IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

SPECIAL SECTION ON CLOUD - FOG - EDGE COMPUTING IN CYBER-PHYSICAL-SOCIAL SYSTEMS (CPSS)

Received February 27, 2020, accepted March 15, 2020, date of publication March 23, 2020, date of current version March 31, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2981950

Self-Layer and Cross-Layer Bilinear Aggregation for Fine-Grained Recognition in Cyber-Physical-Social Systems

YINGQIONG PENG[1,3], YUXIA SONG[2], WEIJI HUANG[3], HONG DENG[1,4], YINGLONG WANG[2], QI CHEN[2], MUXIN LIAO[2], AND JING HUA[4]

[1]Colleges and Universities of Jiangxi Province for Key Laboratory of Information Technology in Agriculture, Jiangxi Agricultural University, Nanchang 330045, China
[2]College of Computer and Information Engineering, Jiangxi Agricultural University, Nanchang 330045, China
[3]Academic Affairs Office, Jiangxi Agricultural University, Nanchang 330045, China
[4]Software Institute, Jiangxi Agricultural University, Nanchang 330045, China

Corresponding author: Jing Hua (15124521@qq.com)

This work was supported in part by the National Science Foundation of China under Grant 61861021, in part by the Science and Technology Plan Project of Jiangxi Provincial Department of Education under Grant GJJ190208, and in part by the Jiangxi Agricultural University Graduate Special Fund Innovative Project under Grant NDYC2019-S008.

ABSTRACT Cyber-Physical-Social Systems (CPSS) integrates cyber, physical and social spaces together, which makes our lives more convenient and intelligent by providing personalized service. In this paper, we will provide CPSS service for fine-grained recognition. Fine-grained visual recognition is a hot but challenging research in computer vision that aims to recognize object subcategories. The reason why it is challenging is that it extremely depends on the subtle discriminative features of local parts. Recently, some bilinear feature based methods were proposed, and the experimental results show state-of-the-art performance. However, most of them neglect the spatial relationships of part-region feature among multiple layers. In this paper, a novel approach of Self-layer and Cross-layer Bilinear Aggregation(SCBA) is proposed for fine-grained recognition. Firstly, a self-layer bilinear feature fusion module is proposed to model the spatial relationship of feature at the same layer. Secondly, we propose a cross-layer bilinear feature fusion module to capture the inter-layer interreaction of information to boost the ability of feature representation. In summary, the method we proposed not only can learn the correlations among different layers but the same layer, which makes it efficient and the experimental results show that it achieves state-of-the-art accuracy on three common fine-grained image datasets.

INDEX TERMS Cross-layer bilinear, fine-grained recognition, self-layer bilinear.

I. INTRODUCTION

With the deepening of application of network, especially, the internet plus, big data, cloud computing, internet of things, information and physical systems are further integrated, the network and human society are seamlessly integrated, forming a more complex system that integrates Human, machine and information. We called it Cyber-Physical-Social Systems (CPSS). Until now, CPSS has been applied to many fields [1]–[6]. In this paper, we realize fine-grained image classification in CPSS. Fine-grained image recognition is also called subcategory classification, which is a classical research topic in computer vision. However, because of those uncertainties, including occlusion, illumination, pose, complex background, and etc, leading to large variance of the same subcategory and high similarity of different subcategory in fine-grained images (see Figure 1). Thus, fine-grained image recognition is a more challenging task than ordinary classification. To overcome difficulties mentioned above, some part-based methods were proposed [7]–[9]. These methods depend upon manually labeled information, such as bounding boxes and part localization. Although they can improve the classification accuracy, it is time consuming and labor extensive. With the mature and development of technology, some weakly-supervised [10]–[13] based methods were proposed, which only rely on the class labels. They can localize the key regions and extract discriminative features automatically via

The associate editor coordinating the review of this manuscript and approving it for publication was Xiaokang Wang.

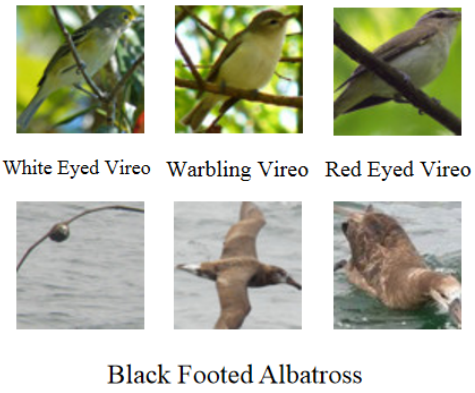


FIGURE 1. Comparison of inter- and intra-class on CUB-200-2011 dataset. The images in first row belong to different class, and the images in second row are the same class.

convolution neural network(CNN). In a CNN, the shallow layer extracts low-level features, such as texture and detailed features, the deep layer extracts high-level features, such as contour, shape and key features. Consequently, the layer goes deeper, the features are more representative. Thus, some methods [14] fuse the output of the last convolution layer with concatenation, sum pooling, etc. for better feature representation. But there are some limitations that these methods:

- neglect the activations of intermediate layer, which may be useful for classification.
- overlook the relationships among different layers that can strengthen the ability of feature representation.

Considering these limitations mentioned above, some methods [15]–[17] propose bilinear pooling to obtain more powerful feature representations. Some of them capture the correlations among different layers by multiplying the features from two different layers, some obtain correlationship of every location on the feature map from the same layer by self-channel multiplication. However, these methods consider the relationships among different layers and the same layer separately. To address this issue, we come up with a novel approach named Self-layer and Cross-layer Bilinear Aggregation(SCBA) to extract the self-channel bilinear features and cross-channel bilinear features among multiple layers. Firstly, we design a module named self-layer bilinear feature fusion to model the spatial relationships at the single layer. Secondly, we propose a cross-layer bilinear feature fusion module to mine the inter-layer interaction information. Finally, we can construct a powerful representation by fusing self-layer and cross-layer bilinear features. In summary, our contributions are briefed as follows:

- We exploit a plain but valid self-layer and cross-layer bilinear feature representation method that simultaneously obtains the self-channel relationships in single layer and inter-layer interaction of features among multiple layers.
- We propose Self-layer and Cross-layer Bilinear Aggregation (SCBA) model to strengthen the representation ability of bilinear features.
- We have performed extensive experiments on three fine-grained image datasets and the experimental results show the superiority of our approach.
- We provide CPSS service for fine-grained recognition, which can solve the task more conveniently and intelligently.

The rest of this paper is arranged as follows. Section II introduces the related work. Section III presents the proposed method. Section IV shows the experiments on three datasets and the result analysis, followed by conclusion in Section V.

II. RELATED WORK

In this section, we will introduce the recent methods from two interesting perspectives that related to our work, containing weakly-supervised fine-grained image recognition and feature fusion.

A. WEAKLY-SUPERVISED FINE-GRAINED IMAGE RECOGNITION

Fine-grained classification is an active research topic in the field of computer vision. The called weakly-supervised based methods is the mainstream in the future, which only use class labels without any bounding boxes or part annotations. This reduces some money, time and labor at some extent. Generally speaking, the aims of fine-grained classification methods are high accuracy and small computational cost. In order to get better performance, Chen *et al.* [18] exploited semantic guided attention mechanism to learn more discriminative regions at each level by incorporating the predicted score vector of the higher level as prior knowledge. Wang *et al.* [19] proposed to add supervision information to filters for optimizing discriminative part detectors and further localizing the key regions. Pang *et al.* [20] firstly proposed a hybird part localization method to generate accurate part proposals, and then, updated the segmentation outputs and the part proposals iteratively for better foreground segmentation. Yang *et al.* [21] exploited self-supervision learning mechanism to locate the informative regions without any annotations in a end-to-end fashion. However, these methods mentioned above always rely on discriminative features, and neglect other part features, which may be harmful for classification results. Thus, Ge *et al.* [22] put forward a novel method to learn complementary features, not only the discriminative features. Actually, feature extraction and feature representation both have contributions to final classification results. Therefore, some methods perform bilinear pooling operation to get more powerful representation, which calculates the second-order statistics of local features that can obtain better feature representation. For example, Li *et al.* [23] proposed fine-brunch and coarse-brunch to obtain different level bilinear features respectively, followed by softmax loss layers with semantic information from hierarchical labels. Although these methods can enhance the feature representation ability, and further improve the classification accuracy, while neglect the size of parameters. Consequently, some methods were proposed to reduce dimension and parameters. Gou *et al.* [24] proposed sub-matrix square-root layer to normalize the

output of the convolution layer before bilinear pooling to depress the feature dimension. Lebedev *et al.* [25] proposed a light-weight Network that consists of a convolution network and a non-parameter classifier, leading to less computational cost.

B. FEATURE FUSION

Feature fusion can be defined as aggregating all extracted local features into one compact feature. In practice, constructing a powerful feature representation plays an important role in computer vision tasks. Some methods combine with the other modal information to learn better features. For example, Zheng *et.al.* [26] proposed a novel method that combines the contextual information with multiviewpoint depth images to construct multiviewpoint context-aware representation for scene classification. Kim *et al.* [27] exploited tri-modal information to produce confidence of the disparity for stereo confidence estimation. However, these methods neglect the spatial relationship among features. Thus, some recent works attempt to excavate spatial information among channels and different convolution layer. For example, Chen and Li [28] exploited CA-Fuse module and level-wise supervision to capture complement among different modal and level respectively for RGB-D Salient Object Detection. Huang *et al.* [29] came up with a framework that combines saliency maps corresponding to different channels generated independently via an adaptive uncertainty weighting approach for saliency detection. The methods above are only pay attention on spatial information, while the methods below consider the problem of multiscale features fusion. Chen *et al.* [30] proposed a new method named MGMR to refine the initial segmentation mask provided by pre-processing for RGB-D perception. Hu *et al.* [31] came up with a new network to learn discriminative features from different scale images and then aggregated them to obtain multiscale feature representation for Plant Leaf Recognition. However, the methods mentioned above focus on feature representation ability and neglect the number of model parameters and dimension of feature representation. Li *et al.* [32] introduced a new structure to aggregate multiscale deep features to enhance feature representation ability and speed up experiment process for real-time semantic segmentation. To address these problem, Yu and Salzmann [33] proposed a parametric compression strategy to produce more compact representations than previous compression tactics. Gao *et al.* [34] proposed NDDR layer to fuse single-task features by layerwise feature fusion for multitask feature learning. Sindagi and Patel *et al.* [35] proposed a new method to combine multiscale information at multiple levels and employed a principled way to increase the effectiveness of this fusion method. These methods were widely applied to computer visual tasks and achieves state-of-the-art performance. For further works, we may exploit these methods to optimize model for better result.

According to the brief introduction of previous methods above, we can find that though these methods take various measures to extract finer-grained features and improve the ability of feature representation, they neglect some not discriminative but useful features and computational cost, leading to the decrease of generalization ability of the method and dimension explosion. In this paper, we are not just focusing on discriminative features, and project high dimension feature to lower dimension for less computational cost.

III. PROPOSED APPROACH

In this section, we develop a self-layer and cross-layer bilinear aggregation(SCBA) model to overcome those limitations mentioned above. Firstly, we introduce the architecture of the SCBA model in Sect. A. Then, based on bilinear pooling, we describe the general formulation of self-layer bilinear feature fusion to jointly represent features from the same convolution layer in Sect. B. Next, we design a cross-layer bilinear feature fusion method to jointly represent features from different convolution layers in Sect. C. Finally, we can obtain complementary and fine-grained features by fusing these features, which are conducive to boost the ability of feature representaion.

A. SCBA MODEL ARCHITECTURE

In this subsection, we introduce our SCBA model architecture, which is able to represent features with their spatial relationships by self-layer and cross-layer bilinear feature fusion. SCBA model contains three modules, i.e., a resnet module for extracting feature maps, a self-layer bilinear feature fusion module for obtaining the spatial relationships from the same convolution layer, and a cross-layer bilinear feature fusion module for building inter-layer interactions from the different convolution layers.

B. SELF-LAYER BILINEAR FEATURE

As we all know, subcategory recognition tasks often have alike feature and can only be recognized by subtle differences in local-region features. Bilinear pooling is an effective method on fine-grained recognition task to capture pairs of characteristic relationships. However, most models based on bilinear pooling only concentrate on obtaining the features representation from individual convolution layer while entirely neglecting other layers of information. The features of single convolution layer are incomplete because each object part has multiple attributes which are keys to discriminate subcategories.

Practically in most cases, we should completely expect multiple factors of part feature to recognize the category for an input image. Thus, to extract the detailed information of fine-grained image, we propose a multilayer self-layer bilinear feature fusion approach that can obtain self-channel relationship of features, and fuse features of multiple convolution layers to get rich and representative feature.

Accordingly, assuming that the output of the convolution layers is a high-dimensional feature map with a dimension of $c \times h \times w$, where c, h and w indicate the number of channel, height, and width respectively. We reshape this feature into a matrix with a shape of $c \times hw$, which is denoted as $X \in R^{c \times hw}$. Then, we further integrate spatial relationship into

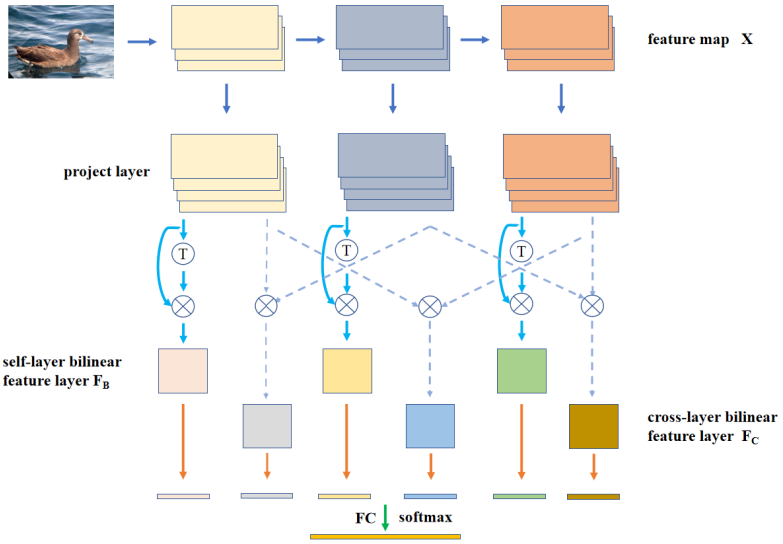


FIGURE 2. Illustration of our SCBA model for fine-grained recognition. The picture on the top left corner is the input image, followed by the features of different convolution layers in the CNN. First, the features from different layers are expended into a high-dimensional space by independent linear mapping to capture attributes of different object parts and then integrated by dot production to model the inter-layer and intra-layer interaction of part attributes. After that the high-dimensional features are compressed into compact ones by performing sum pooling.

feature maps by conducting dot production over X and X^T, the formulation of self-layer bilinear feature is defined by:

$$B = XX^T \tag{1}$$

where the $B \in R^{c\times c}$. Actually, We can transform the equation to $B = \sum_{i=1}^{hw} X_i X_i^T$, the $X_i \in R^{c\times 1}$ means the feature at location i, so $X_i X_i^T$ indicates the correlation matrix of feature X_i and B is the sum of those correlation matrices.

Based on this, to further learn rich representation feature, we fuse multilayer self-layer bilinear features, the formulation can be defined as:

$$F_B = f_b(B_l, B_m, B_n) \tag{2}$$

where $F_B \in R^{c\times c}$, and f_b is a function contains a reshape operation and a concatenation operation. B_l, B_m, B_n present l_{th}, m_{th}, n_{th} self-layer bilinear features in self-layer bilinear layer respectively.

It is worth noting that the features from different convolution layers are expanded into high-dimensional space by independent linear mapping to capture attributes of different object parts. It is expected that the convolution activations and project activations encode global and local feature of object respectively. It is highly in accordance with the human coarse-to-fine mechanism: human always see the global "evil" of object before making out the detailed features.

C. CROSS-LAYER BILINEAR FEATURE

Multilayer self-layer bilinear feature fusion introduced in Sect. B is presentative and valid, as it obtains better representation ability than general bilinear feature while maintaining the same training parameters. This inspires us to exploit multilayer bilinear feature interactions to capture the discriminative information in local region. Based on this, we enlarge the cross-layer bilinear feature fusion among multiple layers to integrate more other convolution layers, moreover improving the ability of features representation. In this subsection, we design a cross-layer bilinear feature fusion approach to involve more features from different convolution layer by conducting dot production over X and Y^T. The formulation of cross-layer bilinear feature is defined by:

$$C = XY^T \tag{3}$$

where $C \in R^{c\times c}$ is the cross-layer bilinear feature, which contains inter-layer interaction of features. Specifically, X and Y are the feature maps from different convolution layers.

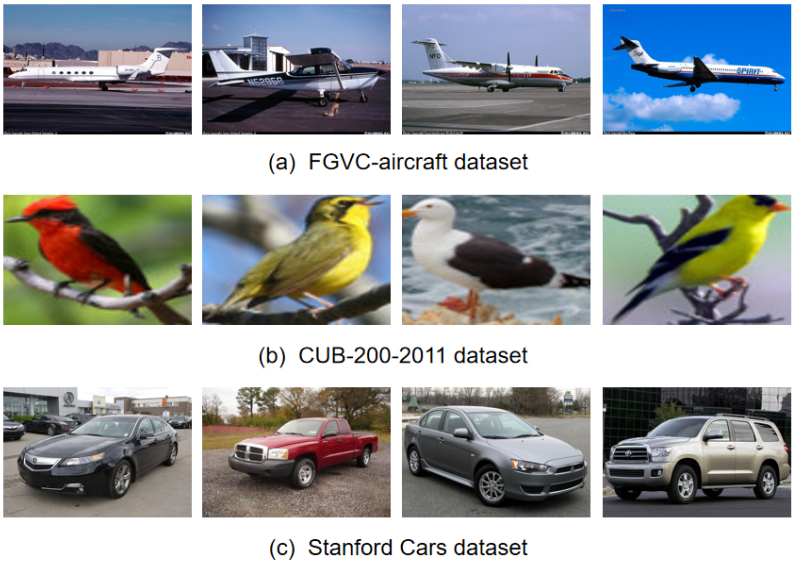
Based on this, to further learn complementary features from the intermediate convolution layers for better performance, we fuse multiple cross-layer bilinear features, the formulation can be defined as:

$$F_C = f_c(C_l, C_m, C_n) \tag{4}$$

where $F_C \in R^{c\times c}$, and f_c is the same as f_b. C_l, C_m, C_n present l_{th}, m_{th}, n_{th} cross-layer bilinear features in cross-layer bilinear layer respectively. The overall flow chart of the SCBA model is shown in Figure 2.

IV. EXPERIMENT AND ANALYSIS

In this section, We conduct experiments on the SCBA model and evaluated the fine-grained recognition performance of the model. Firstly, three experimental datasets and implementation details of SCBA model are introduced in Sect. A. Then, in Sect. B, the effectiveness of each module was verified by the SCBA model configuration experiments. Finally, The comparison results with state-of-the-art models are presented in Sect. C.



(a) FGVC-aircraft dataset

(b) CUB-200-2011 dataset

(c) Stanford Cars dataset

FIGURE 3. Some images of three datasets.

TABLE 1. Summary statistics of datasets.

Datasets	Category	Training	Testing
CUB-200-2011 [36]	200	5994	5794
Stanford Cars [37]	196	8144	8041
FGVC-Aircraft [38]	100	6667	3333

A. DATASET AND IMPLEMENTATION DETAIL

1) DATASET

We have carried out experiments on three usual datasets and summarize the specific statistics that contain the number of category and data segmentation of training and testing in Table 1. Note that we do not use annotation information other than the category labels in our experiments.

What's more, we show some images of three datasets in Figure 3, including CUB-200-2011, Stanford Cars and FGVC-aircraft dataset.

2) IMPLEMENTATION DETAIL

We evaluate our SCBA model using ResNet-50 [39] baseline model which pretrained on ImageNet [40] dataset. It is worth noting that our SCBA model can also be implemented to other models, such as VGG-16 [41] and GoogleNet [42]. The input image size is 448×448. We finetune the entire network using SGD optimizer with momentum of 0.9 and weight decay of 5×10^{-4}, and the number of batch size is 16, and initial learning rate is 10^{-3}. All experiments are implemented using open-source toolbox Pytorch [43] and executed on a server using Tesla V100 GPUs. The source code will provide at https://github.com/seabearlmx/SCBA.

B. CONFIGURATION OF SCBA MODEL

We focus on three layers in ResNet-50, including *relu*5_1, *relu*5_2 and *relu*5_3, since they contain more subtle information of part compared with shallower layers. Then, we project d varies from 2048 to 512 as decreasing d leads to lesser computational expense for our server and SCBA is suitable for $d = 512$. Thus, $d = 512$ is used for SCBA in our following experiments considering the memory size of GPUs.

TABLE 2. Classification accuracy of different feature fusion on CUB-200-2011 dataset.

Method	Accuracy
CC model	86.32%
CL model	87.43%
SCBA model	88.26%

Then, quantitative experiments are conducted on CUB-200-2011 [36] dataset to explore the influential factors of feature fusion. We respectively consider for self-layer (SL) and cross-layer (CL) bilinear feature and fusing them (SCBA). The results show in Table 2 that demonstrate that the performance improvement of the model is basically due to the combination of self-layer and cross-layer. As the SCBA already show the best performance, thus we apply SCBA model in all the experiments in Sect. C.

C. COMPARISON WITH STATE-OF-THE-ART METHODS

The recognition accuracy is shown in Table 3. The table is divided into three parts: the first part lists the name of some weakly supervised methods; the second part lists the name of datasets; the third part lists accuracy of those methods in datasets respectively.

1) RESULTS ON CUB-200-2011

The CUB [36] dataset provides a rich of annotations, but the only annotation we use is category label. In Table 3, we can see that SCBA achieves better result compared with others weakly supervised approaches, even some supervised methods, e.g., Mask-CNN [44] and HSnet [45], which proves the validity of our model. Compared with HSE [18] which used hierarchical semantic embedding to learn stronger representation of fine-grained feature, we improve relative accuracy with 0.16% by our SCBA. We even surpass TASN [46] and DCL [47] which were the state-of-the-art weakly supervised models recently proposed, with 0.36% and 0.46% relative accuracy improves, respectively. Compared with baseline model, containing B-CNN [15], Improved B-CNN [48] and

TABLE 3. Comparison results on CUB-200-2011 dataset. The "✓" means the method used annotations during training or testing (the same below).

Method	Anno	Accuracy
Part RCNN [8]	✓	76.37%
SPDA-CNN [51]	✓	85.1%
B-CNN [15]	✓	85.1%
PN-CNN [9]	✓	85.4%
Mask-CNN [44]	✓	87.3%
HSnet [45]	✓	87.5%
TASN [46]		87.9%
DCL [47]		87.8%
HSE [18]		88.1%
NTS [21]		87.5%
HBP [16]		87.15%
GP [12]		85.8%
PC [50]		86.87%
KP [52]		86.2%
MAMC [13]		86.5%
B-CNN [15]		84.0%
Improved B-CNN [48]		85.8%
Boost CNN [49]		86.2%
SCBA(ours)		88.26%

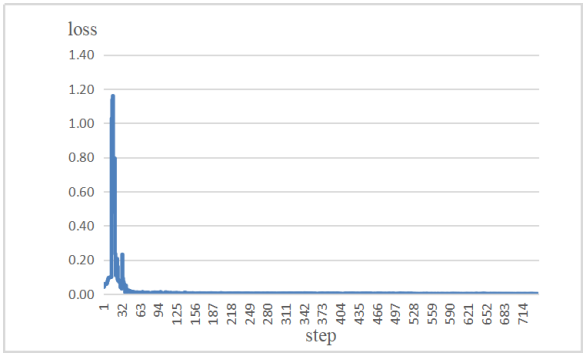


FIGURE 4. The training loss on CUB-200-2011 dataset.

Boost CNN [49], the advantage that we have achieved is primarily the benefit from the channel relationship of feature and the feature fusion of multiple layers. We also surpass GP [12], PC [50], and MAMC [13] which respectively exploited a novel pooling method to solve visual burstiness phenomenon of bilinear pooling method, used a novel optimization procedure to reduce overfitting, and applied the multi-attention multi-class constraint to regulate multiple object parts among different input images. Although HBP [16] described similar approach to obtain inter-layer feature interaction. Our approach can get better accuracy since feature fusion of multilayer self-layer and cross-layer bilinear feature. Note that SCBA outperforms SL and CL, which means that our approach can extract the rich information by fusing multiple layers feature.

We carried out some experiments on CUB-200-2011 dataset. The line chart shows the fluctuation of loss during training, the drift of loss during testing, and the fluctuation of accuracy during testing in Figure 4, Figure 5, Figure 6 respectively.

Figure 4, 5 and 6 show that the model trained nearly 37 epoch (20 steps per epoch), and the training loss suddenly increased to about 1.15, and then decreased to approach

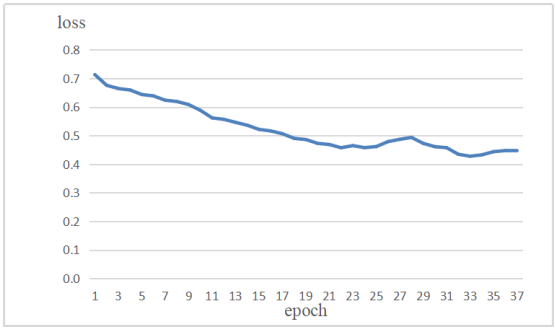


FIGURE 5. The testing loss on CUB-200-2011 dataset.

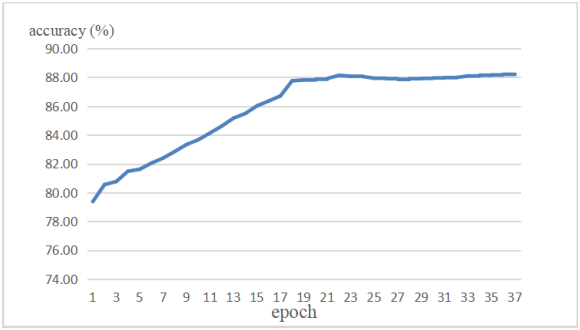


FIGURE 6. The testing accuracy on CUB-200-2011 dataset.

TABLE 4. Comparison results on Stanford Cars dataset.

method	Anno	Accuracy
Gosselin et al. [55]	✓	82.7%
Girshick et al. [56]	✓	88.4%
FCAN [53]	✓	91.3%
B-CNN [15]	✓	91.3%
Krause et al. [54]	✓	92.6%
HSnet [45]	✓	93.9%
TASN [46]		93.2%
DCL [47]		94.5%
NTS [21]		93.9%
HBP [16]		93.7%
GP [12]		92.8%
KP [52]		92.4%
PC [50]		93.43%
MAMC [13]		93%
B-CNN [15]		90.6%
Improved B-CNN [48]		92.0%
Boost CNN [49]		92.1%
SCBA(ours)		94.32%

0 with small fluctuation. The loss of testing stage was decreasing all the time before 25 epoch and fluctuated between 0.4 and 0.5 after that and fell to 0.45 eventually. Different from the training stage, the loss of testing stage was unstable and fluctuated greatly. The accuracy of the model on the testing set was increasing before 18 epoch and subsequently remained stable, ranging from 0.8776 to 0.8826.

2) RESULTS ON STANFORD CARS

We all know that different parts of car are different and complementary, so the positioning of objects and parts plays an important role in recognition task. In Table 4, our SCBA achieves the best result among those supervised

TABLE 5. Comparison results on FGVC-Aircraft dataset.

Method	Anno	Accuracy
B-CNN [15]	✓	84.1%
MG-CNN [57]	✓	86.6%
MDTP [58]	✓	88.4%
DCL [47]		93%
NTS [21]		91.4%
HBP [16]		90.3%
GP [12]		89.8%
PC [50]		89.24%
B-CNN [15]		86.9%
KP [52]		86.9%
Improved B-CNN [48]		88.5%
Boost CNN [49]		88.5%
SCBA(ours)		91.47%

based methods, such as HSnet [45] (93.9%), FCAN [53] (91.3%), Krause et.al [54] (92.6%), and those weakly-supervised based models that achieves the-state-of-art performance, such as NTS [21] (93.9%), TASN [46] (93.2%), HBP [16] (93.7%), PC [50] (93.43%) and etc, except DCL [47] method. Although our SCBA is less than DCL [47] about 0.18%, our SCBA is better than DCL [47] on CUB Birds [36].

3) RESULTS ON FGVC-AIRCRAFT

Because of subtle differences, different aircraft models can be difficult to recognize, for example, by computing the number of windows may be able to discriminate them. The results are similar to those in Table 5 that our SCBA only less than DCL [47] about 1.53% but better than others.

V. CONCLUSION

In this paper, we propose SCBA model to fuse multilayer features for fine-grained recognition, which combines multilayer self-layer and cross-layer bilinear features to learn powerful feature representation. The proposed model can be trained in an end-to-end fashion without the need for bounding box/part annotations. Experimental results on birds, cars and airplanes demonstrate the validity of our model. What's more, we achieve fine-grained recognition in CPSS. However, the proposed method has some limitations, such as expensive computation cost, not effective enough in feature representing, etc. In the future, we will expand our research in two directions, i.e., how to effectively integrate multiscale features to learn rich fine-grained representation, and how to effectively reduce feature dimensionality to decline heavy computational cost.

REFERENCES

[1] X. Wang, L. T. Yang, X. Xie, J. Jin, and M. J. Deen, "A cloud-edge computing framework for cyber-physical-social services," *IEEE Commun. Mag.*, vol. 55, no. 11, pp. 80–85, Nov. 2017.

[2] L. T. Yang, X. Wang, X. Chen, L. Wang, R. Ranjan, X. Chen, and M. J. Deen, "A multi-order distributed HOSVD with its incremental computing for big services in cyber-physical-social systems," *IEEE Trans. Big Data*, early access, Apr. 9, 2018, doi: 10.1109/TBDATA.2018.2824303.

[3] J. Zhou, J. Yan, K. Cao, Y. Tan, T. Wei, M. Chen, G. Zhang, X. Chen, and S. Hu, "Thermal-aware correlated two-level scheduling of real-time tasks with reduced processor energy on heterogeneous MPSoCs," *J. Syst. Archit.*, vol. 82, pp. 1–11, Jan. 2018.

[4] J. Zhou, J. Sun, X. Zhou, T. Wei, M. Chen, S. Hu, and X. S. Hu, "Resource management for improving soft-error and lifetime reliability of real-time MPSoCs," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 38, no. 12, pp. 2215–2228, Dec. 2019.

[5] J. Zhou, X. S. Hu, Y. Ma, J. Sun, T. Wei, and S. Hu, "Improving availability of multicore real-time systems suffering both permanent and transient faults," *IEEE Trans. Comput.*, vol. 68, no. 12, pp. 1785–1801, Dec. 2019.

[6] X. Wang, L. T. Yang, Y. Wang, X. Liu, Q. Zhang, and M. J. Deen, "A distributed tensor-train decomposition method for cyber-physical-social services," *ACM Trans. Cyber-Phys. Syst.*, vol. 3, no. 4, pp. 1–15, Oct. 2019.

[7] N. Zhang, R. Farrell, F. Iandola, and T. Darrell, "Deformable part descriptors for fine-grained recognition and attribute prediction," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 729–736.

[8] N. Zhang, J. Donahue, R. Girshick, and T. Darrell, "Part-based R-CNNs for fine-grained category detection," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Springer, 2014, pp. 834–849.

[9] S. Branson, G. Van Horn, S. Belongie, and P. Perona, "Bird species categorization using pose normalized deep convolutional nets," 2014, *arXiv:1406.2952*. [Online]. Available: http://arxiv.org/abs/1406.2952

[10] L. Wu and Y. Wang, "Where to focus: Deep attention-based spatially recurrent bilinear networks for fine-grained visual recognition," 2017, *arXiv:1709.05769*. [Online]. Available: http://arxiv.org/abs/1709.05769

[11] T. Chen, L. Lin, R. Chen, Y. Wu, and X. Luo, "Knowledge-embedded representation learning for fine-grained image recognition," 2018, *arXiv:1807.00505*. [Online]. Available: http://arxiv.org/abs/1807.00505

[12] X. Wei, Y. Zhang, Y. Gong, J. Zhang, and N. Zheng, "Grassmann pooling as compact homogeneous bilinear pooling for fine-grained visual classification," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 355–370.

[13] M. Sun, Y. Yuan, F. Zhou, and E. Ding, "Multi-attention multi-class constraint for fine-grained image recognition," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 805–821.

[14] G. Huang, Z. Liu, L. V. D. Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4700–4708.

[15] T.-Y. Lin, A. RoyChowdhury, and S. Maji, "Bilinear CNN models for fine-grained visual recognition," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1449–1457.

[16] C. Yu, X. Zhao, Q. Zheng, P. Zhang, and X. You, "Hierarchical bilinear pooling for fine-grained visual recognition," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 574–589.

[17] Y. Gao, O. Beijbom, N. Zhang, and T. Darrell, "Compact bilinear pooling," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 317–326.

[18] T. Chen, W. Wu, Y. Gao, L. Dong, X. Luo, and L. Lin, "Fine-grained representation learning and recognition by exploiting hierarchical semantic embedding," 2018, *arXiv:1808.04505*. [Online]. Available: http://arxiv.org/abs/1808.04505

[19] Y. Wang, V. I. Morariu, and L. S. Davis, "Learning a discriminative filter bank within a CNN for fine-grained recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4148–4157.

[20] C. Pang, H. Yao, X. Sun, S. Zhao, and Y. Zhang, "Exploring part-aware segmentation for fine-grained visual categorization," *Multimedia Tools Appl.*, vol. 77, no. 23, pp. 30291–30310, Dec. 2018.

[21] Z. Yang, T. Luo, D. Wang, Z. Hu, J. Gao, and L. Wang, "Learning to navigate for fine-grained classification," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 420–435.

[22] W. Ge, X. Lin, and Y. Yu, "Weakly supervised complementary parts models for fine-grained image classification from the bottom up," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3034–3043.

[23] X. Li, C. Yang, S. Chen, C. Zhu, and X.-C. Yin, "Semantic bilinear pooling for fine-grained recognition," 2019, *arXiv:1904.01893*. [Online]. Available: http://arxiv.org/abs/1904.01893

[24] M. Gou, F. Xiong, O. Camps, and M. Sznaier, "MoNet: Moments embedding network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3175–3183.

[25] V. Lebedev, A. Babenko, and V. Lempitsky, "Impostor networks for fast fine-grained recognition," 2018, *arXiv:1806.05217*. [Online]. Available: http://arxiv.org/abs/1806.05217

[26] Y. Zheng, H. Ye, L. Wang, and J. Pu, "Learning multiviewpoint context-aware representation for RGB-D scene classification," *IEEE Signal Process. Lett.*, vol. 25, no. 1, pp. 30–34, Jan. 2018.

[27] S. Kim, S. Kim, D. Min, and K. Sohn, "LAF-Net: Locally adaptive fusion networks for stereo confidence estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 205–214.
[28] H. Chen and Y. Li, "Progressively complementarity-aware fusion network for RGB-D salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3051–3060.
[29] K. Huang, C. Zhu, and G. Li, "Saliency detection by adaptive channel fusion," *IEEE Signal Process. Lett.*, vol. 25, no. 7, pp. 1059–1063, Jul. 2018.
[30] C. Chen, H. Huang, C. Chen, Z. Zheng, and H. Cheng, "Multi-scale guided mask refinement for coarse-to-fine RGB-D perception," *IEEE Signal Process. Lett.*, vol. 26, no. 2, pp. 217–221, Feb. 2019.
[31] J. Hu, Z. Chen, M. Yang, R. Zhang, and Y. Cui, "A multiscale fusion convolutional neural network for plant leaf recognition," *IEEE Signal Process. Lett.*, vol. 25, no. 6, pp. 853–857, Jun. 2018.
[32] H. Li, P. Xiong, H. Fan, and J. Sun, "DFANet: Deep feature aggregation for real-time semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 9522–9531.
[33] K. Yu and M. Salzmann, "Statistically-motivated second-order pooling," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 600–616.
[34] Y. Gao, J. Ma, M. Zhao, W. Liu, and A. L. Yuille, "NDDR-CNN: Layerwise feature fusing in multi-task CNNs by neural discriminative dimensionality reduction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3205–3214.
[35] V. Sindagi and V. Patel, "Multi-level bottom-top and top-bottom feature fusion for crowd counting," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1002–1012.
[36] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The Caltech-UCSD birds-200-2011 dataset," California Inst. Technol., Pasadena, CA, USA, Tech. Rep. CNS-TR-2011-001, 2011.
[37] J. Krause, M. Stark, J. Deng, and L. Fei-Fei, "3D object representations for fine-grained categorization," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, Dec. 2013, pp. 554–561.
[38] S. Maji, E. Rahtu, J. Kannala, M. Blaschko, and A. Vedaldi, "Fine-grained visual classification of aircraft," 2013, *arXiv:1306.5151*. [Online]. Available: http://arxiv.org/abs/1306.5151
[39] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[40] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.
[41] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556
[42] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.
[43] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in PyTorch," 2017, *arXiv:1407.3867*. [Online]. Available: https://arxiv.org/abs/1407.3867
[44] X.-S. Wei, C.-W. Xie, and J. Wu, "Mask-CNN: Localizing parts and selecting descriptors for fine-grained image recognition," 2016, *arXiv:1605.06878*. [Online]. Available: http://arxiv.org/abs/1605.06878
[45] M. Lam, B. Mahasseni, and S. Todorovic, "Fine-grained recognition as HSnet search for informative image parts," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2520–2529.
[46] H. Zheng, J. Fu, Z.-J. Zha, and J. Luo, "Looking for the devil in the details: Learning trilinear attention sampling network for fine-grained image recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5012–5021.
[47] Y. Chen, Y. Bai, W. Zhang, and T. Mei, "Destruction and construction learning for fine-grained image recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5157–5166.
[48] T.-Y. Lin and S. Maji, "Improved bilinear pooling with CNNs," 2017, *arXiv:1707.06772*. [Online]. Available: http://arxiv.org/abs/1707.06772
[49] M. Moghimi, S. Belongie, M. Saberian, J. Yang, N. Vasconcelos, and L.-J. Li, "Boosted convolutional neural networks," in *Proc. Brit. Mach. Vis. Conf.*, 2016, pp. 1–24.
[50] A. Dubey, O. Gupta, P. Guo, R. Raskar, R. Farrell, and N. Naik, "Pairwise confusion for fine-grained visual classification," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 70–86.
[51] H. Zhang, T. Xu, M. Elhoseiny, X. Huang, S. Zhang, A. Elgammal, and D. Metaxas, "SPDA-CNN: Unifying semantic part detection and abstraction for fine-grained recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1143–1152.
[52] Y. Cui, F. Zhou, J. Wang, X. Liu, Y. Lin, and S. Belongie, "Kernel pooling for convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2921–2930.
[53] X. Liu, T. Xia, J. Wang, Y. Yang, F. Zhou, and Y. Lin, "Fully convolutional attention networks for fine-grained recognition," 2016, *arXiv:1603.06765*. [Online]. Available: http://arxiv.org/abs/1603.06765
[54] J. Krause, H. Jin, J. Yang, and L. Fei-Fei, "Fine-grained recognition without part annotations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 5546–5555.
[55] P.-H. Gosselin, N. Murray, H. Jégou, and F. Perronnin, "Revisiting the Fisher vector for fine-grained classification," *Pattern Recognit. Lett.*, vol. 49, pp. 92–98, Nov. 2014.
[56] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 580–587.
[57] D. Wang, Z. Shen, J. Shao, W. Zhang, X. Xue, and Z. Zhang, "Multiple granularity descriptors for fine-grained categorization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2399–2406.
[58] Y. Wang, J. Choi, V. I. Morariu, and L. S. Davis, "Mining discriminative triplets of patches for fine-grained classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1163–1172.

YINGQIONG PENG was born in 1978. She is currently an Associate Professor. She is mainly engaged in agricultural informatization, computer vision, and pattern recognition. From 2012 to 2016, she was in charge of the research project of rice nitrogen nutrition diagnosis based on computer vision technology with the Jiangxi Provincial Science and Technology Department. Since 2013, she has been in charge of the research project on the key technology of nondestructive testing of umbilical orange quality based on bioelectric characteristics with the Jiangxi Provincial Science and Technology Department. Since 2014, she has been in charge of the Study on the relationship between electrical impedance variation and freshness of umbilical orange with Jiangxi Agricultural University.

YUXIA SONG was born in 1996. She is currently pursuing the master's degree. She is mainly engaged in deep learning and computer vision.

WEIJI HUANG was born in 1982. He is currently an Assistant Research Fellow. He is mainly engaged in deep learning and computer vision.

HONG DENG was born in Duchang, Jiangxi, in 1977. He is currently an Associate Professor. He is mainly engaged in agricultural informatization and image processing.

YINGLONG WANG was born in 1970. He is currently a Professor and a Master's Supervisor. His main research interests include knowledge discovery and intelligent systems, and agricultural intelligent information systems.

QI CHEN was born in 1982. She is currently an Associate Professor. She is mainly engaged in deep learning and image processing.

MUXIN LIAO was born in 1994. He is currently pursuing the master's degree. He is mainly engaged in deep learning and computer vision.

JING HUA was born in 1985. She is currently an Associate Professor. She is mainly engaged in semi-supervised learning, pattern recognition, and biomedical information coding,