Discriminative Region Enhancing and Suppression Network for Fine-Grained Visual Categorization

Guanhua Wu, Cheng Pang(✉), Rushi Lan, Yilin Zhang, and Pingping Zhou

Guilin University of Electronic Technology, Guilin 541004, China
pangcheng3@guet.edu.cn

Abstract. Attention mechanisms are intensively devoted to local feature abstraction for fine-grained visual categorization. A limitation of attention-based methods is that they focus on salient region mining and feature extraction, while ignoring the ability to incorporate discriminative and complementary features from other parts of the image. In order to address this issue, we introduce a novel network known as the Discriminative Region Enhancing and Suppression Network (DRESNet). This network efficiently extracts a wide range of diverse and complementary features, thereby enhancing the final representation. Specifically, a plug-and-play salient region diffusion (SRD) module is proposed to explicitly enhances the salient features extracted by any backbone network. The SRD module can adaptively adjust the weights of regions and redirect attention to other non-discriminative regions to generate different complementary features. The proposed discriminative region enhancing and suppression network is free from bounding boxes or part annotations and can be trained end-to-end. Our proposed method demonstrates competitive performance on three fine-grained classification benchmark datasets, as supported by extensive experimental results. Additionally, it is compatible with widely used frameworks currently in use.

Keywords: Fine-grained Visual Categorization · Attention mechanism · Region enhancing and suppression

1 Introduction

Fine-grained visual categorization (FGVC) has garnered growing research interest [2–4] in recent years, driven by its promising applications in diverse real-world scenarios such as intelligent retail [24], intelligent transportation [11], and conservation [1]. FGVC aims to recognizing images belonging to multiple subcategories within the same category, e.g., different species of birds [23], models of cars [13] and aircrafts [18]. As different subcategory objects share a similar physical structure (i.e., all kinds of birds have a head, wings, and a tail), and could be only be distinguished by subtle local regions, as shown in Fig. 1. Due to

H. Lu et al. (Eds.): ACPR 2023, LNCS 14408, pp. 93–105, 2023.
https://doi.org/10.1007/978-3-031-47665-5_8

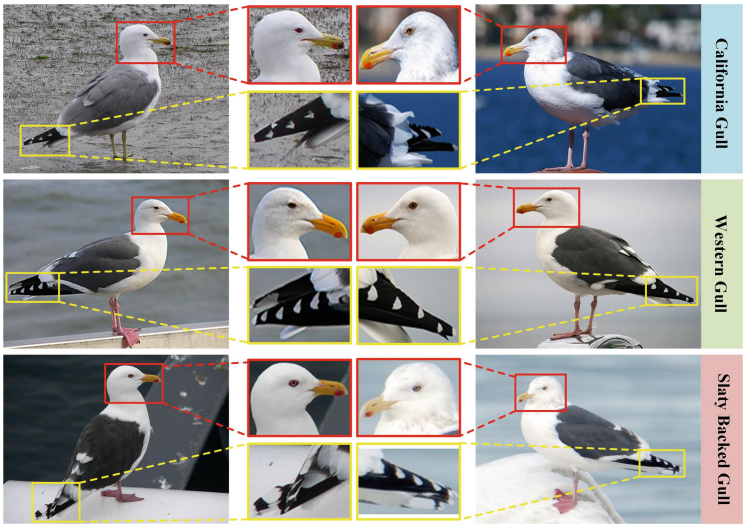

Fig. 1. Samples belonging to the same sub-category share a similar physical structure (i.e., all kinds of birds have a head, wings, and a tail) and appearance, making it infeasible to visually distinguish between them.

low inter-class variances, it further increases the difficulty of fine-grained image classification.

To extract discriminative features for classification, a key process is to focus on the feature representations of different parts of the object. Some previous works [14,25,29,31] rely mainly on predefined bounding boxes and part annotations to locate the distinguishable regions, and then extract part-specific features for fine-grained classification. However, these hand-craft annotations are not optimal for FGVC, the collection of which can be very costly. In recent years, weakly supervised learning using labels in image-level has become the mainstream method for FGVC. Some recent methods [7,22,26,33] attempt to locate the distinguishable regions and learn effective feature representations using attention mechanisms, channel clustering, and other techniques, without requiring bounding box/part annotations. Although these methods are effective, there are still potential limitations: 1) attention-based networks tend to focus on globally salient features while ignore other local discriminative features which potentially carry complementary information for the salient ones; 2) part-based sampling methods are easily affected by the number and size of the sampled parts. These pre-defined parameters greatly limit the effectiveness and flexibility of the model. Therefore, how to effectively extract diverse salient features and how to integrate reasonably these features into the final representation are worthy of discussion for the fine-grained classification task.

In this paper, we propose a discriminative region enhancing and suppression network (DRESNet) to address the above limitations by generating a set

of integral fine-grained features including the globally salient features and their complementary features. More specifically, DRESNet does not focus on how to capture accurate distinguishable parts, but enhances and suppresses the salient parts in the feature map in an adaptive way and then forces the following network to mine other discriminative regions containing potentially complementary features. Some observations show that the most salient region of an image is tend to be noticed by the attention models in the first time, then the feature abstraction of this region is enhanced while that of other regions is suppressed. As a result, visual cues from suppressed features may be absent in the learned features. However, a set of integral features consisting of both globally salient features and other local features are crucial for FGVC tasks [6,9].

The proposed DRESNet consists of a feature extraction backbone network and a salient region diffusion (SRD) module. By inserting salient region diffusion module into various stages of the backbone network, it efficiently extracts various potential features. SRD module enhances and suppresses features to obtain part-specific representations. Note that, in the suppression operation, SRD module does not require additional complex hyperparameters to suppress salient regions information. Instead, it effectively expands the discriminative region through a learnable way (suppress excessive feature expression of salient regions while encourage feature expression from adjacent non-salient regions). We demonstrate hat the feature learning of the backbone networks could be substantially improved in this way.

Our contributions are summarized as follows: (1) We propose a salient region diffusion (SRD) module, which enhances the prominent features of the network. Additionally, through an adaptive learning-based suppression process, SRD can compel the network to learn more complementary features. (2) Our proposed region enhancing and suppression network (DRESNet) achieves competitive results on three benchmark fine-grained classification datasets.

2 Related Work

Fine-grained visual categorization (FGVC) aims to identify visually similar subcategories within the same basic category, e.g., different species of birds [23], models of cars [13] and aircraft [18]. Currently, research on weakly supervised fine-grained visual categorization methods mainly focuses on three aspects: fine-grained feature learning, discriminative region localization, and visual attention mechanisms.

2.1 Fine-Grained Feature Learning

In order to more accurately describe the subtle differences between fine-grained categories, Lin et al. [15] proposed a bilinear model consisting of two feature extractors. This model adopts a translation-invariant approach to extract features from different parts of the image and fuse them, thereby enhancing the

ability to learn fine-grained features. However, the use of bilinear features generated by the outer product results in extremely high dimensions, which increases the computational complexity. To address this issue, Gao et al. [8] attempted to approximate the second-order statistics of the original bilinear pooling operation by applying Tensor Sketch to reduce the dimensionality of bilinear features. Kong et al. [12] used a low-rank approximation for the covariance matrix and further learned low-rank bilinear classifiers, which significantly reduced the computation time and effective number of parameters. In addition, Yu et al. [27] proposed a cross-layer bilinear pooling method to integrate multiple cross-layer bilinear features and capture part relationships in the features from inter layers.

2.2 Discriminative Region Localization

Localization-based methods capture discriminative semantic parts of fine-grained objects and then construct intermediate representations corresponding to these parts for final classification. Fu et al. [7] proposed a recurrent attention convolutional neural network, which recursively learns discriminative regions region-based feature representation at multiple scales in a mutually reinforced way. Yang et al. [26] utilized self-supervision attention mechanism to effectively locate regions in images that contain more semantic features. In [34], a part proposal network generates multiple local attention maps and s part rectification network learns rich part-specific features. Zhang et al. [28] proposed a multi-scale learning network, which predicts the position of objects and local regions information through the attention object location module and attention part proposal module.

2.3 Visual Attention Mechanisms

In fine-grained image classification tasks, introducing visual attention mechanisms helps to capture subtle inter-class differences in the image. Ding et al. [6] proposed a selective sparse sampling learning method that learns a set of sparse attention for obtaining discriminative and complementary regions. Similar to [6], TASN [33] regards the regions with high responses in the attention map as informative parts, and proposes a trilinear attention sampling network to learn fine-grained feature representations. Through attention-based samplers, TASN can re-sample the focused regions in the image to emphasize fine-grained details. Sun et al. [22] proposed a one-squeeze multi-excitation module to extract features from multiple attention regions and proposed a multi-attention multi-class constraint to enhance the correlation of attention features. Zhang et al. [30] introduced a progressively enhancing strategy by highlighting important regions through class activation maps.

3 Method

In visual attention models, we observe that the learning of features in salient regions may hinder its further feature abstraction in non-salient regions. As

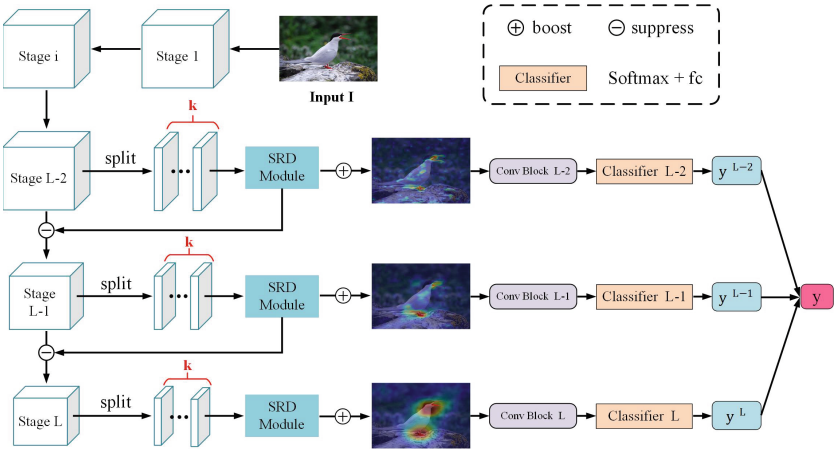

Fig. 2. Overview of the discriminative region enhancing and suppression network (DRESNet). Any backbone networks can be used coupled with the proposed salient region diffusion (SRD) module, for further mining complementary features for FGVC. The Conv Block consists of two convolutional layers followed by a max-pooling layer. ⊕ represents boosting feature and ⊖ represents suppressing feature.

a result, the complementary information and discriminative cues being beneficial to FGVC are eliminated. To tackle this issue, we present discriminative region enhancing and suppression network (DRESNet) that mine diverse potential useful features that be masked by the other salient features stage-by-stage, and each stage integrates different feature embedding for the last discriminative fine-grained representation. In addition, it can be easily implemented on various convolutional neural networks, such as VGG [21] and ResNet [10]. The DRESNet network structure, illustrated in Fig. 2.

3.1 Salient Region Diffusion Module

The method base on attention model tends to highlight the most distinct feature regions. However, such mechanism hurts the further exploration of the rich information from other regions in an image. Thus, we propose a simple yet effective salient region diffusion (SRD) module to encourage the diffusion of the attention from salient regions to more other regions, obtaining discriminative and complementary features.

As illustrated in Fig. 3, suppose $X \in R^{C\times W\times H}$ is the feature map of an image extracted from backbone network, where C, W, and H indicate the number of channels, width, and height of the feature map, respectively. In the enhancing step, SRD module simply splits X evenly into k vertical groups along the width dimension. This clearly indicates that the feature map will be composed of these k multiple local features, where each group is represented as $X_i \in R^{C\times(W/k)\times H}, i = [1, 2, ..., k]$. Then, the importance of each group is cal-

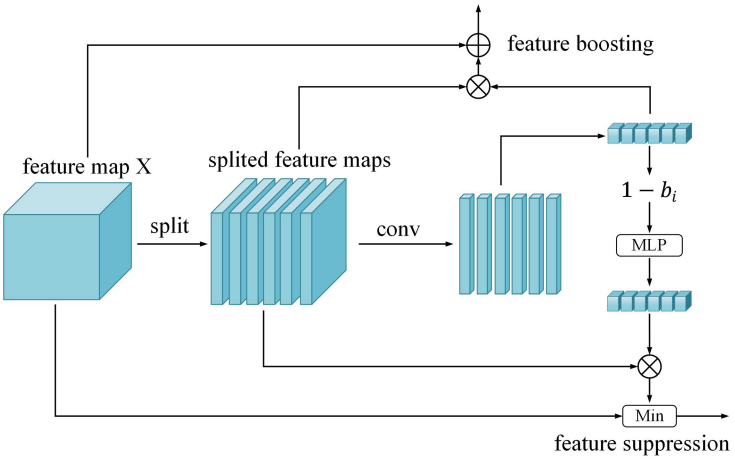

Fig. 3. The proposed Salient Region Diffusion (SRD) module diffuses the visual attention from salient region to more other region in an image, obtaining discriminative and complementary features. The SRD module explicitly divides the feature map into multiple groups along the width dimension, facilitating the extraction of corresponding enhancement and suppression factors.

culated using a convolutional layer (Conv) followed by a Rectified Linear Unit (ReLU) operation.

$$e_i = ReLU(Conv(X_i)) \tag{1}$$

where $e_i = [e_1, e_2, ..., e_k] \in R^{(1\times(W/k)\times H)}$ denotes the spatial attention score of the i-th group of feature in X. $Conv$ is a 1×1 kernel sized convolutional layer, where the output channel is set to 1. This configuration enables us to study the importance of the corresponding area. ReLU is applied to remove negative activation values. Afterwards, we apply the global average pooling (GAP) operation to e_i, followed by the application of the softmax function to map the resulting values to the range [0, 1] for normalization:

$$b_i = \frac{exp(GAP(e_i))}{\sum_{j=1}^{k} exp(e_j)} \tag{2}$$

$B = [b_1, b_2, ..., b_k] \in R$ represents the importance of the region. For input feature map X, we obtain enhanced feature F_e by enhancing its most salient part.

$$F_e = X + B \otimes X \tag{3}$$

where $\otimes$ denotes element-wise multiplication. F_e represents the enhanced features obtained through the enhanced steps of the SRD module in different network output stages.

$$F_p = Conv\_Block(F_e) \tag{4}$$

Here $Conv\_Block$ represents the combination of two convolution layers and a global average pooling. The F_p is then fed into the classifier for prediction, obtaining the class probability of the input image.

SRD module suppresses the most salient features to extract complementary attention features. More specifically, in the suppression operation, we do not set a fixed hyperparameter as the upper bound of x and regard it as a starting point to be suppressed. We believe that the larger the value in the enhancement factor b_i, the more important the corresponding group feature is, and the greater the degree of suppression of this feature. Using the learning ability of the fully connected layer, SRD module adaptively suppresses discriminative regions. Formally, the output of the learnable suppression factor is:

$$s_i = \sigma(MLP(1 - b_i)) \tag{5}$$

$S = [s_1, s_2, ..., s_k]$ represents the suppression factor of group i features. The SRD module can improve the learning ability of the network with the help of the nonlinear mapping and powerful fitting ability of multilayer perceptron (MLP). σ is the sigmoid function that can normalizes the suppression factor. Since the parameters in MLP are trained with the classification target, the learnable SRD module can adaptively suppress the discriminative region without greatly impairing its extraction ability.

Finally, we obtain the suppressed feature X_s by calculating the minimum value between the original feature X and $S \otimes X$. The suppressed feature X_s is then passed to the following network to obtain other salient features.

$$X_s = Min(X, S \otimes X) \tag{6}$$

Min denotes the application of element-wise minimum operation on X and $S \otimes X$ to suppress the discriminative region.

3.2 Loss Function

Our framework is illustrated in Fig. 2. Based on single branch DRESNet, we can extract multiple complementary discriminative features from multiple stages of the network. In this paper, we use ResNet50 as a feature extractor, which has S stages (i.e., $L = 5$). With the increase of the number of layers, the features extracted by ResNet50 can more abstract and represent higher level semantic information. In DRESNet, our goal is to optimize the feature representation in the s-th stage. To this end, the SRD module is inserted into the last three stages of ResNet50. The part-specific representation F_p corresponding to different stagescan be obtained from Eq. (4). During the training phase, the classification loss for the last three stages (i.e., $i = L-2, L-1, L$) of DRESNet can be formulated as follows:

$$L_i = -y^T \log(y_i), y_i = classifier_i(F_i) \tag{7}$$

where y is the ground-truth label corresponding to the input image. $y_i \in R^C$ represents the predicted probability of the i-th stage, where C is the number of

classes. Therefore, The overall optimization objective of DRESNet during the training phase is:

$$L_{total} = \sum_{i=1}^{T} L_i \tag{8}$$

Where $T = 3$ represents the number of salient features enhanced by the SRD module in the last three stages of the network. During inference, we calculate the sum of prediction scores for all enhanced part-specific features to obtain the final prediction result.

4 Experiments

4.1 Dataset and Implementation Details

We conducted experiments on three benchmarked datasets for fine-grained classification, including CUB-200-2011 [23], Stanford Cars [13], and FGVC-Aircraft [18]. Table 1 provides detailed statistical data on the number of categories and the standard train-test split. In this paper, we evaluate categorization performance using Top-1 accuracy as the metric.

Table 1. The statistics information of the three widely used Fine-Grained Visual Categorization datasets.

Dataset	Category	Train	Test
CUB-200-2011 [23]	200	5994	5974
FGVC-Aircraft [18]	100	6667	3333
Stanford Cars [13]	196	8144	8041

For all experiments, we utilized pre-trained VGG16 [21] and ResNet50 [10] models on ImageNet [5] as backbone networks for feature extraction from input images. Then, SRD module is inserted into the last three output feature of ResNet50's $res3\_4$, $res4\_6$ and $res5\_3$ (VGG16's $relu3\_3$, $relu4\_3$ and $relu5\_3$). To ensure fair comparison, we maintained the same resolution as other methods. During the training phase, the input image was resized to a fixed size of 550×550 and subsequently randomly cropped to 448×448, accompanied by random horizontal flipping. During the testing stage, the input image was resized to 550×550 and cropped from the center to 448×448. According to the standards in the literature, we set up stochastic gradient descent (SGD) optimizer with a momentum of 0.9 and weight decay of $1e^{-5}$ to optimize the training process. The model trained for 200 epochs, with the learning rate of the backbone initialized at 0.0002 and the other newly added layers set to 0.002. During training, the learning rate was adjusted using the cosine annealing strategy [16]. The batch size is 16. Our model was implemented in Pytorch and trained end-to-end on a single GTX 2080Ti GPU, without any bounding box or part annotation.

4.2 Experimental Results

We compared the proposed DRESNet with state-of-the-art methods on three popular benchmarked fine-grained datasets: CUB-200-2011 [23], Stanford Cars [13], and FGVC Aircraft [18]. The results are shown in Table 2, with the top and second-best values highlighted in **bold** and underline, respectively. Our method achieved significant performance improvements on these three datasets compared to existing techniques.

Table 2. Comparison results on CUB-200-2011, FGVC-Aircraft and Stanford Cars datasets. (The top-1 accuracy is reported in %)

Method	Backbone	CUB-200-2011	Stanford Cars	FGVC-Aircraft
BCNN [15]	VGG16	84.1	91.3	84.1
CBP [8]	VGG16	84.1	91.3	84.1
LRBP [12]	VGG16	84.2	90.9	87.3
HBP [27]	VGG16	**87.1**	**93.7**	90.3
RA-CNN [7]	VGG16	85.3	92.5	88.1
MA-CNN [32]	VGG16	86.5	92.8	89.9
Ours	VGG16	87.0	93.6	**91.6**
Cross-X [17]	ResNet50	87.7	94.6	92.6
S3N [6]	ResNet50	88.5	94.5	**93.0**
CIN [9]	ResNet50	87.5	94.1	92.6
MAMC [22]	ResNet50	86.2	92.8	-
NTS [26]	ResNet50	87.5	93.9	91.4
TASN [33]	ResNet50	87.9	93.8	-
MOMN [19]	ResNet50	88.3	93.2	90.3
API [35]	ResNet50	87.7	**94.8**	**93.0**
MGE-CNN [30]	ResNet50	88.5	93.9	-
Ours	ResNet50	**89.3**	94.5	92.7

Results on CUB200-2011 [23]: CUB-200-2011 [23] is the most challenging benchmark in fine-grained image classification tasks, and our models based on VGG16 and ResNet50 have achieved state-of-the-art performance on this dataset. Compared with fine-grained feature learning methods BCNN [15], CBP [8], and LRBP [12], our method outperforms them by 5.2%, 5.3%, and 5.1%, respectively. MA-CNN [32], TASN [33], and S3N [6] capture subtle inter-class differences in images by introducing visual attention mechanisms. Due to the introduction of feature enhancement and suppression, our proposed method achieves superior accuracy compared to them to varying degrees.

Results on Stanford Cars [13]: This dataset contains more images than CUB-200-2011, but the car images present less structure variations than the

bird images. The proposed methods consistently wins MOMN [19], MGE-CNN [30], CIN [9], and TASN [33] to varying degrees. Cross-X [17] performs robust multi-scale feature learning by using relationships between different images and between different network layers. The accuracy of Cross-X [17] is 0.1% higher than our proposed method. Unlike other methods that use a single image, API [35] learns to recognize each image in a pair by adaptively considering feature priorities and pairs them for comparison step by step. It's accuracy is 0.3% higher than our proposed method.

Results on FGVC-Aircraft [18]**:** Our method achieved competitive performance on the FGVC-Aircraft dataset [18]. Compared to RA-CNN [7] and MA-CNN [32], our method outperforms them by 4.6% and 2.8% respectively. Based on ResNet50, our model's performance is 0.1%, 1.3%, and 2.4% higher than CIN [9], NTS [26], and MOMN [19] respectively. S3N [6] uses sparse attention and selective sampling to capture diverse and discriminative details of parts without losing contextual information. In this dataset, the top-1 accuracy of S3N [6] is slightly higher than our method.

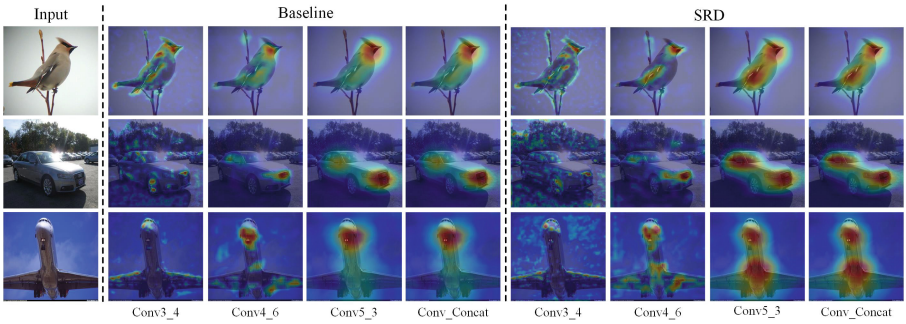

Fig. 4. Visualizations of activation maps in different layers. Feature maps obtained with the proposed SRD have learnt more discriminative body parts of the target. These parts need not to be the most salient features but help to distinguish the confusing sub-categories.

4.3 Visualization Analysis

To better understand the enhancing and suppressing effects of the SRD module on features, we visualized activation maps of ResNet50 [10] with and without the SRD module on three benchmark datasets for fine-grained classification tasks, and the results are shown in Fig. 4. The activation maps were obtained by averaging activation values over the channel dimension of the given feature map. Based on ResNet50, we applied the Grad-CAM [20] algorithm to visualize the $Conv3\_4$, $Conv4\_6$, $Conv5\_3$, and $Conv\_concat$ feature maps of ResNet50 [10] on the three validation sets. We concatenate the $Conv3\_4$,

$Conv4\_6$, $Conv5\_3$ feature map attention pooled by different middle feature maps to get a refined $Conv\_concat$ feature map. By comparing the heatmaps of the baseline and DRESNet on $Conv3$ and $Conv\_concat$, we can demonstrate that the SRD module plays a crucial role in enhancing the baseline and extracting additional complementary features. Taking the bird in Fig. 4 as an example, in the absence of SRD configuration, the baseline network only focuses on the salient head region while disregarding equally significant wing segments. The visualization experiments demonstrate the capability of SRDs for mining multiple different discriminative object parts.

5 Conclusion

In this paper, we propose a new discriminative region enhancing and suppression network for weakly supervised fine-grained image classification, focusing on how to extract the most salient features and complementary attention features. Specifically, we introduce a salient region diffusion (SRD) module, which can be considered as a significant dropout scheme, enabling the network to adaptively mine potentially important information at different levels of importance. Visualization of the feature maps demonstrates that the SRD mines multiple complementary discriminative object parts in different network levels. Our proposed network demonstrates impressive performance on the three most difficult FGVC datasets, outperforming the majority of attention-based methods due to our redesigned attention mechanism. Furthermore, it exhibits exceptional efficacy in handling non-rigid targets featuring multiple body joints.

Acknowledgement. This work is partially supported by the Guangxi Science and Technology Project (2021GXNSFBA220035, AD20159034), the Open Funds from Guilin University of Electronic Technology, Guangxi Key Laboratory of Image and Graphic Intelligent Processing (GIIP2208) and the National Natural Science Foundation of China (61962014).

References

1. Aggrawal, P., Anand, S.: Computer vision applications in wildlife conservation (2020)
2. Angelova, A., Zhu, S.: Efficient object detection and segmentation for fine-grained recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 811–818 (2013)
3. Branson, S., Beijbom, O., Belongie, S.: Efficient large-scale structured learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1806–1813 (2013)
4. Branson, S., Van Horn, G., Belongie, S., Perona, P.: Bird species categorization using pose normalized deep convolutional nets. arXiv preprint arXiv:1406.2952 (2014)
5. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: a large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255. IEEE (2009)

6. Ding, Y., Zhou, Y., Zhu, Y., Ye, Q., Jiao, J.: Selective sparse sampling for fine-grained image recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6599–6608 (2019)
7. Fu, J., Zheng, H., Mei, T.: Look closer to see better: recurrent attention convolutional neural network for fine-grained image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4438–4446 (2017)
8. Gao, Y., Beijbom, O., Zhang, N., Darrell, T.: Compact bilinear pooling. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 317–326 (2016)
9. Gao, Y., Han, X., Wang, X., Huang, W., Scott, M.: Channel interaction networks for fine-grained image categorization. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 10818–10825 (2020)
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
11. Khan, S.D., Ullah, H.: A survey of advances in vision-based vehicle re-identification. Comput. Vis. Image Underst. **182**, 50–63 (2019)
12. Kong, S., Fowlkes, C.: Low-rank bilinear pooling for fine-grained classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 365–374 (2017)
13. 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)
14. Lin, D., Shen, X., Lu, C., Jia, J.: Deep LAC: deep localization, alignment and classification for fine-grained recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1666–1674 (2015)
15. Lin, T.Y., RoyChowdhury, A., Maji, S.: Bilinear CNN models for fine-grained visual recognition. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1449–1457 (2015)
16. Loshchilov, I., Hutter, F.: SGDR: stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983 (2016)
17. Luo, W., et al.: Cross-X learning for fine-grained visual categorization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 8242–8251 (2019)
18. Maji, S., Rahtu, E., Kannala, J., Blaschko, M., Vedaldi, A.: Fine-grained visual classification of aircraft. arXiv preprint arXiv:1306.5151 (2013)
19. Min, S., Yao, H., Xie, H., Zha, Z.J., Zhang, Y.: Multi-objective matrix normalization for fine-grained visual recognition. IEEE Trans. Image Process. **29**, 4996–5009 (2020)
20. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-CAM: visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 618–626 (2017)
21. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
22. Sun, M., Yuan, Y., Zhou, F., Ding, E.: Multi-attention multi-class constraint for fine-grained image recognition. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11220, pp. 834–850. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01270-0_49

23. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The Caltech-UCSD Birds-200-2011 dataset (2011)
24. Wei, X.S., Cui, Q., Yang, L., Wang, P., Liu, L.: RPC: a large-scale retail product checkout dataset. arXiv preprint arXiv:1901.07249 (2019)
25. Wei, X.S., Xie, C.W., Wu, J., Shen, C.: Mask-CNN: localizing parts and selecting descriptors for fine-grained bird species categorization. Pattern Recogn. **76**, 704–714 (2018)
26. Yang, Z., Luo, T., Wang, D., Hu, Z., Gao, J., Wang, L.: Learning to navigate for fine-grained classification. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) Computer Vision – ECCV 2018. LNCS, vol. 11218, pp. 438–454. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01264-9_26
27. Yu, C., Zhao, X., Zheng, Q., Zhang, P., You, X.: Hierarchical bilinear pooling for fine-grained visual recognition. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11220, pp. 595–610. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01270-0_35
28. Zhang, F., Li, M., Zhai, G., Liu, Y.: Multi-branch and multi-scale attention learning for fine-grained visual categorization. In: Lokoč, J., et al. (eds.) MMM 2021. LNCS, vol. 12572, pp. 136–147. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-67832-6_12
29. Zhang, H., et al.: SPDA-CNN: unifying semantic part detection and abstraction for fine-grained recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1143–1152 (2016)
30. Zhang, L., Huang, S., Liu, W., Tao, D.: Learning a mixture of granularity-specific experts for fine-grained categorization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 8331–8340 (2019)
31. Zhang, N., Donahue, J., Girshick, R., Darrell, T.: Part-based R-CNNs for fine-grained category detection. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8689, pp. 834–849. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10590-1_54
32. Zheng, H., Fu, J., Mei, T., Luo, J.: Learning multi-attention convolutional neural network for fine-grained image recognition. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 5209–5217 (2017)
33. Zheng, H., Fu, J., Zha, Z.J., Luo, J.: Looking for the devil in the details: learning trilinear attention sampling network for fine-grained image recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5012–5021 (2019)
34. Zheng, H., Fu, J., Zha, Z.J., Luo, J., Mei, T.: Learning rich part hierarchies with progressive attention networks for fine-grained image recognition. IEEE Trans. Image Process. **29**, 476–488 (2019)
35. Zhuang, P., Wang, Y., Qiao, Y.: Learning attentive pairwise interaction for fine-grained classification. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 13130–13137 (2020)