AUTO-VIEW CONTRASTIVE LEARNING FOR FEW-SHOT IMAGE RECOGNITION

Anonymous authors
Paper under double-blind review

ABSTRACT

Few-shot learning aims to recognize new classes with few annotated instances within each category. Recently, metric-based meta-learning approaches have shown the superior performance in tackling few-shot learning problems. Despite their success, existing metric-based few-shot approaches often fail to push the fine-grained sub-categories apart in the embedding space given no fine-grained labels. This may result in poor generalization to fine-grained sub-categories, and thus affects model interpretation. To alleviate this problem, we introduce contrastive loss into few-shot classification for learning latent fine-grained structure in the embedding space. Furthermore, to overcome the drawbacks of random image transformation used in current contrastive learning in producing noisy and inaccurate image pairs (i.e., views), we develop a learning-to-learn algorithm to automatically generate different views of the same image. Extensive experiments on standard few-shot learning benchmarks and few-shot fine-grained image classification demonstrate the superiority of our method.

1 INTRODUCTION

Few-shot learning has been widely studied to recognize unseen classes with limited samples for each novel class (Li et al., 2006; Finn et al., 2017; Lee et al., 2019; Tseng et al., 2020). Recently, metric-based meta-learning methods have attracted extensive attention in image classification due to their superior performance and simplicity (Vinyals et al., 2016; Snell et al., 2017; Sung et al., 2018). For making inference, these methods compare the similarity between the feature embedding of query images and that of a few labeled images of each class. This therefore requires learning a flexible encoder, which can map the data points with similar semantics in the input space to locate closely in the embedding space. Meanwhile, those data with different semantic meanings in the input space should disperse in the embedding space. Accordingly, a new sample from the novel class can be recognized directly through a simple distance metric within the learned embedding space. Indeed, the performance of recognizing novel classes in metric-based meta-learning extremely relies on the learned embedding space.

Despite the success of recognizing novel classes, existing metric-based few-shot approaches often fail to push the fine-grained sub-categories apart in the embedding space given no fine-grained labels in training. For illustration, we merge nine different sub-categories of dogs in the miniImageNet dataset into a coarse-grained class as a new label "dog" to train the Prototypical Network(PN) (Snell et al., 2017) without changing other classes. As shown in Figure 1(a), we visualize the features of the input data of three fine-grained sub-categories of "dog" using t-SNE. It is clearly revealed that features of fine-grained classes learned by the Prototypical Network cannot be separated without further label information. This means that these methods often do not generalize well to fine-grained sub-categories. Since labeling the fine-grained sub-categories requires strong expertise, the generalization ability to unseen fine-grained sub-categories is of critical importance.

In this paper, we try to alleviate this problem by seeking self-supervised learning to learn the fine-grained structure without given corresponding label information. As a powerful self-supervised representation learning paradigm, Contrastive Learning (van den Oord et al., 2018; Chen et al., 2020; He et al., 2020) has outperformed over even supervised learning in many situations. The key insight of contrastive learning is to contrast semantically similar (positive) with dissimilar (negative) pairs of data points. Nice theoretical results for contrastive learning has been given in Saunshi et al.

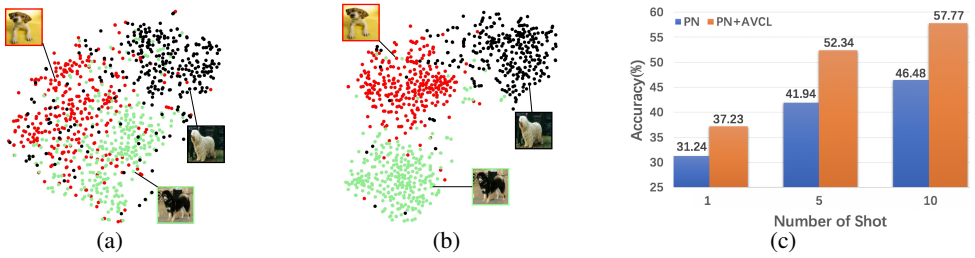

Figure 1: We merge nine fine-grained sub-categories of dog into one class as a new coarse-grained label "dog" for training. (a) and (b): T-SNE visualization of feature vectors extracted from three out of nine fine-grained classes of dogs in miniImageNet using Prototypical Network and our proposed method, respectively. (c): Evaluation results on the fine-grained Stanford Dogs dataset (Khosla et al., 2011).

(2019), by hypothesizing that semantically similar points are often sampled from the same latent class. Hence, contrastive learning has the potential to bring closer the representations of the same latent class and separate those of different latent class. Therefore, we introduce a contrastive loss into few-shot classification and learn latent fine-grained structures in the embedding space, which helps to cluster samples with similar representations to form similar sub-categories.

A critical issue in contrastive learning is generating a pair of semantically similar representations(views) of the same image for contrasting with those dissimilar ones in the embedding space. However, due to the limited number of training images, random image transformation may generated poor positive pairs with more substantial noises and less concept-relevant information when directly applied to few-shot learning. This may therefore make contrastive learning fail to learn fine-grained structure. To effectively improve fine-grained structure learning in the few shot learning setting, we propose auto-view contrastive learning(AVCL) for metric-based meta-learning. Specifically we replace random image transformation of contrastive learning with spatial transformer network(STN) (Jaderberg et al., 2015), a learned module that allows flexible spatial manipulation of images, and develop a learning-to-learn algorithm to adaptively generate different views of the same image. In detail, the parameters of STN are optimized through the contribution of the contrastive loss to few-shot image recognition.

To verify that our proposed approach can improve the generalization ability of unseen fine-grained sub-categories without corresponding fine-grained label information, we follow the experiment setting of the coarse-grained class "dog" for training and further test the learned model on Stanford Dogs, a fine-grain dataset containing 120 fine-grained dog classes (Khosla et al., 2011). As shown in Figure 1(b,c), our proposed method can learn a better embedding space and significantly improve the test accuracy on the Stanford Dogs dataset. For instance, our method obtains an improvement of 10.40% on the 5-way 5-shot task over the Prototypical Network.

We summarize the main contributions as follows:

- We firstly introduce contrastive loss into few-shot classification for learning the fine-grained structure given no corresponding label information. This can make models learn better embedding space and improve the generalization capability of learned models to unseen fine-grained sub-categories.
- We propose a learning-to-learn auto-view learning approach for contrastive learning to tackle the few-shot learning problems, which can keep classification-relevant information of input data intact in views to learn fine-grained structures effectively. Extensive experiments on standard few-shot learning benchmarks demonstrate the superiority of our method. In particular, compared to contrastive learning with random views, our AVCL method obtains 2.21% and 2.73% performance gains on miniIMageNet under the 5-way 1-shot and 5-way 5-shot settings, respectively.
- Since different sub-categories are distinguished by subtle and local differences, fine-grained image recognition is more difficult than standard few-shot image recognition. Extensive

experiments on fine-grained image recognition is conducted to testify the effectiveness and advantages of our method.

2 RELATED WORKS

Few-shot image recognition Few-shot image recognition was first proposed by (Li et al., 2006), with the aim to solve the problem of classifying novel categories with few labeled images per class. Nowadays, two types of meta-learning methods are the mainstream methods to address this problem. One is gradient-based method that empower the model with ability to rapidly fine-tune to novel classes with limited labeled images (Finn et al., 2017; Ravi & Larochelle, 2017; Rusu et al., 2019; Bertinetto et al., 2019; Lee et al., 2019). The other is metric-based method, which makes predictions based on a similar metric in a learned feature space between images with and without labels. Common similar metric used in this method includes cosine similarity (Vinyals et al., 2016), Euclidean similarity (Snell et al., 2017), relation module (Sung et al., 2018), and graph neural network (Satorras & Estrach, 2018).

In our work, we primarily consider improving the performance of metric-based meta-learning methods, especially Prototypical Network (Snell et al., 2017). Recently, there are many methods proposed to improve the ability of metric-based meta-learning by constraining the structure of the feature space. Li et al. (2020a) introduce an extra margin loss that leverages external content information, e.g., pre-trained word embeddings, to generate adaptive margin between classes. This leads the feature space to have better semantic structure. Contrastively, our work does not import external information and utilizes semantic information extracted purely from images themselves. Works most relevant to ours are Sundermeyer et al. (2018); Gidaris et al. (2019) which utilize self-supervised tasks to improve few-shot learning. However, their self-supervised pretext tasks are fixed at the training stage. Our proposed framework can progressively change the view of contrastive learning under a learn-to-learn paradigm.

Contrastive Learning Contrastive learning is one of the most popular methods for unsupervised visual representation learning. View transformations and contrastive loss are two key parts of contrastive learning. This framework attains representations by optimizing contrastive loss that maximize agreement between transformed views of the same image and minimize agreement between transformed views of different images. Contrastive learning was first proposed by (Hadsell et al., 2006). Recently, Wu et al. (2018) consider instance discrimination and use noise contrastive loss to learn representations. Contrastive multiview Coding (Tian et al., 2019) utilizes contrastive learning to attain representations on multi-view setting. SimCLR (Chen et al., 2020) summarizes a standard framework of contrastive learning and shows the effect of different random view transformations. However, the performance of SimCLR relies on large batch size. Momentum Contrast (MoCo) (He et al., 2020) is proposed to alleviate this problem by constructing a queue to preserve immediate preceding samples. In our work, due to the relatively small batch size of metric-based meta-learning, we adopt MoCo in our model. Tian et al. (2020) also point out the importance of view transformation for contrastive learning, and propose to learn view transformation via information bottleneck principle under unsupervised and semi-supervised settings. Our work employs a learn-to-learn framework to automatically learn the view for better performance for few-shot classification.

3 PROPOSED METHOD

3.1 PROBLEM SETTING

A few-shot classification task is characterized as N-way, K-shot only if the model adapt to classify new data after seeing K examples from each of the N classes. Meta-learning algorithms imitate this setting in each iteration by randomly sampling N classes and their corresponding images from a base dataset $\mathbf{D}_B$ to construct a pseudo few-shot task or episode T. Specifically, the input of each episode T can be divided into two sets: (1) Observable support set $\mathbf{D}_S = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{N \times K}$, formed by randomly selecting K samples from each of the N classes and (2) Unseen query set $\mathbf{D}_Q = \{(\mathbf{x}_j, \mathbf{y}_j)\}_{j=1}^{M}$, containing other M samples from the same N categories. Given $\mathbf{D}_S$, the pseudo task is employed by making predictions of $\mathbf{D}_Q$. Then the labels of query set are used to compute classification loss to guide the update of the model.

3.2 MODEL DESCRIPTION

Figure 2 illustrates the framework of our proposed method. Two complementary classification tasks are employed simultaneously to learn the main encoder $F_\theta(\cdot)$, which is the key component that maps the input into a feature space. One path is metric-based meta-learning, which utilizes explicit label information to regularize the feature space. Another path is contrastive prediction task, which is a self-supervised instance-level classification task. This task is designed to identify latent fine-grained structure in the feature space by aggregating the representations of the same latent class and separating those of different latent classes at the same time.

In an episode T, sampled images in $\mathbf{D}_S$ and $\mathbf{D}_Q$ are used in the two paths. In the metric-based meta-learning path, the main encoder $F_\theta(\cdot)$ first maps all images into the feature space, and then all support features in the same class C_i aggregate into one vector h_i. Typically, this is accomplished by averaging all support features, i.e. $\mathbf{h}_i = \frac{1}{K}\sum_{(x,y)\in\mathbf{D}_S} \mathbb{1}_{[y=i]} F_\theta(x)$. It is followed by computing the similarities between query features and aggregated features in each class. The final classification loss $\mathcal{L}^{meta}$ is defined as average cross entropy between true labels and predictions based on similarities. This can be formulated as:

$$\mathcal{L}^{\text{meta}}(\mathbf{D}_S, \mathbf{D}_Q, \theta) = -\frac{1}{M} \sum_{(x,y)\in\mathbf{D}_Q} \log \frac{e^{\text{sim}(F_\theta(x), h_y))}}{\sum_{i=1}^{N} e^{\text{sim}(F_\theta(x), h_i)}}, \tag{1}$$

where sim denotes a similarity metric.

The core idea of the contrastive path is the process of producing two views from one image which lie in the same latent space, i.e. containing similar semantic contents. This process is accomplished by two differentiable auto-view modules $\mathcal{G}_{\gamma_1}(\cdot)$ and $\mathcal{G}_{\gamma_2}(\cdot)$ parameterized by γ_1 and γ_2, respectively. Each of them is a spatial transformer network(STN) (Jaderberg et al., 2015) that allows flexible semantics-invariant spatial manipulation of images, including cropping, one dimensional scaling, As for translation, image deformation and proportional shrinkage, see appendix B for more details of STN. They are applied to each image $\mathbf{x}_i$, producing two views $\mathbf{x}_i^{(1)}$ and $\mathbf{x}_i^{(2)}$. These views are then mapped into feature space by the main encoder $F_\theta(\cdot)$ and momentum encoder $F_\omega(\cdot)$, respectively. As mentioned in He et al. (2020), the momentum encoder's parameter ω is a moving average of θ, which makes two encoders behave similar. Given $F_\theta(\mathbf{x}_i^{(1)})$, the contrastive loss aims to identify $F_\omega(\mathbf{x}_i^{(2)})$ in thousands of features $\{F_\omega(\mathbf{x}_k^{(2))}\}_{k\neq i}$, and can be formulated as:

$$\mathcal{L}^{\text{con}}(\mathbf{D}_S, \mathbf{D}_Q, \omega, \theta, \gamma) = -\sum_{x\in\mathbf{D}_S\cup\mathbf{D}_Q} \log \frac{e^{\text{sim}(F_\theta(\mathbf{x}^{(1)}), F_\omega(\mathbf{x}^{(2)}))}}{\sum_{j=1}^{r} e^{\text{sim}(F_\theta(\mathbf{x}^{(1)}), F_\omega(\mathbf{x}_j^{(2)})))}}, \tag{2}$$

where r denotes the number of negative samples, and γ = $[\gamma_1, \gamma_2]$ denotes the parameters of auto-view modules. By minimizing $\mathcal{L}^{\text{con}}$ w.r.t θ, we force the main encoder $F_\theta(\cdot)$ to map views of one image which are semantically similar into closer points in the feature space, thus constructs a better fine-grained semantic structure.

3.3 LEARNING STRATEGY

The optimization of our model during each iteration contains two stages. We denote $\theta^t, \omega^t, \gamma^t$ as parameters and $\mathbf{D}_S^t$, $\mathbf{D}_Q^t$ as support set and query set during iteration t, respectively. We first update two encoders based on the meta loss $\mathcal{L}^{\text{meta}}$ and contrastive loss $\mathcal{L}^{\text{con}}$:

$$\theta^{t+1} = \theta^t - \alpha\nabla_{\theta^t}(\mathcal{L}^{\text{meta}}(\mathbf{D}_S^t, \mathbf{D}_Q^t, \theta^t) + \beta\mathcal{L}^{\text{con}}(\mathbf{D}_S^t, \mathbf{D}_Q^t, \omega^t, \theta^t, \gamma^t)), \tag{3}$$

$$\omega^{t+1} = \epsilon\omega^t + (1-\epsilon)\theta^{t+1}, \tag{4}$$

where β denotes the regularization hyperparameter weighting two losses, α denotes the learning rate of θ, and ϵ denotes momentum coefficient that controls the chasing speed of the momentum encoder $F_\omega(\cdot)$. Note that the value of θ^{t+1} relies on γ^t via the contrastive loss $\mathcal{L}^{\text{con}}$. This fact is crucial for the update of γ^t in the next stage.

The updating criterion for the auto-view modules $G_{\gamma_1}(\cdot)$ and $G_{\gamma_2}(\cdot)$ is to improve the positive effect of contrastive loss for meta-learning. Keeping this in mind, we use the updated encoder $F_{\theta^{t+1}}(\cdot)$ to

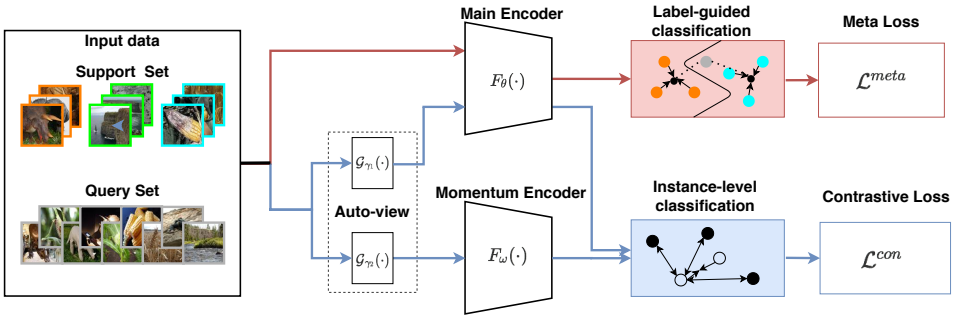

Figure 2: A brief flow diagram of our model at training stage. Our framework consists of two main tasks. The red arrow is metric-based meta-learning, while blue arrow depicts instance-level contrastive learning. At meta-test stage, only the main encoder $F_\theta(\cdot)$ is held for evaluation under few-shot setting.

Table 1: Details of datasets used in experiments

Dataset	Images	Classes	Train-val-test	Resolution(after resize)
miniImageNet	60000	100	64-16-20	80×80
CUB-200-2011	11788	200	100-50-50	80×80
Cars	16185	196	98-49-49	80×80
Places	73000	365	183-91-91	80×80
Plantae	47242	200	100-50-50	80×80

compute meta loss again on the same task, and cast the loss as evaluation of update quality in the first stage. Then the loss is used to guide the update of the auto-view modules via gradient descent:

$$\gamma^{t+1} = \gamma^t - \eta \nabla_{\gamma^t} \mathcal{L}^{\text{meta}}(\mathbf{D}_S^t, \mathbf{D}_Q^t, \theta^{t+1}), \tag{5}$$

where η is the learning rate of the auto-view modules. We mention again that θ^{t+1} can be cast as a function of γ^t. Thus the loss depends on γ^t through the computation graph of θ^{t+1} in first stage. The update of γ^t adjusts produced views towards better update of encoder $F_\theta(\cdot)$ in the first stage. This indicates that learned views further pushes positive effect of contrastive loss for meta-learning.

4 Experimental Results

In this section, we conduct experiments to demonstrate the effectiveness of our method. Our proposed method is evaluated on standard few-shot learning benchmarks and few-shot fine-grained image recognition. We also conduct ablative study about the auto-view module in our learning framework.

4.1 Experimental Setup

We conduct 5-way 5-shot and 5-way 1-shot classification for all datasets. The metric-based meta-learning method adopted in our model is Prototypical Network, one of the state-of-the-art metric-based meta-learning methods for few-shot learning.

Datasets We follow the general few-shot image recognition settings and evaluate our method on two benchmarks: **miniImageNet** (Vinyals et al., 2016) and **CUB-200-2011** (Welinder et al., 2010). miniImageNet is selected from the well-known ImageNet(Russakovsky et al., 2015) dataset. CUB-200-2011 was originally proposed for fine-grained bird classification. To further validate that the feature space benefits from a better fine-grained semantic structure, we evaluate our approach on three datasets with fine-grained categorization: **Cars**(Krause et al., 2013), **Places**(Zhou et al., 2018) and **Plantae**(Horn et al., 2018). We follow the dataset split in (Tseng et al., 2020). We list image number, class number, splits and images resolution of all datasets in Table 1.

Table 2: Comparative results for 5-way classification on miniImageNet. Average accuracies on the meta-test set with 95 confidence interval are reported. † denotes methods using external text information. ‡ denotes result reported in (Gidaris et al., 2019).

Model	backbone	1-shot	5-shot
MAML (Finn et al., 2017)	Conv4	48.70 ± 1.84	63.11 ± 0.92
Matching Network (Vinyals et al., 2016)	Conv4	43.56 ± 0.84	55.31 ± 0.73
Relation Networks (Sung et al., 2018)	Conv4	50.44 ± 0.82	65.32 ± 0.70
IDeMe-Net (Chen et al., 2019b)	ResNet18	59.14 ± 0.86	74.63 ± 0.74
LwoF (Gidaris & Komodakis, 2018)	WRN-28-10	60.06 ± 0.14	76.39 ± 0.11
wDAE-GNN (Gidaris & Komodakis, 2019)	WRN-28-10	61.07 ± 0.15	76.75 ± 0.11
PPA (Qiao et al., 2018)	WRN-28-10	59.60 ± 0.41	73.74 ± 0.19
Prototypical Network(PN)‡ (Snell et al., 2017)	WRN-28-10	55.85 ± 0.48	68.72 ± 0.36
PN+TRAML† (Li et al., 2020a)	ResNet12	60.31 ± 0.48	$\mathbf{77.94 \pm 0.57}$
PN+rot+jigsaw (Su et al., 2020)	Resnet18	-	76.6
SEN PN (Kampffmeyer & Jenssen, 2020)	WRN-16-6	-	72.3
PN+rotation (Gidaris et al., 2019)	WRN-28-10	58.28 ± 0.49	72.13 ± 0.38
PN + AVCL(ours)	WRN-28-10	$\mathbf{61.75 \pm 0.43}$	$\mathbf{77.19 \pm 0.51}$

Auto-view transformation architectures We use 4-layer convolutional nets with 64 channels as localisation networks in STN modules that receive images of size 80×80 and output 4-dimensional vectors. The vectors are used as parameters of two diagonal affine transformations which apply to images and produce two views. When implementing without auto-view module, we replace the module with a random cropping function.

Feature extractor architectures In all our experiments, the main encoder and momentum encoder share the same architecture. Following prior work (Gidaris & Komodakis, 2019; Gidaris et al., 2019), we use a 2-layer Wide Residual Network(WRN-28-10) that outputs 640-dimensional feature vectors after global pooling given images(or views) of size 80×80. This feature space is directly used for metric-based meta-learning, but will be further mapped by a 2-layer mlp project head to a 128-dimentional hidden space for contrastive learning.

Training details At training, each minibatch contains 4 tasks, and classes for each task are randomly selected from training set. The query set contains 4 samples during meta-training and 16 samples during meta-testing. All samples from query set and support set are used for computing contrastive loss. All learnable components of our model are trained for 60 epochs by SGD optimizer with Nesterov momentum 0.9 and weight decay 0.0005. The learning rate for STN was set to 0.00001, the same magnitude as in the original paper. The learning rate for other parts of model was initially set to 0.1, and then changed to 0.01 and 0.001 at epochs 20 and 40, respectively. Moreover, regularization hyperparameter β was set to 2.0. We use a queue containing 63000 negative samples for contrastive learning. Momentum coefficient ϵ for updating momentum encoder was set to 0.999, following (He et al., 2020).

4.2 Evaluation on Benchmarks

Below we report comparative results on two benchmarks for FSL: MiniImageNet and CUB in Table 2 and 3. In Table 2 we divide methods into two groups and compare them with our proposed method, respectively. The first group contains recent comparative few-shot learning methods. The second group contains baseline method(Prototypical Network) and other methods that aim at improving it. Analysis from the results, we can find that: (1)Our method consistently improves the baseline method (prototypical Network). For instance, our model boosts performance of Prototypical Network on miniImageNet by 5.90% and 8.47% under the 1-shot and 5-shot settings, respectively. This verifies that our method can indeed improve model performance by refining fine-grained semantic structure of the feature space . (2)Our model outperforms recent comparable few-shot learning methods and also outperforms other approaches that aim at improving Prototypical Network. Moreover, we achieve competitive performance with the method using external text information (PN+TRAML). This further gives evidence of the superiority of our learned feature space for FSL.

Table 3: Comparative results for 5-way classification on CUB. Average accuracies on the meta-test set with 95 confidence interval are reported.

Model	backbone	1-shot	5-shot
AFHN (Li et al., 2020b)	ResNet18	70.53 ± 1.01	83.95 ± 0.63
FEAT (Ye et al., 2018)	ResNet 12	68.87 ± 0.22	82.90 ± 0.15
MAML (Chen et al., 2019a)	ResNet34	67.28 ± 1.08	83.47 ± 0.59
cosine classifier (Chen et al., 2019a)	ResNet34	68.00 ± 0.83	84.50 ± 0.51
Relationnet (Chen et al., 2019a)	ResNet34	66.20 ± 0.99	82.30 ± 0.58
DEML (Ye et al., 2018)	ResNet50	66.95 ± 1.06	77.11 ± 0.78
PN (Snell et al., 2017)	WRN-28-10	66.08 ± 0.54	78.79 ± 0.23
PN+AVCL(ours)	WRN-28-10	$\mathbf{71.21 \pm 0.43}$	$\mathbf{85.08 \pm 0.36}$

Table 4: Results for 5-way few-shot classification on three fine-grained datasets: Cars, Places and Plantae. Average accuracies on the meta-test set are reported.

	Cars		Places		Plantea	
Model	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
PN	62.11	75.83	62.19	76.67	53.59	67.98
PN+AVCL	**76.93**	**87.04**	**64.50**	**78.96**	**59.80**	**75.37**

4.3 Few-shot Fine-grained Image Recognition

Fine-grained categories are distinguished by subtle and local semantic differences, which makes few-shot fine-grained classification more difficult. We experimentally show that such difficulty can be largely addressed by our method. Table 4 presents 5-way mean accuracy on three datasets with fine-grained categories: Cars, Places and Plantae. It can be observed that our method improved performance of Prototypical Network by a large margin under both 5-shot and 1-shot settings. For instance, our method obtains 14.83% and 11.21% gains under 1-shot and 5-shot settings on Cars, respectively. This verifies that metric-based meta-learning benefits from better fine-grained semantic structure learnt by our method.

4.4 Evaluation of auto-view learning

In Fig. 3 we show four distinct types of views produced by our auto-view module. Aside from local-to-local and global-to-local views which can also be accomplished by random cropping, our auto-view module additionally allows one dimension scaling, Translation transformation, image deformation and proportional shrinkage. This flexibility can enrich semantics-invariant transformations applied to the images, forcing the encoder to extract essential content of the image. This allows the samples in the feature space to be distributed according to their semantics. Thus images from the same novel category can be mapped to close points in the feature space, which greatly improves generalization capability.

We quantitatively evaluate the effect of our auto-view module on five datasets in Table 5. PN+CL denotes models that replace auto-view modules with random cropping functions, which is the same as in (Chen et al., 2020; He et al., 2020). The results show that our auto-view module can indeed improve the quality of views, thus reach a better performance. Compared to random views, our AVCL method obtains 2.21% and 2.73% performance gains under the 5-way 1-shot and 5-shot settings on miniIMageNet, respectively. We additionally show training and test errors during training on miniImageNet in Figure 4. It can be observed that the curves of training error are similar, while the curves of test errors are different. While contrastive regularization helps the model generalize better, our auto-view module further improves it. This strongly supports our motivation that encoding fine-grained semantic contents can help metric-based meta-learning generalize better to novel classes in FSL.

Table 5: Quantitative evaluation of auto-view module. PN+CL denotes model with traditional random-view contrastive learning.

	Model	miniImageNet	CUB	Cars	Places	Plantae
1-shot	PN+CL	59.54	70.45	71.91	63.15	54.19
	PN+AVCL	**61.75**	**71.21**	**76.93**	**64.50**	**59.80**
5-shot	PN+CL	74.46	82.67	84.38	**79.43**	72.43
	PN+AVCL	**77.19**	**85.08**	**87.04**	78.96	**75.37**

Figure 3: Four types of views learned by our method: (a) Local-to-local, (b) Global-to-local, (c) One-dimension Scaling, and (d) Proportional zooming.

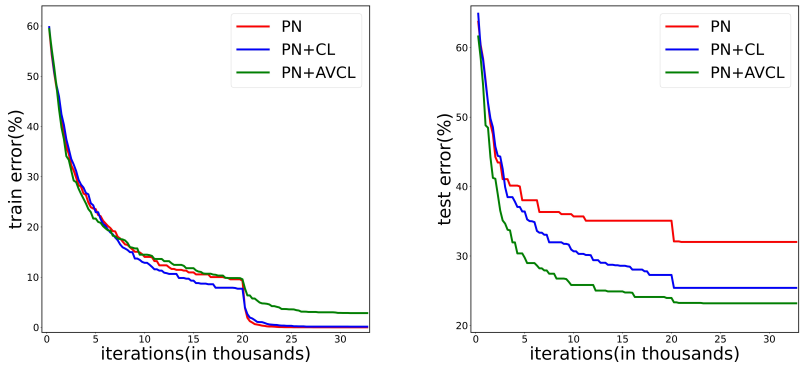

Figure 4: Training errors (left) and test errors (right) on miniImageNet. The auto-view module significantly decreases test errors, while keeping not overfit to the training set.

5 Conclusion

In this paper, we propose auto-view contrastive learning to improve few-shot image recognition. In particular, we design a learning-to-learn algorithm to adaptively learn the views. We carry out two paths of tasks, one is label-guided metric-based meta-learning, another is instance-level classification for exploring fine-grained semantic structure of feature space. Extensive experiments on benchmarks demonstrate that our method effectively boosts performance of metric-based meta-learning.

References

Luca Bertinetto, João F. Henriques, Philip H. S. Torr, and Andrea Vedaldi. Meta-learning with differentiable closed-form solvers. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. *CoRR*, abs/2002.05709, 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 *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019a.

Zitian Chen, Yanwei Fu, Yu-Xiong Wang, Lin Ma, Wei Liu, and Martial Hebert. Image deformation meta-networks for one-shot learning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 8680–8689. Computer Vision Foundation / IEEE, 2019b.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1126–1135. PMLR, 2017.

Spyros Gidaris and Nikos Komodakis. Dynamic few-shot visual learning without forgetting. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pp. 4367–4375. IEEE Computer Society, 2018.

Spyros Gidaris and Nikos Komodakis. Generating classification weights with GNN denoising autoencoders for few-shot learning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 21–30. Computer Vision Foundation / IEEE, 2019.

Spyros Gidaris, Andrei Bursuc, Nikos Komodakis, Patrick Pérez, and Matthieu Cord. Boosting few-shot visual learning with self-supervision. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pp. 8058–8067. IEEE, 2019.

Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2006), 17-22 June 2006, New York, NY, USA*, pp. 1735–1742. IEEE Computer Society, 2006.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. Momentum contrast for unsupervised visual representation learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pp. 9726–9735. IEEE, 2020.

Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alexander Shepard, Hartwig Adam, Pietro Perona, and Serge J. Belongie. The inaturalist species classification and detection dataset. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pp. 8769–8778. IEEE Computer Society, 2018.

Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. Spatial transformer networks. In Corinna Cortes, Neil D. Lawrence, Daniel D. Lee, Masashi Sugiyama, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pp. 2017–2025, 2015.

Davide Roverso Kampffmeyer and Robert Jenssen. Sen: A novel feature normalization dissimilarity measure for prototypical few-shot learning networks. 2020.

Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Li Fei-Fei. Novel dataset for fine-grained image categorization. In *First Workshop on Fine-Grained Visual Categorization, IEEE Conference on Computer Vision and Pattern Recognition*, Colorado Springs, CO, June 2011.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *2013 IEEE International Conference on Computer Vision Workshops, ICCV Workshops 2013, Sydney, Australia, December 1-8, 2013*, pp. 554–561. IEEE Computer Society, 2013.

Kwonjoon Lee, Subhransu Maji, Avinash Ravichandran, and Stefano Soatto. Meta-learning with differentiable convex optimization. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 10657–10665. Computer Vision Foundation / IEEE, 2019.

Aoxue Li, Weiran Huang, Xu Lan, Jiashi Feng, Zhenguo Li, and Liwei Wang. Boosting few-shot learning with adaptive margin loss. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pp. 12573–12581. IEEE, 2020a.

Fei-Fei Li, Robert Fergus, and Pietro Perona. One-shot learning of object categories. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28(4):594–611, 2006. doi: 10.1109/TPAMI.2006.79.

Kai Li, Yulun Zhang, Kunpeng Li, and Yun Fu. Adversarial feature hallucination networks for few-shot learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pp. 13467–13476. IEEE, 2020b.

Siyuan Qiao, Chenxi Liu, Wei Shen, and Alan L. Yuille. Few-shot image recognition by predicting parameters from activations. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pp. 7229–7238. IEEE Computer Society, 2018.

Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Fei-Fei Li. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.*, 115(3):211–252, 2015.

Andrei A. Rusu, Dushyant Rao, Jakub Sygnowski, Oriol Vinyals, Razvan Pascanu, Simon Osindero, and Raia Hadsell. Meta-learning with latent embedding optimization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.

Victor Garcia Satorras and Joan Bruna Estrach. Few-shot learning with graph neural networks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.

Nikunj Saunshi, Orestis Plevrakis, Sanjeev Arora, Mikhail Khodak, and Hrishikesh Khandeparkar. A theoretical analysis of contrastive unsupervised representation learning. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5628–5637. PMLR, 2019.

Jake Snell, Kevin Swersky, and Richard S. Zemel. Prototypical networks for few-shot learning. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pp. 4077–4087, 2017.

Jong-Chyi Su, Subhransu Maji, and Bharath Hariharan. When does self-supervision improve few-shot learning? *CoRR*, abs/1910.03560, 2020.

Martin Sundermeyer, Zoltan-Csaba Marton, Maximilian Durner, Manuel Brucker, and Rudolph Triebel. Implicit 3d orientation learning for 6d object detection from RGB images. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss (eds.), *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part VI*, volume 11210 of *Lecture Notes in Computer Science*, pp. 712–729. Springer, 2018.

Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip H. S. Torr, and Timothy M. Hospedales. Learning to compare: Relation network for few-shot learning. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pp. 1199–1208. IEEE Computer Society, 2018.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. *CoRR*, abs/1906.05849, 2019.

Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning. *CoRR*, abs/2005.10243, 2020.

Hung-Yu Tseng, Hsin-Ying Lee, Jia-Bin Huang, and Ming-Hsuan Yang. Cross-domain few-shot classification via learned feature-wise transformation. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748, 2018.

Oriol Vinyals, Charles Blundell, Tim Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. In Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pp. 3630–3638, 2016.

Peter Welinder, Steve Branson, Takeshi Mita, Catherine Wah, Florian Schroff, Serge Belongie, and Pietro Perona. Caltech-ucsd birds 200. *Technical Report CNS-TR-2010-001, California Institute of Technology*, 2010.

Zhirong Wu, Yuanjun Xiong, Stella X. Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pp. 3733–3742. IEEE Computer Society, 2018.

Han-Jia Ye, Hexiang Hu, De-Chuan Zhan, and Fei Sha. Learning embedding adaptation for few-shot learning. *CoRR*, abs/1812.03664, 2018.

Bolei Zhou, Àgata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(6): 1452–1464, 2018.

A LEARNING PROCEDURE

The pseudo code of our learning procedure is shown in Algorithm1,

Algorithm 1 Auto-View Contrastive Learning(AVCL) for metric-based meta-learning

Require: Base dataset $\mathbf{D}_B$, learning rate α and η, weight hyperparameter β, momentum coefficient ϵ, and maximum iteration number $t_{\max}$
Random initialization for $\theta^0, \omega^0, \gamma^0$
for $t = 0$ to $t_{\max}$ **do**
  /* Sample tasks */
  Randomly sample N classes from $\mathbf{D}_B$.
  Randomly sample K images from each class in $\mathbf{D}_B$ to form $\mathbf{D}_S^t$
  Randomly sample other M images from the same N classes in $\mathbf{D}_B$ to form $\mathbf{D}_Q^t$
  /* First forward pass */
  Using θ^t and ω^t to compute $\mathcal{L}^{\text{meta}}$ and $\mathcal{L}^{\text{con}}$ through Eq. (1) and Eq. (2)
  /* Optimize main encoder $F_\theta(\cdot)$, project head $g_\theta(\cdot)$ and momentum encoder $F_\omega(\cdot)$ */
  Update (θ^t, ω^t) to $(\theta^{t+1}, \omega^{t+1})$ through Eq. (3) and Eq.(4), and retain computational graph.
  /* Second forward pass */
  Using θ^{t+1} to compute $\mathcal{L}^{\text{meta}}$ through Eq. (1)
  /* Optimize spatial transformation module $G_{\gamma_1}(\cdot)$ and $G_{\gamma_2}(\cdot)$ */
  Update (γ_1^t, γ_2^t) to $(\gamma_1^{t+1}, \gamma_2^{t+1})$ through Eq. (5)
end for

Table 6: Effects of the β value in AVCL on model performances.

	5-way Acc.	
	1-shot	5-shot
β = 0.5	58.84 $\pm$ 0.39%	74.21 $\pm$ 0.55%
β = 1.0	59.45 $\pm$ 0.36%	75.64 $\pm$ 0.71%
β = 2.0	61.75 $\pm$ 0.43%	77.19 $\pm$ 0.51%
β = 5.0	58.73 $\pm$ 0.50%	74.37 $\pm$ 0.77%

B SPATIAL TRANSFORMER NETWORKS

In spatial transformer networks, the input source image $\mathbf{x}^s$ is first fed into a localisation net $G_\gamma(\cdot)$ and outputs six affine transformation parameters. This parameters form a 2×3 matrix which defines a affine transformation mapping each pixel coordinates (u_i^t, v_i^t) in the output $\mathbf{x}^t$ to a source coordinates (u_i^s, v_i^s) in the input. In our setting, we contrain the matrix to be diagonal so as to avoid skewing which could possibly change the semantics of images:

$$\begin{pmatrix} u_i^s \\ v_i^s \end{pmatrix} = \tau_\lambda(u_t^s, v_t^s) = \begin{bmatrix} \lambda_{11} & 0 & \lambda_{13} \\ 0 & \lambda_{22} & \lambda_{23} \end{bmatrix} \begin{pmatrix} u_i^t \\ v_i^t \\ 1 \end{pmatrix}$$

Finally, the values of each pixels in $\mathbf{x}^t$ is determined by bilinear interpolation at their corresponding coordinates in the source images, called differentiable image sampling.

C MOMENTUM CONTRAST

Different from the standard framework in SimCLR (Chen et al., 2020) ,momentum contrast framework introduces a queue q preserving negative samples and a momentum encoder $F_\omega(\cdot)$, to alleviate the problem of need for very large batch size for contrastive learning. In each iteration, immediate preceding features in the queue encoded by $F_\omega(\cdot)$ could be reused as negative samples to compute the contrastive loss. At the end of each iteration, features of current mini-batch is enqueued to the queue, and earliest features in the queue are removed. The update of the encoder $F_\omega(\cdot)$ is intractable by back-propagation. To maintain consistency, MoCo updates ω as a moving average of the main encoder's parameter θ, as shown in eq. (4).

D EFFECT OF REGULARIZATION HYPERPARAMETER

We perform ablation study w.r.t. the regularization hyperparameter β which controls the magnitude of contrastive loss. Table 6 shows the accuracies of 5-way few-shot learning on miniImageNet. We can observe that when β is small, the contrastive loss cannot thoroughly explore the semantics inside data, thus cannot boost the performance much. When the value of β is s too large, the accuracy also decreases. This implies that supervised information is somewhat ignored, which is still important.