the above description we can see that the key is to learn a good feature extractor from base dataset $\mathcal{D}^{base}$ and meanwhile obtain a robust classifier with limited data in support set $\mathcal{S}$. To this end, we propose Part Discovery and Augmentation Network (PDA-Net), which consists of a Part Discovery Network (PDN) to learn effective representations, and a Part Augmentation Network (PAN) to augment few support examples with relevant part features. The whole framework is shown in Figure 2.

4.1 Part Discovery Network

Our Part Discovery Network (PDN) is a part-based self-supervised learning model. We will introduce how to select parts and training details as follows.

Extracting Multiple Parts

Without part labels, we generate parts by randomly cropping a given image $\boldsymbol{x}$ into n patches $\{\boldsymbol{x}_i^{\boldsymbol{p}}\}_{i=1}^n$. Meanwhile, a larger crop $\boldsymbol{x}^{\boldsymbol{g}}$ is obtained to serve as global context. Random transformations are further applied to increase data diversity. To get latent part representations $\{\boldsymbol{h}_i^{\boldsymbol{p}}\}_{i=1}^n$ and global representation $\boldsymbol{h}^{\boldsymbol{g}}$, a convolution neural network f is applied followed by global average pooling and a MLP projection head.

Selecting Discriminative Part

Since we generate multiple parts with random crop, there are inevitably some crops that belong to background. Directly matching these crops to the global view will create bias towards backgrounds and hurt the generalization of the learned representations. To solve this problem, we develop an effective strategy to select the most discriminative part. Given a set of negative samples $\mathcal{D}^- = \{\boldsymbol{h}_i^{\boldsymbol{g}-}\}_{i=1}^{N^-}$ that are exclusive with the input image $\boldsymbol{x}$, we define a sample-set distance metric $d(\boldsymbol{h}_i^{\boldsymbol{p}}, \mathcal{D}^-)$ in the feature space, which indicates the distance between a part and all negative samples. We select the part with the maximum distance as the most discriminative one, therefore excluding background crops and less informative ones. The strategy is formulated as

$$\boldsymbol{h}^{\boldsymbol{p}} := \boldsymbol{h}_{i^*}^{\boldsymbol{p}}, \quad where\ i^* = \arg\max_i d(\boldsymbol{h}_i^{\boldsymbol{p}}, \mathcal{D}^-) \tag{1}$$

where $\boldsymbol{h}^{\boldsymbol{p}}$ is the selected part. The rationale of this strategy is that the discriminative part should be able to distinguish the original image from others, so its distance to other images should be very large.

The distance between $\boldsymbol{h}_i^{\boldsymbol{p}}$ and $\mathcal{D}^-$ can be defined in many forms. One could use the minimum distance between $\boldsymbol{h}_i^{\boldsymbol{p}}$ and all samples in $\mathcal{D}^-$. However, some similar images may exist in $\mathcal{D}^-$, so the minimum distance is severely affected by these similar images and can not reflect the true distance to most negative samples. Here we choose the mean distance to calculate $d(\boldsymbol{h}_i^{\boldsymbol{p}}, \mathcal{D}^-)$, which considers more on the samples of other classes. $d(\boldsymbol{h}_i^{\boldsymbol{p}}, \mathcal{D}^-)$ is calculated as:

$$d(\boldsymbol{h}_i^{\boldsymbol{p}}, \mathcal{D}^-) = \frac{1}{|\mathcal{D}^-|} \sum_{\boldsymbol{h}^{\boldsymbol{g}-} \in \mathcal{D}^-} -s(\boldsymbol{h}_i^{\boldsymbol{p}}, \boldsymbol{h}^{\boldsymbol{g}-}) \tag{2}$$

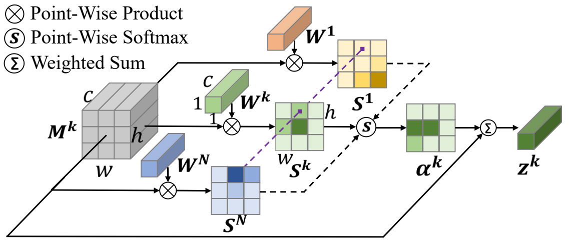

Figure 3: Illustration of C^{2}AM. We omit bias item for simplicity.

where $|\mathcal{D}^-|$ is the number of negative samples in this set, and s is the cosine similarity.

Training

With selected discriminative parts, we train the PDN with a contrastive loss, which is formulated as

$$\mathcal{L}_{ct} = -\log \frac{\exp(s(\boldsymbol{h^p},\boldsymbol{h^g})/\tau)}{\exp(s(\boldsymbol{h^p},\boldsymbol{h^g})/\tau+\sum_{\boldsymbol{h^{g-}}\in\mathcal{D}^-}\exp(s(\boldsymbol{h^p},\boldsymbol{h^{g-}})/\tau)} \tag{3}$$

where τ denotes a temperature hyper-parameter.

To get a large negative set, we follow [He *et al.*, 2020] to organize $\mathcal{D}^-$ as a queue and use a momentum encoder to get consistent negative representations. The momentum encoder is an exponential moving average version of the feature extractor and MLP head. Its parameter θ_m is updated as

$$\theta_m \longleftarrow m\theta_t + (1-m)\theta_m \tag{4}$$

where m is a momentum hyper-parameter, and θ_t denotes the parameters of the feature extractor and MLP head at training step t.

4.2 Part Augmentation Network

Our Part Augmentation Network (PAN) aims to augment support set with relevant part features from base dataset. To this end, we first retrieve extra images from base dataset, then create a Class-Competitive Attention Map (C^{2}AM) to guide the relevant part selection from the retrieved images, and finally refine target classifier with these selected parts.

Retrieving Extra Images

Since the size of $\mathcal{D}^{base}$ is very large, we propose a simple but effective strategy to select a small number of very similar images from $\mathcal{D}^{base}$, which are more likely to contain relevant parts. Specifically, we first train a linear classifier $p(y|\boldsymbol{z};\boldsymbol{W},\boldsymbol{b})$ on support set $\mathcal{S}$, where $\boldsymbol{z}$ denotes a feature vector, $\boldsymbol{W}$ and $\boldsymbol{b}$ are weight matrix and bias, respectively. Then, we employ this classifier to classify each image in $\mathcal{D}^{base}$ as:

$$\hat{y} = \arg\max_i p(y=i|\boldsymbol{z}=GAP(\boldsymbol{M})) \tag{5}$$

where GAP is a global average pooling operator and $\boldsymbol{M}$ denotes the feature map extracted by the learned feature extractor f in Section 4.1

Among the images of class k in $\mathcal{D}^{base}$, we keep the N_a images with the highest classification probability as the retrieval results. The feature maps of these N_a images are denoted as $\mathcal{A}_k = \{\boldsymbol{M_i^k}\}_{i=1}^{N_a}$.

Class-Competitive Attention Map

To further extract relevant parts from retrieved feature maps $\mathcal{A}_k$, we propose a novel CAM-based [Zhou *et al.*, 2016] attention mechanism, Class-Competitive Attention Map (C^{2}AM), illustrated in Figure 3. Given a feature map $\boldsymbol{M^k} \in \mathcal{A}_k$, we obtain a class attention map $\boldsymbol{S^k}$ that indicates its relevance to class k at each spatial location:

$$\boldsymbol{S^k}(i,j) = \boldsymbol{W^k}\boldsymbol{M^k}(i,j) + b^k \tag{6}$$

where $\boldsymbol{S^k}(i,j)$ and $\boldsymbol{M}^k(i,j)$ are the classification score for class k and the feature vector at location (i,j), respectively. $\boldsymbol{W^k}, b^k$ are the classifier weight vector and bias for class k, which have been learned in $p(y|\boldsymbol{z};\boldsymbol{W},\boldsymbol{b})$.

Although the class attention map $\boldsymbol{S^k}$ is very useful to locate class-specific image regions or object parts, it still contains some parts that have high classification scores for all classes. These parts provide less information for classification, and should be further inhibited. We perform softmax over the classification score vector $\boldsymbol{S}(i,j)$ at each spatial location, which provides a class competitive mechanism to highlight the parts which only have higher score for class k. The class-competitive attention map $\boldsymbol{\alpha^k}$ for class k is as follows:

$$\boldsymbol{\alpha^k}(i,j) = \frac{exp(\boldsymbol{S^k}(i,j))}{\sum_c^{|\mathcal{C}_{novel}|} exp(\boldsymbol{S^c}(i,j))} \tag{7}$$

With this revised attention map $\boldsymbol{\alpha^k}$, we can extract more relevant part features $\boldsymbol{z^k}$ to augment the class k, which is calculated by the weighted sum of feature map $\boldsymbol{M^k}$:

$$\boldsymbol{z^k} = \frac{\sum_{i,j}\boldsymbol{\alpha^k}(i,j)\boldsymbol{M^k}(i,j)}{\sum_{i,j}\boldsymbol{\alpha^k}(i,j)} \tag{8}$$

The retrieved feature set for class k is now updated as $\mathcal{A}_k = \{\boldsymbol{z_i^k}\}_{i=1}^{N_a}$. We denote $\mathcal{A} = \cup_k \mathcal{A}_k$ as the set of augmented part features for all classes.

Refining Target Classifier

With augmented part features, we can now refine the initial classifier with both support set $\mathcal{S}$ and augmented set $\mathcal{A}$ to get a more robust classifier. Since augmented features are not from the exactly same classes as support samples, we adopt the label smoothing technique [Szegedy *et al.*, 2016] to prevent overfitting on the augmented features. Specifically, we convert an one-hot label y into a smoothed probability distribution $\boldsymbol{p_y}$:

$$\boldsymbol{p_y}(k) = \begin{cases} 1-\epsilon & ,k=y \\ \frac{\epsilon}{|\mathcal{C}_{novel}|-1} & ,k\neq y \end{cases} \tag{9}$$

where $\boldsymbol{p_y}(k)$ is the probability of the class k, and ϵ is a hyper-parameter in range $(0,1)$. Finally, the total loss for refining the classifier is:

$$\mathcal{L}_{cls} = \sum_{\boldsymbol{z}\in\mathcal{S}} -\log p(y|\boldsymbol{z}) + \lambda \sum_{\boldsymbol{z}\in\mathcal{A}} KL(p(\cdot|\boldsymbol{z}),\boldsymbol{p_y}) \tag{10}$$

where the first term is cross entropy loss for samples in support set $\mathcal{S}$, and the second term is K-L divergence between predicted probability distribution $p(\cdot|\boldsymbol{z})$ and smoothed ground-truth distribution $\boldsymbol{p_y}$ for augmented features.

			*mini*ImageNet 5-way		*tiered*ImageNet 5-way	
Setting	Method	Backbone	1-shot	5-shot	1-shot	5-shot
Supervised	IDeMe-Net [Chen *et al.*, 2019]	ResNet-12	59.14±0.86	74.63±0.74	-	-
	MetaOptNet [Lee *et al.*, 2019]	ResNet-12	62.64±0.61	78.63±0.46	65.99±0.72	81.56±0.53
	Distill [Tian *et al.*, 2020]	ResNet-12	64.82±0.60	82.14±0.43	71.52±0.69	86.03±0.49
	Finetune [Dhillon *et al.*, 2020]	WRN-28-10	57.73±0.62	78.17±0.49	66.58±0.70	85.55±0.48
	LEO [Rusu *et al.*, 2019]	WRN-28-10	61.76±0.08	77.59±0.12	66.33±0.05	81.44±0.09
	CC+rot [Gidaris *et al.*, 2019]	WRN-28-10	62.93±0.45	79.87±0.33	70.53±0.51	84.98±0.36
	Align [Afrasiyabi *et al.*, 2020]	WRN-28-10	**65.92±0.60**	**82.85±0.55**	**74.40±0.68**	**86.61±0.59**
Unsupervised	CACTUs [Hsu *et al.*, 2019]	Conv4	39.90±n/a	53.97±n/a	-	-
	UMTRA [Khodadadeh *et al.*, 2019]	Conv4	39.93±n/a	50.73±n/a	-	-
	Rot [Gidaris *et al.*, 2019]	WRN-28-10	43.43±n/a	60.86±n/a	-	-
	GdBT2 [Khoi and Sinisa, 2020]	SN-GAN	48.28±0.77	66.06±0.70	47.86±0.79	67.70±0.75
	MoCo [Tian *et al.*, 2020]	ResNet-50	54.19±0.93	73.04±0.61	-	-
	CMC [Tian *et al.*, 2020]	ResNet-50	56.10±0.89	73.87±0.65	-	-
	PDA-Net (Ours)	ResNet-50	**63.84±0.91**	**83.11±0.56**	**69.01±0.93**	**84.20±0.69**

Table 1: Comparison with prior work on *mini*ImageNet and *tiered*ImageNet. Accuracy is reported with 95% confidence intervals.

5 Experiments

5.1 Datasets

***mini*ImageNet.** *mini*ImageNet is a standard benchmark for few-shot learning proposed by [Vinyals *et al.*, 2016]. It is a subset of the ImageNet [Russakovsky *et al.*, 2015] and contains 100 classes and 600 examples for each class. We follow the protocol in [Ravi and Larochelle, 2017] to use 64 classes for training, 16 classes for validation and 20 classes for test.

***tiered*ImageNet.** *tiered*ImageNet [Ren *et al.*, 2018] is a larger subset of ImageNet and contains 608 classes and 1000 images in each class. Theses classes are grouped into 34 higher categories, where 20 categories (351 classes) for training, 6 categories (97 classes) for validation and 8 categories (160 classes) for test. The large semantic difference between categories makes it more challenging for few-shot learning.

5.2 Implementation Details

For PDN, we transform input images with random crop, horizontal flip, color jitter and Gaussian blur. The crop scales for part view and global view are in range of (0.05, 0.14) and (0.14, 1), respectively. The number of cropped parts is set as $n = 6$ by cross validation. The size of $\mathcal{D}^-$ is 1024 and 10240 for *mini*ImageNet and *tiered*ImageNet, respectively. We use standard ResNet-50 as backbone and resize images to 224×224. We set hyper-parameter $m = 0.999$ and $\tau = 0.2$. We adopt SGD optimizer with cosine learning rate decay. The learning rate is 0.015 for *mini*ImageNet and 0.03 for *tiered*ImageNet.

For PAN, we retrieve $N_a = 1024$ extra images for each class. The label smoothing hyper-parameter ϵ is 0.2 for 1-shot and 0.7 for 5-shot. The loss weight λ is set as 1. The classifier is trained with Adam and the learning rate is 0.001. Following the similar setting to [Vinyals *et al.*, 2016], We evaluate our method on 600 test episodes.

5.3 Comparison with Prior Work

In Table 1, we compare our method with both supervised and unsupervised few-shot learning methods. Overall, our method achieves the best performance under unsupervised setting, and is comparable with state-of-the-art supervised methods.

PDN w/o Select	PDN	PAN w/o C^{2}AM	PAN	Acc.
				57.12
✓				59.32
✓			✓	61.74
	✓			61.77
	✓	✓		62.64
	✓		✓	**63.84**

Table 2: Ablation study on *mini*ImageNet in 5-way 1-shot.

Our PDA-Net significantly outperforms unsupervised meta-learning [Hsu *et al.*, 2019] by 23.91% and 29.14% on *mini*ImageNet in 1-shot and 5-shot, respectively. Compared with the most related method which also employs MoCo for representation learning [Tian *et al.*, 2020], we achieve 9.65% and 10.07% improvement in 1-shot and 5-shot on *miniImageNet*. On the larger tieredImageNet, our method largely outperforms a recent GAN-based method, GdBT2 [Khoi and Sinisa, 2020], by 21.15% and 16.50% in 1-shot and 5-shot, respectively. The significant improvements over the compared unsupervised methods verify the effectiveness of our PDA-Net.

Furthermore, our method outperforms two supervised meta-learning methods, LEO [Rusu *et al.*, 2019] and MetaOptNet [Lee *et al.*, 2019], on both datasets in both 1-shot and 5-shot. Compared with the supervised transfer learning method, Distill [Tian *et al.*, 2020], the performance of our PDA-Net only drops 0.98% on *mini*ImageNet in 1-shot, which is acceptable considering the expensive label budget. Alignment [Afrasiyabi *et al.*, 2020] uses base labels to detect the most related class, which achieves better performance in 1-shot, but is comparable with ours in 5-shot.

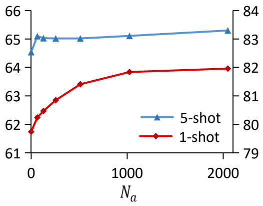

Figure 4: Accuracy with different number of augmented features in PAN on *mini*ImageNet.

n	1-shot	5-shot
2	57.84	76.38
4	60.98	80.10
6	**63.84**	**83.11**
8	63.56	82.58

Table 3: Accuracy with different number of cropped parts in PDN on *mini*ImageNet.

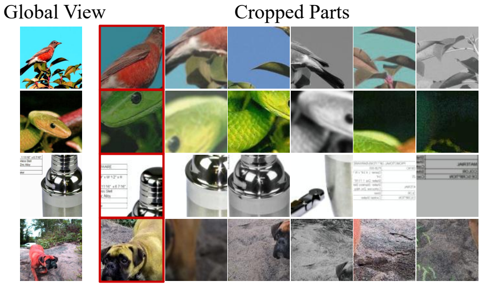

Figure 5: Visualization of the global view and cropped parts.

5.4 Further Analysis

Ablation Study. We evaluate the effectiveness of each key component of PDA-Net on *mini*ImageNet in 1-shot, and show the results in Table 2. The baseline is constructed by ablating part discovery and part augmentation, thus only taking as input two global views for contrastive learning. PDN w/o Select means that we use all the 6 parts for contrastive learning, while PAN w/o C^{2}AM means that we use class activation map instead [Zhou *et al.*, 2016]. Compared with the baseline, PDN w/o Select obtains 2.20% performance gain, which verifies the advantage of part-based representation learning. PDN can get further improvement via selecting the most discriminative part. Better results are achieved by adding PAN to PDN or PDN w/o Select, which illustrates the effectiveness of our part augmentation strategy. Compared with CAM in [Zhou *et al.*, 2016], our C^{2}AM can obtain better performance, demonstrating the effectiveness of our proposed class competitive mechanism.

Number of Crops. In PDN, we use random crop to get n parts. Here we evaluate the impact of number of cropped parts on representation learning. As shown in Table 3, the performance rapidly increases with the growing number of cropped parts, because it is more likely to contain discriminative parts with more cropped parts. The performance gets saturated after reaching 6, which indicates that our selection strategy is an effective way to extract discriminative parts.

Number of Augmented Features. In PAN, we retrieve N_a extra images to augment each novel class. Here we experiment with different N_a and show the results in Figure 4. It can be seen that the benefit of augmented features is more significant for 1-shot than 5-shot. For 1-shot, the performance grows rapidly as N_a increases and finally gets saturated after $N_a = 1024$, which is because relevant features in base classes are limited. With a larger base dataset, we can infer that our method can get better results.

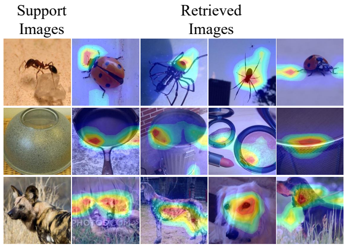

Figure 6: Visualization of retrieved images and their C^{2}AMs.

5.5 Visualization

To explore the discriminative part selection, we visualize the cropped parts and their corresponding global view in Figure 5. It should be noted that these images are all randomly transformed, so they may have different appearances even from the same input image. We sort the cropped parts based on the distances to negative samples and select the parts with the largest distances within red boxes. For example, the main body of a bird and the head of a dog are selected in the first and fourth rows, respectively. We can see that as the distances decrease, the parts are more likely to belong to background and contain less information about target object.

To illustrate the rationality of part augmentation strategy, we visualize the retrieved unlabeled images and their class-competitive attention maps in Figure 6. We can see that the retrieved images are usually similar to support images. More interestingly, C^{2}AM can locate class-specific image regions which are very relevant to the objects in support set. For example, the antennae and legs of the ladybug are highlighted in the retrieved image, which are very similar to those of the ant in support image. In contrast, its shell attracts less attention due to the distinctive appearance from the ant.

6 Conclusion

In this paper, we present a Part Discovery Network, which can extract effective prior knowledge for few-shot learning from a flat collection of unlabeled images. Furthermore, a Part Augmentation Network is proposed to augment support examples with relevant parts, which can mitigate overfitting and lead to better classification boundaries. The experimental results demonstrate that our method significantly outperforms previous unsupervised meta-learning methods and achieves comparable accuracy with state-of-the-art supervised methods.

Acknowledgements

This work is supported by National Natural Science Foundation of China (61976214, 61633021, 61721004 , 61836008), Beijing Municipal Natural Science Foundation (4214075), and Science and Technology Project of SGCC Research on feature recognition and prediction of typical ice and wind disaster for transmission lines based on small sample machine learning method.

References

[Afrasiyabi *et al.*, 2020] Arman Afrasiyabi, Jean-Franccois Lalonde, and Christian Gagné. Associative alignment for few-shot image classification. In *ECCV*, 2020.

[Caron *et al.*, 2020] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *NeurIPS*, 2020.

[Chen *et al.*, 2019] Zitian Chen, Yanwei Fu, Yu-Xiong Wang, Lin Ma, Wei Liu, and Martial Hebert. Image deformation meta-networks for one-shot learning. *CVPR*, 2019.

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

[Dhillon *et al.*, 2020] Guneet Singh Dhillon, Pratik Chaudhari, Avinash Ravichandran, and Stefano Soatto. A baseline for few-shot image classification. In *ICLR*, 2020.

[Doersch *et al.*, 2015] Carl Doersch, Abhinav Gupta, and Alexei A. Efros. Unsupervised visual representation learning by context prediction. *ICCV*, 2015.

[Doersch *et al.*, 2020] Carl Doersch, Ankush Gupta, and Andrew Zisserman. Crosstransformers: Spatially-aware few-shot transfer. *NeurIPS*, 2020.

[Felzenszwalb *et al.*, 2009] Pedro Felzenszwalb, Ross Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. In *TPAMI*, 2009.

[Finn *et al.*, 2017] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *ICML*, 2017.

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

[Gidaris *et al.*, 2019] Spyros Gidaris, Andrei Bursuc, Nikos Komodakis, Patrick Pérez, and Matthieu Cord. Boosting few-shot visual learning with self-supervision. In *ICCV*, 2019.

[He *et al.*, 2020] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. Momentum contrast for unsupervised visual representation learning. *CVPR*, 2020.

[Hsu *et al.*, 2019] Kyle Hsu, Sergey Levine, and Chelsea Finn. Unsupervised learning via meta-learning. In *ICLR*, 2019.

[Khodadadeh *et al.*, 2019] Siavash Khodadadeh, Ladislau Boloni, and Mubarak Shah. Unsupervised meta-learning for few-shot image classification. In *NeurIPS*, 2019.

[Khoi and Sinisa, 2020] Nguyen Khoi and Todorovic Sinisa. A self-supervised gan for unsupervised few-shot object recognition. *ICPR*, 2020.

[Lee *et al.*, 2019] Kwonjoon Lee, Subhransu Maji, A. Ravichandran, and Stefano Soatto. Meta-learning with differentiable convex optimization. *CVPR*, 2019.

[Ouali *et al.*, 2020] Yassine Ouali, Céline Hudelot, and Myriam Tami. Spatial contrastive learning for few-shot classification. *arXiv*, 2020.

[Peng *et al.*, 2019] Zhimao Peng, Zechao Li, Junge Zhang, Yan Li, Guo-Jun Qi, and Jinhui Tang. Few-shot image recognition with knowledge transfer. In *ICCV*, 2019.

[Ravi and Larochelle, 2017] Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. In *ICLR*, 2017.

[Ren *et al.*, 2018] Mengye Ren, Eleni Triantafillou, Sachin Ravi, Jake Snell, Kevin Swersky, Joshua B. Tenenbaum, Hugo Larochelle, and Richard S. Zemel. Meta-learning for semi-supervised few-shot classification. In *ICLR*, 2018.

[Russakovsky *et al.*, 2015] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *IJCV*, 2015.

[Rusu *et al.*, 2019] Andrei A. Rusu, Dushyant Rao, Jakub Sygnowski, Oriol Vinyals, Razvan Pascanu, Simon Osindero, and Raia Hadsell. Meta-learning with latent embedding optimization. In *ICLR*, 2019.

[Schwartz *et al.*, 2018] Eli Schwartz, Leonid Karlinsky, Joseph Shtok, Sivan Harary, Mattias Marder, Abhishek Kumar, Rogerio Feris, Raja Giryes, and Alex Bronstein. Delta-encoder: an effective sample synthesis method for few-shot object recognition. In *NeurIPS*, 2018.

[Singh *et al.*, 2012] Saurabh Singh, Abhinav Gupta, and Alexei A Efros. Unsupervised discovery of mid-level discriminative patches. In *ECCV*. Springer, 2012.

[Snell *et al.*, 2017] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *NeurIPS*, 2017.

[Szegedy *et al.*, 2016] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. *CVPR*, 2016.

[Tian *et al.*, 2020] 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 *ECCV*, 2020.

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

[Wang *et al.*, 2018] Yu-Xiong Wang, Ross Girshick, Martial Hebert, and Bharath Hariharan. Low-shot learning from imaginary data. In *CVPR*, 2018.

[Zhang *et al.*, 2016] Richard Zhang, Phillip Isola, and Alexei A. Efros. Colorful image colorization. In *ECCV*, 2016.

[Zhou *et al.*, 2016] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *CVPR*, 2016.