Journal of Network Intelligence
Taiwan Ubiquitous Information
©2021 ISSN 2414-8105 (Online)
Volume **6**, Number **2**, May **2021**

Multi-Scale Discriminative Regions Attention Network for Fine-grained Vehicle Classification

Wen-Zhong Rong

College of Computer Science and Engineering
Shandong University of Science and Technology
Qingdao 266590, P.R. China
sdhtrwz@163.com

Jin Han*

College of Computer Science and Engineering
Shandong University of Science and Technology
Qingdao 266590, P.R. China
shnk123@163.com

Ying-Hao Cai*

The State Key Laboratory of Management and Control for Complex Systems
Institute of Automation, Chinese Academy of Sciences
Beijing 100190, P.R. China
yinghao.cai@ia.ac.cn

Gen Liu

College of Computer Science and Engineering
Shandong University of Science and Technology
Qingdao 266590, P.R. China
lg97@sdust.edu.cn

*Corresponding Authors: Jin Han(shnk123@163.com), Ying-Hao Cai(yinghao.cai@ia.ac.cn)

Received January 2021; revised March 2021

Abstract. *Fine-grained vehicle classification is a challenging task in computer vision due to the low intra-class variance. Some methods have been developed to improve the accuracy of fine-grained vehicle classification by improving the ability of discriminative features extraction, but there is still room for further improvement in the localization accuracy of vehicle discriminative regions. Based on deep convolutional neural networks, we focus on finding a more efficient structure that pays more attention to the discriminative image regions to enhance the ability of fine-grained vehicle classification. We propose a novel Multi-Scale Discriminative Regions Attention Network (MS-DRAN), which extracts feature maps through ResNet-50 backbone network and generates multi-scale feature maps by a Feature Pyramid Network (FPN). Then, MS-DRAN generates discriminative regions attention maps on the multi-scale feature maps. The attention maps from shallow layer perform pixel-level multiplication with the feature maps from deeper layer. By this way, the network gradually extracts a more discriminating feature maps for classification. We also design a multi-task loss function to associate the classification results for each scale and optimize the network parameter in training. Lastly, we validate the MS-DRAN on Stanford Cars-196 dataset and CompCars dataset and achieves* 94.3% *and* 98.1% *in accuracy, respectively.*

Keywords: fine-grained classification; vehicle classification; discriminative feature map; multi-scale attention map; attention mechanism.

1. **Introduction.** With the development of society, transportation security and social safety have received considerable attention, which leads a great deal of research in vehicle classification [1–5], vehicle detection [6–11], Vehicle instance segmentation [12–14] and fine-grained vehicle classification [15–18]. Benefitting from the rapid development of convolutional neural networks (CNNs) [19, 20], research fields mentioned above have made significant progresses, the performance of many applications have been significantly improved. However, there still exits challenges in these fields. For fine-grained vehicle classification task, improving the classification network to extract finer discriminative features and make it more focused on vehicle discriminative regions is the critical issue to improve classification accuracy. We focus on fine-grained vehicle classification task, which aims to identify the make, model, and year of vehicles, such as Volkswagen, Passat, 2015. Due to the low inter-class variance of vehicles in appearance, generic classification networks [21, 22] are already struggling to achieve satisfactory accuracy. The reason is that these classification networks do not have an effective mechanism to notice the discriminative regions of the image which is critical to the fine-grained classification. Therefore, this paper provides an in-depth study on the problem of how to effectively make the network automatically generate higher attention to discriminative regions, so that the network can better extract discriminative features and improve fine-grained vehicle classification accuracy.

Before the deep learning era, image classification was dominated by the traditional image classification methods. Lim et al. [23] use Gabor filters to extract features for car recognition. Support Vector Machine method (SVM) [24] is one of the machine learning methods based on statistic theory. Based on SVM, Sinatra et al. [25] use PCA dimensionality reduction method to reduce the dimensionality of the features of the vehicle image, and use SVM to classify the vehicle image. Mehran et al. [26] proposed a vehicle recognition method based on features of vehicle rear part. The method first detects the rear part of the vehicle, then extracts features, and finally classifies the vehicle using a hybrid dynamic Bayesian network. Arróspide et al. [27] use Histogram of Oriented Gradient (HOG) features for vehicle classification. Vehicle classification methods based on traditional methods mostly use SIFT and HOG features, these hand-designed features are cumbersome, poorly robust, and difficult to apply to fine-grained classification.

With the successful application of deep convolutional neural networks [21, 28–31], some fine-grained vehicle classification works based on CNNs proposed some novel approaches and made remarkable progress. Hu et al. [32] proposed a spatially weighted pooling strategy that considerable improves the robustness and effectiveness of the feature representation of most dominant DCNNs. In [33], the authors proposed an iterative discrimination CNN (ID-CNN) based on selective multi-convolutional region (SMCR) feature extraction, and achieved state-of-the-art performance on the Stanford Cars-196 dataset [34]. Ma et al. [18] proposed a channel max pooling (CMP) which divides the feature maps within each sub-group into a new one and reduces the number of parameters. They achieved advanced 97.89% in accuracy on the CompCars dataset [35]. These works focus on improving the structure of the networks to extract more discriminative features and achieve effective improvements. There are also some works that attempt to improve the loss function for fine-grained classification. In [36] the authors proposed a generalized large-margin softmax (L-Softmax) loss which explicitly compacts the intra-class and separates the inter-class between features. Wen et al. [37] proposed center loss function which learns a center for deep features of each class and penalizes the distances between the deep features and their corresponding class centers. In [17], the authors proposed a Dual Cross-Entropy Loss built on CE loss and improved the classification performance on small-sample datasets [38, 39].

The critical issue for fine-grained vehicle classification is whether the classification network can draw more attention to the discriminative regions of the vehicle image. We believe that although the inter-class variance in fine-grained vehicle classification is low, the classification network can still accomplish accurate fine-grained classification by identifying certain discriminative regions of vehicle appearance. Inspired by the novel idea in [40] which maps the predicted class score back to the previous convolutional layer to generate the class activation maps (CAMs) that highlight the class-specific discriminative regions. We propose a novel Multi-Scale Discriminative Regions Attention Network (MS-DRAN) which generates discriminative regions attention maps on multi-scale feature maps under the guidance of the gradients from scores of each category. The attention maps from shallow layer are down-sampled to half of its original size and perform element-level multiplication with the feature maps from deeper layer. Specifically, a ResNet-50 [41] backbone network and a Feature Pyramid Network (FPN) [42] are used to extract feature pyramid maps. At each scale of the feature pyramid maps, we perform Global Average Pooling (GAP) [43] to get a K-dimensional vector, and then a full connection layer and a K-way Softmax layer are performed to output the categorical probabilities. The discriminative regions attention maps are generated under the guidance of the gradients from scores of each category at 8x and 16x down-sampled feature maps. To facilitate description, we call the discriminative regions attention maps as the attention maps. The attention maps from the shallow layer (8x down-sampled scale) are down-sampled to half of its original size and perform pixel-level multiplication with the feature maps from the deeper layer (16x down-sampled scale). To eliminate the perturbation to the feature maps caused by the multiplication operation, an additional normalization operation is performed on the attention maps. We also improve the loss function by adding an additional inter-class ranking loss term to enforce the finer-scale (or the deeper-scale) to generate more confident predictions. We train and validate MS-DRAN on Stanford Cars-196 [34] and CompCars [35] fine-grained vehicle classification datasets, and the experimental results demonstrate the effectiveness of our proposed method.

The main contribution of this paper is to propose a novel method to generate attention maps for discriminative regions, and to make the responses of attention maps to discriminative regions progressively more accurate by using multi-scale pyramidal feature maps. Compared with other methods, the MS-DRAN proposed in this paper does not require additional vehicle component annotations to enable the network to accurately locate vehicle discriminative component regions. We also design a sensible visualization experiment to map the attention map generated by the network to the original image, which visualizes the distribution characteristics of the regions of interest when the network classifies vehicles at a fine-grained level.

2. **Related Work.**

2.1. **Fine-grained Classification Methods.**

The task of fine-grained classification is to identify the subclass under the broad category. Research topics related to fine-grained classification include identification of different species of birds [44–47], dogs [48], flowers [49], vehicles [15–18,32–35], aircraft [50,51], and etc. Some early approaches [52–55] did not use any part annotations, but their performance has been eclipsed by methods developed to explicitly take advantages of the structure present in fine-grained classes [56–60]. In [34, 61], the authors tried to reconstruct a 3-D representation of the target by using 2-D part annotations for fine-grained classification. However, part annotations for large-scale datasets are very labor-intensive, which limits the applications of fine-grained classification algorithms in other domains.

Some works aimed to find localized regions or parts in an unsupervised way. Zhang et al. [62] proposed to learn a set of part detectors by analyzing filter responses from CNN that respond to specific patterns consistently in an unsupervised way. Methods [63–65] proposed to zoom in on discriminative local regions to improve the performance of fine-grained classification. Zheng et al. [66] proposed a multiple attention convolutional neural network which uses channel grouping network to locate the discriminative partial regions.

2.2. **Convolutional Networks.** Since AlexNet [67] won the 2012 ImageNet competition, Convolutional networks have evolved considerably. VGGNet [21] uses small-scale convolutional kernels and pooling kernels to construct deep architecture for extraction of features at low spatial resolution. In [29], the authors introduced a structure known as inception block which allows abstraction of features at different spatial scales. In 2015, He et al. [30] proposed ResNet which leads to the conception of skip connections. Afterward, this concept was used by most of the succeeding networks, such as Inception-ResNet [68], ResNeXt [41], etc. In the past few years, some works focus on the improvement architectural design of the network, such as channel boosting [69], attention mechanism [70], etc

3. **Method.** Although vehicles produced by the same make are stylistically very uniform, there are still many details that differ from one model to another, and if these details are observed, the specific model of the vehicle can be identified, as shown in Figure 1, experienced experts can accurately identify the make, model, and year of the vehicle based on some of its partial appearances (e.g., headlights, taillights, grille, etc.).

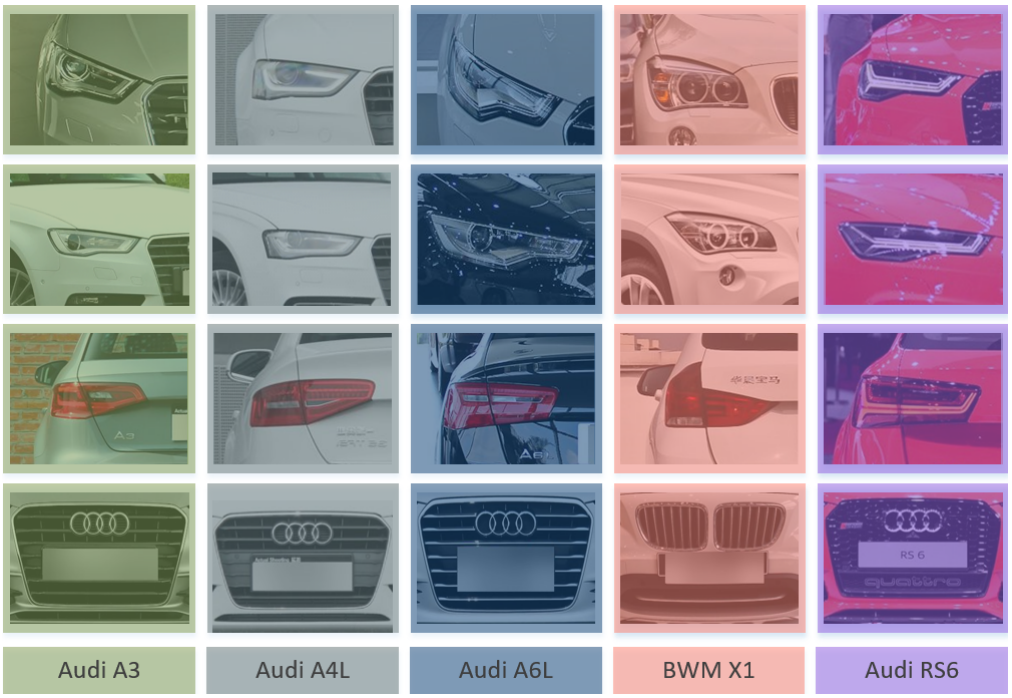

FIGURE 1. Each column from left to right is the local appearance of Audi 3, Audi A4L, Audi A6L, BMW X1, and Audi RS6. Each row from top to bottom is the front view of the front lights, the oblique view of the front lights, the front view of the tail lights, and the grilles. It is easy for humans to identify the vehicle type by their local appearance

To categorize vehicle types at a fine-grained level, it is necessary to equip the classification networks with the capability to extract discriminative features. We propose the

Multi-Scale Discriminative Regions Attention Network (MS-DRAN, as shown in Figure 2), which aims at supervising multi-scale discriminative regions attention maps when we train the network for the task of fine-grained vehicle classification. In this way, the network gradually generates increasingly accurate predictions from shallow feature maps to deep feature maps. The proposed MS-DRAN predicts categorical scores at each scale of the pyramid feature maps, and the prediction from the last scale is the final prediction. Gradients from the scores of each category at the first and the second scale are used to generate discriminative regions attention maps, which will be described in detail in Section 3.2. The discriminative regions attention maps help the network pay more attention to the discriminative regions, which is helpful for fine-grained vehicle classification.

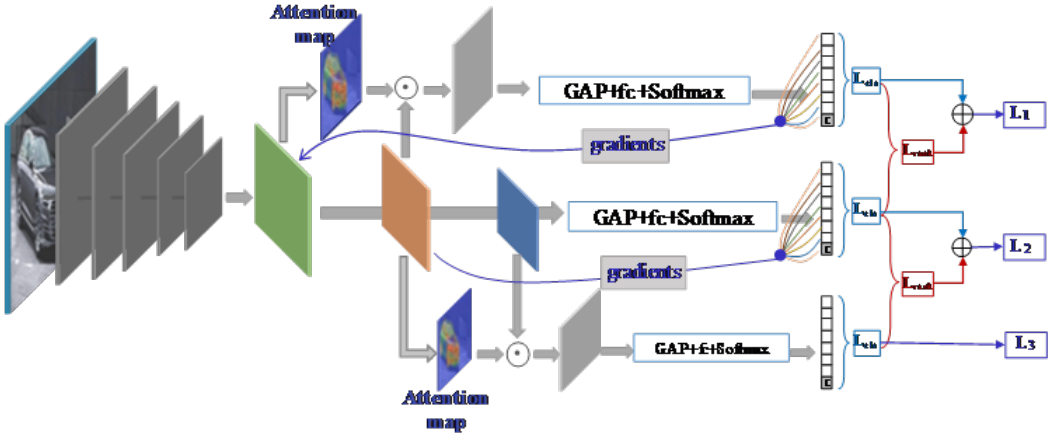

FIGURE 2. Framework of the Multi-Scale Discriminative Regions Attention Network (MS-DRAN).

3.1. **Class-Discriminative Feature Maps.** Convolutional features naturally retain spatial information that would be lost in fully-connected layers, which limits the sensitivity of the classification networks to discriminative regions. Even though some approaches attempt to improve the loss function or perform attention mechanism on feature maps, these methods still struggle to eliminate the destruction of spatial information by the fully-connected layers. To solve this issue, we design an architecture to generate class-discriminative maps.

Let $\mathrm{F}^{c}_{disc} \in \mathbb{R}^{u \times v \times k}$ be the class-discriminative feature maps of category c, where u and v denote the width and height of the class-discriminative feature map, and k denotes the number of categories. The score for category c, y^c is the output of the fully-connected layer. The class-discriminative feature map is calculated as

$$F^{c}_{disc} = Leaky\mathrm{ReLU}\left(\frac{\partial y^{c}}{\partial A_{S}}\right) \tag{1}$$

A_S is the last convolutional layer of the pyramid feature map at scale S. Consider that an increase or decrease in the gradient of each pixel on the class-discriminative feature map has an effect on y^c, we apply a Leaky ReLU to the gradient of the score for classes. Without performing the Leaky ReLU, pixels with negative gradients will introduce too much interference in global average pooling (GAP).

3.2. **Discriminative Regions Attention Maps.** The discriminative regions feature maps generated at a specific scale assign weight to each channel of the feature map. For the k category predictions, a total of k groups of the discriminative regions feature maps

are derived. We use $F_S \in \mathbb{R}^{\frac{u}{2} \times \frac{v}{2} \times k}$ to represent the attention maps on the scale S. F_S is calculated as

$$\mathrm{F}_S = Conv\left(\sum_{x,y} w^c \mathrm{A}_S^c\left(x,y\right)\right) \quad c \in (1,k) \tag{2}$$

w^c is calculated as

$$w^c = GAP\left(\mathrm{F}_{disc}^c\right) \quad c \in (1,k) \tag{3}$$

We perform global average pooling on the class-discriminative feature maps for each group to derive k vectors with dimension k. Followed by a convolutional layer, the attention map is down-sampled to half of its original size. The attention map generation details are shown in Figure 3.

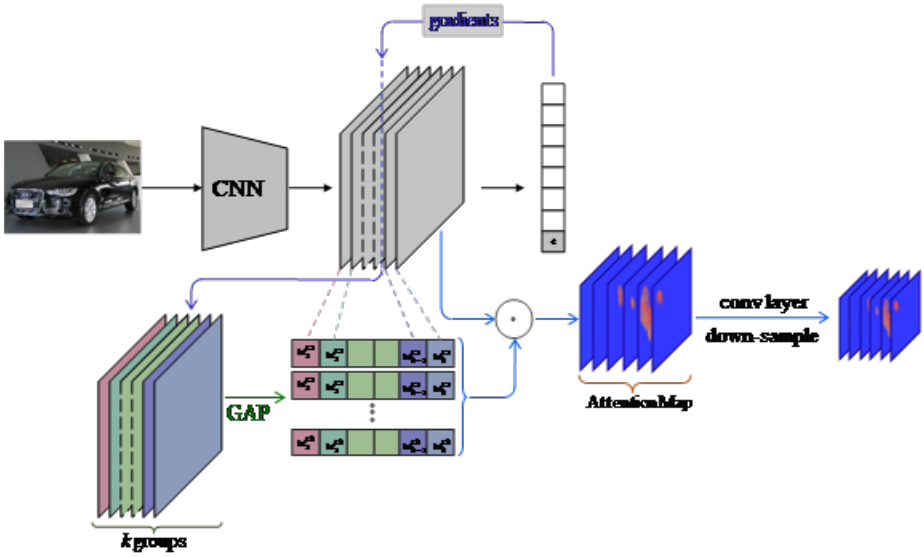

FIGURE 3. Details of the attention map generation.

3.2.1. *Prediction Head.* Given an input image $\mathbf{X}$, the CNN backbone network and FPN extracts deep representations at multi-scale that are represented as

$$\mathrm{A}_S = \mathrm{W}_o * \mathrm{X} \tag{4}$$

where $*$ denotes a set of operations of convolution, pooling and activation, and W_o denotes the overall parameters of the CNN backbone network and FPN. For facilitating the description, we denote the feature maps of 8x, 16x, and 32x downsampling by $\mathrm{A}_1, \mathrm{A}_2$, and A_3. As shown in Figure 2, there are total three prediction heads for categorical predictions. The probability distribution $\mathbf{p}$ over fine-grained categories is calculated as

$$\mathrm{p}_S = f\left(A'_S\right) \tag{5}$$

where $f(.)$ represents a set of operations of GAP, fully-connected layer, and softmax. A'_S represents the last convolutional feature map before the prediction head

$$\begin{cases} A'_1 = A_1 \\ A'_2 = F_1{}^c(x,y) A_2{}^c(x,y) \\ A'_3 = F_2{}^c(x,y) A_3{}^c(x,y) \end{cases} \quad c \in (1,k) \tag{6}$$

The proposed MS-DRAN is optimized by two types of supervision, i.e. intra-scale classification loss and inter-scale pairwise ranking loss. The loss function for an image $\mathbf{X}$ is defined as

$$L(\mathbf{X}) = \sum_{S=1}^{3} \left(L_{cls}\left(\mathrm{Y}_S, \mathrm{Y}^*\right)\right) + \sum_{S=1}^{2} \left(L_{rank}\left(y_S^c, y_{S+1}^c\right)\right) \tag{7}$$

where Y_S and Y^* denote the prediction over categories at scale S and the ground truth label over categories, respectively. L_{cls} represents classification loss, which is a focal-loss form as

$$L_{cls} = -\left(1 - y_S^c\right)^{\gamma} \log y_S^c \tag{8}$$

y_S^c is the prediction probability on the correct category label c. L_{rank} is the ranking loss which is given by

$$L_{rank}\left(y_S^c, y_{S+1}^c\right) = max\left\{0, y_S^c - y_{S+1}^c + m\right\} \tag{9}$$

The ranking loss can enforce the finer-scale (or the deeper-scale) to generate more confident predictions. We set the $m = 0.005$ by default.

4. **Experiment.** In this section, we first introduce the datasets and our experimental settings. Then, the evaluation metrics are presented. Finally, we report the results on the benchmark datasets and show the analysis.

4.1. **Datasets.** To evaluate the proposed MS-DRAN, we comprehensively evaluate the MS-DRAN on the CompCars dataset [35]and Stanford Cars-196 dataset [34], which have been widely applied to evaluating the performance of fine-grained vehicle classification. CompCars dataset for fine-grained vehicle classification contains 30955 vehicle images from 431 models and it is split into two parts in a $50\% - 50\%$ ratio for training and testing, respectively. Stanford Cars-196 dataset contains 16185 images from 196 car models. The detailed statistics of the category numbers and the dataset partition are summarized in Table 1.

Table 1. The statistics of fine-grained vehicle classification datasets.

Datasets	Category	Training split	Test split
Stanford Cars-196	196	8144	8041
CompCars	431	36456	15627

4.2. **Evaluation Metrics.** The performance of the fine-grained vehicle classification is evaluated by overall accuracy, which is calculated as

$$Overall\ Accuracy = \frac{\sum_{c=1}^{k} P(c)}{N} \tag{10}$$

where $P(c)$ is the accurate prediction times of class c, N is the quantity of samples.

4.3. **End-to-end Training.** The experiment uses the Pytorch deep learning framework, which is carried out on the AMD Ryzen 3600X CPU@4.4GHz with 16GB RAM, and uses the NVIDIA GTX1080Ti GPU for acceleration. Before the training, the backbone network of the proposed MS-DRAN is pre-trained on ImageNet [71]. We use ResNet-50 as the backbone network. The training process spans 100 epochs, and the SGD-with momentum optimization method is adopted. The learning rate is set to 10^{-1} for the first 50 epochs, and divided by 10 at 65 and 90 epochs. The momentum is set to 0.9.

4.4. **Baselines.** We use the method in [72], ResNet-50, SE-ResNet [73], VGG19, and DenseNet161-CMP [18] as our baselines. The structure settings of ResNet-50 and VGG19 are described in below.

1)The vehicle classification network based on VGG19 consists of two parts, the backbone network and classifier. The numbers of the hidden nodes for the fully-connected layer are 256 for Stanford Cars-196 and 512 for CompCars, respectively. A P-way softmax layer is adopted, where P is the number of categories in the dataset.

2)The vehicle classification network based on ResNet-50 is similar as VGG19 based classification network. The fully-connected layer has 256 nodes for Stanford Cars-196 and 512 for CompCars, respectively, and a P-way softmax layer is adopted. In addition, we construct another classifier with a GAP layer and fully-connected layer, which we call it ResNet-50-GAP. The ResNet-50, SE-ResNet, VGG19, and ResNet-GAP are trained on each car dataset. For the training set, we use the SGD-with momentum. The hyper-parameters are same as the proposed MS-DRAN. For the method in [72] and DenseNet161-CMP, we use the validation results from the papers.

4.5. **Results and Analysis.**

4.5.1. *Comparison with Baselines.* Table 2 shows the recognition results on each fine-grained vehicle dataset. On the Stanford Cars-196 dataset, the VGG19 network achieves 90.6% in accuracy on Cars-196 dataset and 92.5% on CompCars dataset. Due to the greater number of parameters and deeper network structure, ResNet-50 achieves 90.8% in accuracy on Cars-196 dataset and 94.0% on CompCars dataset, which outperforms 0.2% and 1.5% than VGG19. With the same backbone network, ResNet-50-GAP and SE-ResNet outperform ResNet-50 by 0.5% and 1.3% on Cars-196 dataset, respectively, and outperform ResNet-50 by 0.3% and 0.2% on CompCars dataset, respectively. Method in [72] makes the CNN streams to automatically learn to attend critical object parts via a deep attention-based spatially recursing model, and achieves 93.4% in accuracy on Cars-196 dataset. The DenseNet161-CMP learns more robust discriminative features by the so-called channel max pooling (CMP) and achieves 93.7% in accuracy on Cars-196 dataset and 97.9% on CompCars dataset. It can be seen that the accuracy of fine-grained vehicle classification can be significantly improved by leveraging discriminative features. The MS-DRAN we proposed achieves the highest accuracy in Table 2 on both the Stanford Cars-196 dataset and CompCars dataset, with accuracies of 94.3% and 98.1%, respectively.

Table 2. Comparison of recognition results on Stanford Cars-196 dataset and CompCars dataset

Methods	Overall Accuracy	
	Stanford Cars-196	CompCars
VGG19	90.6	92.5
ResNet-50	90.8	94.0
ResNet-50-GAP	91.3	94.3
SE-ResNet	92.1	94.2
Wu et al. [72]	93.4	83.0
DenseNet161-CMP	93.7	97.9
MS-DRAN (ours)	94.3	98.1

4.5.2. *Ablation Studies.* In this section, we conduct ablation studies to verify the effectiveness of the components of the proposed MS-DRAN separately. We use the classification network composed of ResNet-50 backbone network, fully-connected layer, and a P-way

Softmax layer as the baseline. All the ablation studies are carried out on CompCars dataset. As shown in Table 3, the baseline classification network achieves 94.6% accuracy. We add a GAP layer in front of the fully-connected layer, the performance of the network increased. The accuracy has increased by 0.7%. By adding a FPN, the network outputs classification results at each of the three scales, the accuracies are 94.6%, 95.1%, and 95.5% for 8x, 16x, and 32x downsampling feature maps, respectively. At the last scale, the accuracy improves by only 0.2%. Then, we validate the proposed MS-DRAN without the additional constraint L_{rank} in the loss function, and the network achieves 97.8% in accuracy. With the constraint L_{rank}, the MS-DRAN ultimately achieves an accuracy of 98.1%.

TABLE 3. Ablation study of the MS-DRAN on CompCars dataset.

model						Overall accuracy
ResNet-50	Fc+softmax	GAP	FPN	Attention map	L_{rank}	
✓	✓					94.6
✓		✓				95.3
✓		✓	✓			95.5
✓		✓	✓	✓		97.8
✓		✓	✓	✓	✓	98.1

4.5.3. *Visualization and Discussion.* We show qualitative results of the discriminative regions that are highlighted in Figure 4. We chose the last scale of the classification pipeline for visualization. For ease of description, we use $\mathrm{A}_3\,(x, y) \in \mathbb{R}^{u_3 \times v_3 \times k}$ to represent the element on position (x, y) of the last feature map at the last scale. The score from the correct category S_c, we obtain

$$S_c = \sum_k w_k^c \frac{\sum_{x,y} A_3^k\,(x, y)}{u_3 \times v_3} \tag{11}$$

We define M_c as the discriminative regions for class c, where each element of M_c is given by

$$M_c\,(x, y) = \sum_k w_k^c \mathrm{A}_3\,(x, y) \tag{12}$$

We then use the Nearest-neighbor algorithm to amplify M_c to the original image size and visualize M_c and the original image together. After training, the highlight regions tend to be located in the discriminative parts such as lights, vehicle logo, grille, etc. Therefore, the proposed MS-DRAN has significantly beneficial effects on discriminative feature extraction and location, which makes the network obtain better performance.

FIGURE 4. Visualization of the discriminative regions. We map each spatial location in the attention map back to the patch in the original image. The results show that the proposed MS-DRAN pays more attention to the discriminative regions of the vehicle.

5. **Conclusion.** In this paper, we propose a novel multi-scale discriminative region attention network (MS-DRAN) for fine-grained vehicle classification. The proposed MS-DRAN generates discriminative regions attention maps on multi-scale feature maps under the guidance of the gradients from scores of each category. The attention maps from shallow layer are down-sampled to half of its original size and perform pixel-level multiplication with the feature maps from the deeper layer. With such a design, the classification network gradually extracts more refined and robust discriminative features. The validation on Stanford Cars-196 dataset and CompCars dataset shows that the proposed MS-DRAN achieves state-of-the-art performance. In the future it may be illuminating to deploy our method on other high-level tasks than categorization, such as fine-grained vehicle detection.

Acknowledgment. This work is supported by National Natural Science Foundation of China (U1913201), The Key Project of Shandong Provincial Natural Science Foundation (ZR2020KE023), Collaborative Education Project of Ministry of Education (201901055015), Postgraduate Education Quality Improvement Project of Educational Commission of Shandong Province of China (ADYAL17034), and Excellent Teaching Team Support Project of Shandong University of Science and Technology (JXTD20170503).

REFERENCES

[1] X. MA and W.E.L. GRIMSON. Edge-based rich representation for vehicle classification, *In Tenth IEEE International Conference on Computer Vision (ICCV'05)*, Volume 1, pp. 1185-1192, 2005.

[2] P. DAIGAVANE, P.R. BAJAJ and M. DAIGAVANE. Vehicle detection and neural network application for vehicle classification, *In 2011 International Conference on Computational Intelligence and Communication Networks*, pp. 758-762, 2011.

[3] X. MEI and H. LING, Robust visual tracking and vehicle classification via sparse representation, *IEEE transactions on pattern analysis and machine intelligence*, vol. 33, no. 11: pp. 2259-2272, 2011.

[4] D. ZHAO, Y. CHEN and L. LV, Deep reinforcement learning with visual attention for vehicle classification. *IEEE Transactions on Cognitive and Developmental Systems*,vol. 9, no. 4: pp. 356-367, 2016.

[5] Z. LUO, F. BRANCHAUD-CHARRON, C. LEMAIRE, J. KONRAD, S. LI, A. MISHRA, A. ACHKAR, J. EICHEL, and P.-M. JODOIN, MIO-TCD: A new benchmark dataset for vehicle classification and localization. *IEEE Transactions on Image Processing*,vol. 27, no. 10: pp. 5129-5141, 2018.

[6] M. VARGAS, S. TORAL, F. BARRERO, and J. MILLA. An enhanced background estimation algorithm for vehicle detection in urban traffic video. *In 2008 11th International IEEE Conference on Intelligent Transportation Systems*, pp. 784-790, 2008.

[7] V.D. NGUYEN, T.T. NGUYEN, D.D. NGUYEN, S.J. LEE, and J.W. JEON, A fast evolutionary algorithm for real-time vehicle detection. *IEEE Transactions on Vehicular Technology*,vol. 62, no. 6: pp. 2453-2468, 2013.

[8] L. WEN, D. DU, Z. CAI, Z. LEI, M.-C. CHANG, H. QI, J. LIM, M.-H. YANG, and S. LYU, UA-DETRAC: A new benchmark and protocol for multi-object detection and tracking. *arXiv preprint*, arXiv:1511.04136, 2015.

[9] Q. FAN, L. BROWN and J. SMITH. A closer look at Faster R-CNN for vehicle detection.*In 2016 IEEE intelligent vehicles symposium* (IV), pp. 124-129, 2016.

[10] L. WANG, Y. LU, H. WANG, Y. ZHENG, H. YE, and X. XUE. Evolving boxes for fast vehicle detection. *In 2017 IEEE international conference on multimedia and Expo (ICME)*, pp. 1135-1140, 2017.

[11] J. SANG, Z. WU, P. GUO, H. HU, H. XIANG, Q. ZHANG, and B. CAI, An improved YOLOv2 for vehicle detection.*sensors*,vol. 18, no. 12: pp. 4272, 2018.

[12] L. MOU and X.X. ZHU, Vehicle instance segmentation from aerial image and video using a multitask learning residual fully convolutional network. *IEEE Transactions on Geoscience and Remote Sensing*,vol. 56, no. 11: pp. 6699-6711, 2018.

[13] G. HUANG, Y. MA and Y. YU. Vehicle segmentation from remote sensing images using the small object segmentation convolutional network. *In 2017 4th International Conference on Systems and Informatics (ICSAI)*, pp. 1292-1296, 2017.

[14] N. AUDEBERT, B. LE SAUX and S. LEFèVRE, Segment-before-detect: Vehicle detection and classification through semantic segmentation of aerial images. *Remote Sensing*,vol. 9, no. 4: pp. 368, 2017.

[15] J. FANG, Y. ZHOU, Y. YU, and S. DU, Fine-grained vehicle model recognition using a coarse-to-fine convolutional neural network architecture. *IEEE Transactions on Intelligent Transportation Systems*,vol. 18, no. 7: pp. 1782-1792, 2016.

[16] S. YU, Y. WU, W. LI, Z. SONG, and W. ZENG, A model for fine-grained vehicle classification based on deep learning. *Neurocomputing* ,vol. 257: pp. 97-103, 2017.

[17] X. LI, L. YU, D. CHANG, Z. MA, and J. CAO, Dual cross-entropy loss for small-sample fine-grained vehicle classification. *IEEE Transactions on Vehicular Technology*,vol. 68, no. 5: pp. 4204-4212, 2019.

[18] Z. MA, D. CHANG, J. XIE, Y. DING, S. WEN, X. LI, Z. SI, and J. GUO, Fine-grained vehicle classification with channel max pooling modified CNNs. *IEEE Transactions on Vehicular Technology*,vol. 68, no. 4: pp. 3224-3233, 2019.

[19] Y.L. CUN, B. BOSER, J. DENKER, D. HENDERSON, R. HOWARD, W. HUBBARD, and L. JACKEL. Handwritten digit recognition with a back-propagation network. *In Advances in Neural Information Processing Systems*, 1990.

[20] Y. LECUN and L. BOTTOU, Gradient-based learning applied to document recognition. *Proceedings of the IEEE*,vol. 86, no. 11: pp. 2278-2324, 1998.

[21] K. SIMONYAN and A. ZISSERMAN, Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv*:1409.1556, 2014.

[22] M.D. ZEILER and R. FERGUS. Visualizing and understanding convolutional networks.*In European conference on computer vision*, pp. 818-833, 2014.
[23] T.R. LIM and A.T. GUNTORO. Car recognition using Gabor filter feature extraction. *In Asia-Pacific Conference on Circuits and Systems*, pp. 451-455, 2002.
[24] C. CORTES and V. VAPNIK, Support-vector networks.*Machine learning*,vol. 20, no. 3: pp. 273-297, 1995.
[25] B. CUI, T. XUE and K. YANG. Vehicle recognition based on support vector machine. *In 2008 International Symposium on Intelligent Information Technology Application Workshops*, pp. 443-446, 2008.
[26] M. KAFAI and B. BHANU, Dynamic Bayesian networks for vehicle classification in video. *IEEE Transactions on Industrial Informatics*,vol. 8, no. 1: pp. 100-109, 2011.
[27] J. ARRóSPIDE, L. SALGADO and M. CAMPLANI, Image-based on-road vehicle detection using cost-effective histograms of oriented gradients. *Journal of Visual Communication and Image Representation*,vol. 24, no. 7: pp. 1182-1190, 2013.
[28] A. KRIZHEVSKY, I. SUTSKEVER and G.E. HINTON. Imagenet classification with deep convolutional neural networks. *In Advances in neural information processing systems*, pp. 1097-1105, 2012.
[29] C. SZEGEDY, W. LIU, Y. JIA, P. SERMANET, S. REED, D. ANGUELOV, D. ERHAN, V. VANHOUCKE, and A. RABINOVICH. Going deeper with convolutions.*In Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1-9, 2015.
[30] K. HE, X. ZHANG, S. REN, and J. SUN. Deep residual learning for image recognition. *In Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770-778, 2016.
[31] G.-Y. KANG, Z.-Q. LU and Z.-M. LU, Lightweight Human Pose Estimation Network and Angle-based Action Recognition.*Journal of Network Intelligence*,vol. 5, no. 4: pp. 240-249,
[32] Q. HU, H. WANG, T. LI, and C. 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.
[33] Y. TIAN, W. ZHANG, Q. ZHANG, G. LU, and X. WU. Selective multi-convolutional region feature extraction based iterative discrimination CNN for fine-grained vehicle model recognition. *In 2018 24th International Conference on Pattern Recognition (ICPR)*, pp. 3279-3284, 2018.
[34] J. KRAUSE, M. STARK, J. DENG, and L. FEI-FEI. 3d object representations for fine-grained categorization. *In Proceedings of the IEEE international conference on computer vision workshops*, pp. 554-561, 2013.
[35] L. YANG, P. LUO, C. CHANGE LOY, and X. TANG. A large-scale car dataset for fine-grained categorization and verification.*In Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3973-3981, 2015.
[36] W. LIU, Y. WEN, Z. YU, and M. YANG. Large-margin softmax loss for convolutional neural networks. *In ICML*, pp. 7, 2016.
[37] Y. WEN, K. ZHANG, Z. LI, and Y. QIAO. A discriminative feature learning approach for deep face recognition.*In European conference on computer vision*, pp. 499-515, 2016.
[38] L.-J. LI and L. FEI-FEI. What, where and who? classifying events by scene and object recognition. *In 2007 IEEE 11th international conference on computer vision*, pp. 1-8, 2007.
[39] B.C. RUSSELL, A. TORRALBA, K.P. MURPHY, and W.T. FREEMAN, LabelMe: a database and web-based tool for image annotation. *International journal of computer vision*,vol. 77, no. 1-3: pp. 157-173, 2008.
[40] B. ZHOU, A. KHOSLA, A. LAPEDRIZA, A. OLIVA, and A. TORRALBA. Learning deep features for discriminative localization.*In Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2921-2929, 2016.
[41] S. XIE, R. GIRSHICK, P. DOLLáR, Z. TU, and K. HE. Aggregated residual transformations for deep neural networks. *In Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1492-1500, 2017.
[42] T.-Y. LIN, P. DOLLáR, R. GIRSHICK, K. HE, B. HARIHARAN, and S. BELONGIE. Feature pyramid networks for object detection. *In Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2117-2125, 2017.
[43] M. LIN, Q. CHEN and S. YAN, Network in network. *arXiv preprint arXiv*:1312.4400, 2013.
[44] J. KRAUSE, H. JIN, J. YANG, and L. FEI-FEI. Fine-grained recognition without part annotations. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5546-5555, 2015.

[45] Z. AKATA, S. REED, D. WALTER, H. LEE, and B. SCHIELE. Evaluation of output embeddings for fine-grained image classification. *In Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2927-2936, 2015.

[46] T. XIAO, Y. XU, K. YANG, J. ZHANG, Y. PENG, and Z. ZHANG. The application of two-level attention models in deep convolutional neural network for fine-grained image classification. *In Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 842-850, 2015.

[47] D. LIN, X. SHEN, C. LU, and J. JIA. 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.

[48] A. KHOSLA, N. JAYADEVAPRAKASH, B. YAO, and F.-F. LI. Novel dataset for fine-grained image categorization: Stanford dogs. *In Proc. CVPR Workshop on Fine-Grained Visual Categorization (FGVC)*, 2011.

[49] M.-E. NILSBACK and A. ZISSERMAN. Automated flower classification over a large number of classes.*In 2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pp. 722-729, 2008.

[50] S. MAJI, E. RAHTU, J. KANNALA, M. BLASCHKO, and A. VEDALDI, Fine-grained visual classification of aircraft. *arXiv preprint arXiv*:1306.5151, 2013.

[51] T.-Y. LIN, A. ROYCHOWDHURY and S. MAJI. Bilinear cnn models for fine-grained visual recognition. *In Proceedings of the IEEE international conference on computer vision*, pp. 1449-1457, 2015.

[52] S. YANG, L. BO, J. WANG, and L.G. SHAPIRO. Unsupervised template learning for fine-grained object recognition. *In Advances in neural information processing systems*, pp. 3122-3130, 2012.

[53] B. YAO, A. KHOSLA and L. FEI-FEI. Combining randomization and discrimination for fine-grained image categorization.*In CVPR 2011*, pp. 1577-1584, 2011.

[54] B. YAO, G. BRADSKI and L. FEI-FEI. A codebook-free and annotation-free approach for fine-grained image categorization. *In 2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3466-3473, 2012.

[55] K. DUAN, D. PARIKH, D. CRANDALL, and K. GRAUMAN. Discovering localized attributes for fine-grained recognition. *In 2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3474-3481, 2012.

[56] J. DENG, J. KRAUSE and L. FEI-FEI. Fine-grained crowdsourcing for fine-grained recognition. *In Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 580-587, 2013.

[57] N. ZHANG, R. FARRELL, F. IANDOLA, and T. DARRELL. Deformable part descriptors for fine-grained recognition and attribute prediction. *In Proceedings of the IEEE International Conference on Computer Vision*, pp. 729-736, 2013.

[58] N. ZHANG, J. DONAHUE, R. GIRSHICK, and T. DARRELL. Part-based R-CNNs for fine-grained category detection. *In European conference on computer vision*, pp. 834-849, 2014.

[59] S. BRANSON, G. VAN HORN, P. PERONA, and S.J. BELONGIE. Improved Bird Species Recognition Using Pose Normalized Deep Convolutional Nets. *In BMVC*, pp. 7, 2014.

[60] T. BERG and P.N. BELHUMEUR. Poof: Part-based one-vs.-one features for fine-grained categorization, face verification, and attribute estimation. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 955-962, 2013.

[61] R. FARRELL, O. OZA, N. ZHANG, V.I. MORARIU, T. DARRELL, and L.S. DAVIS. Birdlets: Subordinate categorization using volumetric primitives and pose-normalized appearance. *In 2011 International Conference on Computer Vision*, pp. 161-168, 2011.

[62] X. ZHANG, H. XIONG, W. ZHOU, W. LIN, and Q. TIAN. Picking deep filter responses for fine-grained image recognition. *In Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1134-1142, 2016.

[63] X. LIU, T. XIA, J. WANG, Y. YANG, F. ZHOU, and Y. LIN, Fully convolutional attention networks for fine-grained recognition. *arXiv preprint arXiv*:1603.06765, 2016.

[64] B. ZHAO, X. WU, J. FENG, Q. PENG, and S. YAN, Diversified visual attention networks for fine-grained object classification. *IEEE Transactions on Multimedia*,vol. 19, no. 6: pp. 1245-1256, 2017.

[65] J. FU, H. ZHENG and T. 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*, pp. 4438-4446, 2017.

[66] H. ZHENG, J. FU, T. MEI, and J. LUO. 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.

[67] C. SZEGEDY, S. IOFFE, V. VANHOUCKE, and A. ALEMI, Inception-v4, inception-resnet and the impact of residual connections on learning. *arXiv preprint arXiv*:1602.07261, 2016.

[68] A. KHAN, A. SOHAIL and A. ALI, A new channel boosted convolutional neural network using transfer learning. *arXiv preprint arXiv*:1804.08528, 2018.

[69] F. WANG, M. JIANG, C. QIAN, S. YANG, C. LI, H. ZHANG, X. WANG, and X. TANG. Residual attention network for image classification. *In Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3156-3164, 2017.

[70] O. RUSSAKOVSKY, J. DENG, H. SU, J. KRAUSE, S. SATHEESH, S. MA, Z. HUANG, A. KARPATHY, A. KHOSLA, et al.,*Imagenet large scale visual recognition challenge. International journal of computer vision*,vol. 115, no. 3: pp. 211-252, 2015.

[71] L. WU, Y. WANG, X. LI, and J. GAO, Deep attention-based spatially recursive networks for fine-grained visual recognition. *IEEE transactions on cybernetics*,vol. 49, no. 5: pp. 1791-1802, 2018.

[72] J. HU, L. SHEN and G. SUN. Squeeze-and-excitation networks. *In Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7132-7141, 2018.

[73] A. AREFI, A.M. MOTLAGH, K. MOLLAZADE, and R.F. TEIMOURLOU, Recognition and localization of ripen tomato based on machine vision. *Australian Journal of Crop Science*,vol. 5, no. 10: pp. 1144, 2011.