2021 IEEE International Conference on Multimedia and Expo (ICME) | 978-1-6654-3864-3/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICME51207.2021.9428223

SELECTIVE, STRUCTURAL, SUBTLE: TRILINEAR SPATIAL-AWARENESS FOR FEW-SHOT FINE-GRAINED VISUAL RECOGNITION

Heng Wu[1], *Yifan Zhao*[1], *Jia Li*[1,2,*]

[1]State Key Laboratory of Virtual Reality Technology and Systems, SCSE, Beihang University
[2]Peng Cheng Laboratory, Shenzhen, China
{wuheng, zhaoyf, jiali}@buaa.edu.cn

ABSTRACT

Few-shot learning aims to recognize the novel categories from a few examples. However, most of the existing approaches usually focus on general image classification and fail to handle subtle differences between images. To alleviate this issue, we propose a trilinear spatial-awareness network for few-shot-grained visual recognition, called S3Net, which is composed of a spatial selection module, structural pyramid descriptor, and subtle difference mining module. Specifically, we first build the global relation to strengthen the features by spatial selection module. The structural pyramid descriptor then constructs a multi-scale representation for enhancing the rich contextual information by exploiting different receptive fields in the same feature layer. Furthermore, a similarity loss based on local descriptors and a global classification loss is design to help the network learn discrimination capability by handling subtle differences in confused or near-duplicated samples. Extensive experiments on 4 few-shot fine-grained benchmarks demonstrate that our proposed approach is effective and outperforms state-of-the-art models by large margins.

Index Terms— few-shot learning, fine-grained classification, spatial selection, structural pyramid, subtle difference mining

1. INTRODUCTION

Few-shot learning has recently attracted extensive research attention due to its similarity to human perception way on novel concepts in real-world scenarios. Given a few annotated samples, the goal of few-shot learning is to find a generalized feature embedding when facing novel testing samples. With the development of the convolutional neural networks, few-shot learning [1, 2, 3, 4] has made significant breakthroughs in general object recognition tasks. However, learning to recognize fine-grained objects under few-shot settings is still less explored and the process of the human-annotation usually needs expert knowledge of the corresponding field, which are labour- and time-consuming.

Most of the existing works [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12] in few-shot learning mainly consists of two groups: nearest neighbor based approaches [1, 2, 3, 4, 5, 6, 7] and attention mechanism based approaches [8, 9, 10, 11, 12]. The first group attempts to calculate the similarity scores of nearest neighbor samples by designing an effective distance function loss to optimize the model. For example, Prototype Net [4] constructs the center of each class of the support set to compare the samples of its nearest neighbor query set by employing euclidean distance. Relation Net [1] extracts the tensor features to calculate a relation score loss by concatenating the pair features. Li *et al.* [3] further explore the role of local descriptors to align k-nearest neighbor local features and propose Deep Nearest Neighbor Neural Network (DN4). Although these works achieve significant successes in few-shot learning tasks, features extracted from the standard network usually bring in confusing information due to the ignoring of dominative object enhancement.

The second group focuses on enhancing the discriminative feature representation by paying attention to the dominative region. For example, Wu *et al.* [10] introduce a dual correlation attention mechanism and a deformable feature extractor to obtain the intensified features, where these features concatenate together and learn their relationship score by position-aware relation network. Wei *et al.* [8] first tackle the few-shot fine-grained problem by jointing a piecewise mapping technique and bilinear feature learning module. Zhu *et al.* [9] propose to utilize a multi-attention module to highlight the dominative information for few-shot fine-grain image classification. Despite their utilization of attention mechanisms, the contextual understanding of spatial relations, especially the multi-scale representation, is less explored.

To address this problem, in this paper, we investigate three different aspects for exploiting spatial information in fine-grained few-shot learning. Based on this investigation, we proposed a trilinear spatial-awareness network in Fig. 1, namely S3Net for selecting, constructing, and finding subtle differences in spatial dimension for few-shot features. In our proposed S3Net, we first propose to strengthen the content-aware features with a spatial selection module, which builds

∗ Jia Li is the corresponding author.

978-1-6654-3864-3/21/$31.00 ©2021 IEEE

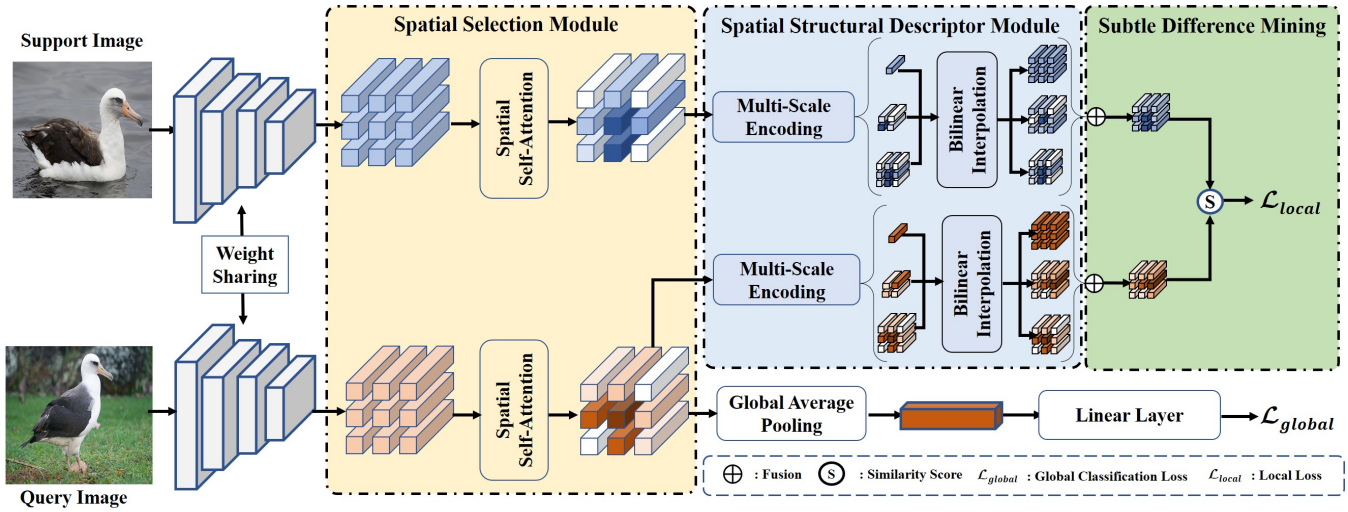

Fig. 1. The framework of our proposed S3Net approach mainly contains three components. 1) Spatial Selection Module: The module is applied to strengthen the spatial information and suppress confusing information. 2) Spatial Structural Descriptor Module: The module can construct rich representation by structural descriptor based on multi-scale features. 3) Subtle Difference Mining: The subtle differences enable the model to mine and learn discrimination capability.

global relations in the spatial dimension. Second, with this spatial-selected feature, we exploit the pyramid pooling encoder for constructing structural descriptors for fine-grained feature measurements, which is inspired by [13, 14] but enhancing the rich representation for global features. Our structural descriptor builds a multi-scale understanding by transforming different receptive fields in the same feature layer. Last but not least, different from the conventional metric learning methods [4] in few-shot learning, we propose to learn different prototypes by exploiting local descriptors with a similarity-based cosine loss and cooperate global classification loss. This helps the network discrimination capability of handling subtle differences for confused or near-duplicated samples. In this way, the proposed approach achieves state-of-the-art results on four few-shot fine-grained benchmarks, *i.e.*, Stanford Dogs [15], Stanford Cars [16], CUB-200-2010 [17], and CUB-200-2011 [18].

Main contributions of our proposed approach are summarized as follows:

• We propose a trilinear spatial-awareness network to address the problem of few-shot fine-grained visual recognition.

• We design a spatial structural descriptor module that can encode and fuse the features of multi-scale information.

•We conduct extensive experiments on four few-shot fine-grained benchmarks and validate that our approach is effective and outperforms state-of-the-art models.

2. THE APPROACH

Consider a support set $\mathcal{S}$, which contains N labeled samples $\{\mathbf{X}_n\}_{n=1}^N \in \mathbb{R}^{C\times W\times H}$ of C channels, width W and height H and corresponding categorical labels $\{y_n\}_{n=1}^N \in [1, 2, \cdots, c]$. The feature maps $\mathbf{X}_n$ are extracted from a standard embedding network. The goal of few-shot learning is to learn embedding parameter Θ and determine the category of samples of the query set $\mathcal{Q}$.

2.1. Spatial Selection Module

The feature maps extracted from the backbone usually employ the receptive field with uniform sliding in the local window. The factor makes it difficult to select fine-grained spatial features. Inspired by the non-local structure [19] for object detection tasks, we attempt to build the global spatial-awareness relationship, with the response at a spatial feature map as a weighted sum of all spatial feature maps, to form an attention map. The constructing attention map can be as a spatial amplifier, which selects effective regional features and suppresses the confusing regional features.

Following this thought, for each regional descriptor $\mathbf{x}_n^m$ $(m = 1, 2, \cdots, M)$ of $\mathbf{X}_n$, its global relationship features with self-attention can be defined as:

$$\widetilde{\mathbf{x}}_n^m = \frac{1}{C}\sum_{j=1}^{M}(\mathcal{F}(\mathbf{x}_n^m)^T, \mathcal{G}(\mathbf{x}_n^j))\mathcal{I}(\mathbf{x}_n^j)^T, \tag{1}$$

where $\mathcal{F}$, $\mathcal{G}$ and $\mathcal{I}$ denote a linear embedding function. C indicates the normalization factor. Further, by employing Eq 1,

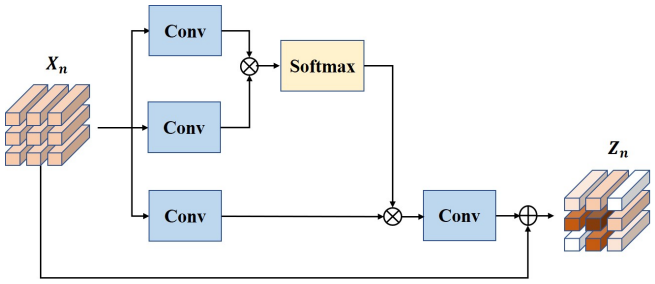

Fig. 2. Spatial self-attention module. Conv represents the 1×1 convolution.

the strengthening features is represented as:

$$\mathbf{z}_n^m = W_{\mathbf{z}} \widetilde{\mathbf{x}}_n^m + \mathbf{x}_n^m, \tag{2}$$

where $W_{\mathbf{z}}$ is the weight matrix parameter to be learned. Therefore, we aggregate each regional feature $\mathbf{z}_n^m$ by Eq. 2 to get the spatial selection features $\mathbf{Z}_n = \{\mathbf{z}_n\}_{m=1}^M \in \mathbb{R}^{C\times W\times H}$, as shown in Fig. 2.

2.2. Structural Pyramid Descriptor

Classical feature representation in few-shot learning usually applies single-scale representation to encode global features [4] or tensor features [3]. Different from the single-scale representation, we inspect and find that multi-scale representation plays an important role in visual comprehension [13, 14]. Keeping this in our mind, we design a structural pyramid descriptor to construct contextual information.

Specifically, we first exploit the spatial pyramid pooling [13] to encode $\mathbf{Z}_n$ (*e.g.*, $1 \times 1, 2 \times 2, ..., P \times P$ scales) to obtain the multi-scale features $\{\mathbf{S}_p\}_{p=1}^P$. The multi-scale features can maintain the structure information in the image. Generally speaking, each region on large-scale features contains more rich information than small-scale features due to the utilization of larger receptive fields. Therefore, for the few-shot fine-grained visual recognition task, how to incorporate both small- and large-scale features in one unified embedding is key. To address this issue, we leverage bilinear interpolation to magnify the large-scale features to the maximal small-scale features of the same local region. This allows the large-scale features to play a uniform encoding effect with the small-scale and to attach more importance. It has the form

$$\widetilde{\mathbf{S}}_p = \mathcal{B}(\mathbf{S}_p), \tag{3}$$

where $\mathcal{B}$ represents the bilinear interpolation operation. These features $\widetilde{\mathbf{S}}_p$ then are aggregated as the dense features $\widetilde{\mathbf{S}}_n$ by

$$\widetilde{\mathbf{S}}_n = \sum_{p=1}^{P} \widetilde{\mathbf{S}}_p. \tag{4}$$

As a result, based on the spatial selection module, dense features constructed from the structural pyramid descriptor incorporate rich and benefit feature representation.

2.3. Subtle Difference Mining

The loss function is crucial for the training model, which enables the feature representations to acquire the corresponding peculiarities. Most of the existing approaches [5, 4] usually calculate the similarity scores by the global features and ignore the local difference mining. To alleviate this issue, we measure the similarity scores between dense features, which encode the fine-grained information, and mine the pair-wise subtle difference for each regional feature representation due to fusion of small- and large-scale features based on spatial selection enhancement.

Specifically, the center $\mathbf{C}_n \in \mathbb{R}^{C\times W\times H}$ of each category of support set can be expressed as follow

$$\mathbf{C}_n = \frac{1}{|\mathcal{S}_n|} \sum_{(\widetilde{\mathbf{S}}_n, y_n)\in\mathcal{S}_n} \widetilde{\mathbf{S}}_n, \tag{5}$$

where $\mathcal{S}_n$ is the set of images labeled with categories n. For the features $\{\mathbf{X}_i \in \mathbb{R}^{C\times W\times H}\}_{i=1}^Q$ extracted from network in query set $\mathcal{Q}$, the local loss (*i.e.*, few-shot loss) between the features $\mathbf{X}_i$ and the center $\mathbf{C}_n$ can be represented by

$$\mathcal{L}_{local} = -\sum_{i=1}^{Q} \log \frac{\exp(-\mathcal{H}(\mathbf{C}_n, \mathcal{K}(\mathcal{J}(\mathbf{X}_i))))}{\sum_{n'=1}^{c} \exp(-\mathcal{H}(\mathbf{C}_{n'}, \mathcal{K}(\mathcal{J}(\mathbf{X}_i))))}, \tag{6}$$

where $\mathcal{J}$ and $\mathcal{K}$ express the spatial self-attention operation and structural pyramid descriptor module. $\mathcal{H}$ is the metric distance function (*e.g.*, cosine distance). Moreover, the global loss with a cosine linear classifier is utilized to assist local loss. Note that different from the dot product, cosine linear classifier between the features and the weight parameter for each category can increase the inter-class variations and reduce the intra-class variations. It is defined by

$$\mathcal{L}_{global} = -\sum_{i=1}^{Q} \log \frac{\exp(\cos(W, GAP(\mathbf{X}_i)))}{\sum_{i'=1}^{c} \exp(\cos(W, GAP(\mathbf{X}_{i'})))}, \tag{7}$$

where $\cos$ and GAP are the cosine distance and global average pooling. W is the weight parameter of each category. Therefore, the cooperative loss is represented as follow

$$\mathcal{L}_{total} = \mathcal{L}_{global} + \lambda \mathcal{L}_{local}, \tag{8}$$

where λ is the hyperparameter, which maintains the balance both the $\mathcal{L}_{global}$ and $\mathcal{L}_{local}$. During the test stage, the local loss $\mathcal{L}_{local}$ only is used to calculate the similarity scores to classifier the novel query set.

Table 1. Performance comparison of the state-of-the-art and our work on benchmark datasets. Results are the average accuracies with 95% confidence intervals on 5-way 1-shot and 5-way 5-shot tasks. Bold and underline in each column are marked with the best and runner-up results, respectively.

Method	5-way Acc (%)									
	Stanford Dogs		Stanford Cars		CUB-200-2010		CUB-200-2011		Avg	
	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
Relation Net [1]	43.29±0.46	55.15±0.39	47.79±0.49	60.60±0.41	58.99±0.52	71.20±0.40	–	–	50.02	62.32
ConvaMNet [2]	49.10±0.76	63.04±0.65	56.65±0.86	71.33±0.62	52.42±0.76	63.76±0.64	–	–	52.72	66.04
DN4 [3]	45.73±0.76	66.33±0.66	61.51±0.85	**89.60±0.44**	53.15±0.84	**81.90±0.60**	–	–	53.46	79.28
LRPABN [12]	45.72±0.75	60.94±0.66	60.28±0.76	73.29±0.58	–	–	–	–	53.00	67.12
Matching Net [5]	35.80±0.99	47.50±1.03	34.80±0.98	44.70±1.03	45.30±1.03	59.50±1.01	–	–	38.63	50.57
Prototypical Net [4]	37.59±1.00	48.19±1.03	40.90±1.01	52.93±1.03	37.36±1.00	45.28±1.03	–	–	38.61	48.80
MAML [6]	44.84±0.31	58.61±0.30	47.25±0.30	61.11±0.29	58.13±0.36	71.51±0.30	–	–	50.07	63.74
adaCNN [20]	42.16±0.43	54.12±0.39	41.88±0.40	49.87±0.37	56.76±0.50	61.05±0.44	–	–	46.93	55.01
GNN [7]	46.98±0.98	62.27±0.95	55.85±0.97	71.25±81.53	51.83±0.98	63.69±0.94	–	–	51.55	64.74
MattML [9]	54.84±0.53	71.34±0.38	66.11±0.54	82.80±0.28	–	–	66.29±0.56	80.34±0.30	62.41	78.16
Ours	**63.56±0.49**	**77.54±0.35**	**71.19±0.50**	84.40±0.34	**64.27±0.50**	78.02±0.38	**72.30±0.51**	**84.23±0.33**	**67.83**	**81.05**

3. EXPERIMENTS

3.1. Experiment Setting

Dataset. We use four datasets (*i.e.*, Stanford Dogs [15], Stanford Cars [16], CUB-200-2010 [17], and CUB-200-2011 [18]). Stanford Dogs contains 120 subcategories (70 categories for training, 20 categories for validation, and 30 categories for testing) of dogs with 20,580 images. Stanford Cars includes 196 subcategories (130 categories for training, 17 categories for validation, and 49 categories for testing) of cars with 16,185 images. CUB-200-2011, consisting of 200 subcategories (130 categories for training, 20 categories for validation, and 50 categories for testing) of birds with 16,185 images, is the extended version of CUB-200-2010 with 6033 images. For the four benchmarks, we follow [3] to split them.
Network architectures. Following the previous works [9], we adopt the standard feature extraction network $Conv4$. This contains four convolutional modules, each of which consists of a convolutional layer with 3×3 size followed by batch normalization layer and ReLU layer. Besides, a 2×2 max-pooling is appended for the first two convolutional modules.
Implementation details. All experiments are implemented by Pytorch with one GTX1080Ti GPU. SGD is adopted as an optimizer with an initial learning rate of 0.1, which decreases to 0.006 at 60 epoch and then times 0.2 every 10 epoch. Each epoch contains 1200 episodes. During the test, 2000 episodes are randomly sampled from datasets and the accuracy with 95% confidence intervals is reported on these episodes.

3.2. Comparison with the State-of-the-art

To evaluate the superiority of our approach, we conduct an comparison with the state-of-the-art models on benchmark datasets. The experimental results are reported in Tab. 1. As can be seen in Tab. 1, compared with the previous approaches, our approach achieves the highest performance on four few-shot fine-grained datasets. Specifically, for 5-way 1-shot task on benchmarks, we can increase by 17.81%, 14.37%, 29.22%, and 5.42% on average over Relation Net [1], DNN [3], Prototype Net [4], MattML [9]. For 5-way 5-shot task, we also can gain by 18.73%, 1.77%, 32.25%, and 2.89 % improvement with a large margin on average over Relation Net [1], DNN [3], Prototype Net [4], MattML [9]. The results mean that our approach enables the network to better extract the subtle discriminant features, which are a benefit for the similarity measure of fine-grained images.

Table 2. Ablation studies of the linear classifier.

Dataset	Method	5-way Acc (%)	
		1-shot	5-shot
Stanford Dogs	BL	57.35±0.49	72.74±0.37
	BL++	**60.48±0.49**	**74.27±0.37**
Stanford Cars	BL	61.73±0.49	80.42±0.33
	BL++	**66.91±0.48**	**81.68±0.32**
CUB-200-2011	BL	62.21±0.49	78.20±0.37
	BL++	**68.46±0.50**	**82.06±0.35**

3.3. Ablation Analysis

Impact of linear classifier. We study the linear classifier on the few-shot fine-grained recognition, as shown in Tab. 2. 'BL' and 'BL++' are the baseline and baseline++, where the baseline is the local loss and linear classifier which baseline++ is the local loss and cosine linear classifier. From Tab. 2, we obverse that 'BL++' achieves higher performance than 'BL' in 5-way 1-shot and 5-shot tasks. This means that the model benefits from a cosine classifier, which can better

Table 3. Ablation studies of the spatial self-attention module.

Dataset	BL++	SSA	5-way Acc (%)	
			1-shot	5-shot
Stanford Dogs	✓		54.34±0.44	71.11±0.38
	✓	✓	**63.56±0.49**	**77.54±0.35**
Stanford Cars	✓		65.53±0.48	81.21±0.36
	✓	✓	**71.19±0.50**	**84.23±0.33**
CUB-200-2011	✓		53.14±0.43	73.12±0.38
	✓	✓	**72.30±0.51**	**84.40±0.34**

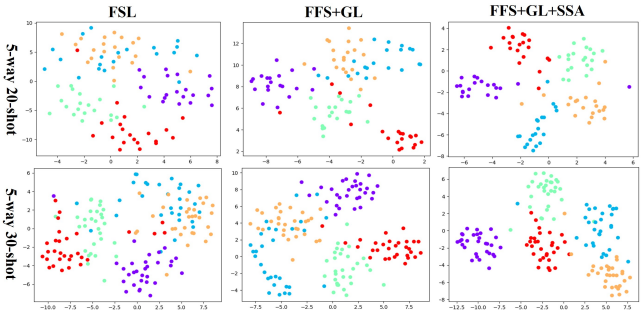

Fig. 3. Illustration of dimension reduction in Cub-200-2011 dataset. The images are randomly sampled in a 5-way 20-shot and 30-shot tasks and each color indicates different classes.

enhance the fine-grained features of the subcategory.

Impact of spatial self-attention module. Attention mechanism is utilized to focus on the region of interest. We evaluate the influence of spatial self-attention module for few-shot fine-grained image classification in Tab. 3. 'SSA' is the spatial self-attention module. As can be seen in Tab. 3, the combination of 'BL++' and SSA obtain improvement with a large margin in both 5-way 1-shot and 5-way 5-shot tasks. We also conduct visualization by CAM [21] and t-sen [22] for the two modules, as presented in Fig. 3 and Fig. 4. The corrected features by the spatial self-attention module demonstrate the ability to grasp the structure information.

Impact of spatial pyramid pooling. Spatial pyramid pooling can keep spatial information by pooling in different spatial bins. Based on baseline++ and spatial attention, we discuss the effect of improved spatial pyramid pooling for few-shot visual classification, as shown in Tab. 4. Four grid scales (1×1, 7×7, 10×10, 21×21) are exploited for spatial pyramid pooling block. 'GAP' and 'BI' mean global average pooling and bilinear interpolation operation, respectively. From Tab. 4, we observe that 'SPP+BI' obtains the best performance in both 5-way 1-shot and 5-shot tasks. The reason may that the spatial pyramid aggregation reduces confusing information and enhances the utilization of effective features.

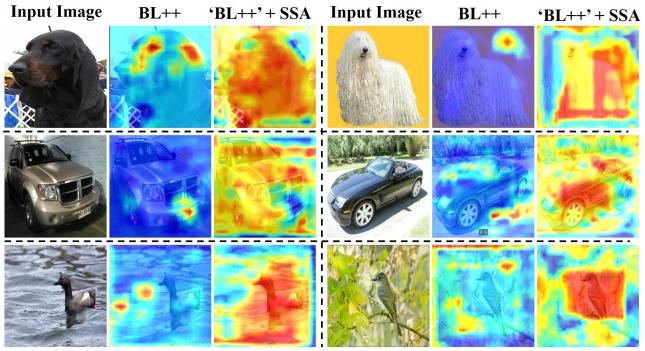

Fig. 4. Visual illustration of feature maps from the last layer of the model on few-shot benchmark datasets.

Table 4. Ablation studies of the spatial pyramid pooling.

Dataset	Method	5-way Acc (%)	
		1-shot	5-shot
Stanford Dogs	GAP	63.53±0.49	77.52±0.35
	SPP	63.52±0.49	77.51±0.35
	SPP+BI+GAP	63.52±0.49	77.51±0.35
	SPP+BI	**63.56±0.49**	**77.54±0.35**
Stanford Cars	GAP	71.02±0.50	84.06±0.33
	SPP	71.09±0.50	84.10±0.33
	SPP+BI+GAP	71.13±0.50	84.12±0.33
	SPP+BI	**71.19±0.50**	**84.23±0.33**
CUB-200-2011	GAP	72.28±0.51	84.18±0.34
	SPP	72.22±0.51	84.19±0.34
	SPP+BI+GAP	72.15±0.51	84.20±0.34
	SPP+BI	**72.30±0.51**	**84.40±0.34**

4. CONCLUSION

In this paper, we propose a trilinear spatial-awareness network (namely S3Net) to overcome the challenge of few-shot fine-grained recogintion. In S3Net, a spatial selection module first is utilized to strengthen the content-aware features by building the spatial global relation. We then construct the multi-scale features by a structural pyramid for enhancing rich representation in global features. Finally, we design a local similarity loss by the subtle differences and cooperate a global loss to learn the discriminative features in an end-to-end manner. Extensive experiments on four benchmarks show the effectiveness and superiority of our proposed approach and achieve state-of-the-art results.

5. ACKNOWLEDGEMENT

This work was supported by grants from National Natural Science Foundation of China (No.61922006) and Baidu Academic Collaboration Program.

6. REFERENCES

[1] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales, "Learning to compare: Relation network for few-shot learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1199–1208.

[2] Wenbin Li, Jinglin Xu, Jing Huo, Lei Wang, Yang Gao, and Jiebo Luo, "Distribution consistency based covariance metric networks for few-shot learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, vol. 33, pp. 8642–8649.

[3] Wenbin Li, Lei Wang, Jinglin Xu, Jing Huo, Yang Gao, and Jiebo Luo, "Revisiting local descriptor based image-to-class measure for few-shot learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7260–7268.

[4] Jake Snell, Kevin Swersky, and Richard Zemel, "Prototypical networks for few-shot learning," in *Advances in neural information processing systems*, 2017, pp. 4077–4087.

[5] Qi Cai, Yingwei Pan, Ting Yao, Chenggang Yan, and Tao Mei, "Memory matching networks for one-shot image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4080–4088.

[6] Chelsea Finn, Pieter Abbeel, and Sergey Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," *arXiv preprint arXiv:1703.03400*, 2017.

[7] Victor Garcia and Joan Bruna, "Few-shot learning with graph neural networks," *arXiv preprint arXiv:1711.04043*, 2017.

[8] Xiu-Shen Wei, Peng Wang, Lingqiao Liu, Chunhua Shen, and Jianxin Wu, "Piecewise classifier mappings: Learning fine-grained learners for novel categories with few examples," *IEEE Transactions on Image Processing*, vol. 28, no. 12, pp. 6116–6125, 2019.

[9] Yaohui Zhu, Chenlong Liu, and Shuqiang Jiang, "Multi-attention meta learning for few-shot fine-grained image recognition," in *Twenty-Ninth International Joint Conference on Artificial Intelligence and Seventeenth Pacific Rim International Conference on Artificial Intelligence*, 2020, pp. 1090–1096.

[10] Ziyang Wu, Yuwei Li, Lihua Guo, and Kui Jia, "Parn: Position-aware relation networks for few-shot learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 6659–6667.

[11] Fusheng Hao, Fengxiang He, Jun Cheng, Lei Wang, Jianzhong Cao, and Dacheng Tao, "Collect and select: Semantic alignment metric learning for few-shot learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 8460–8469.

[12] Huaxi Huang, Junjie Zhang, Jian Zhang, Jingsong Xu, and Qiang Wu, "Low-rank pairwise alignment bilinear network for few-shot fine-grained image classification," *IEEE Transactions on Multimedia*, 2020.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 9, pp. 1904–1916, 2015.

[14] Yingying Zhu, Jiong Wang, Lingxi Xie, and Liang Zheng, "Attention-based pyramid aggregation network for visual place recognition," in *Proceedings of the 26th ACM international conference on Multimedia*, 2018, pp. 99–107.

[15] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Fei-Fei Li, "Novel dataset for fine-grained image categorization: Stanford dogs," in *Proc. CVPR Workshop on Fine-Grained Visual Categorization (FGVC)*, 2011, vol. 2.

[16] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei, "3d object representations for fine-grained categorization," in *Proceedings of the IEEE international conference on computer vision workshops*, 2013, pp. 554–561.

[17] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona, "Caltech-UCSD Birds 200," Tech. Rep. CNS-TR-2010-001, California Institute of Technology, 2010.

[18] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie, "The caltech-ucsd birds-200-2011 dataset," 2011.

[19] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He, "Non-local neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7794–7803.

[20] Tsendsuren Munkhdalai, Xingdi Yuan, Soroush Mehri, and Adam Trischler, "Rapid adaptation with conditionally shifted neurons," in *International Conference on Machine Learning*, 2018, pp. 3664–3673.

[21] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba, "Learning deep features for discriminative localization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2921–2929.

[22] Laurens van der Maaten and Geoffrey Hinton, "Visualizing data using t-sne," *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.