*LETTER* **Communicated by Xiu-Shen Wei**

# Fine-Grained 3D-Attention Prototypes for Few-Shot Learning

**Xin Hu**
*dr.huxin711@foxmail.com*
**Jun Liu**
*liukeen@mail.xjtu.edu.cn*
*National Engineering Lab for Big Data Analytics and School of Computer Science and Technology, Xi'an Jiaotong University, Xi'an, Shaanxi 710049, China*

**Jie Ma**
*dr.majie@foxmail.com*
**Yudai Pan**
*pyd418@foxmail.com*
*School of Computer Science and Technology and Shaanxi Province Key Laboratory of Satellite and Terrestrial Network Tech. R&D, Xi'an Jiaotong University, Xi'an, Shaanxi 710049, China*

**Lingling Zhang**
*zhanglling@xjtu.edu.cn*
*Shaanxi Province Key Laboratory of Satellite and Terrestrial Network Tech. R&D, Xi'an Jiaotong University, Xi'an, Shaanxi 710049, China*

**In the real world, a limited number of labeled finely grained images per class can hardly represent the class distribution effectively. Due to the more subtle visual differences in fine-grained images than simple images with obvious objects, that is, there exist smaller interclass and larger intraclass variations. To solve these issues, we propose an end-to-end attention-based model for fine-grained few-shot image classification (AFG) with the recent episode training strategy. It is composed mainly of a feature learning module, an image reconstruction module, and a label distribution module. The feature learning module mainly devises a 3D-Attention mechanism, which considers both the spatial positions and different channel attentions of the image features, in order to learn more discriminative local features to better represent the class distribution. The image reconstruction module calculates the mappings between local features and the original images. It is constrained by a designed loss function as auxiliary supervised information, so that the learning of each local feature does not need extra annotations. The label distribution module is used to predict the label distribution of a given unlabeled sample, and we use the local features to represent the image features for classification. By conducting comprehensive experiments on Mini-ImageNet**

*Neural Computation* **32, 1664–1684** (2020)
© 2020 Massachusetts Institute of Technology
https://doi.org/10.1162/neco_a_01302

Downloaded from http://direct.mit.edu/neco/article-pdf/32/9/1664/1865020/neco_a_01302.pdf by guest on 29 September 2021

**and three fine-grained data sets, we demonstrate that the proposed model achieves superior performance over the competitors.**

## 1 Introduction

Few-shot learning (FSL) refers to the ability to learn a new concept by accessing only one or a few examples. Currently, it focuses mainly on generic images of simple and obvious objects. For example, the two species *dog* and *monkey* have obvious differences in appearance. On MatchingNet (Vinyals et al., 2016) and Protonet (Snell, Swersky, & Zemel, 2017), those classifiers are trained by learning the mappings from global image features into label space.

However, images in the real world are far more complicated than generic ones. Thus, recognizing fine-grained classes such as bird species, plane models, or flower types is currently an active topic. Given a *dog*, if analyzing its fine-grained local features of the *ears*, *mouth*, and *legs*, we can learn that it is a *dog* and also a *Corky* (see Figure 1, top).

In addition, there is small interclass variation and large intraclass variation in fine-grained images. Consider the two bird species, Brandt cormorant and pelagic cormorant (see Figure 1, bottom). Distinguishing those two species is quite challenging because of the subtle visual differences: in their beaks and wings are almost the same (yellow and purple circles). In some images, the same species also have great differences due to their postures and positions (red and blue dotted rectangles).

Although there are some classic studies of few-shot learning in fine-grained settings (Li, Wang et al., 2019; Wei, Wang, Liu, Shen, and Wu, 2019; Wertheimer & Hariharan, 2019), two problems remain to be solved in the existing methods. One problem is that a limited number of images per class can not represent the class distribution effectively, and image-level features are difficult to reflect the subtle visual differences between fine-grained images. The other is that expensive supervision is needed in some existing methods (e.g., human-annotated bounding box/part annotations; Zhang et al., 2019), to obtain local features for fine-grained images.

To address these issues we propose an end-to-end attention-based model for fine-grained few-shot image classification (AFG). It is mainly composed of a feature learning module, an image reconstruction module, and a label distribution module. The feature learning module, with a designed *3D-attention* mechanism, considers different spatial positions and channels of image features in order to learn more discriminative local features to represent class distribution better. The image reconstruction module maps the local features into the original images to promote the learning effect of the previous module. The label distribution module calculates the similarity of several labeled samples and an unlabeled one in order to predict the label

Downloaded from http://direct.mit.edu/neco/article-pdf/32/9/1664/1865020/neco_a_01302.pdf by guest on 29 September 2021

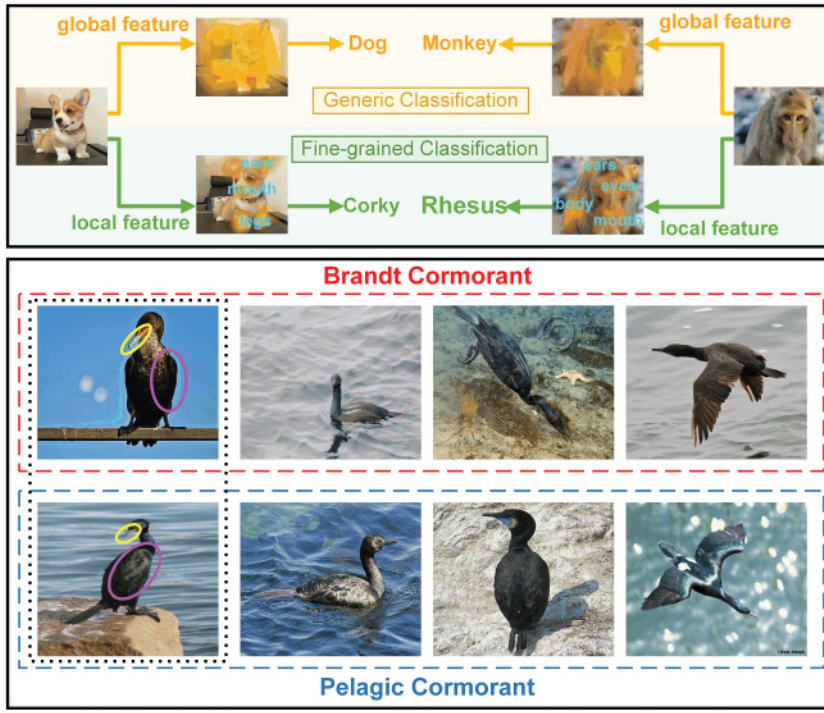


Figure 1: (Top) Comparison of the targets of generic classification and finely grained classification. (Bottom) Illustration of smaller interclass and larger intraclass variations in finely grained images. The red and blue dotted boxes show two bird species. There exists larger intraclass variation along each row. The content of the black dotted box indicates the smaller interclass variation owing to the differences in the pattern on their beaks and wings.

distribution of the unlabeled sample. Our major contributions are summarized as follows:

1. We discuss few-shot learning in a challenging fine-grained setting and propose a novel attention-based model, AFG, where three main modules are combined to learn the suitable local features for fine-grained images without additional annotations.
2. We devise a novel 3D-Attention mechanism. It treats the entire image as a set of local descriptors, each of which describes the discriminative feature of a particular spatial position. In addition, a channel attention function is proposed to calculate the attention distribution of semantic information along the channel dimensions.

Downloaded from http://direct.mit.edu/neco/article-pdf/32/9/1664/1865020/neco_a_01302.pdf by guest on 29 September 2021

3. We propose an image reconstruction module constrained by a loss function as auxiliary supervised information. It makes full use of the limited local descriptors to reconstruct the original image in order to indirectly improve the performance of the feature learning module.
4. To validate the AFG, we conduct experiments on Mini-ImageNet and three fine-grained benchmark data sets: CubBirds (Wah, Branson, Welinder, Perona, & Belongie, 2011), StanfordDogs (Khosla, Jayadevaprakash, Yao, & Li, 2011), and StanfordCars (Krause, Stark, Deng, & Fei-Fei, 2013). Experimental results show that our model significantly outperforms over the competitors.

The remainder of the letter is organized as follows. Section 2 describes various existing methods in few-shot learning and fine-grained image recognition. In section 3, we briefly introduce the task scenario and propose the novel attention-based model, AFG, to solve the fine-grained few-shot learning problem. Section 4 gives the experimental details for our model. Conclusions and the future work are provided in section 5.

## 2 Related Work

The key point of fine-grained few-shot image classification is to transfer common knowledge from seen fine-grained samples to unseen ones. The combination of few-shot learning and fine-grained recognition can help in learning more discriminative local features with limited samples.

**2.1 Few-Shot Learning.** Few-shot learning has become more and more attractive in deep learning scenarios (Snell et al., 2017; Sung et al., 2018; Cao, Wang, & Brown, 2016) in the fields of neural language processing and computer vision (Miller, Wang, & Kesidis, 2019; Vyškovskỳ, Schwarz, & Kašpárek, 2019; Tang, Ma, Kong, & Li, 2019). The main methods are data-augment-based, metric-learning-based, and meta-learning-based.

*2.1.1 Data-Augment-Based Methods.* These methods directly extend the number of images with transformation, such as cropping (Qi, Brown, & Lowe, 2018), reflecting (Kozerawski & Turk, 2018), and flipping (Shyam, Gupta, & Dukkipati, 2017). Alfassy et al. (2019) proposed combining pairs of given examples in feature space, so that the resulting synthesized feature vectors will correspond to examples whose label sets are obtained through certain set operations (intersection, union, and subtraction) on the label sets of the corresponding input pairs. Because the augmentation is based on original samples, given some transformation rules, there are few differences between a constructed new sample and the original one, that is, the combination choices are great limited. There are also several generating methods. Edraki and Qi (2018) proposed a regularized loss-sensitive GAN model with proven distributional consistency and generalizability to generate real

Downloaded from http://direct.mit.edu/neco/article-pdf/32/9/1664/1865020/neco_a_01302.pdf by guest on 29 September 2021

data. Schwartz et al. (2018) used a variant of the autoencoder to capture the intraclass differences between two classes in the latent space and then transferred class distributions from training to novel classes. However, a rigorous assumption of this work is that intraclass variances can always be generalized to new classes.

*2.1.2 Metric-Learning-Based Methods.* These methods mainly learn sample distance in the image feature space to represent the similarity between each other (Satorras & Estrach, 2018; Koch, Zemel, & Salakhutdinov, 2015; Snell et al., 2017; Li, Xu et al., 2019; Sung et al., 2018). Vinyals et al. (2016) proposed an episode training strategy and introduced a contextual mechanism by using an attention-LSTM model when computing the cosine similarity between samples. Koch et al. (2015) proposed a SiameseNet based on computing the pairwise samples distance. It can be used to solve one-shot problems by *k*-nearest neighbors classification. Snell et al. (2017) built a ProtoNet by taking the mean of each class as its "prototype to learn a representation, and it provided significant improvements over Vinyals et al. (2016) by using Euclidean distance instead of cosine. These methods are simple and intuitive, but defining an effective measure of sample distance remains an open question.

*2.1.3 Meta-Learning-Based Methods.* These methods aim to train a meta-learner to learn meta-knowledge, for example, initial parameters, that are excellent in various tasks (Thrun, 1998; Vilalta & Drissi, 2002; Andrychowicz et al., 2016; Zhang et al., 2020, 2018). For instance, model-agnostic meta-learning (MAML) (Finn, Abbeel, & Levine, 2017) is a meta-learning model or learning a better network initialization, and it is trained with a gradient descent procedure. Li, Zhou, Chen, and Li (2017) developed a new SGD-like meta-learner meta-SGD, which can learn all ingredients of an optimizer, namely, initialization, update direction, and learning rate in an end-to-end manner. Ravi and Larochelle (2017) adopted an LSTM-based meta-learner as an optimizer to train another classifier, as well as learn a task-common initialization for this classifier. The limitation of these kinds of methods is that a large number of different classes are required to train a classifier in order to be better generalized to other tasks.

However, previous few-shot learning tended to focus on generic rather than fine-grained images, which have subtle visual differences from each other.

**2.2 Fine-Grained Image Recognition.** Fine-grained image recognition is a challenging problem due to smaller interclass and larger intraclass variations between samples. Three main types of methods are commonly used now: (Branson, Van Horn, Belongie, Perona, & Tech, 2014; Fu, Zheng, & Mei, 2017; Huang, Xu, Tao, & Zhang, 2016; Zhang, Wei et al., 2016; Zheng, Fu, Mei, & Luo, 2017) mainly include three types: supervised methods, weakly supervised methods and unsupervised methods.

Downloaded from http://direct.mit.edu/neco/article-pdf/32/9/1664/1865020/neco_a_01302.pdf by guest on 29 September 2021

*2.2.1 Supervised Methods.* These methods use extra manual labeling information such as bounding box and local positions, and not just category labels. Based on manually labeled part annotations, Huang et al. (2016) proposed a novel part-stacked CNN architecture that consists of a fully convolutional network to locate object parts and a two-stream classification network to encode object-level and part-level cues simultaneously. Wei, Xie, Wu, and Shen (2018) proposed a fully convolutional network to locate the discriminative parts with part annotations for fine-grained images. These kinds of methods require labor-intensive annotations, which limit their usability.

*2.2.2 Weakly-Supervised Methods.* These methods rely solely on category labels to complete classification, which is a major trend in fine-grained image recognition. Lin, RoyChowdhury, and Maji (2015) designed a bilinear structure to compute the pairwise feature interactions, which can better model subtle visual differences. Because of the dimension explosion of bilinear structure, Gao, Beijbom, Zhang, and Darrell (2016) applied Tensor Sketch to approximate the second-order statistics to reduce the feature dimension. Zheng, Fu, Zha, and Luo (2019) designed a trilinear attention sampling network (TASN) to learn fine-grained features in an efficient teacher-student manner.

*2.2.3 Unsupervised Methods.* Wei, Luo, Wu, and Zhou (2017) proposed the selective convolutional descriptor aggregation (SCDA) to localize the main object in fine-grained image. Then the selected (localized) descriptors were aggregated to produce a short feature vector for a fine-grained image. Xie, Wang, Zhang, and Tian (2015) proposed a fine-grained image search that returns not only near-duplicate but also fine-grained results. It formulated the fine-grained image search problem by constructing a new database and defining an evaluation method. Wei, Wang, Liu, Shen, and Wu (2019) proposed an end-to-end trainable network for the fine-grained few-shot learning task (FSFG). By considering the special structure of bilinear CNN features, it decomposed the exemplar-to-classifier mapping into a set of more attainable "part"-to-"part classifier" mappings.

Compared with previous studies, the proposed AFG takes the relationship between each local descriptor into account under few-shot learning and designs a constraint strategy to promote its performance.

## 3 The Proposed Model

In this section, we first present the problem scenario in fine-grained few-shot learning. Then we demonstrate the three modules of AFG in the following sections.

Downloaded from http://direct.mit.edu/neco/article-pdf/32/9/1664/1865020/neco_a_01302.pdf by guest on 29 September 2021

**3.1 Problem Definition.** Considering the $N$-way $K$-shot learning task, the goal is to classify several unlabeled fine-grained samples *query set* ($\mathbb{Q}$) into their true classes giving a *support set* ($\mathbb{S}$). $\mathbb{S}$ consists of $N$ different image classes with $K$-labeled fine-grained training samples per class.

Although a classifier can learn when $\mathbb{S}$ is used, image-level features are difficult to reflect the subtle visual differences between fine-grained images. Such a classifier always results in overfitting. Therefore, we use an auxiliary training set like that in previous studies (Snell et al., 2017; Sung et al., 2018; Satorras & Estrach, 2018) to improve an $N$-way $K$-shot classification performance by learning more discriminative local features. This set with a large number of labeled training samples is called *base set*, while the target few-shot learning task is verified on the *novel set*. Note that the classes in the base set and the novel set are disjoint.

The model is trained using episodes (Vinyals et al., 2016). In each episode we randomly select $N$ classes with $K$ samples per class from the base set to act as $\mathbb{S}_{base} = \{(X_i, Y_i)\}_{i=1}^{N\times K}(Y_i \in \{1, 2, \ldots, N_{base}\})$. The remainders of those $N$ classes with $M$ samples per class compose the $\mathbb{Q}_{base} = \{(\hat{X}_i, \hat{Y}_i)\}_{i=1}^{N\times M}(\hat{Y}_i \in \{1, 2, \ldots, N_{base}\})$. Here, $X(\hat{X})$ is an image, and $Y(\hat{Y})$ is its corresponding one-hot label.

During training, tens of thousands of episodes will be constructed to train the proposed AFG. In the test stage, with novel support set $\mathbb{S}_{novel}$, the learned model can be directly used to classify each image in $\mathbb{Q}_{novel}$.

**3.2 AFG Model.** The overall architecture of the proposed AFG is summarized in Figure 2. It mainly consists of three components: a feature learning module, an image reconstruction module, and a label distribution module.

The feature learning module first feeds $S$ into the feature extractor $\mathbb{F}(\cdot)$ and adopts a 3D-Attention mechanism to obtain $m$ more representative local attentions. Then the dot product of original feature maps and the learned attentions can be used to calculate more discriminative local features $\mathbf{x}_t$, each of which is called a *prototype* of the image. We concatenate all those local prototypes to represent the global image features that are suitable for fine-grained image classification.

The image reconstruction module is a designed deconvolution network to obtain the mappings between local prototypes and the original images. We can restore each local prototype $\mathbf{x}_t$ to the corresponding local image $R_t^{img}$. By combining each $R_t^{img}$, this module generates a global reconstructed image $\tilde{X}$.

The label distribution module is mainly used to calculate the similarity between several labeled samples $\mathbb{S}$ and an unlabeled one $\mathbb{Q}$ to predict the label distribution of $\mathbb{Q}$. In this letter, we use the cosine similarity measurement. The $\mathbb{Q}$ label is predicted as the label with the largest similarity. The

Downloaded from http://direct.mit.edu/neco/article-pdf/32/9/1664/1865020/neco_a_01302.pdf by guest on 29 September 2021

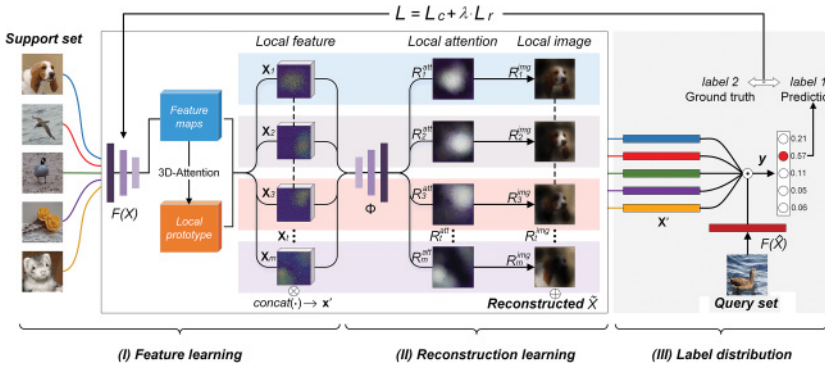


Figure 2: Overview architecture of our proposed AFG model, consists mainly of three components. (I) The first component (the feature learning module) is used to learn more discriminative semantic features with the 3D-Attention mechanism. ⊗ indicates concatenating local features to represent image features. (II) The second component (the image reconstruction module) reconstructs the original image by local features. ⊕ indicates the superposition method to compute the reconstructed image. (III) The third component (the label distribution module) represents the calculation of label distribution for the $Q$ sample by using the cosine distance of the feature vector between $S$ samples and the $Q$ sample. ⊙ indicates the similarity measure.

difference between the prediction and the ground truth is used as the loss to optimize the proposed AFG.

We illustrate the first two modules in sections 3.2.1 and 3.2.2 and the designed loss functions in section 3.2.3.

*3.2.1 Discriminative Feature Learning.* Many studies have proposed to learn the localization of parts by using extra annotations of the bounding box (Huang et al., 2016; Lin, Shen, Lu, & Jia, 2015; Zhang, Xu et al., 2016; Zhang, Donahue, Girshick, & Darrell, 2014). Acquiring rich annotations is quite labor-consuming, so we employ the local descriptors (Li, Wang et al., 2019; Hu, Wang, Yang, & Nie, 2019) to represent global image features without unnecessary human efforts. Because we consider both the spatial positions and different channel attentions of the image features, we call it a *3D-Attention* mechanism (see Figure 3). The details are as follows.

*Spatial attention* mainly defines which prototypes of the image should be noticed. We first feed an input image $X$ into the feature extractor $\mathbb{F}(\cdot)$, which contains only convolutional layers but no fully connected layer, and then it generates the needed deep feature $\mathbf{x}$ (tensor) as $\mathbb{F}(X) = \mathbf{x} \in \mathbb{R}^{w \times h \times c}$, where $w \times h$ is the number of descriptors and $c$ is the channel numbers of each

Downloaded from http://direct.mit.edu/neco/article-pdf/32/9/1664/1865020/neco_a_01302.pdf by guest on 29 September 2021

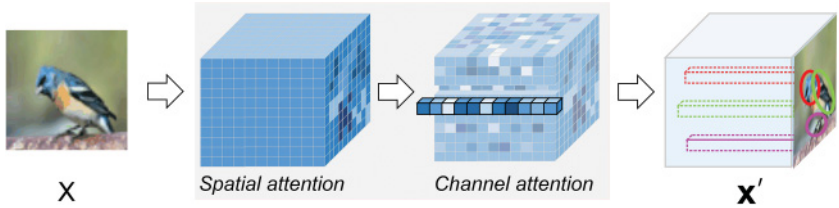


Figure 3: Discriminative feature learning with 3D-Attention.

descriptor. In our experiments, each image corresponds to 100 descriptors, and each descriptor has 1024 channels.

Although a descriptor can correspond to a spatial position, it is usually difficult to express rich part information by a single one. In order to better represent every prototype for fine-grained images, AFG learns representative *local prototypes* with the attention module for an image. First, we feed the learned $\mathbf{x}$ into spatial attention function $\psi(\cdot)$, which consists of two convolution blocks and a softmax function. Then it outputs $m$ discriminative local part attentions, which is given as equation 3.1. In this way, our model can filter out complex and independent background and focus on meaningful foreground targets:

$$\psi(\mathbf{x}) = [\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_t, \ldots, \mathbf{a}_m], \forall t : \mathbf{a}_t \in \mathbb{R}^{1\times c}. \tag{3.1}$$

*Channel attention* is used to learn the different attentions along channel dimensions. We design a function $\mathbb{CA}(\cdot)$, which mainly contains a convolution block and a sigmoid function $\sigma(x) = \frac{1}{1+e^{-x}}$ to produce attention distributions for this dimension. The final discriminative feature is calculated by equation 3.2. First, we calculate a prototype $\mathbf{x}_t$, which is the dot product $\odot$ of the discriminative local attention $\mathbb{CA}(\mathbf{a}_t)$ and the original image features $\mathbb{F}(X)$. *Concat* is a concatenate operation to splice $m$ prototypes (where $t$ takes a value of 1 to $m$) when representing the global feature $\mathbf{x}'$ of the image:

$$\begin{aligned} \mathbf{x}_t &= \mathbb{CA}(\mathbf{a}_t) \odot \mathbb{F}(X), t \in [1:m], \\ \mathbf{x}' &= concat_{t:1\to m}(\mathbf{x}_t). \end{aligned} \tag{3.2}$$

Each channel of the deep features represents different semantic information. According to human behaviors, when analyzing the bird species, we need to observe only a bird's head, body, foot, and color. That is, we focus on just a few key prototypes instead of all of them, which can help us learn the category label of the object.

*3.2.2 Image Reconstruction.* Figure 4 shows the image reconstruction module, which consists of three deconvolutional neural networks. It feeds

Downloaded from http://direct.mit.edu/neco/article-pdf/32/9/1664/1865020/neco_a_01302.pdf by guest on 29 September 2021

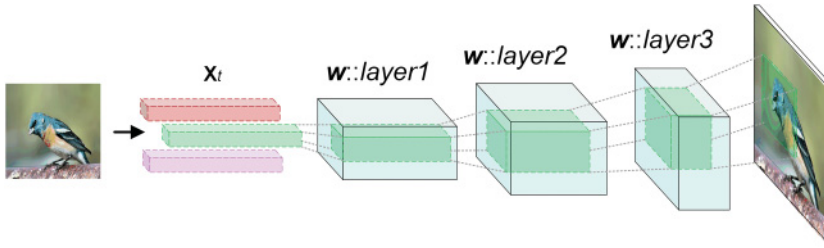


Figure 4: Diagrammatic sketch for image reconstruction.

the discriminative prototypes from the previous module as input to get the response between $\mathbf{x}_t$ and the original image $X$. This process is represented as follows:

$$R_t^{att} = \sigma(\Phi(\mathbf{x}_t; \mathbf{W}_t, \mathbf{b}_t)). \tag{3.3}$$

$\Phi$ is a deconvolution module, and $\mathbf{W}_t$ and $\mathbf{b}_t$ are the weights and bias, respectively, of a deconvolutional network. Moreover, the final output with sigmoid function $\sigma(x)$ is to generate the attention scores $R_t^{att}$ of the prototypes on the entire original image. Then the reconstructed part images $R_t^{img}$ can be acquired by adopting a dot production of $R_t^{att}$ and the original image $X$, shown in equation 3.4. We then sum all of the part images to represent the reconstructed image $\tilde{X}$:

$$\tilde{X} = \sum_{t=1}^{m} R_t^{img} = \sum_{t=1}^{m} R_t^{att} \odot X. \tag{3.4}$$

The image reconstruction module encourages the network to generate local original images that are as realistic as possible. In other words, minimizing the difference between the reconstructed image and the original input image can be used indirectly as supervised information for feature learning, without extra annotations of parts. The specific constraint of this module is detailed in the loss function in the next section.

*3.2.3 Objective Function.* Given a query image $\hat{X}$, our model computes a label distribution $\hat{y}$ by weighted summation over $Y_j$ with support images $X_j$ as follows:

$$P(\hat{y}|\hat{X}, S_{base}) = \sum_{j=1}^{N\times K} sim(c(\hat{\mathbf{x}}', \mathbf{x}_j')) \cdot Y_j, \tag{3.5}$$

Downloaded from http://direct.mit.edu/neco/article-pdf/32/9/1664/1865020/neco_a_01302.pdf by guest on 29 September 2021

where $\mathbf{x}'_j, Y_j$ are, respectively, discriminative semantic features of support images and corresponding labels. We firstly use *cosine* as the weight measurement function $c(\cdot)$ to estimate the distance between query image features $\hat{\mathbf{x}}'$ and $\mathbf{x}'_j$. Then we apply the softmax function $sim(\cdot)$ over the cosine distance $c$.

To stabilize the training of our model, we optimize an objective function $L(\cdot)$, which consists of a classification loss $L_c$ and a reconstruction loss $L_r$:

$$L = L_c + \lambda \cdot L_r. \tag{3.6}$$

$\lambda$ is used to balance the importance of two loss functions, which will enforce the feature representation of samples and improve the classification performance.

Classification loss $L_c$ classifies unlabeled images into corresponding categories. The proposed AFG adopts cross-entropy loss, shown in equation 3.7, where $\hat{Y}$ is the ground truth one-hot label and $P$ is the label distribution calculated by equation 3.5. $\mathbb{A}$ is an indicator variable, $\mathbb{A}[\hat{Y}_i = Y_j] = 1$, if the label of sample $i$ in $Q$ is equal to the label of sample $j$ in $S$; otherwise, $\mathbb{A}[\hat{Y}_i \neq Y_j] = 0$:

$$L_c = -\frac{1}{M \times N} \sum_{i=1}^{M} \sum_{j=1}^{N} \mathbb{A}[\hat{Y}_i = Y_j] \cdot \log P. \tag{3.7}$$

Reconstruction loss $L_r$ encourages the network to minimize the discrepancy between the composite image $\tilde{X}$ and the original input image $X$. As shown in equation 3.8, we use mean-square error (MSE) to constrain the similarity from width ($W$), height ($H$), and channel ($C$) dimensions, respectively. This constraint not only guarantees the appearance consistency between learned part images and the original images, but also maintains the logical relationship among semantic prototypes:

$$L_r = \frac{1}{W \times H \times C} \sum_{i=1}^{W} \sum_{j=1}^{H} \sum_{c'=1}^{C} \| \tilde{X}_{i,j,c'} - X_{i,j,c'} \|_2. \tag{3.8}$$

## 4 Experiments

In this section, we first describe the data sets and experimental settings. Then we present results on four benchmark data sets. Finally, we evaluate the effectiveness of our proposed model.

**4.1 Data Sets.** Our experiments are conducted on Mini-ImageNet and three fine-grained data sets: Cub Birds, Stanford Dogs, and Stanford Cars:

Downloaded from http://direct.mit.edu/neco/article-pdf/32/9/1664/1865020/neco_a_01302.pdf by guest on 29 September 2021

**Mini-ImageNet** (Vinyals et al., 2016). This data set consists of 60,000 color images of size 84×84 pixels with 100 classes, each having 600 examples. For our experiments, we use the splits introduced in Wei et al. (2019), in order to directly compare with these baselines. The data set is divided into 64 training, 16 validation, and 20 testing classes.

**Cub Birds** (Wah et al., 2011). We use CUB-200-2011, which contains 11,788 images of 200 bird species in the experiments, which we split into 130, 20, and 50 classes for training, validation, and testing.

**Stanford Dogs** (Khosla et al., 2011). This data set contains 20,580 images with 120 classes, which is a benchmark data set for a fine-grained recognition task like CubBirds. We use 70, 20, and 30 classes for training, validation, and testing, respectively.

**Stanford Cars** (Krause et al., 2013). The original download StanfordCars is split into 8144 training and 8041 testing images. In our experiments, we divide all 196 classes into 130 for training, 17 for validation, and 49 for testing.

We resize all images of the last three data sets to 84×84 pixels in order to facilitate a fair comparison with these baselines for few-shot fine-grained image classification.

### 4.2 Experimental Implementation.

*4.2.1 Baselines.* In our experiments, we select various state-of-the-art models as competitors:

*Metric-learning-based models*. We compare the proposed AFG to four classic metric-learning based models in few-shot learning: MatchingNet (Vinyals et al., 2016), ProtoNet (Snell et al., 2017), RelationNet (Sung et al., 2018), and GNN (Satorras & Estrach, 2018), which are trained by the Conv64 as their embedding module.

*Meta-learning-based models*. As we adopt meta-learning strategy for training, we pick three high-reference meta-learning models for reference: model-agnostic meta-learning (MAML; Finn et al., 2017), optimiLSTM (Ravi & Larochelle, 2017), and R2-D2 (Bertinetto, Henriques, Torr, & Vedaldi, 2019).

*Fine-grained-based models*. Now that few-shot learning in fine-grained scenarios has been studied, we chose some new work published as baselines—few-shot fine-grained image recognition (FSFG; Wei et al., 2019) and Meta-iNat (Wertheimer & Hariharan, 2019) and some other classical models that are validated on fine-grained data sets.

*4.2.2 Settings.* All experiments are conducted around the 5-way 1-shot and 5-shot classification tasks on the above data sets. Taking 5-way 1-shot

Downloaded from http://direct.mit.edu/neco/article-pdf/32/9/1664/1865020/neco_a_01302.pdf by guest on 29 September 2021

Table 1: Mean Accuracy (%) on Mini-ImageNet.

| | | Five-Way on Mini-ImageNet | |
|---|---|---|---|
| Model | Type | One-Shot | Five-Shot |
| Matching Net | Metric | 43.56 | 55.31 |
| ProtoNet | Metric | 49.42 | 68.20 |
| RelationNet | Metric | 50.44 | 65.32 |
| GNN | Metric | 50.33 | 66.41 |
| MAML | Meta | 48.70 | 63.11 |
| OptimiLSTM | Meta | 43.44 | 60.60 |
| R2-D2 | Meta | 49.50 | 65.40 |
| Meta-iNat | – | 49.64 | **69.45** |
| Ours | Metric | **51.03** | 69.14 |

Note: The highest average accuracy of each column is marked in bold.

as an example, there are five support images in one episode. To mimic the testing condition, we randomly sample and construct 300,000 episodes for the training set. Considering the fairness of the comparison experiments, we use three convolution blocks as an embedding network for the feature learning module, each of which consists of a convolutional layer, a batch normalization layer, a LeakyReLU layer, and a max-pooling layer. We train the model with Adam and set the initial learning rate of 0.001, reducing the learning rate by half for every 100,000 episodes. Our model is implemented using the open-source library PyTorch.

**4.3 Experimental Results.** This section presents the average accuracy rates of AFG on the novel classes of the previous four data sets. For each data set, all of our experiments revolve around the same basic tasks. We compare the results with generic few-shot models and few-shot fine-grained models.

*4.3.1 Comparison with Generic Few-Shot Models.* We conduct this experiment on MiniImageNet. Our proposed AFG can learn the prototypes and reconstruct the original image in an end-to-end manner by employing the episode, which indeed obtains superior results:

- As shown in Table 1, our AFG offers significant improvements over baselines. Compared to the metric-learning based methods, AFG gains 7.47%, 1.61%, 0.59%, and 0.7% improvements in the 1-shot setting and 13.83%, 0.94%, 3.82%, and 2.73% in the 5-shot setting respectively. It indicates that we use prototypes instead of complex image-level features like baselines to represent targets, and then our model achieves obvious improvements.

Downloaded from http://direct.mit.edu/neco/article-pdf/32/9/1664/1865020/neco_a_01302.pdf by guest on 29 September 2021

Table 2: The Mean Accuracy (%) on Three Fine-Grained Data Sets.

| | 5-Way | | | | | |
|---|---|---|---|---|---|---|
| | Cub Birds | | Stanford Dogs | | Stanford Cars | |
| Model | 1-Shot | 5-Shot | 1-Shot | 5-Shot | 1-Shot | 5-Shot |
| k-NN | 38.85 | 55.58 | 24.53 | 40.30 | 26.99 | 43.40 |
| SVM | 34.47 | 59.19 | 23.37 | 39.50 | 25.66 | 51.07 |
| SiameseNet[a] | 37.38 | 57.73 | 23.99 | 39.69 | 25.81 | 48.95 |
| SiameseNet[b] | 26.58 | 43.51 | 19.28 | 31.49 | 22.41 | 40.07 |
| Matching Net | 45.30 | 59.50 | 35.80 | 47.50 | 34.80 | 44.70 |
| ProtoNet | 37.36 | 45.28 | 37.59 | 48.19 | 40.90 | 52.93 |
| FSFG | 42.10 | **62.48** | 28.78 | 46.92 | 29.63 | 52.28 |
| Ours | **50.02** | 54.41 | **39.40** | **59.61** | **42.23** | **62.90** |

Notes: The highest average accuracy of each column is marked in bold. [a]Uses fully bilinear pooling representations. [b]Uses compact bilinear pooling.

- We also report the results of the typical meta-learning-based methods. We can see that AFG is still competitive with these methods. Concretely, it has 2.33%, 7.59%, and 1.53% higher performance than that of MAML, optimiLSTM, and R2-D2 in the 1-shot setting and 6.03%, 8.54%, and 3.74% higher results in the 5-shot setting, respectively.
- For the comparison results with Meta-iNat (Wertheimer & Hariharan, 2019), AFG is only 0.31% lower than it in the 5-shot setting. By analyzing the implementation details, the difference is caused by different feature extractors. Meta-iNat uses the pretrained ResNet-50 with stronger feature representation ability, while we use three-layer convolutional neural networks. When AFG adopts ResNet-18, it achieves 72.7% in the 5-shot setting and is 3.25% higher than Meta-iNat.
- Another difference between Meta-iNat and AFG is learning strategy. Meta-iNat uses leave-one-out cross-validation within each batch, abandoning the hard Support/Query split to make full use of all samples for training, while we use episode to mimic the target few-shot setting.

*4.3.2 Compared with Few-Shot Fine-Grained Models.* The results are verified on three fine-grained data sets, and for a convenient and fair comparison, the accuracy of baselines in Table 2 is directly used by those given in (Li, Wang et al., 2019; Wei et al., 2019). Our model achieves competitive results on all experimental settings:

- Table 2 shows that the simple baselines of k-NN and SVM perform even better than other sophisticated neural networks. For the 1-shot setting, the performance of our proposed model is 4.72%, 1.81%, and

Downloaded from http://direct.mit.edu/neco/article-pdf/32/9/1664/1865020/neco_a_01302.pdf by guest on 29 September 2021

Table 3: The Mean Accuracy (%) on Four Benchmark Data Sets to Verify the Effect of the 3D-Attention Mechanism.

| (a) Five-Way on Mini-ImageNet | | | (b) Five-Way on Cub Birds | | |
|---|---|---|---|---|---|
| Model | *AFG-3D* | *AFG* | Model | *AFG-3D* | *AFG* |
| 1-shot | 52.70 | 51.03 **(↓ 1.67)** | 1-shot | 39.74 | 50.02 **(↑ 10.28)** |
| 5-shot | 68.30 | 69.14 **(↑ 0.84)** | 5-shot | 41.13 | 54.41 **(↑ 13.28)** |
| (c) Five-Way on Stanford Dogs | | | (d) Five-Way on Stanford Cars | | |
| Model | *AFG-3D* | *AFG* | Model | *AFG-3D* | *AFG* |
| 1-shot | 25.80 | 39.40 **(↑ 13.6)** | 1-shot | 37.20 | 42.23 **(↑ 5.03)** |
| 5-shot | 51.29 | 59.61 **(↑ 8.32)** | 5-shot | 51.30 | 62.90 **(↑ 11.6)** |

Note: The difference between the two models is marked in bold.

1.33% higher than state-of-the-art methods on each of the three fine-grained data sets, respectively. For the 5-shot setting, our model is improved by 11.42% on Stanford Dogs and 9.97% on Stanford Cars.

- The effect of SiameseNet is generally lower than that of MatchingNet, ProtoNet, and FSFG. This reflects that the training strategy *episode* can better generalize to novel fine-grained categories. The proposed AFG adopts the same strategy to get better results than the models just noted.
- The effect on Cub Birds is not ideal due to the discriminative capacity of the bilinear CNN features in FSFG (Wei et al., 2019). We also compare the three fine-grained data sets and find that Cub Birds is more complex than the other two. Many target *bird* objects make up only a small portion of the entire image, and the movements of birds in the air are diverse. The data sets are analyzed in detail in section 4.3.3.

*4.3.3 Visualization of Reconstruction Module.* Comparing the results of Tables 2 and 3, we can see that the accuracy of AFG on Mini-ImageNet for generic few-shot learning is almost higher than that of the three fine-grained data sets. Especially in the 5-shot setting, the accuracy is 14.7%, 9.5%, and 6.2% higher on Mini-ImageNet than that on Cub Birds, Stanford Dogs, and Stanford Cars, respectively.

The visualization of the reconstructed part images $R_t^{img}$ are shown in Figure 5. $X$ and $\tilde{X}$ represent the original input image and the reconstructed image, respectively. $R_t^{img}$ is acquired by adopting a dot production of $R_t^{att}$ and the original image $X$. The visualization in Figure 5 shows the four prototypes learned by 3D-Attention in the AFG model. That is, the proposed image reconstruction module encourages the network to generate local original images that are as realistic as possible. In other words,

Downloaded from http://direct.mit.edu/neco/article-pdf/32/9/1664/1865020/neco_a_01302.pdf by guest on 29 September 2021

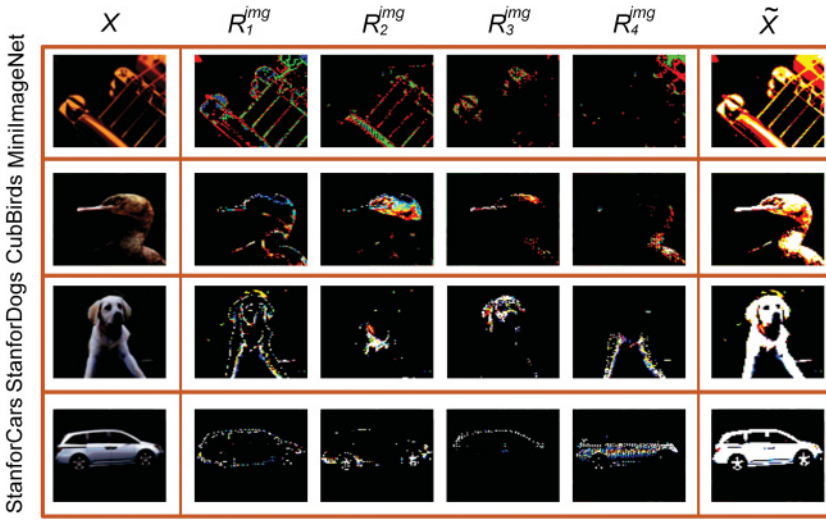


Figure 5: Visualization of attention results of the reconstructed part images in the MiniImageNet and three fine-grained data sets.

minimizing the difference between the reconstructed image and the original input image can be used indirectly as a supervised information for feature learning, without extra part annotations.

**4.4 Ablation Studies.** This section verifies the influence of the feature learning and reconstruction modules in the proposed AFG in detail the balancing hyperparameter $\lambda$ is also substantiated in a wider range.

*4.4.1 Ablation Study of 3D-Attention..* To further verify the effectiveness of the 3D-Attention mechanism, we perform an ablation study by developing a variant of AFG without 3D-Attention, termed *AFG-3D*, which denotes that we focus on all feature descriptors to represent the global image. Thus, the reconstruction module has a slight change, and we reconstruct all 100 descriptors instead of the previous attention-based local descriptors. The results on four benchmark data sets are reported in Table 3.

It shows that in most cases, the 3D-Attention mechanism helps to improve the effect of AFG, especially over three fine-grained data sets. However, it does not improve the performance of AFG in a 1-shot setting on Mini-ImageNet because almost all images in this data sets consist of simple, obvious objects. AFG misses other global information when learning local prototypes, which results in a slight difference with AFG-3D. In other words, the 3D-Attention mechanism is better suited to fine-grained image classification.

Downloaded from http://direct.mit.edu/neco/article-pdf/32/9/1664/1865020/neco_a_01302.pdf by guest on 29 September 2021

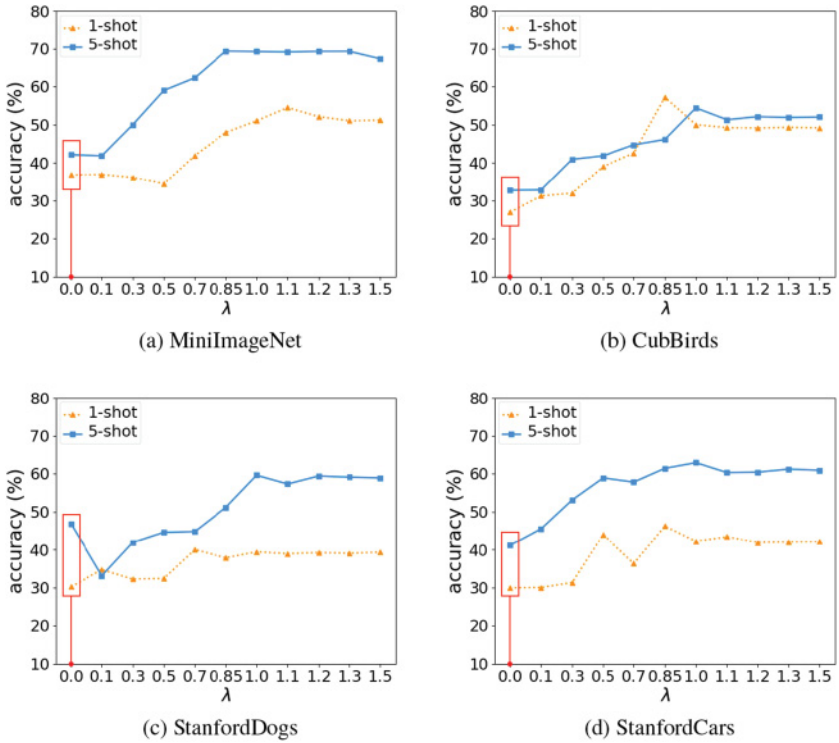


Figure 6: Line chart analyzing hyperparameter $\lambda$ on MiniImageNet and three fine-grained data sets. When $\lambda = 0$, the proposed AFG outputs the results without reconstruction constraint (red signs).

*4.4.2 Ablation Study of Reconstruction Constraint.* As far as we know, few researchers who study few-shot methods have used the idea of reconstruction to supervise the features learning. In our experiments, $\lambda$ plays an important role in balancing the effect of feature learning and image reconstruction. In order to verify the effectiveness of the reconstruction module, we set $\lambda = 0$ to denote the original model without it. From Figure 6 (red signs), we can see that when AFG loses the constraint of reconstruction, the performance on four data sets is not ideal.

We also analyze multiple values of $\lambda$ in [0, 1.5]. Figure 6 shows the performance of different values of $\lambda$ on four data sets:

- For Mini-ImageNet, when the value of $\lambda$ is approximately equal to 1.0, the performance of 1-shot and 5-shot learning does not cause substantial floating.

Downloaded from http://direct.mit.edu/neco/article-pdf/32/9/1664/1865020/neco_a_01302.pdf by guest on 29 September 2021

- For the other three fine-grained data sets, when the value of $\lambda$ ranges from 1.1 to 1.15, the performance of our model is almost better than that in other $\lambda$ values.
- A comparison of shows that the experimental results of the generic and fine-grained data sets, the latter requires stronger reconstruction constraints to learn more discriminative features.

## 5 Conclusion

In this letter, we propose a novel attention-based model AFG for few-shot learning, which takes different classes as input with one or a few examples and then learns an effective classifier to address more challenging problems in fine-grained setting. The proposed model does not need any extra annotations of the parts and can be trained in an end-to-end manner. The major contributions of our model are a feature learning module and an image reconstruction module. The 3D-Attention mechanism in the feature learning module promotes learning more discriminative prototypes to better represent class distribution. The image reconstruction module is constrained by a designed loss function to act as auxiliary supervised information, so that each local part does not require extra annotations.

In the future, we will be conducting research in two directions: integrating semisupervised or unsupervised methods into a few-shot fine-grained classification tasks and transferring the proposed model to the field of education with more abstract semantics.

## Acknowledgments

This work was supported by the National Key Research and Development Program of China (2018YFB1004500), the National Natural Science Foundation of China (61532004, 61532015, 61672418, and 61672419), the Innovative Research Group of the National Natural Science Foundation of China (61721002), the Innovation Research Team of the Ministry of Education (IRT_17R86), and the Project of China Knowledge Center for Engineering Science and Technology.

## References

Alfassy, A., Karlinsky, L., Aides, A., Shtok, J., Harary, S., Feris, R., & Bronstein, A. M. (2019). Lasso: Label-set operations networks for multi-label few-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 6548–6557). Piscataway, NJ: IEEE.

Andrychowicz, M., Denil, M., Gomez, S., Hoffman, M. W., Pfau, D., Schaul, T., & De Freitas, N. (2016). Learning to learn by gradient descent by gradient descent. In D. D. Lee, M. S. Sugiyama, U. V. Luxburg, L. Guyon, & R. Garnett (Eds.), *Advances in neural information processing systems*, *29* (pp. 3981–3989). Red Hook, NY: Curran.

Downloaded from http://direct.mit.edu/neco/article-pdf/32/9/1664/1865020/neco_a_01302.pdf by guest on 29 September 2021

Bertinetto, L., Henriques, J., Torr, P., & Vedaldi, A. (2019). Meta-learning with differentiable closed-form solvers. In *Proceedings of the International Conference on Learning Representations*.

Branson, S., Van Horn, G., Belongie, S., Perona, P., & Tech, C. (2014). Bird species categorization using pose normalized deep convolutional nets. *Image*, *70*.

Cao, J., Wang, B., & Brown, D. (2016). Similarity based leaf image retrieval using multiscale r-angle description. *Information Sciences*, *374*, 51–64.

Edraki, M., & Qi, G.-J. (2018). Generalized loss-sensitive adversarial learning with manifold margins. In *Proceedings of the European Conference on Computer Vision* (pp. 87–102). Berlin: Springer.

Finn, C., Abbeel, P., & Levine, S. (2017). Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, *70* (pp. 1126–1135).

Fu, J., Zheng, H., & Mei, T. (2017). Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 4438–4446). Piscataway, NJ: IEEE.

Gao, Y., Beijbom, O., Zhang, N., & Darrell, T. (2016). Compact bilinear pooling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 317–326). Piscataway, NJ: IEEE.

Hu, H., Wang, R., Yang, X., & Nie, F. (2019). Scalable and flexible unsupervised feature selection. *Neural Computation*, *31*(3), 517–537.

Huang, S., Xu, Z., Tao, D., & Zhang, Y. (2016). Part-stacked CNN for fine-grained visual categorization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1173–1182). Piscataway, NJ: IEEE.

Khosla, A., Jayadevaprakash, N., Yao, B., & Li, F.-F. (2011). Novel dataset for fine-grained image categorization: Stanford dogs. In *Proceedings of the CVPR Workshop on Fine-Grained Visual Categorization* (vol. *2*). Piscataway, NJ: IEEE.

Koch, G., Zemel, R., & Salakhutdinov, R. (2015). Siamese neural networks for one-shot image recognition. In *Proceedings of the International Conference on Machine Learning* (vol. *2*).

Kozerawski, J., & Turk, M. (2018). CLEAR: Cumulative learning for one-shot one-class image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 3446–3455). Piscataway, NJ: IEEE.

Krause, J., Stark, M., Deng, J., & Fei-Fei, L. (2013). 3D object representations for fine-grained categorization. In *Proceedings of the IEEE International Conference on Computer Vision Workshops* (pp. 554–561). Piscataway, NJ: IEEE.

Li, W., Wang, L., Xu, J., Huo, J., Gao, Y., & Luo, J. (2019). Revisiting local descriptor based image-to-class measure for few-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 7260–7268). Piscataway, NJ: IEEE.

Li, W., Xu, J., Huo, J., Wang, L., Gao, Y., & Luo, J. (2019). Distribution consistency based covariance metric networks for few-shot learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, *33* (pp. 8642–8649). Cambridge, MA: MIT Press

Li, Z., Zhou, F., Chen, F., & Li, H. (2017). Meta-SGD: *Learning to learn quickly for few shot learning*. arXiv:1707.09835

Downloaded from http://direct.mit.edu/neco/article-pdf/32/9/1664/1865020/neco_a_01302.pdf by guest on 29 September 2021

Lin, D., Shen, X., Lu, C., & Jia, J. (2015). Deep LAC: Deep localization, alignment and classification for fine-grained recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1666–1674). Piscataway, NJ: IEEE.

Lin, T.-Y., RoyChowdhury, A., & Maji, S. (2015). Bilinear CNN models for fine-grained visual recognition. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 1449–1457). Piscataway, NJ: IEEE.

Miller, D., Wang, Y., & Kesidis, G. (2019). When not to classify: Anomaly detection of attacks (ADA) on DNN classifiers at test time. *Neural Computation*, *31*(8), 1624–1670.

Qi, H., Brown, M., & Lowe, D. G. (2018). Low-shot learning with imprinted weights. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 5822–5830). Piscataway, NJ: IEEE.

Ravi, S. & Larochelle, H. (2017). Optimization as a model for few-shot learning. In *Proceedings of the International Conference on Learning Representations*.

Satorras, V. G., & Estrach, J. B. (2018). Few-shot learning with graph neural networks. In *Proceedings of the International Conference on Learning Representations.*

Schwartz, E., Karlinsky, L., Shtok, J., Harary, S., Marder, M., Kumar, A., & Bronstein, A., (2018). Delta-encoder: An effective sample synthesis method for few-shot object recognition. In S. Bengio, H. Wallach, H. Larichelle, K. Grauman, N. Cesa-Bianchi, & R. Garnett (Eds.), *Advances in neural information processing systems*, *31* (pp. 2845–2855). Red Hook, NY: Curran.

Shyam, P., Gupta, S., & Dukkipati, A. (2017). Attentive recurrent comparators. In *Proceedings of the 34th International Conference on Machine Learning*, *70* (pp. 3173–3181).

Snell, J., Swersky, K., & Zemel, R. (2017). Prototypical networks for few-shot learning. In L. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems*, *30* (pp. 4077–4087). Red Hook, NY: Curran.

Sung, F., Yang, Y., Zhang, L., Xiang, T., Torr, P. H., & Hospedales, T. M. (2018). Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1199–1208). Piscataway, NJ: IEEE.

Tang, X.-L., Ma, W.-C., Kong, D.-S., & Li, W. (2019). Semisupervised deep stacking network with adaptive learning rate strategy for motor imagery EEG recognition. *Neural Computation*, *31*(5), 919–942.

Thrun, S. (1998). Lifelong learning algorithms. In S. Thrun& L. Pratt (Eds.), *Learning to learn* (pp. 181–209). Berlin: Springer.

Vilalta, R., & Drissi, Y. (2002). A perspective view and survey of meta-learning. *Artificial Intelligence Review*, *18*(2), 77–95.

Vinyals, O., Blundell, C., Lillicrap, T., Kavukcuogly, K., & Wierstra, D. (2016). Matching networks for one shot learning. In D. D. Lee, M. S. Sugiyama, U. V. Luxburg, L. Guyon, & R. Garnett (Eds.),*Advances in neural information processing systems*, *29* (pp. 3630–3638. Red Hook, NY: Curran.

Vyškovskỳ, R., Schwarz, D., & Kašpárek, T. (2019). Brain morphometry methods for feature extraction in random subspace ensemble neural network classification of first-episode schizophrenia. *Neural Computation*, *31*(5), 897–918.

Downloaded from http://direct.mit.edu/neco/article-pdf/32/9/1664/1865020/neco_a_01302.pdf by guest on 29 September 2021

Wah, C., Branson, S., Welinder, P., Perona, P., & Belongie, S. (2011). *The CalTech-UCSD birds-200-2011 dataset* (Technical Report CNS-TR-2011-001). Pasadena, CA: CalTech.

Wei, X., Luo, J., Wu, J., & Zhou, Z. (2017). Selective convolutional descriptor aggregation for fine-grained image retrieval. *IEEE Transactions on Image Processing*, *26*(6), 2868–2881.

Wei, X.-S., Wang, P., Liu, L., Shen, C., & Wu, J. (2019). Piecewise classifier mappings: Learning fine-grained learners for novel categories with few examples. *IEEE Transactions on Image Processing*, *28*(12), 6116–6125.

Wei, X.-S., Xie, C.-W., Wu, J., & Shen, C. (2018). Mask-CNN: Localizing parts and selecting descriptors for fine-grained bird species categorization. *Pattern Recognition*, *76*, 704–714.

Wertheimer, D., & Hariharan, B. (2019). Few-shot learning with localization in realistic settings. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 6558–6567). Piscataway, NJ: IEEE.

Xie, L., Wang, J., Zhang, B., & Tian, Q. (2015). Fine-grained image search. *IEEE Transactions on Multimedia*, *17*(5), 636–647.

Zhang, H., Xu, T., Elhoseiny, M., Huang, X., Zhang, S., Elgammal, A., & Metaxas, D. (2016). SPDA-CNN: Unifying semantic part detection and abstraction for fine-grained recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1143–1152). Piscataway, NJ: IEEE.

Zhang, L., Chang, X., Liu, J., Luo, M., Wang, S., Ge, Z., & Hauptmann, A. G. (2020). *Zstad: Zero-shot temporal activity detection*. arXiv:2003.05583

Zhang, L., Liu, J., Luo, M., Chang, X., & Zheng, Q. (2018). Deep semisupervised zero-shot learning with maximum mean discrepancy. *Neural Computation*, *30*, 1–22.

Zhang, L., Liu, J., Luo, M., Chang, X., Zheng, Q., & Hauptmann, A. G. (2019). Scheduled sampling for one-shot learning via matching network. *Pattern Recognition*, *96*, 106962.

Zhang, N., Donahue, J., Girshick, R., & Darrell, T. (2014). Part-based r-CNNS for fine-grained category detection. In *Proceedings of the European Conference on Computer Vision* (pp. 834–849). Berlin: Springer.

Zhang, Y., Wei, X.-S., Wu, J., Cai, J., Lu, J., Nguyen, V.-A., & Do, M. N. (2016). Weakly supervised fine-grained categorization with part-based image representation. *IEEE Transactions on Image Processing*, *25*(4), 1713–1725.

Zheng, H., Fu, J., Mei, T., & Luo, J. (2017). Learning multi-attention convolutional neural network for fine-grained image recognition. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 5209–5217). Piscataway, NJ: IEEE.

Zheng, H., Fu, J., Zha, Z.-J., & Luo, J. (2019). Looking for the devil in the details: Learning trilinear attention sampling network for fine-grained image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 5012–5021). Piscataway, NJ: IEEE.

Received February 4, 2020; accepted April 23, 2020.

Downloaded from http://direct.mit.edu/neco/article-pdf/32/9/1664/1865020/neco_a_01302.pdf by guest on 29 September 2021