# A NOVEL PART FEATURE INTEGRATION AND FUSION METHOD FOR FINE-GRAINED VEHICLE RECOGNITION

*Ping Wang, Yijie Cao, Lei Lu*[⋆]

School of Information and Communications Engineering, Xi'an Jiaotong University

## ABSTRACT

In this paper, we propose a novel light-weight feature integration and fusion method to enhance the discriminative ability of deep convolutional features for the task of fine-grained vehicle recognition. The proposed method is built on the deep convolutional layers from which the discriminative part features could be integrated and fused accordingly. More specifically, a basic feature integration module is adopted to integrate the feature maps of deep convolutional layers into groups in each of which the related discriminative parts are assembled together. Then a fusion module follows to model the coarse-to-fine relationship of the part features and further ensure the integrity and effectiveness of the part features. We conduct comparison experiments on public dataset, and the results show that the proposed method achieves comparable performance with state-of-the-art algorithms.

***Index Terms*—** Fine-grained, feature fusion, feature integration

## 1. INTRODUCTION

Fine-grained Vehicle Recognition is an essential research branch of fine-grained object recognition and has gained significant attention in the past few years [1, 2]. With the recognition of the finer-level attributes of vehicles, the availability and effectiveness in the areas of intelligent transportation, security monitoring, law enforcement, access control systems and etc., could be significantly enhanced. It will improve the robustness and reliability of vehicle-related applications by providing much more specific vehicle information which could not only be used to identify specific vehicle models but also benefits to many real-life applications.

Though researchers have achieved remarkable progress, there are still several challenges existed in this ares, and the main challenges are the inter-class similarity and intra-class variations problems. These challenges are mainly caused by the minor difference between the vehicle models. Since the minor difference between the subordinate categories usually lied in some local regions of the object, adopting the discriminative features from the informative part regions could better address these problems.

⋆Corresponding author

After several years of study, the deep learning-based approach has become the essential scheme adopted by the researchers, and many efficient deep neural networks have been developed to solve this task. To gain further insight into the deep networks and acquire more finer-level feature representation, the deep features from the backbone networks are worthy of further investigation, in which the object-level information has already been well-modeled. In general, the backbone networks for the general-purpose object recognition task are undergoing repaid iteration and development, and their performance on feature extraction and representation have already achieved promising results. However, they tend to focus more on the global object-level features than the finer-level part features which are more suitable for the fine-grained recognition tasks. Therefore, further efforts should be made to explore the deep convolutional features based on which to extract more local discriminative features that are suitable for fine-grained target recognition.

In this paper, we propose light-weight feature integration and fusion modules which are on the basis of the object-level features extracted by the typical CNN backbones. Theses modules could integrate and fuse object-level features from deep convolutional layers, and focus on the features of the discriminative part regions through the learning process. The remainder of this paper is organized as follows. In Section 2, we briefly review the related works on fine-grained object recognition. The proposed method will be detailed in Section 3. In Section 4, we introduce the implementation details of the extensive experiments. The conclusions and future work are presented in Section 5.

## 2. RELATED WORK

The related studies could be summarized into three categories: part detector-based methods, attention-based methods and part learning-based methods.

**Part detector-based methods.** For the part detector-based methods, part annotations are indispensable and have shown considerable effective for fine-grained object recognition. Zhang et al. [3] used the DPM (Deformable Parts Model) as part detectors and proposed two pose-nomalized feature descriptors on the basis of well-labeled part regions. Similarly, Liao et al. [4] also employed the DPM as part detectors and performed the recognition accordingly. Be-

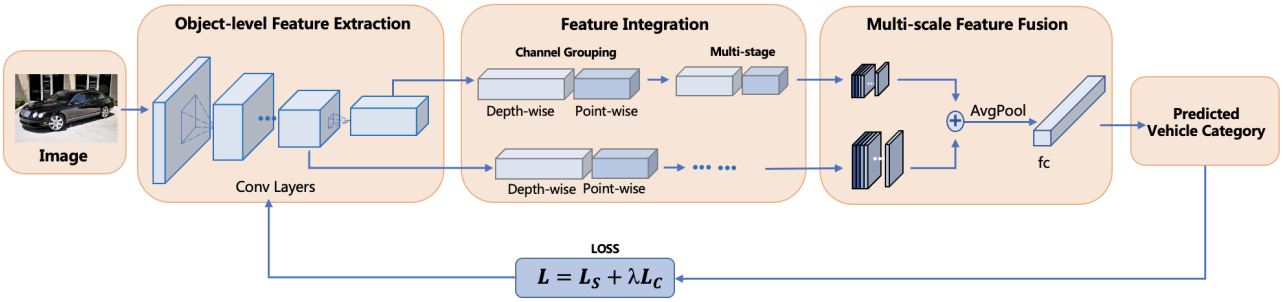


**Fig. 1**. The framework of the proposed feature integration and fusion modules. The feature integration module will extract the discriminative part features accordingly from the object-level features of the backbone network, and then the feature fusion module follows to further enhance the part-level features.

sides traditional part detector, Krause et al. [5] detected the part regions by co-segmentation and alignment. Though part annotations are claimed not required, they did improve the segmentation quality considerably. Lam et al. [6] proposed the HSnet to search informative part regions with the help of part annotations. Wei et al. [7] proposed a selective convolutional descriptor aggregation (SCDA) method to locate the concerned target regions.

**Attention-based methods.** Considering the enormous cost of precise part annotations, many researchers try to detect the discriminative part regions without annotations. The key feature of these methods is to utilize the attention mask to assist the extraction of discriminative features. On the basis of the visual attention mechanism, many works have been done to study the task-specific part regions. Various attention mask generation strategies have been investigated on the basis of the feed-forward network structure. Hu et al. [8] proposed the Spatially Weighted Pooling (SWP) layer connected after the last convolutional layer, based on which a set of predefined learnable attention masks were generated directly. Ma et al. [9] proposed a Channel Max pooling (CMP) method to select the maximum feature value along the channel side. Rodriguez et al. [10] proposed a modular attention architecture, which was deployed to different layers of the convolutional layer to generate part feature scores.

**Part learning-based methods.** This kind of method puts more emphasis on the optimization and improvement of the neural network components, and extracts the concerned features in the learning process of the networks [11–14]. In [11], the authors observed that some filter channels could respond to specific patterns consistently. Therefore, they first generated a set of region proposals and resized them to a fixed size. Then, these proposals were sent to the convolutional layer to generate the response value for each channel. The filter channels were grouped according to the response value of each channel and served as part detectors. Zheng et al. [12] adopted the fully-connected layer to rate the filter channels. Xiang et al. [13] took each channel of the feature maps generated by the truncated CNN as the part detector and proposed a part assembling method to assemble closely related parts.

From the analysis of the previous work, it could be concluded that the attention-based methods and part learning-based methods tackle the task of fine-grained object recognition in similar ways. They both learn discriminative part regions or part features in a weak supervised way, and will change the network structure in different ways to produce task-specific networks. In this paper, to further explore the discrimination ability of the higher-layer features derived from the deep networks, we propose a general feature integration and fusion module for the fine-grained vehicle recognition. This module is built on the object-level features of the backbone networks and will efficiently reuse the high-level features to produce discriminative part features for the fine-grained recognition tasks.

## 3. PROPOSED METHOD

In this section, we introduce the proposed fine-grained feature integration and fusion module for fine-grained vehicle recognition, which contains feature integration, multi-scale feature fusion and classification. The detailed structure of the proposed method is illustrated in Fig. 1. The typical CNN backbone takes image as input to generate the feature maps representing the global visual cues of the concerned object, then proposed module follows to refine the feature extraction process to achieve the discriminative part features.

### 3.1. Object-level Feature Extraction

The deep Convolutional Neural Network (CNN) has achieved great performance on some general-purpose object recognition tasks. In these networks, the feature maps of higher convolutional layers tend to focus more on the target object, and contain abundant discriminative and informative features of the concerned object. This implies the high-level feature maps have well-represented the global and subtle part-level information of the target object, and on top of which will facilitate the extraction of more discriminative part features of the object. For that reason, some of the recent works take the feature maps of higher-layer as part and global feature extractor [13].

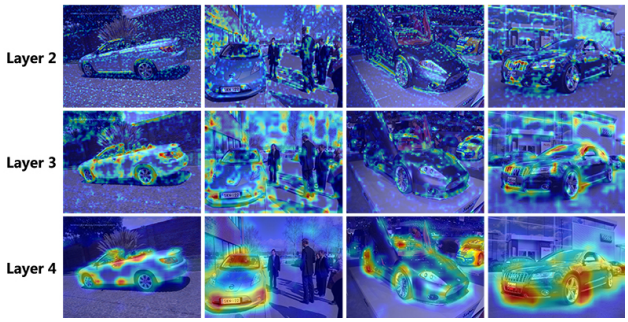


**Fig. 2**. Visualization of the higher-layer feature maps of the pre-trained CNN models (ResNet50). Each row shows feature maps from different convolutional layers, and convolutional layers from top to down are 2, 3 and 4, respectively.

To show the effectiveness of the higher-layer feature maps, the Grad-CAM technique is adopted to interpret the feature maps visually, which uses the gradient of network back propagation to calculate the weight of each channel of the feature map. In Fig. 2, we select four images from Stanford Cars to visualize the results of feature maps from different layers, and feature maps demonstrated on each columns from top to down are from layers of 2,3 and 4. It can be seen that, the background region is much more concentrated at lower layer (layer 2 and 3) feature map and gradually the vehicle object is concentrated as the the layer increases. At layer 4, although some main discriminative parts of the vehicle (e.g. grill, light and lamps parts in the frontal vehicle) are not given much weight, the vehicle object itself has obtained the most feature weight.

For fine-grained recognition task, more weights should be given to the discriminative local part regions so that the minor differences among various classes could be differentiated efficiently. Therefore, we need to further assemble these feature maps and guide them to focus more on the discriminative part regions.

### 3.2. Part-level Feature Integration

Generally, the visual differences between two specific vehicle models might in part regions (in one or some of the component-based part regions). Since the higher-layer feature maps can not represent those minor differences very well, in this study, instead of using the detector-based method to focus on the part-level features, we extract the discriminative part features based on the object-level cues of the higher-layer feature maps.

In this step, we need to integrate the feature maps of specific layers into groups, in each of which the discriminative parts that have strong relationship between each other are expected to be integrated together. For this purpose, we adopt a basic channel grouping module, which consist of channel-wise and cross-channel operations to emphasize discriminative part features progressively and accordingly. The channels of the feature maps contain the response to different parts of the target object as well as some background regions.

Firstly, we adopt the operation of depth-wise separable convolution as a channel-level guide to establish a clear response relationship to each discriminative parts of the vehicle. Assume that the chosen $i_{th}$ object-level Conv features with a dimension of $c \times h \times w$ can be denoted as $F_i \in R^{c \times h \times w}$ , where $c$, $h$ and $w$ indicate the number of channels, height and width respectively. We apply depth-wise separable convolutions to enhance the object-level feature map, whose convolution kernels can be treated as detectors of different discriminative parts, that is ,one kernel of depth-wise convolution kernel takes charge of one channel. Since each group of the feature maps are assumed to represent different part regions, they should complement each other and have their own specificity. Then a $1 \times 1$ point-wise convolution is employed for the purpose of cross-channel feature decorrelation and redundancy reduction. The channel-wise and cross-channel operations for the higher-layer feature maps form the basic channel grouping module, which will enhance the part-related features selectively and refine the feature specificity of each channels groups.

### 3.3. Multi-scale Feature Fusion

In practice, multi-stage channel grouping modules are employed to enable the networks to gradually model the relationship of features channels. However, during the process of feature abstraction of deep networks, it is inevitable for some discriminative features to get lost or lose their proper weights. Therefore, to retain the discriminative features, we combine the object-level features of the last two layers, that is, the features from the last two convolution layers of the backbone. The reason for choosing these layers for feature fusion is that the global and subtle part-level information of the object are well extracted in these layers. As shown in Fig. 1, the feature maps from the last two layers are restructured through the multi-stage channel grouping module, then we up-sample the feature map of the last layer, and add these two feature maps derived from different object-level point by point, so that a more representative and discriminative feature representation is constructed.

After the feature fusion process, the output features are fed into a global average pooling (GAP) and a fully-connected layer follows to conduct the final feature classification. Besides, the center loss [15] is adopted to train the networks. In the experiment, we find that using one GAP layer and one fully connected layer to replace all the original fully connected layers could not only greatly reduces the amount of network parameters but also improves the prediction performance.

## 4. EXPERIMENT

In this section, a set of experiments has been performed to evaluate the proposed feature integration and fusion modules. Firstly, ablation studies were carried out to test the effectiveness of the proposed module. Then, the proposed modules

**Table 1**. Ablation study of the proposed modules(ResNet101)

| Global Feature Extraction | Feature integration | Multi-Scale Fusion | Accuracy |
|---|---|---|---|
| ✓ | ✗ | ✗ | 90.3% |
| ✓ | ✓ | ✗ | 92.6% |
| ✓ | ✓ | ✓ | 94.1% |

were tested on top of different backbone networks. Besides, we compared the recognition performance of the proposed module with that of some state-of-the-art methods. All the experiments were performed based on the fine-grained vehicle dataset, Stanford Cars.

### 4.1. Parameters and Settings

In the experiment, ResNet101 pre-trained on ImageNet is used as the object-level feature extraction module. We take MSRA to initialize parameters of other modules. The global learning rate is initialized as 0.01 and drop it by a factor of 10 after each 30 training epochs. We use a weight decay of 0.0001 and a momentum of 0.9. After the deep models converge adequately, we take the model with the best performance on test set as the optimal model.

### 4.2. Results and Analysis

Firstly, we tested the group number $K$ for the feature Integration module and the layers $C_i$ to be fused in the multi-scale fusion module. The performance results show that when $K$ is set to $512$, and the grouping results of the $4_{th}$ and $5_{th}$ convolutional layers are fused, this parameters setting achieve the best performance on the Stanford Cars.

Then, we conducted an ablation study to test the performance of the proposed integration and fusion modules, using ResNet101 as object-level feature extraction module on Stanford Cars dataset. As shown in Table 1, we first tested the performance of ResNet101 without any module added on it. Then the feature integration and the fusion module were added on ResNet101 gradually. The comparison results show that each proposed module has a positive impact on the classification accuracy (from 90.3% to 94.1%), especially the feature integration module.

We also compared the performance of the proposed method with some baseline networks as well as some recent works on the task of fine-grained vehicle recognition. The performance of the compared methods is tabulated in Table 2. As shown in Table 2, the baseline networks achieve the poorer results among all the methods, which mainly focus on the object-level features and is weak in extracting the part features. The part annotation-based methods (i.e., BOT, PA-CNN, FCAN) have achieved comparable performance with that of some attention-based methods including Kernel Pooling, RA-CNN and MA-CNN. Since the annotation-based method need extensive manual work which is not feasible for real-life applications. DCL obtained an accuracy of 94.5% which was inferior to that of the ACNet. The proposed method had acquired a competitive results which was comparable with the best result of 94.6% achieved by the ACNet. The comparison results imply that feature integration and fusion module proposed in this paper could extract the discriminative part features from the object-level features of the backbone networks and thus improve the vehicle recognition ability effectively.

**Table 2**. Performance comparison on Stanford Cars.

| Type | Method | Anno. | Acc. |
|---|---|---|---|
| Baselines | VGG19 | ✗ | 88.8% |
| | InceptionV3 | ✗ | 88.6% |
| | ResNet50 | ✗ | 89.7% |
| | ResNet101 | ✗ | 90.3% |
| Recent works | FCAN* [16] | ✓ | 91.3% |
| | BOT* [17] | ✓ | 92.5% |
| | PA-CNN* [5] | ✓ | 92.8% |
| | ResNet101-SWP* [8] | ✓ | 93.1% |
| | DenseNet161-CMP* [9] | ✓ | 93.7% |
| | B-CNN [18] | ✗ | 91.3% |
| | Kernel Pooling [19] | ✗ | 92.4% |
| | RA-CNN [20] | ✗ | 92.5% |
| | MA-CNN [12] | ✗ | 92.8% |
| | MAMC [21] | ✗ | 93.0% |
| | DCL [22] | ✗ | 94.5% |
| | ACNet [23] | ✗ | 94.6% |
| Ours | proposed method | ✗ | 94.1% |

## 5. CONCLUSION

In this paper, we propose novel light-weight feature integration and fusion modules for the task of fine-grained vehicle recognition. Referring to the idea of channel clustering, we apply multi-stage depth-wise separable convolutions to guide similar channels to jointly express each discriminative local part, multi-scale discriminative features are then fused to construct a more expressive feature representation. We perform comparison experiments on public dataset, and the results show that our method is comparable to state-of-the-art fine-grained vehicle recognition methods.

## 6. ACKNOWLEDGEMENTS

This work was supported by the National Natural Science Foundation of China under Grant No. 62071373 and Shaanxi Province Smart Wireless Network and Ubiquitous Access Innovation Team No. 2021TD-08.

## 7. REFERENCES

[1] Lei Lu and Hua Huang, "A hierarchical scheme for vehicle make and model recognition from frontal images

of vehicles," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 5, pp. 1774–1786, 2018.

[2] Lei Lu and Hua Huang, "Component-based feature extraction and representation schemes for vehicle make and model recognition," *Neurocomputing*, vol. 372, pp. 92–99, 2020.

[3] Ning Zhang, Ryan Farrell, Forrest Iandola, and Trevor Darrell, "Deformable part descriptors for fine-grained recognition and attribute prediction," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 729–736.

[4] Liang Liao, Ruimin Hu, Jun Xiao, Qi Wang, Jing Xiao, and Jun Chen, "Exploiting effects of parts in fine-grained categorization of vehicles," in *2015 IEEE international conference on image processing (ICIP)*. IEEE, 2015, pp. 745–749.

[5] Jonathan Krause, Hailin Jin, Jianchao Yang, and Li Fei-Fei, "Fine-grained recognition without part annotations," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 5546–5555.

[6] Michael Lam, Behrooz Mahasseni, and Sinisa Todorovic, "Fine-grained recognition as hsnet search for informative image parts," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2520–2529.

[7] Xiu-Shen Wei, Jian-Hao Luo, Jianxin Wu, and Zhi-Hua Zhou, "Selective convolutional descriptor aggregation for fine-grained image retrieval," *IEEE Transactions on Image Processing*, vol. 26, no. 6, pp. 2868–2881, 2017.

[8] Qichang Hu, Huibing Wang, Teng Li, and Chunhua Shen, "Deep cnns with spatially weighted pooling for fine-grained car recognition," *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 11, pp. 3147–3156, 2017.

[9] Zhanyu Ma, Dongliang Chang, Jiyang Xie, Yifeng Ding, Shaoguo Wen, Xiaoxu Li, Zhongwei Si, and Jun Guo, "Fine-grained vehicle classification with channel max pooling modified cnns," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 4, pp. 3224–3233, 2019.

[10] Pau Rodríguez, Josep M Gonfaus, Guillem Cucurull, F XavierRoca, and Jordi Gonzalez, "Attend and rectify: a gated attention mechanism for fine-grained recovery," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 349–364.

[11] Xiaopeng Zhang, Hongkai Xiong, Wengang Zhou, Weiyao Lin, and Qi Tian, "Picking deep filter responses for fine-grained image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1134–1142.

[12] Heliang Zheng, Jianlong Fu, Tao Mei, and Jiebo Luo, "Learning multi-attention convolutional neural network for fine-grained image recognition," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 5209–5217.

[13] Ye Xiang, Ying Fu, and Hua Huang, "Global topology constraint network for fine-grained vehicle recognition," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 7, pp. 2918–2929, 2019.

[14] Yifan Zhao, Jia Li, Xiaowu Chen, and Yonghong Tian, "Part-guided relational transformers for fine-grained visual recognition," *IEEE Transactions on Image Processing*, vol. 30, pp. 9470–9481, 2021.

[15] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao, "A discriminative feature learning approach for deep face recognition," in *European conference on computer vision*. Springer, 2016, pp. 499–515.

[16] Xiao Liu, Tian Xia, Jiang Wang, and Yuanqing Lin, "Fully convolutional attention localization networks: Efficient attention localization for fine-grained recognition," *arXiv preprint arXiv:1603.06765*, vol. 1, no. 2, pp. 4, 2016.

[17] Yaming Wang, Jonghyun Choi, Vlad Morariu, and Larry S Davis, "Mining discriminative triplets of patches for fine-grained classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1163–1172.

[18] Tsung-Yu Lin, Aruni RoyChowdhury, and Subhransu Maji, "Bilinear cnn models for fine-grained visual recognition," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1449–1457.

[19] Yin Cui, Feng Zhou, Jiang Wang, Xiao Liu, Yuanqing Lin, and Serge Belongie, "Kernel pooling for convolutional neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2921–2930.

[20] Jianlong Fu, Heliang Zheng, and Tao Mei, "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*, 2017, pp. 4438–4446.

[21] Ming Sun, Yuchen Yuan, Feng Zhou, and Errui Ding, "Multi-attention multi-class constraint for fine-grained image recognition," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 805–821.

[22] Yue Chen, Yalong Bai, Wei Zhang, and Tao Mei, "Destruction and construction learning for fine-grained image recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5157–5166.

[23] Ruyi Ji, Longyin Wen, Libo Zhang, Dawei Du, Yanjun Wu, Chen Zhao, Xianglong Liu, and Feiyue Huang, "Attention convolutional binary neural tree for fine-grained visual categorization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10468–10477.