ASRSNet: Automatic Salient Region Selection Network for Few-Shot Fine-Grained Image Classification

Yi Liao[1][0000−0002−4594−1416], Weichuan Zhang[1][0000−0002−3437−4153], Yongsheng Gao[1][0000−0002−5382−5351]✉, Changming Sun[2][0000−0001−5943−1989], and Xiaohan Yu[1][0000−0001−6186−0520]

[1] Griffith University, Brisbane, Australia
yi.liao2@griffithuni.edu.au, zwc2003@163.com, {yongsheng.gao, xiaohan.yu}@griffith.edu.au
[2] CSIRO Data61 PO Box 76 Epping, NSW 1710, Australia
changming.sun@data61.csiro.au

Abstract. Few-shot learning for image classification aims at predicting unseen classes with only a few images. Recent works, especially the works on few-shot fine-grained image classification (FSFGIC), have achieved great progress. However, most of them neglected the spatial information and computed the distance between a query image and a support image directly, which may cause vagueness because the dominant objects can exist anywhere on images. A promising solution is to locate salient regions from images for discriminative feature representation learning. This paper develops an automatic salient region selection network without the use of a bounding box or part annotation mechanism for locating salient regions from images. Then a weighted average mechanism is introduced for facilitating a neural network to focus on those salient regions, optimizing the network, and performing the FSFGIC tasks. The experimental results on four benchmark datasets demonstrate the effectiveness of the proposed strategy.

Keywords: Few-shot fine-grained image classification · discriminative feature representation learning · automatic salient region selection network · weighted average mechanism

1 Introduction

Few-shot fine-grained image classification (FSFGIC) methods refer to machine learning methods which aim to classify images belonging to subordinate object categories of the same entry-level category with only a few samples. In the last few years, FSFGIC has achieved stable progresses. The learning ability of deep neural networks [26] [8] [30] for recognizing the subtle differences between highly similar objects has been continuously improved. Meanwhile, a large number of fine-grained image datasets (e.g., CUB-200-2010 [4], Standard Cars [15], Aircraft [20], and Plant Disease [25]) have been collected by domain experts using

complex rules to determine the accuracy of different types of object classification methods and to assist researchers to improve the algorithms for achieving better performance on FSFGIC tasks.

Learning discriminative feature representation from images plays a key role on FSFGIC which is used not only to represent training samples but also to construct a classifier for performing FSFGIC tasks. The primal step of discriminative feature representation learning is to locate salient regions from images. Currently, a bounding box or part annotations mechanism [28] [1] [11] [14] is widely applied for locating salient regions and performing object classification using the discriminative information from the selected regions. In this paper, an automatic salient region selection network without the use of a bounding box or part annotations mechanism is designed for learning discriminative feature representations and performing FSFGIC tasks. First, for each image, it is divided it into M parts equally. Then the image and its corresponding M sub-images are sent into a given neural network (e.g., Conv-64F [30]) for training and obtaining feature descriptors. Second, the similarity between the query and support images is measured based on the obtained feature descriptors and they are named as basic similarity. Meanwhile, M similarities between M pairs of sub-images from query and support images are obtained based on their corresponding feature descriptors and named as sub-similarities. Third, if some sub-similarities are larger than the basic similarity, their corresponding sub-images are marked as salient regions and a weighted average mechanism is designed for a neural network to focus on these salient regions, optimizing the network, and performing the FSFGIC tasks. It is worth noting that if sub-similarities are less than the basic similarity in one episode, the designed neural network is optimized by using the basic similarity in this episode.

The main contributions in our proposed method comprise two aspects.

- An automatic salient region selection network is designed for discriminative feature representations learning and performing FSFGIC tasks. This designed network enables the salient regions on images to be detected automatically without the help of bounding box and annotation information.
- A weighed average mechanism is designed for enhancing the effect of discriminative features in the tasks of classifying fine-grained images.
- Experimental evaluation are performed on four public datasets, verifying the effectiveness of the proposed method in various FSFGIC tasks.

2 Related work

In this section, we briefly introduce the existing FSFGIC related methods: fine-grained image classification methods, meta-learning based FSFGIC methods, and metric-learning based FSFGIC methods.

2.1 Fine-grained image classification

Fine-grained image classification (FGIC) is an attractive topic in the computer vision research community. In FGIC, the training samples share the same class space with testing samples. Current regional feature based FGIC approaches try to discover salient regions from images without the help of bounding box and annotation information. Peng et al. [21] proposed an object-part attention model in which potential objects from images were detected and the most discriminative parts of the object were selected as feature representation. The classification accuracy is improved by 1.2% on dataset CUB-200-2010 [31]. Zhang et al. [32] proposed an approach by which all the potential object parts were generated by using selective search method [29] and the most discriminative object part was selected to form image representation according to its importance value computed with the help of Fisher Vector [23]. The approach improved classification accuracy by 3.5% on dataset CUB-200-2010 [31]. Therefore, locating the discriminative regions can boost classification performance in FGIC tasks.

2.2 Meta-learning based FSFGIC methods

Different from FGIC, The class space of training data and the class space of testing data are disjoint in FSFGIC. Meta-learning based FSFGIC is a branch of FSFGIC. Finn et al. [6] proposed a model-agnostic meta-learning (MAML) method by which any model is trained successively twice to obtain two groups of parameters. The new gradient for updating the model is computed using both groups of parameters. Cai et al. [3] proposed a memory matching network (MM-Net) where a contextual learner consisting of multiple bidirectional long-shot term memory (LSTM) [24] is devised to predict the parameters for the embedding network. Sachin et al. [22] proposed an LSTM-based optimizer by combining the standard gradient descent algorithm and the cell state of LSTM [9]. In this way, a novel gradient is obtained by training a LSTM network.

2.3 Metric-learning based FSFGIC methods

The existing metric-learning based FSFGIC methods usually consist of three steps. Firstly, the images including support images and query images are embedded into their image representations by embedding networks (e.g., Conv-64F [30]). Secondly, the distances between each query embeddings and all support embeddings are calculated by employing different distance metrics (e.g., cosine similarity [17], Euclidean distance [27], and Kullback-Leibler(KL) distance [16]). Thirdly, each query image is allocated to the support class according to the closest distance principle.

Snell et al. [27] proposed a prototypical network that employs the Euclidean distance for measuring the similarity between the support image representations and the query image representation. Li et al. [17] proposed a deep nearest neighbour neural network (DN4) using cosine similarity as the measurement. In [16],

an asymmetric Kullback-Leibler (KL) divergence was utilized to measure the relation of distribution between query image and support image. In [18], a covariance metric network is proposed to measure the relation between a query image and support categories. However, the aforementioned methods [27] [17] [16] [18] did not deal with the spatial information discriminatively and calculated similarity distance between a query image and a support class directly without considering the effect of salient regions on images in FSFGIC.

Recently, many few-shot fine-grained learning research works focus on attention mechanisms. For example, Dong et al. [5] presented a novel ATL-Net in which a task adaptive attention module is designed to generate a relation matrix. The relation matrix consists of cosine similarity between each local representation of a query embedding and each local representation of one support class embeddings. All the local representations are processed by a convolution layer and filtered through a threshold. The weights of every image patch are decided by the threshold. However, the threshold is predicted by a multi-layer perceptron (MLP) trained on query embeddings. Therefore, the threshold will vary with various query images and the weights for semantic patches from query image are not stable enough. It should have negative influence on the final classification accuracy. Yan et al. [19] presented a novel method called dense classification in which the weights of each patch are learned through training on auxiliary data. The assumption is that the weights learned on auxiliary data are generic enough to be used for new classes.

Different from the methods above, our proposed method not only has an ability to automatically locate the semantic regions with aid of the comparison between image patch level similarity and image level similarity, but assigns a stable weight to most discriminative regions by using a weighted average mechanism.

3 Methodology

In this section, we first present a brief review of the problem definition of few-shot classification. Then we illustrate how to automatically select salient regions from images. Finally, a weighted average mechanism is designed for a neural network to focus on these salient regions, optimizing the network, and performing the FSFGIC tasks. The overview of the proposed framework for one-shot image classification is shown in Fig. 1

3.1 Problem definition

For an FSFGIC task, the target dataset $\mathcal{D}$ contains two parts: a support set $\mathcal{S}$ and a query set $\mathcal{Q}$. The small support set $\mathcal{S}$ includes C unseen classes, and each of which has K labeled samples. The query set $\mathcal{Q}$ contains J unlabeled samples.

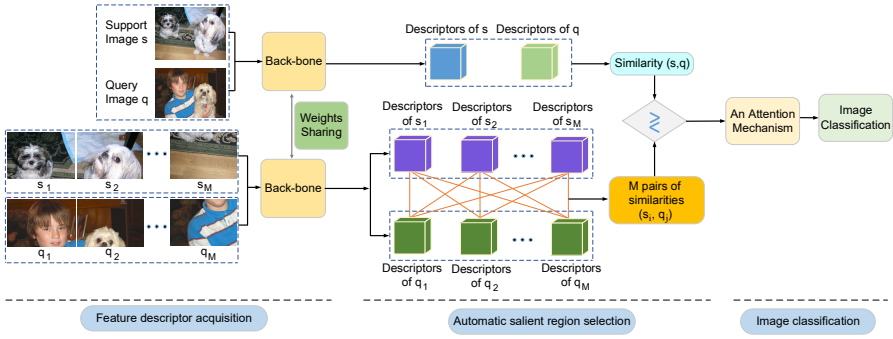

Fig. 1. The overall pipeline of our proposed automatic salient region selection framework. (1) Feature descriptor acquisition: support and query and their corresponding sub-images are sent into backbones for obtaining feature descriptors. (2) Automatic salient region selection: if some sub-similarities are larger than the basic similarity, their corresponding sub-images are marked as salient regions. (3) Image classification: a weighted average mechanism is designed for a neural network to focus on these salient regions and performing FSFGIC tasks.

Displayed equations are centered and set on a separate line.

$$\mathcal{D} = \left\{ \mathcal{S} = \left\{ (x_i, y_i)_{i=1}^{C \times K} \right\} \cup \mathcal{Q} = \left\{ (x_j)_{j=1}^{J} \right\} \right\}, \tag{1}$$

where $\mathcal{S} \cap \mathcal{Q} = \emptyset$, x_i and x_j denote fine-grained samples and $y_i \subset C$ represents the ground truth label of x_i. The goal of FSFGIC is to successfully classify x_j into its corresponding class in C in $\mathcal{S}$. Thus, the problem is denoted as a C-way K-shot task.

It is worth noting that the training samples of each class in FSFGIC are too limited to effectively learn transferable knowledge [33] for performing FSFGIC tasks. Then, an episodic training paradigm [30] with an auxiliary set $\mathcal{A}$, which has similar data distribution with $\mathcal{D}$, is applied to tackle the aforementioned problem as follows

$$\mathcal{A} = \left\{ \mathcal{E} = \left\{ (u_i, v_i)_{i=1}^{N} \right\} \cup \mathcal{F} = \left\{ (u_j, v_j)_{j=1}^{L} \right\} \right\}, \tag{2}$$

where u_i and u_j are fine-grained images, v_i and v_j are their corresponding labels; $\mathcal{E} \cup \mathcal{F} = \emptyset$, $\mathcal{D} \cup \mathcal{A} = \emptyset$. The auxiliary set A contains sufficient classes and labeled samples which are far larger than C and K respectively.

In each round of training, $\mathcal{A}$ is randomly separated into two parts: an auxiliary support set $\mathcal{G} = \left\{ (u_i, v_i)_{i=1}^{C \times K} \right\}$ and an auxiliary query set $\mathcal{H} = \left\{ (u_j, v_j)_{j=1}^{J} \right\}$. With $N >> C \times K$, $\mathcal{E}$ can mimic the composition of $\mathcal{S}$ in each iteration. Then $\mathcal{A}$ is employed to learn prior knowledge for training $\mathcal{S}$.

3.2 Automatic salient regions selection for FSFGIC

Let s ($s \in \mathcal{S}$) and q ($q \in \mathcal{Q}$) denote a support image and a query image respectively. The support image s and query image q are divided into M (here $M = 36$) parts equally. Then sub-images (i.e., s_1, s_2,..., s_M) of support image s and sub-images (i.e., q_1, q_2,..., q_M) of query image q are obtained. Meanwhile, the sub-images and query and support images are resized into the same widths and they are sent into a given backbone (i.e., Conv-64F [30]) for training and obtaining their corresponding descriptors. Typically, after an input image x is sent into a backbone Θ, dhw-dimensional feature descriptors (i.e., $\Theta(x) \in R^{d\times(h\times w)}$) can be obtained where h and w denote the height and width of the feature tensor map respectively, d is the number of filters, and R represents real space. Therefore, the feature descriptors of the support image s, the query image q, the sub-images (i.e., s_1, s_2,...,s_M) of support image s, and the sub-images (i.e., q_1, q_2, ...,q_M) of query image q are obtained as

$$\begin{aligned}
&\Psi^s = [\Theta_1(s), \cdots, \Theta_\tau(s)],\\
&\Psi^q = [\Theta_1(q), \cdots, \Theta_\tau(q)],\\
&\Psi^{s_l} = [\Theta_1(s_l), \cdots, \Theta_\tau(s_l)],\\
&\Psi^{q_l} = [\Theta_1(q_1), \cdots, \Theta_\tau(q_l)],\ l = 1, \cdots, M,
\end{aligned} \tag{3}$$

where τ ($\tau = h \times w$) is the total number of descriptors for each image. In this work, a cosine measure is utilized for calculating the similarity between two different images based on the 1-nearest neighbor method [2]. Then the basic similarity between a query image and a support image is as follows,

$$\begin{aligned}
&\Lambda(\Theta_t(s), \Theta_t(q)) = \frac{\Theta_t(s)^T \Theta_t(q)}{\parallel \Theta_t(s_u) \parallel \cdot \parallel \Theta_t(q_v) \parallel},\\
&\Lambda(\Psi^s, \Psi^s) = \max\left\{\Lambda(\Theta_1(s_u), \Theta_1(q_v)), \cdots, \Lambda(\Theta_\tau(s_u), \Theta_\tau(q_v))\right\},\\
&t = 1, 2, \cdots, \tau.
\end{aligned} \tag{4}$$

The sub-similarities between the sub-images of a support image and the sub-images of a query image are as follows,

$$\begin{aligned}
&\Lambda(\Theta_t(s_u), \Theta_t(q_v)) = \frac{\Theta_t(s_u)^T \Theta_t(q_v)}{\parallel \Theta_t(s_u) \parallel \cdot \parallel \Theta_t(q_v) \parallel},\\
&\Lambda(\Psi^{s_u}, \Psi^{q_v}) = \max\left\{\Lambda(\Theta_1(s_u), \Theta_1(q_v)), \cdots, \Lambda(\Theta_\tau(s_u), \Theta_\tau(q_v))\right\},\\
&t = 1, 2, \cdots, \tau,\ u = 1, 2, \cdots, M,\ v = 1, 2, \cdots, M.
\end{aligned} \tag{5}$$

From Equation (5), an $M \times M$ sub-similarity matrix ζ_1 can be obtained. Then we first find the maximum value in the subsimilarity matrix ζ_1 and its corresponding row i and column j which is denoted as $\eta_1 = \zeta_1(i, j)$. If η_1 is larger than its corresponding basic similarity, the sub-image pair between the i-th sub-image of the support image and the j-th subimage of the query image are marked as a pair of salient regions. Second, we delete the i-th row and j-th column of the matrix ζ_1 and construct a new $(M-1)\times(M-1)$ matrix ζ_2. Then we find the maximum

value in matrix ζ_2 and its corresponding row u and column v which is denoted as $\eta_2 = \zeta_2(u, v)$. If η_2 is also larger than its corresponding basic similarity, we will continue to perform this operation until the maximum value η_{k+1} of the (k + 1)-th sub-similarity matrix is less than the basic similarity. Finally, k pairs of sub-images corresponding to the k sub-similarities are marked as salient regions, and a weighted average β is presented on the k subsimilarities (i.e., η_1, η_2, ..., η_k) for representing the similarity between the support image and the query image as follows

$$\beta(s, q) = \frac{1}{\sum_{i=1}^{k} \eta_i}(\eta_1^2 + \eta_2^2 + \cdots + \eta_k^2). \tag{6}$$

It is worth noting that if η_1 is less than the basic similarity, the similarity between the support and query images is represented by

$$\beta(s, q) = \Lambda(\Psi^s, \Psi^q). \tag{7}$$

Furthermore, the similarity between query image q and class C is calculated which is named as image-to-class similarity measure [16] as follows,

$$\xi(q, C) = \sum_{j=1}^{\tau} \beta(s^j, q), \tag{8}$$

where s^j represents the j-th support image in class C. Then the Adam optimization method [13] with a cross-entropy loss is used to train the whole network for learning the parameters and performing FSFGIC tasks. The detailed process is listed as Algorithm 1.

Algorithm 1: Salient regions selection mechanism

Input: sub-similarity matrix ζ_1, M
Output: $\beta(s, q)$

$k = 0, t = 0, s = 0$;
while $k \leq M$ **do**
  $k = k + 1$;
  $\eta_k = \max(\zeta_k)$;
  if $\eta_k \geq \Lambda(\Psi^s, \Psi^s)$ **then**
    $(i, j) = \text{argmax}(\zeta_k)$;
    η_{k+1} is contructed by i-th row and j-th column are deleted from η_k;
    $t = t + \eta_k$;
    $s = s + \eta_k^2$;
  else
    break;
  end
end
$\beta(s, q) = \frac{t}{s}$;
return $\beta(s, q)$

4 Experiment

4.1 Datasets

Our proposed network is evaluated on four fine-grained datasets, i.e., the Stanford Dogs [12], Stanford Cars [15], CUB-200-2010 [31], and Plant Disease [25] datasets. The Stanford Dogs dataset consists of 120 dog classes with 20,580 samples. The Stanford Cars dataset consists of 196 car classes with 16,185 samples. The CUB-200-2010 dataset consists of 200 bird classes with 6,033 samples. The Plant Disease dataset consists of 38 plant disease classes with 54,306 samples. For fair performance comparisons, we follow the same data split as used in [17] that are illustrated in Table. 1.

Table 1. The class split of four fine-grained datasets. N_{train}, N_{val}, and N_{test} are the numbers of classes in the auxiliary set, validation set, and test set respectively.

Dataset	N_{train}	N_{val}	N_{test}
Stanford Dogs	70	20	30
Stanford Cars	130	17	49
CUB-200-2010	130	20	50
Plant Disease	20	10	8

4.2 Experimental setup

In this work, both the 5-way 1-shot and 5-way 5-shot FSFGIC tasks are performed on the four datasets. We follow the basic feature extraction network (i.e., Conv-64F [30]). Each input image is resized to 84×84. Then we have $h = w = 21$, $d = 64$, and τ=441. Random crop, random color transformations, random horizontal flips, and random rotations are utilized for data augmentation. There are 300,000 episodes which are randomly sampled and constructed for training the proposed models by utilizing the episodic training paradigm [30]. For each episode, 15 query samples per class are randomly selected for the four datasets. The Adam optimization method [13] is utilized for training the models using 30 epochs. The learning rate is initially set as 0.001 and multiplied by 0.5 for every 100,000 episodes. In the testing stage, 600 episodes are randomly constructed from the testing set for obtaining the classification results. The top-1 mean accuracy is employed as the evaluation criteria. The above process is repeated five times and the final mean results are obtained as the classification accuracy for FSFGIC. Meanwhile, the 95% confidence intervals are obtained and reported.

4.3 Performance comparison

The experimental results of eight state-of-the-art metric learning methods (i.e., Matching Net (M-Net) [30], Prototypical Net (P-Net) [27], GNN [7], CovaMNet [18], DN4 [17], PABN+$_{cpt}$ [10], LRPABN$_{cpt}$ [10], and ATL-Net [5]) and the

Table 2. Comparison results on Stanford Dogs and Stanford Cars datasets

Model	5-way Accuracy(%)			
	Stanford Dogs		Stanford Cars	
	1-shot	5-shot	1-shot	5-shot
M-Net [30]	35.80 ± 0.99	47.50 ± 1.03	34.80 ± 0.98	44.70 ± 1.03
P-Net [27]	37.59 ± 1.00	48.19 ± 1.03	40.90 ± 1.01	52.93 ± 1.03
GNN [7]	46.98 ± 0.98	62.27 ± 0.95	55.85 ± 0.97	71.25 ± 0.89
CovaMNet [18]	49.10 ± 0.76	63.04 ± 0.65	56.65 ± 0.86	71.33 ± 0.62
DN4 [17]	45.73 ± 0.76	66.33 ± 0.66	61.51 ± 0.85	89.60 ± 0.44
PABN+$_{cpt}$ [10]	45.65 ± 0.71	61.24 ± 0.62	54.44 ± 0.71	67.36 ± 0.61
LRPABN$_{cpt}$ [10]	45.72 ± 0.75	60.94 ± 0.66	60.28 ± 0.76	73.29 ± 0.58
ATL-Net [5]	54.49 ± 0.92	73.20 ± 0.69	**67.95±0.84**	89.16 ± 0.48
Proposed ASRSNET	**54.97±0.88**	**73.21±0.57**	64.33 ± 0.75	**91.15±0.78**

proposed method on the four datasets are summarized in Table 2 and Table 3. It is worth noting that the accuracies of the seven other methods are also tested on the same feature extraction network (i.e., Conv-64F [30]). For three fine-grained datasets, i.e., the Stanford Dogs [12], Stanford Cars [15], CUB-200-2010 [31], we use the officially provided results for all the other methods. For the Plant Disease [25] dataset, we utilize the codes provided to test their corresponding results. Because the codes for PABN+$_{cpt}$ [10] and LRPABN$_{cpt}$ [10] are not provided, we leave them blank on the Plant Disease dataset.

It can be found from Table 2 and Table 3 that the proposed method gets steady and notable improvements on almost all FSFGIC tasks. For the 5-way 5-shot task, the proposed method achieves the best performance on four fine-grained datasets. For the 5-way 1-shot task, the proposed method also achieves the best performance on Standford Dogs, CUB-200-2010, and Plant Disease and achieves the second best performance on Standford Cars. For the 5-way 1-shot and 5-way 5-shot FSFGIC tasks on the CUB-200-2010 dataset, our proposed method achieves 41.56%, 71.65%, 23.73%, 22.33%, 20.65%, 1.22%, 0.78%, and 5.28% improvements and 38%, 81.34%, 28.92%, 28.78%,0.25%, 9.90%, 0.95%, and 6.56% improvements over M-Net, P-Net, GNN, CovaMNet, DN4, PABN+$_{cpt}$, LRPABN$_{cpt}$, and ATL-Net respectively. Such improvements demonstrate the ability of ASRSNET for effectively highlighting the feature representation in salient regions in images and making the similarity measure between the samples within the same class larger and the similarity measure between samples from different classes smaller with limited training samples.

5 Conclusion

Current few-shot metric-based learning methods have improved classification accuracy greatly in FSFGIC tasks. However, they computed similarity distance between a query image and a support class directly without considering the ef-

Table 3. Comparison results on Plant Disease and CUB-200-2010 datasets.

Model	5-way Accuracy(%)			
	Plant Disease		CUB-200-2010	
	1-shot	5-shot	1-shot	5-shot
M-Net [30]	62.93 ± 0.94	80.55 ± 0.93	45.30 ± 1.03	59.50 ± 1.01
P-Net [27]	64.97 ± 0.85	82.73 ± 0.91	37.36 ± 1.00	45.28 ± 1.03
GNN [7]	69.85 ± 0.91	88.69 ± 0.79	51.83 ± 0.98	63.69 ± 0.64
CovaMNet [18]	70.72 ± 0.89	88.92 ± 0.81	52.42 ± 0.76	63.76 ± 0.64
DN4 [17]	72.47 ± 0.76	90.68 ± 0.44	53.15 ± 0.84	81.90 ± 0.60
PABN+$_{cpt}$ [10]	-	-	63.36 ± 0.80	74.71 ± 0.60
LRPABN$_{cpt}$ [10]	-	-	63.23 ± 0.77	76.06 ± 0.58
ATL-Net [5]	72.18 ± 0.92	90.11 ± 0.65	60.91 ± 0.91	77.05 ± 0.67
Proposed ASRSNET	**73.58±0.87**	**91.76±0.79**	**64.13±0.85**	**82.11±0.72**

fect of salient regions on images. Such image-level based similarity resulted in a vagueness problem because the dominant objects can exist anywhere on images. In this paper, a novel automatic salient region selection network without the use of a bounding box or part annotation mechanism is proposed for obtaining salient region pairs from query and support images, aiming to locate the more discriminative regions for improving the representation ability of neural networks trained by few samples. Meanwhile, it donates an advanced approach to alleviating vagueness problem for the FSFGIC research community. Furthermore, a weighted average mechanism is designed for facilitating a neural network to focus on those salient regions, optimizing the network, and enhance the classification accuracy for FSFGIC tasks. The effectiveness of our proposed method has been demonstrated through experiments on four benchmark fine-grained datasets.

References

1. 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)
2. Boiman, O., Shechtman, E., Irani, M.: In defense of nearest-neighbor based image classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1–8 (2008)
3. Cai, Q., Pan, Y., Yao, T., Yan, C., Mei, T.: Memory Matching Networks for One-Shot Image Recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4080–4088 (2018)
4. Cathrine wah, S.B., Peter Welinder, P.P., Belongie, S.: The caltech-ucsd birds-200-2011 dataset. California Institute of Technology (2011)
5. Dong, C., Li, W., Huo, J., Gu, Z., Gao, Y.: Learning task-aware local representations for few-shot learning. In: Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence. pp. 716–722 (2021)

6. Finn, C., Abbeel, P., Levine, S.: Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. In: Proceedings of International Conference on Machine Learning. pp. 1126–1135 (2017)
7. Garcia, V., Bruna, J.: Few-shot learning with graph neural networks. In: Proceedings of the International Conference on Learning Representations (2018)
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, 2016. pp. 770–778 (2016)
9. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Computation **9**(8), 1735–1780 (1997). https://doi.org/10.1162/neco.1997.9.8.1735
10. Huang, H., Zhang, J., Zhang, J., Xu, J., Wu, Q.: Low-rank pairwise alignment bilinear network for few-shot fine-grained image classification. IEEE Transactions on Multimedia **23**, 1666–1680 (2021). https://doi.org/10.1109/TMM.2020.3001510
11. Huang, S., Xu, Z., Tao, D., Zhang, Y.: Part-stacked cnn for fine-grained visual categorization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1173–1182 (2016)
12. Khosla, A., Jayadevaprakash, N., Yao, B., Li, F.F.: Novel dataset for fine-grained image categorization: Stanford dogs. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshop on Fine-Grained Visual Categorization (2011)
13. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: Proceedings of the International Conference on Learning Representations (2015)
14. Krause, J., Jin, H., Yang, J., Fei-Fei, L.: Fine-grained recognition without part annotations. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5546–5555 (2015)
15. Krause, J., Stark, M., Deng, J., Fei-Fei, L.: 3D object representations for fine-grained categorization. In: Proceedings of the IEEE International Conference on Computer Vision Workshops. pp. 554–561 (2013)
16. Li, W., Wang, L., Huo, J., Shi, Y., Gao, Y., Luo, J.: Asymmetric distribution measure for few-shot learning. arXiv preprint arXiv:2002.00153 (2020)
17. Li, W., Wang, L., Xu, J., Huo, J., Gao, Y., Luo, J.: 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 (2019)
18. Li, W., Xu, J., Huo, J., Wang, L., Gao, Y., Luo, J.: Distribution consistency based covariance metric networks for few-shot learning. In: Proceedings of the Association for the Advancement of Artificial Intelligence. pp. 8642–8649 (2019)
19. Lifchitz, Y., Avrithis, Y., Picard, S., Bursuc, A.: Dense classification and implanting for few-shot learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 9258–9267 (2019)
20. Maji, S., Rahtu, E., Kannala, J., Blaschko, M., Vedaldi, A.: Fine-grained visual classification of aircraft. ArXiv Preprint ArXiv:1306.5151 (2013)
21. Peng, Y., He, X., Zhao, J.: Object-part attention model for fine-grained image classification. IEEE Transactions on Image Processing **27**(3), 1487–1500 (2017)
22. Ravi, S., Larochelle, H.: Optimization as a Model for Few-Shot Learning. In: Proceedings of International Conference on Learning Representations (2017)
23. Sánchez, J., Perronnin, F., Mensink, T., Verbeek, J.: Image classification with the fisher vector: Theory and practice. International Journal of Computer Vision **105**(3), 222–245 (2013)
24. Schuster, M., Paliwal, K.K.: Bidirectional recurrent neural networks. IEEE Transactions on Signal Processing **45**(11), 2673–2681 (1997). https://doi.org/10.1109/78.650093

25. Sharma, S.R.: Plant disease. https://www.kaggle.com/saroz014/plant-disease (2018)
26. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: Proceedings of the International Conference on Learning Representations. pp. 770–784 (2015)
27. Snell, J., Swersky, K., Zemel, R.S.: Prototypical networks for few-shot learning. In: Conference on Neural Information Processing Systems. pp. 4077–4087 (2017)
28. Sun, X., Xv, H., Dong, J., Zhou, H., Chen, C., Li, Q.: Few-shot learning for domain-specific fine-grained image classification. IEEE Transactions on Industrial Electronics **68**(4), 3588–3598 (2020)
29. 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)
30. Vinyals, O., Blundell, C., Lillicrap, T., Wierstra, D., et al.: Matching networks for one shot learning. Conference on Neural Information Processing Systems **29**, 3630–3638 (2016)
31. Welinder, P., Branson, S., Mita, T., Wah, C., Schroff, F., Belongie, S., Perona, P.: Caltech-UCSD birds 200. California Institute of Technology (2010)
32. 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)
33. Zhangy, W., Liuy, X., Xue, Z., Gao, Y., Sun, C.: Ndpnet: A novel non-linear data projection network for few-shot fine-grained image classification. arXiv preprint arXiv:2106.06988 (2021)