Fine-Grained Visual Categorization using Meta-Learning Optimization with Sample Selection of Auxiliary Data

Yabin Zhang[0000−0002−2179−3973], Hui Tang[0000−0001−8856−9127], and Kui Jia*

School of Electronic and Information Engineering,
South China University of Technology, Guangzhou, China
{zhang.yabin,eehuitang}@mail.scut.edu.cn, kuijia@scut.edu.cn

Abstract. Fine-grained visual categorization (FGVC) is challenging due in part to the fact that it is often difficult to acquire an enough number of training samples. To employ large models for FGVC without suffering from overfitting, existing methods usually adopt a strategy of pre-training the models using a rich set of auxiliary data, followed by fine-tuning on the target FGVC task. However, the objective of pre-training does not take the target task into account, and consequently such obtained models are suboptimal for fine-tuning. To address this issue, we propose in this paper a new deep FGVC model termed MetaFGNet. Training of MetaFGNet is based on a novel regularized meta-learning objective, which aims to guide the learning of network parameters so that they are optimal for adapting to the target FGVC task. Based on MetaFGNet, we also propose a simple yet effective scheme for selecting more useful samples from the auxiliary data. Experiments on benchmark FGVC datasets show the efficacy of our proposed method.

Keywords: Fine-grained visual categorization · Meta-learning · Sample selection

1 Introduction

Fine-grained visual categorization (FGVC) aims to classify images of subordinate object categories that belong to a same entry-level category, e.g., different species of birds [27, 26, 3] or dogs [9]. The visual distinction between different subordinate categories is often subtle and regional, and such nuance is further obscured by variations caused by arbitrary poses, viewpoint change, and/or occlusion. Subordinate categories are leaf nodes of a taxonomic tree, whose samples are often difficult to collect. Annotating such samples also requires professional expertise, resulting in very few training samples per category in existing FGVC datasets [26, 9]. FGVC thus bears problem characteristics of few-shot learning.

Most of existing FGVC methods spend efforts on mining global and/or regional discriminative information from training data themselves. For example,

* Corresponding author

state-of-the-art methods learn to identify discriminative parts from images of fine-grained categories either in a supervised [33, 32] or in a weakly supervised manner [29, 6, 36, 35, 17, 37]. However, such methods are approaching a fundamental limit since only very few training samples are available for each category. In order to break the limit, possible solutions include identifying auxiliary data that are more useful for (e.g., more related to) the FGVC task of interest, and also better leveraging these auxiliary data. These solutions fall in the realm of domain adaptation or transfer learning [15].

A standard way of applying transfer learning to FGVC is to fine-tune on the target dataset a model that has been pre-trained on a rich set of auxiliary data (e.g., the ImageNet [22]). Such a pre-trained model learns to encode (generic) semantic knowledge from the auxiliary data, and the combined strategy of pre-training followed by fine-tuning alleviates the issue of overfitting. However, the objective of pre-training does not take the target FGVC task of interest into account, and consequently such obtained models are suboptimal for transfer.

Inspired by recent meta-learning methods [5, 25, 19] for few-shot learning, we propose in this paper a new deep learning method for fine-grained classification. Our proposed method is based on a novel *regularized meta-learning objective* for training a deep network: the regularizer aims to learn network parameters such that they can encode generic or semantically related knowledge from auxiliary data; the meta-learning objective is designed to guide the process of learning, so that the learned network parameters are optimal for adapting to the target FGVC task. We term our proposed FGVC method as MetaFGNet for its use of the meta-learning objective. Figure 1 gives an illustration. Our method can effectively alleviate the issue of overfitting, as explained in Section 3.

An important issue to achieve good transfer learning is that data in source and target tasks should share similar feature distributions [15]. If this is not the case, transfer learning methods usually learn feature mappings to alleviate this issue. Alternatively, one may directly identify source data/tasks that are more related to the target one. In this work, we take the later approach and propose a simple yet very effective scheme to select more useful samples from the auxiliary data. Our scheme is naturally admitted by MetaFGNet, and only requires a forward computation through a trained MetaFGNet for each auxiliary sample, which contrasts with a recent computationally expensive scheme used in [7]. In this work, we investigate ImageNet [22], a subset of ImageNet and a subset of L-Bird [11] as the sets of auxiliary data. For the L-Bird subset, for example, our scheme can successfully remove noisy, semantically irrelevant images. Experiments on the benchmark FGVC datasets of CUB-200-2011 [26] and Stanford Dogs [9] show the efficacy of our proposed MetaFGNet with sample selection of auxiliary data. Our contributions are summarized as follows.

- We propose a new deep learning model, termed MetaFGNet, for fine-grained classification. Training of MetaFGNet is based on a novel *regularized meta-learning objective*, which aims to guide the learning of network parameters so that they are optimal for adapting to the target FGVC task (cf. Section 3).

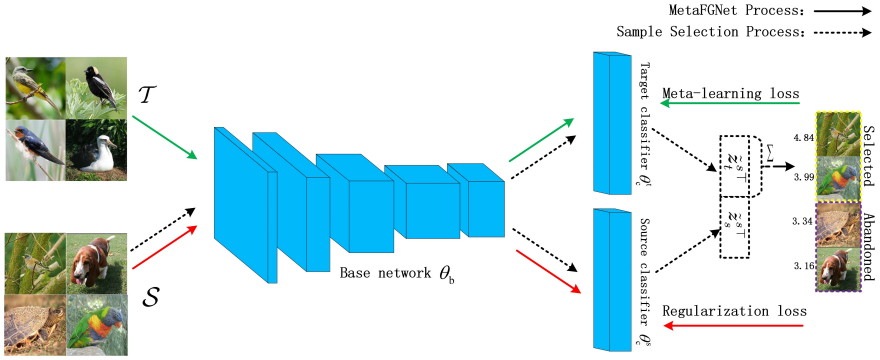


Fig. 1. Illustrations of MetaFGNet with regularized meta-learning objective (solid line) and the process of sample selection from auxiliary data (dashed line).

- Our proposed MetaFGNet admits a natural scheme to perform sample selection from auxiliary data. Given a trained MetaFGNet, the proposed scheme only requires a forward computation through the network to produce a score for each auxiliary sample (cf. Section 4). Such scores can be used to effectively select semantically related auxiliary samples (or remove noisy, semantically irrelevant ones).
- We present intensive comparative studies on different ways of leveraging auxiliary data for the target FGVC task. Experiments on the benchmark CUB-200-2011 [26] and Stanford Dogs [9] datasets also show the efficacy of our proposed method. In particular, our result on Stanford Dogs is better than all existing ones with a large margin. Based on a better auxiliary dataset, our result on CUB-200-2011 is better than those of all existing methods even when they use ground-truth part annotations (cf. Section 5).

2 Related works

In this section, we briefly review recent fine-grained classification methods, in particular those aiming for better leveraging auxiliary data, and also meta learning methods for the related problem of few-shot learning. We present brief summaries of these methods and discuss their relations and differences with our proposed one.

Fine-grained visual categorization State-of-the-art FGVC methods usually follow the pipeline that first discovers discriminative local parts from images of fine-grained categories, and then utilizes the discovered parts for classification. For example, Lam *et al.* [13] search for discriminative parts by iteratively evaluating and generating bounding box proposals with or without the supervision of ground-truth part annotations. Based on off-the-shelf object proposals [24], part detectors are learned in [35] by clustering subregions of object proposals. In

[6], a hierarchical three-level region attention mechanism is proposed that is able to attend to discriminative regions, where region discrimination is measured by classification probability. Multiple part attention maps are generated by clustering and weighting the spatially-correlated channels of the convolutional feature maps in [37].

There exist FGVC methods [20, 2, 18, 4, 14, 28] that process a whole image instead of local parts, yet their results are generally worse than those of part based methods. Another line of methods push the state of the art by identifying and leveraging auxiliary data beyond the ImageNet. In particular, the method of [11] sets an astonishing baseline on CUB-200-2011 simply by pre-training a standard deep model using a huge auxiliary set of web images that are obtained by using subordinate categories of bird as search keywords; note that such obtained auxiliary images are quite noisy in terms of their category labels. Xie *et al.* [30] propose to augment the fine-grained data with a large number of auxiliary images labeled by hyper-classes; these hyper-classes are some attributes that can be annotated more easily than the fine-grained category labels, so that a large number of images labeled with attributes can be easily acquired; by joint training the model for hyper-class classification and FGVC, the performance of FGVC gets improved. Instead of searching for more semantically relevant auxiliary data from the Internet, Ge and Yu [7] propose to refine the ImageNet images by comparing them with those in the training set of the target FGVC task, using low-level features (e.g., the Gabor filter responses); such a refined ImageNet is then used to jointly train a model with training images of the FGVC task.

All the above methods use auxiliary data either to pre-train a model, or to jointly train a model with training images of the target FGVC task. In contrast, our proposed MetaFGNet uses a regularized meta-learning objective that can make full use of the auxiliary data, while at the same time making the obtained model optimal for a further adaptation to the target FGVC task. We also compare our training objective with that of joint training technically in Section 3 and empirically in Section 5.

Few-shot learning via meta learning Meta learning aims to learn experience from history and adapt to new tasks with the help of history knowledge. Few-shot learning is one of its applications. [10] trains a siamese neural network for the task of verification, which is to identify whether input pairs belong to the same class; once the verification model is trained, it can be used for few- or one-shot learning by calculating the similarity between the test image and the labelled images. [25] realizes few-shot learning with a neural network which is augmented with external memories; it uses two embeddings to map the images to feature space and the classification is obtained by measuring the cosine distance in the feature space; the embedding of the test images can be modified by the whole support set through a LSTM attention module, which makes the model utilize the support set more reasonably and effectively. In [19], SGD is replaced by a meta-LSTM that can learn an update rule for training networks. Finn *et al.* [5] propose a meta learning method termed MAML, which trains a meta model in a multi-task fashion. Different from the problem setting of MAML,

which learns meta models from a set of training tasks that are independent of the target task, our proposed MetaFGNet involves directly the target task into the training objective.

3 The proposed MetaFGNet

For a target FGVC of interest, suppose we have training data $\mathcal{T} = \{(\mathbf{x}_i^t, \mathbf{y}_i^t)\}_{i=1}^{|\mathcal{T}|}$, where each pair of $\mathbf{x}_i^t$ and $\mathbf{y}_i^t$ represents an input image and its one-hot vector representation of class label. We also assume that a set of auxiliary data (e.g., the ImageNet) is available that contains images different from (but possibly semantically related to) the target data $\mathcal{T}$. Denote the auxiliary data as $\mathcal{S} = \{(\mathbf{x}_i^s, \mathbf{y}_i^s)\}_{i=1}^{|\mathcal{S}|}$. As illustrated in Figure 1, our proposed MetaFGNet is based on a deep neural network consisting of two parallel classifiers of fully-connected (FC) layers that share a common base network. The two classifiers are respectively used for $\mathcal{T}$ and $\mathcal{S}$. We correspondingly denote parameters of the two classifiers as θ_c^s and θ_c^t, and denote those of the base network collectively as θ_b, which contains parameters of layer weights and bias. For ease of subsequent notations, we also denote the parameters of target and source model as $\theta^t = (\theta_b, \theta_c^t)$ and $\theta^s = (\theta_b, \theta_c^s)$ respectively.

In machine learning, $\mathcal{T}$ is usually sampled i.i.d. from an underlying (unknown) distribution $\mathcal{D}$. To learn a deep model θ^t, one may choose an appropriate loss function $L(\mathbf{x}^t, \mathbf{y}^t; \theta^t)$, and minimize the following expected loss over $\mathcal{D}$

$$\min_{\theta^t} \mathbf{E}_{(\mathbf{x}^t, \mathbf{y}^t) \sim \mathcal{D}} \left[L(\mathbf{x}^t, \mathbf{y}^t; \theta^t) \right]. \tag{1}$$

In practice, however, minimizing the above objective is infeasible since the underlying distribution $\mathcal{D}$ is unknown. As an alternative, one chooses to minimize the following empirical loss to learn θ^t

$$\min_{\theta^t} \frac{1}{|\mathcal{T}|} L\left(\mathcal{T}; \theta^t\right) = \frac{1}{|\mathcal{T}|} \sum_{i=1}^{|\mathcal{T}|} L\left(\mathbf{x}_i^t, \mathbf{y}_i^t; \theta^t\right). \tag{2}$$

As discussed in Section 1, fine-grained classification bears problem characteristics of few-shot learning, and its training set $\mathcal{T}$ is usually too small to well represent the underlying distribution $\mathcal{D}$. Thus directly minimizing the empirical loss (2) causes severe overfitting. In the literature of fine-grained classification, this issue is usually addressed by pre-training the model θ^t using an auxiliary set of data $\mathcal{S}$ (e.g., the ImageNet), and then fine-tuning it using $\mathcal{T}$. Note that this strategy alleviates overfitting in two aspects: (1) pre-training gives the model a good initialization that has learned (generic) semantic knowledge from $\mathcal{S}$; and (2) fine-tuning itself reduces overfitting via early stop of training. In other words, one may understand the strategy of fine-tuning as imposing *implicit regularization* on the learning of θ^t. Alternatively, one may apply *explicit regularization* to (2), resulting in the following general form of regularized loss minimization

$$\min_{\theta^t} \frac{1}{|\mathcal{T}|} L\left(\mathcal{T}; \theta^t\right) + R(\theta^t). \tag{3}$$

The auxiliary set $\mathcal{S}$ can be used as an instantiation of the regularizer, giving rise to the following *joint training method*

$$\min_{\theta_b,\theta_c^t,\theta_c^s} \frac{1}{|\mathcal{T}|} L\left(\mathcal{T};\theta_b,\theta_c^t\right) + \frac{1}{|\mathcal{S}|} R\left(\mathcal{S};\theta_b,\theta_c^s\right), \tag{4}$$

where regularization is only imposed on parameters θ_b of the base network. By leveraging $\mathcal{S}$, the joint training method (4) could be advantageous over fine-tuning since network training has a chance to converge to a more mature, but not overfitted, solution. Based on a similar deep architecture as in Figure 1, the joint training method (4) is used in a recent work of fine-grained classification [7]. The choice of the auxiliary set $\mathcal{S}$ also matters in (4). Established knowledge from the literature of transfer learning [15] suggests that $\mathcal{S}$ should ideally have similar distribution of feature statistics as that of $\mathcal{T}$, suggesting that a refinement of $\mathcal{S}$ could be useful for better regularization.

3.1 A meta-learning objective for MetaFGNet

Inspired by recent meta learning methods [5, 25, 19] that learn a meta model from a set of training few-shot learning tasks, we propose in this paper a meta-learning objective for the target fine-grained classification task $\mathcal{T}$. Instead of using the loss $L(\mathcal{T};\theta^t)$ directly as in (3), the meta-learning objective is to *guide the optimization of* θ^t so that the obtained θ^t can fast adapt to the target task via a second process of fine-tuning. Suppose the fine-tuning process achieves

$$\theta^t \leftarrow \theta^t + \triangle(\theta^t), \tag{5}$$

where $\triangle(\theta^t)$ denotes the amount of parameter update. The problem nature of few-shot learning suggests that fine-tuning should be a fast process: a small number of (stochastic) gradient descent steps may be enough to learn effectively from $\mathcal{T}$, and taking too many steps may result in overfitting. One-step gradient descent can be written as

$$\triangle(\theta^t) = -\eta \frac{1}{|\mathcal{T}|} \nabla_{\theta^t} L(\mathcal{T};\theta^t), \tag{6}$$

where η denotes the step size. Based on (6), we write our proposed *regularized meta-learning objective* for fine-grained classification as

$$\min_{\theta_b,\theta_c^t,\theta_c^s} \frac{1}{|\mathcal{T}|} L\left(\mathcal{T};\theta^t - \eta \frac{1}{|\mathcal{T}|} \nabla_{\theta^t} L(\mathcal{T};\theta^t)\right) + \frac{1}{|\mathcal{S}|} R\left(\mathcal{S};\theta^s\right). \tag{7}$$

Our proposed meta-learning objective can also be explained from the perspective of reducing effective model capacity, and can thus achieve additional alleviation of overfitting apart from the effect of the regularizer $R(\mathcal{S};\theta_b,\theta_c^s)$, in which the regularization is achieved by base parameters updating from auxiliary data.

Remarks Both our proposed MetaFGNet and the meta-learning methods [5, 19] contain loss terms of meta-learning, which guide the trained model to be able

to fast adapt to a target task. We note that our method is for a problem setting different from those of [5, 19], and consequently is the objective (7): they learn meta models from a set of training tasks and subsequently use the learned meta model for a new target task; here training set $\mathcal{T}$ of the target task is directly involved in the main learning objective.

3.2 Training algorithm

Solving the proposed objective (7) via stochastic gradient descent (SGD) involves computing gradient of a gradient for the first term, which can be derived as

$$\nabla_{\theta^{t'}} \frac{1}{|\mathcal{T}_j|} L\left(\mathcal{T}_j; \theta^{t'}\right) \left[\mathbf{I} - \eta \frac{1}{|\mathcal{T}_i|} \left(\frac{\partial^2 L(\mathcal{T}_i; \theta^t)}{\partial (\theta^t)^2}\right)\right], \tag{8}$$

where $\mathcal{T}_i$ and $\mathcal{T}_j$ denote mini-batches of $\mathcal{T}$, and $\theta^{t'} = \theta^t - \eta \frac{1}{|\mathcal{T}_i|} \nabla_{\theta^t} L(\mathcal{T}_i; \theta^t)$. Hessian matrix is involved in (8), computation of which is supported by modern deep learning libraries [1, 16]. In this work, we adopt the Pytorch [16] to implement (8), whose empirical computation time is about 0.64s per iteration (batchsize = 32) when training MetaFGNet on a GeForce GTX 1080 Ti GPU. Training of MetaFGNet is given in Algorithm 1.

Algorithm 1 Training algorithm for MetaFGNet

Require: $\mathcal{T}$: target train data; $\mathcal{S}$: auxiliary train data
Require: η, α: hyperparameters of step size
1: initialize θ_b, θ_c^t, θ_c^s
2: **while** not done **do**
3: Sample mini-batches $\mathcal{T}_i$, $\mathcal{S}_i$ from $\mathcal{T}$, $\mathcal{S}$
4: Evaluate:
$[\triangle(\theta_b; \mathcal{S}_i), \triangle(\theta_c^s; \mathcal{S}_i)] = \frac{1}{|\mathcal{S}_i|} \nabla_{\theta^s} R\left(\mathcal{S}_i; \theta^s\right)$
$[\triangle(\theta_b; \mathcal{T}_i), \triangle(\theta_c^t; \mathcal{T}_i)] = \frac{1}{|\mathcal{T}_i|} \nabla_{\theta^t} L(\mathcal{T}_i; \theta^t)$
5: Compute adapted parameters with SGD:
$\theta^{t'} = \theta^t - \eta \frac{1}{|\mathcal{T}_i|} \nabla_{\theta^t} L(\mathcal{T}_i; \theta^t)$
6: Sample another mini-batch $\mathcal{T}_j$ from $\mathcal{T}$
7: Evaluate:
$[\triangle(\theta_b; \mathcal{T}_j), \triangle(\theta_c^t; \mathcal{T}_j)] = \nabla_{\theta^{t'}} \frac{1}{|\mathcal{T}_j|} L\left(\mathcal{T}_j; \theta^{t'}\right) \left[\mathbf{I} - \eta \frac{1}{|\mathcal{T}_i|} \left(\frac{\partial^2 L(\mathcal{T}_i; \theta^t)}{\partial (\theta^t)^2}\right)\right]$
8: Update:
$\theta_b \leftarrow \theta_b - \alpha[\triangle(\theta_b; \mathcal{S}_i) + \triangle(\theta_b; \mathcal{T}_j)]$
$\theta_c^t \leftarrow \theta_c^t - \alpha \triangle(\theta_c^t; \mathcal{T}_j)$
$\theta_c^s \leftarrow \theta_c^s - \alpha \triangle(\theta_c^s; \mathcal{S}_i)$
9: **end while**

4 Sample selection of auxiliary data using the proposed MetaFGNet

Established knowledge from domain adaptation suggests that the auxiliary set $\mathcal{S}$ should ideally have a similar distribution of feature statistics as that of the target set $\mathcal{T}$. This can be achieved either via transfer learning [15], or via selecting/refining samples of $\mathcal{S}$. In this work, we take the second approach and propose a simple sample selection scheme that is naturally supported by our proposed MetaFGNet (and in fact by any deep models with a two-head architecture as in Figure 1).

Given a trained MetaFGNet, for each auxiliary sample $\mathbf{x}^s$ from $\mathcal{S}$, we compute through the network to get two *prediction vectors* $\mathbf{z}_s^s$ and $\mathbf{z}_t^s$, which are respectively the output vectors of the two classifiers (before the softmax operation) for the source and target tasks. Length of $\mathbf{z}_s^s$ (or $\mathbf{z}_t^s$) is essentially equal to category number of the source task (or that of the target task). To achieve sample selection from the auxiliary set $\mathcal{S}$, we take the approach of assigning a score to each $\mathbf{x}^s$ and then ranking scores of all auxiliary samples. The score of $\mathbf{x}^s$ is computed as follows: we first set negative values in $\mathbf{z}_s^s$ and $\mathbf{z}_t^s$ as zero; we then concatenate the resulting vectors and apply L2 normalization, producing $\tilde{\mathbf{z}}^s = [\tilde{\mathbf{z}}_s^{s\top}, \tilde{\mathbf{z}}_t^{s\top}]^\top$; we finally compute the score for $\mathbf{x}^s$ as

$$O^s = \tilde{\mathbf{z}}_t^{s\top} \cdot \mathbf{1}, \tag{9}$$

where $\mathbf{1}$ represents a vector with all entry values of 1. A specified ratio of top samples can be selected from $\mathcal{S}$ and form a new set of auxiliary data. Rationale of the above scheme lies in that auxiliary samples that are more semantically related to the target task would have higher responses in the target classifier, and consequently would have higher values in $\tilde{\mathbf{z}}_t^{s\top}$.

Experiments in Section 5 show that such a sample selection scheme is effective to select images that are semantically more related to the target task and improve performance of fine-grained classification. Some high-scored and low-scored samples in the auxiliary data are also visualized in Figure 3.

5 Experiments

5.1 Datasets and implementation details

CUB-200-2011 The CUB-200-2011 dataset [26] contains 11,788 bird images. There are altogether 200 bird species and the number of images per class is about 60. The significant variations in pose, viewpoint, and illumination inside each class make this task very challenging. We adopt the publicly available split [26], which uses nearly half of the dataset for training and the other half for testing.
Stanford Dogs The Stanford Dogs dataset [9] contains 120 categories of dogs. There are 12,000 images for training and 8,580 images for testing. This dataset is also challenging due to small inter-class variation, large intra-class variation, and cluttered background.

ImageNet Subset The ImageNet Subset contains all categories of the original ImageNet [22] except the 59 categories of bird species, providing more realistic auxiliary data for CUB-200-2011 [26]. Note that almost all existing methods on CUB-200-2011 use the whole ImageNet dataset as the auxiliary set.

L-Bird Subset The original L-Bird dataset [11] contains nearly 4.8 million images which are obtained by searching images of a total of 10,982 bird species from the Internet. The dataset provides urls of these images, and by the time of our downloading the dataset, we only manage to get 3.3 million images from effective urls. To build the dataset of L-Bird Subset, we first choose bird species/classes out of the total 10,982 ones whose numbers of samples are beyond 100; we then remove all the 200 bird classes that are already used in CUB-200-2011, since the L-Bird Subset will be used as an auxiliary set for CUB-200-2011; we finally hold out 1% of the resulting bird images as a validation set, following the work of [11]. The final auxiliary L-Bird Subset contains 3.2 million images.

Remarks on the used datasets We use the ImageNet Subset, the ImageNet ILSVRC 2012 training set [22], or the L-Bird Subset as the set of auxiliary data for CUB-200-2011, and use the ImageNet ILSVRC 2012 training set as the set of auxiliary data for Stanford Dogs.

Implementation details Many existing convolutional neural networks (CNNs), such as AlexNet [12], VGGNet [23], or ResNet [8], can be used as backbone of our MetaFGNet. In this work, we use the pre-activation version of the 34-layer ResNet [8] in our experiments, *which can achieve almost identical performance on ImageNet with a batch normalization powered VGG16*. To adapt any of them for MetaFGNet, we remove its last fully-connected (FC) layer and keep the remaining ones as the base network of MetaFGNet, which are shared by the auxiliary and target data as illustrated in Figure 1. Two parallel FC layers of classifiers are added on top of the base network which are respectively used for the meta-learning objective of the target task and the regularization loss of the auxiliary task. The MetaFGNet adapted from the 34-layer ResNet is used for both the ablation studies and the comparison with the state of the art. For a fair comparison with existing FGVC methods, base network is pre-trained on ImageNet for all experiments reported in this paper. When using ImageNet Subset or ImageNet as the auxiliary data, we start from the 60th epoch pre-trained model, mainly for a quick comparison with baseline methods. When using L-Bird Subset as the auxiliary data, we employ the released pre-trained model from [8]. The architectural design of our MetaFGNet is straightforward and simple; in contrast, most of existing FGVC methods [32, 6, 13] adopt more complicated network architectures in order to exploit discrimination of local parts with or without use of ground-truth part annotations.

During each iteration of SGD training, we sample one mini-batch of auxiliary data for the regularization loss, and two mini-batches of target data for the meta-learning loss (cf. Algorithm 1 for respective use of the two mini-batches). Each mini-batch includes 256 images. We do data augmentation on these images according to [8]. In experiments using ImageNet Subset or ImageNet as the auxiliary data, the learning rate (α in Algorithm 1) starts from 0.1 and is divided

by 10 after every 10 epochs; we set momentum as 0.9 and weight decay as 0.0001; the meta learning rate (η in Algorithm 1) starts from 0.01 and is divided by 10 after every 10 epochs, in order to synchronize with the learning rate; the experiments end after 30 training epochs, which gives a total of the same 90 epochs as that of a pre-trained model. When using L-Bird Subset as the auxiliary data, the experiments firstly fine-tune an ImageNet pre-trained model on L-Bird Subset for 32 epochs, and then train our MetaFGNet for 8 epochs starting from the 24th epoch fine-tuned model; the learning rate and meta learning rate are divided by 10 respectively after 4 and 6 epochs; other settings are the same as in experiments using ImageNet or ImageNet Subset as the auxiliary data. Given parameters of such trained MetaFGNets and re-initialized target classifiers, we fine-tune (θ_b, θ_c^t) of them on the target data for another 160 epochs, which is the same for all comparative methods. We do sample selection from the auxiliary data as the way described in Section 4, *using the trained MetaFGNets before fine-tuning*. For the auxiliary sets of ImageNet Subset, ImageNet and L-Bird Subset, we respectively use 50%, 6%, and 80% of their samples as the selected top samples. Note that such ratios are empirically set and are suboptimal. After sample selection, we use the remained auxiliary samples to form a new auxiliary set, and train and fine-tune a MetaFGNet again from the MetaFGNet that have been trained using the original auxiliary datasets.

5.2 Comparison with alternative baselines

The first baseline method (referred as "Fine-tuning" in tables reported in this subsection) simply fine-tunes on the target dataset a model that has been pre-trained on the ImageNet Subset or ImageNet, of which the latter is typically used in existing FGVC methods. The second baseline (referred as "Joint training" in tables reported in this subsection) uses a joint training approach of the objective (4). The third baseline (referred as "Fine-tuning L-Bird Subset" in tables reported in this subsection) firstly fine-tunes the ImageNet pre-trained 34-layer ResNet model on L-bird Subset, and then fine-tunes the resulting model on CUB-200-2011. Experiments in this subsection are based on a MetaFGNet adapted from the 34-layer ResNet, for which we refer to it as "MetaFGNet". The baselines of Fine-tuning and Fine-tuning L-Bird Subset use half of the MetaFGNet that contains parameters of (θ_b, θ_c^t). The baseline of Joint training uses the same MetaFGNet structure as our method does.

Tables 1 and 2 report these controlled experiments on the CUB-200-2011 dataset [26]. Using ImageNet Subset or ImageNet as the set of auxiliary data, Fine-tuning gives baseline classification accuracies of 73.5% and 76.8% respectively; the result of Joint training is better than that of Fine-tuning, suggesting the usefulness of the objective (4) for FGVC tasks - note that a recent method [7] is essentially based on this objective. Our proposed MetaFGNet with regularized meta-learning objective (7) achieves a result better than that of Joint training. Our proposed sample selection scheme further improves the results to 75.3% and 80.3% respectively, thus justifying the efficacy of our proposed method. When using L-Bird Subset as the auxiliary set, our method without sample selection

improves the result to 87.2%, showing that a better auxiliary set is essential to achieve good performance on FGVC tasks. Note that L-Bird Subset does not contain the 200 bird species of the CUB-200-2011 dataset. Our method with sample selection further improves the result to 87.6%, confirming the effectiveness of our proposed scheme. Samples of the selected images and abandoned images from three auxiliary datasets are also shown in Section 5.5.

Table 1. Comparative studies of different methods on the CUB-200-2011 dataset [26], using ImageNet Subset as the auxiliary set. Experiments are based on networks adapted from a 34-layer ResNet.

Methods	Auxiliary set	Accuracy (%)
Fine-tuning	ImageNet Subset	73.5
Joint training w/o sample selection	ImageNet Subset	74.5
Joint training with sample selection	ImageNet Subset	75.0
MetaFGNet w/o sample selection	ImageNet Subset	75.0
MetaFGNet with sample selection	ImageNet Subset	75.3

Table 2. Comparative studies of different methods on the CUB-200-2011 dataset [26], using ImageNet [22] or L-Bird Subset as the auxiliary set.

Methods	Auxiliary set	Accuracy (%)
Fine-tuning	ImageNet	76.8
Joint training w/o sample selection	ImageNet	78.8
Joint training with sample selection	ImageNet	79.4
MetaFGNet w/o sample selection	ImageNet	79.5
MetaFGNet with sample selection	ImageNet	80.3
Fine-tuning L-Bird Subset	L-Bird Subset	86.2
MetaFGNet w/o sample selection	L-Bird Subset	87.2
MetaFGNet with sample selection	L-Bird Subset	87.6

In Figure 2, we also plot the training loss curves, using both the ImageNet auxiliary data and the target CUB-200-2011 data, of our method and Joint training, and also their fine-tuning loss curves on the target data. Figure 2 shows that our MetaFGNet converges to a better solution that supports a better fine-tuning than Joint training does.

5.3 Results on the CUB-200-2011

We use the MetaFGNet adapted from a 34-layer ResNet to compare with existing methods on CUB-200-2011 [26]. The most interesting comparison is with

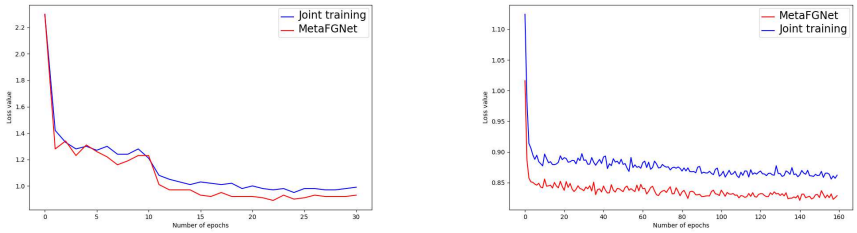


Fig. 2. Left: training loss curves of MetaFGNet and Joint training. Right: fine-tuning loss curves of MetaFGNet and Joint training. The auxiliary and target datasets are ImageNet and CUB-200-2011 respectively. MetaFGNet and Joint training models are adapted from the 34-layer ResNet.

the methods [20, 2, 18, 4, 14, 28, 11] that focus on learning from the whole bird images. In contrast, part based methods [35, 34, 6, 37, 33, 32, 31, 13] enjoy a clear advantage by exploiting discrimination of local parts either in a weakly supervised manner, or in a supervised manner using ground-truth part annotations. Table 3 also shows that our method with L-Bird Subset as auxiliary data outperforms all existing methods even when they use ground-truth part annotations. We also construct our MetaFGNet based on the popular VGGNet. Using L-Bird Subset as the auxiliary set, our MetaFGNet with sample selection gives an accuracy of 87.5%, which also confirms the efficacy of our proposed method.

5.4 Results on the Stanford Dogs

We apply the MetaFGNet to the Stanford Dogs dataset [9], using ImageNet as the auxiliary data. The used MetaFGNet is adapted from a 152-layer ResNet [8], which is the same as the one used in the state-of-the-art method [7]. Table 4 shows the comparative results. Our method outperforms all exiting methods with a large margin. *We note that previous methods use ImageNet as the auxiliary data for the Stanford Dogs task, however, it is inappropriate because the dataset of Stanford Dogs is a subset of ImageNet.* Thus, we remove all the 120 categories of dog images from ImageNet to introduce an appropriate auxiliary dataset for the Stanford Dogs dataset [9]. Based on a 34-layer ResNet [8], simple fine-tuning after pre-training on the resulting ImageNet images gives an accuracy of 69.3%; our MetaFGNet with sample selection improves it to 73.2%.

5.5 Analysis of selected and abandoned auxiliary images

In Fig 3, we qualitatively visualize the selected top-ranked images from ImageNet [22], ImageNet Subset, and L-Bird Subset, and also the abandoned bottom-ranked images respectively from the three auxiliary sets, when using CUB-200-2011 [26] as the target data. It can be observed that for ImageNet and ImageNet

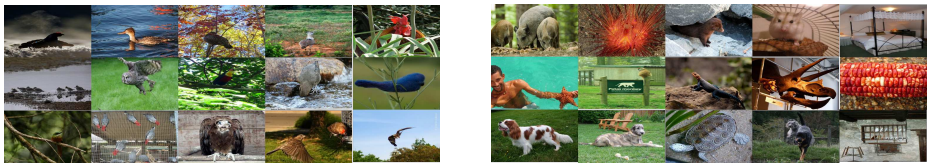
Table 3. Comparison of different methods on the CUB-200-2011 dataset [26]. Methods of 'part supervised' use ground-truth part annotations of bird images. Methods of 'part aware' detect discriminative local parts in a weakly supervised manner. Result of [11] is based on own implementation using the currently available L-Bird Subset. *The method [14] generates high-dimensional, second-order features via bilinear pooling, and the result of [28] is from a multi-scale ensemble model; both of the methods make contributions to image based FGVC complementary to our proposed MetaFGNet.*

Methods	Auxiliary Set	Part Supervision	Acc. (%)
CNNaug-SVM [20]	ImageNet	n/a	61.8
Deep Optimized [2]	ImageNet	n/a	67.1
MsML [18]	ImageNet	n/a	67.9
Deep Metric [4]	ImageNet + Web	n/a	80.7
Bilinear [14]	ImageNet	n/a	84.1
Deep Image [28]	ImageNet	n/a	84.9
Rich Data [11]	L-Bird Subset	n/a	**86.2**
MetaFGNet with sample selection	ImageNet	n/a	80.3
MetaFGNet with sample selection	L-Bird Subset	n/a	**87.6**
Weakly sup. [35]	ImageNet	part-aware	80.4
PDFS [34]	ImageNet	part-aware	84.5
RA-CNN [6]	ImageNet	part-aware	85.3
MA-CNN [37]	ImageNet	part-aware	86.5
Part R-CNN[33]	ImageNet	part supervised	73.9
SPDA-CNN[32]	ImageNet	part supervised	81.0
Webly-sup.[31]	ImageNet + Web	part supervised	84.6
Hsnet[13]	ImageNet	part supervised	87.5

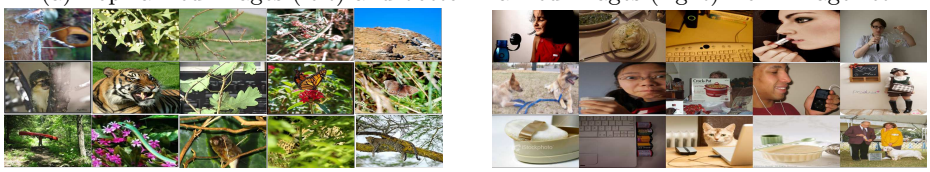
Table 4. Comparison of different methods on the Stanford Dogs dataset [9].

Methods	Auxiliary Set	Part Supervision	Acc. (%)
Weakly sup. [35]	ImageNet	part-aware	80.4
DVAN [36]	ImageNet	part-aware	81.5
MsML[18]	ImageNet	n/a	70.3
MagNet[21]	ImageNet	n/a	75.1
Selective joint training[7]	ImageNet	n/a	90.3
MetaFGNet with sample selection	ImageNet	n/a	**96.7**

Subset, images that are semantically related to the target CUB-200-2011 task are ranked top and selected by our proposed scheme; for L-Bird Subset, noisy images that are irrelevant to the target task are ranked bottom and removed. Quantitatively, when using ImageNet as the auxiliary dataset, 65.3% of the selected auxiliary images belong to the basic-level category of "bird".



(a) Top-ranked images (left) and bottom-ranked images (right) from ImageNet



(b) Top-ranked images (left) and bottom-ranked images (right) from ImageNet Subset

(c) Top-ranked images (left) and bottom-ranked images (right) from L-Bird Subset

Fig. 3. (a) Top-ranked images and bottom-ranked images from ImageNet. (b) Top-ranked images and bottom-ranked images from ImageNet Subset. (c) Top-ranked images and bottom-ranked images from L-Bird Subset. Results are obtained by using MetaFGNet and our sample selection scheme on the CUB-200-2011 dataset [26].

6 Conclusion

In this paper, we propose a new deep learning model termed MetaFGNet, which is based on a novel regularized meta-learning objective that aims to guide the learning of network parameters so that they are optimal for adapting to a target FGVC task. Based on MetaFGNet, we also propose a simple yet effective scheme for sample selection from auxiliary data. Experiments on the benchmark CUB-200-2011 and Stanford Dogs datasets show the efficacy of our proposed method.

Acknowledgment. This work is supported in part by the Program for Guangdong Introducing Innovative and Enterpreneurial Teams (Grant No.: 2017ZT07X183), and the National Natural Science Foundation of China (Grant No.: 61771201).

References

1. Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D.G., Steiner, B., Tucker, P., Vasudevan, V., Warden, P., Wicke, M., Yu, Y., Zheng, X.: Tensorflow: A system for large-scale machine learning. In: 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16). pp. 265–283 (2016), https://www.usenix.org/system/files/conference/osdi16/osdi16-abadi.pdf

2. Azizpour, H., Razavian, A.S., Sullivan, J., Maki, A., Carlsson, S.: From generic to specific deep representations for visual recognition. In: CVPRW DeepVision Workshop, June 11, 2015, Boston, MA, USA. IEEE conference proceedings (2015)
3. Berg, T., Liu, J., Woo Lee, S., Alexander, M.L., Jacobs, D.W., Belhumeur, P.N.: Birdsnap: Large-scale fine-grained visual categorization of birds. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2011–2018 (2014)
4. Cui, Y., Zhou, F., Lin, Y., Belongie, S.: Fine-grained categorization and dataset bootstrapping using deep metric learning with humans in the loop. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1153–1162 (2016)
5. Finn, C., Abbeel, P., Levine, S.: Model-agnostic meta-learning for fast adaptation of deep networks. arXiv preprint arXiv:1703.03400 (2017)
6. Fu, J., Zheng, H., Mei, T.: Look closer to see better: recurrent attention convolutional neural network for fine-grained image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
7. Ge, W., Yu, Y.: Borrowing treasures from the wealthy: Deep transfer learning through selective joint fine-tuning. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI. vol. 6 (2017)
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
9. Khosla, A., Jayadevaprakash, N., Yao, B., Li, F.F.: Novel dataset for fine-grained image categorization: Stanford dogs. In: Proc. CVPR Workshop on Fine-Grained Visual Categorization (FGVC). vol. 2 (2011)
10. Koch, G., Zemel, R., Salakhutdinov, R.: Siamese neural networks for one-shot image recognition. In: ICML Deep Learning Workshop. vol. 2 (2015)
11. Krause, J., Sapp, B., Howard, A., Zhou, H., Toshev, A., Duerig, T., Philbin, J., Fei-Fei, L.: The unreasonable effectiveness of noisy data for fine-grained recognition. In: European Conference on Computer Vision. pp. 301–320. Springer (2016)
12. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems. pp. 1097–1105 (2012)
13. Lam, M., Mahasseni, B., Todorovic, S.: Fine-grained recognition as hsnet search for informative image parts. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2520–2529 (2017)
14. Lin, T.Y., RoyChowdhury, A., Maji, S.: Bilinear cnn models for fine-grained visual recognition. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1449–1457 (2015)
15. Pan, S.J., Yang, Q.: A survey on transfer learning. IEEE Trans. Knowl. Data Eng. **22**(10), 1345–1359 (2010)
16. Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A.: Automatic differentiation in pytorch (2017)
17. Peng, Y., He, X., Zhao, J.: Object-part attention model for fine-grained image classification. IEEE Transactions on Image Processing **27**(3), 1487–1500 (2018)
18. Qian, Q., Jin, R., Zhu, S., Lin, Y.: Fine-grained visual categorization via multi-stage metric learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3716–3724 (2015)
19. Ravi, S., Larochelle, H.: Optimization as a model for few-shot learning (2016)

20. Razavian, A.S., Azizpour, H., Sullivan, J., Carlsson, S.: Cnn features off-the-shelf: an astounding baseline for recognition. In: Computer Vision and Pattern Recognition Workshops (CVPRW), 2014 IEEE Conference on. pp. 512–519. IEEE (2014)
21. Rippel, O., Paluri, M., Dollar, P., Bourdev, L.: Metric learning with adaptive density discrimination. arXiv preprint arXiv:1511.05939 (2015)
22. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. International Journal of Computer Vision **115**(3), 211–252 (2015)
23. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
24. Uijlings, J.R., Van De Sande, K.E., Gevers, T., Smeulders, A.W.: Selective search for object recognition. International journal of computer vision **104**(2), 154–171 (2013)
25. Vinyals, O., Blundell, C., Lillicrap, T., Wierstra, D., et al.: Matching networks for one shot learning. In: Advances in Neural Information Processing Systems. pp. 3630–3638 (2016)
26. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The caltech-ucsd birds-200-2011 dataset (2011)
27. Welinder, P., Branson, S., Mita, T., Wah, C., Schroff, F., Belongie, S., Perona, P.: Caltech-ucsd birds 200 (2010)
28. Wu, R., Yan, S., Shan, Y., Dang, Q., Sun, G.: Deep image: Scaling up image recognition. arXiv preprint arXiv:1501.02876 **7**(8) (2015)
29. Xiao, T., Xu, Y., Yang, K., Zhang, J., Peng, Y., Zhang, Z.: The application of two-level attention models in deep convolutional neural network for fine-grained image classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 842–850 (2015)
30. Xie, S., Yang, T., Wang, X., Lin, Y.: Hyper-class augmented and regularized deep learning for fine-grained image classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2645–2654 (2015)
31. Xu, Z., Huang, S., Zhang, Y., Tao, D.: Webly-supervised fine-grained visual categorization via deep domain adaptation. IEEE transactions on pattern analysis and machine intelligence (2016)
32. Zhang, H., Xu, T., Elhoseiny, M., Huang, X., Zhang, S., Elgammal, A., Metaxas, D.: 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 (2016)
33. Zhang, N., Donahue, J., Girshick, R., Darrell, T.: Part-based r-cnns for fine-grained category detection. In: European conference on computer vision. pp. 834–849. Springer (2014)
34. Zhang, X., Xiong, H., Zhou, W., Lin, W., Tian, Q.: Picking deep filter responses for fine-grained image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1134–1142 (2016)
35. Zhang, Y., Wei, X.S., Wu, J., Cai, J., Lu, J., Nguyen, V.A., Do, M.N.: Weakly supervised fine-grained categorization with part-based image representation. IEEE Transactions on Image Processing **25**(4), 1713–1725 (2016)
36. Zhao, B., Wu, X., Feng, J., Peng, Q., Yan, S.: Diversified visual attention networks for fine-grained object classification. arXiv preprint arXiv:1606.08572 (2016)
37. Zheng, H., Fu, J., Mei, T., Luo, J.: Learning multi-attention convolutional neural network for fine-grained image recognition. In: Int. Conf. on Computer Vision. vol. 6 (2017)