Embedding Pose Information for Multiview Vehicle Model Recognition

Ye Yu, Haitao Liu, Yuanzi Fu, Wei Jia, Jun Yu and Zhisheng Yan

***Abstract*—Vehicle model recognition is a typical fine-grained classification task that has a wide range of application prospects in safe cities and constitutes a research hotspot in the field of computer vision. Vehicles in images can appear at various angles, resulting in large differences in appearance. The existence of "multiviews" renders vehicle model recognition challenging. Recent research on vehicle model recognition has not fully explored the pose information of vehicles in different images, resulting in low model performance. In this study, we use vehicle pose information to solve the multiview vehicle model recognition (MV-VMR) problem and design a convolutional neural network (CNN) model with embedded vehicle pose information, known as the embedding pose CNN (EP-CNN). The proposed model includes two subnetworks: the pose estimation subnetwork (PE-SubNet) and vehicle model classification subnetwork (VMC-SubNet). PE-SubNet extracts the vehicle pose information, including the pose features and vehicle viewpoint. In VMC-SubNet, considering the scale variation of vehicles, an improved squeeze-and-excitation (SE) block, named the MultiSE block is implemented. We embed the vehicle viewpoint into the MultiSE block, which reweighs each channel such that the extracted features elicit different responses to different viewpoints. Subsequently, the pose features and classification features are integrated for classification. Experiments are conducted on the benchmark CompCars web-nature and Stanford Cars datasets. The results demonstrate that the proposed EP-CNN method can achieve higher recognition accuracy than most classic CNN models and several state-of-the-art fine-grained vehicle model classification algorithms. Code has been made available at: https://github.com/HFUT-CV/EP-CNN.**

***Index Terms*—Convolutional neural network, fine-grained classification, vehicle model recognition, pose estimation, scale-aware features.**

I. Introduction

With the continuous development of cities and increasing urbanization, traffic management and public security have become increasingly important. Since expanding the police force is a costly and inefficient solution, the concept of safe cities, which emphasizes scientific and technological methods, has emerged as an important means of solving this problem. Computer vision and pattern recognition are notable technologies in safe cities. In the recent two decades, the recognition of vehicle attributes such as license plates [1] [2], vehicle logos [3] [4] and vehicle types [5] in an image has been extensively studied and applied to traffic flow, road condition, and traffic violation detection systems.

However, these information sources play a limited role in safe cities. For example, vehicle license plates and logos are easily falsified and occluded. Vehicle types provide limited information in the investigation of traffic and security incidents. It is thus necessary to extract finer and richer vehicle feature information from images and videos to facilitate safe cities. Vehicle model recognition (VMR) has emerged as a key task because it can assist the discovery and tracking of traffic violations such as hit-and-run accidents and fake license plates. These factors are of practical significance for maintaining traffic, decreasing the crime rate, and constructing a safe city.

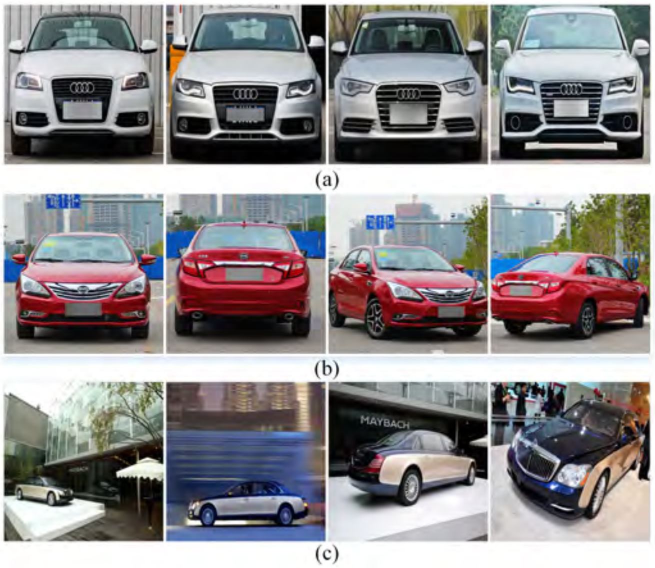


Fig. 1. Various models of vehicles. (a) Left to right: Audi A3L, Audi A4L, Audi A6L, and Audi A7. (b) Vehicle images obtained from different angles. (c) Vehicles appear in different scales.

The term "model" refers to the name used by a manufacturer to market a range of similar cars, such as "Volkswagen Passat, 2017 model" [6]. VMR is generally a challenging problem because although a large number of vehicle models are available, many vehicle models, especially those by the same manufacturer, are only slightly different. For instance,

This work is partly supported by grants of the National Natural Science Foundation of China (Nos. 61906061 and 62076086), Provincial Key Research and Development Program of Anhui Province (Nos. 201904d07020010, 202104a05020007 and 202004d07020008) and CAAI-Huawei MindSpore Open Fund (CAAIXSJLJJ-2021-016B).

Ye Yu, Haitao Liu, Yuanzi Fu and Wei Jia are with the Key Laboratory of Knowledge Engineering with Big Data, Ministry of Education, Hefei University of Technology, Hefei 230009, China; School of Computer Science and Information, Hefei University of Technology, Hefei 230009, China; and Anhui Province Key Laboratory of Industry Safety and Emergency, Hefei 230009, China.

Jun Yu (Corresponding author) is with the Department of Automation, University of Science and Technology of China, Hefei 230026, China; and the National Engineering Laboratory for Speech and Language Information Processing, Hefei 230026, China.

Zhisheng Yan is with the Department of Information Science and Technology, School of Computing, at George Mason University, Fairfax, VA 22030, USA.

in a medium city, the number of vehicle models on the road may be greater than 2,000. As shown in Fig. 1(a), Audi A3L, Audi A4L, Audi A6L, and Audi A7 are different vehicle models with similar appearances, and it is difficult to visually recognize their differences.

Another notable challenge for VMR is that vehicle images captured by different city surveillance systems are obtained from different angles. Thus, one vehicle model may appear completely different in different images (Fig. 1(b)). A vehicle is a rigid body, and its front, side and rear structures often exhibit substantial differences in appearance. When captured at random angles, the same vehicle exhibits different poses, which further complicates VMR. The vehicle's varying scale in the captured images poses another challenge (Fig. 1(c)). Small-scale vehicles tend to have subtle features compared to the complex background information, which is not conducive to recognition. In contrast, large-scale vehicles occupy almost the entire image. Such multiscale variation in vehicles render their recognition difficult.

The VMR task based on images captured from different angles is referred to as multiview VMR (MV-VMR). Recent efforts toward VMR do not satisfactorily address the MV-VMR problem. Part-based methods [7] [8] aim to detect distinct vehicle parts and differentiate different vehicle models. However, these methods require manual annotation of vehicle parts and are infeasible for large city-scale datasets. In contrast, attention-based methods [9]–[12] exploit the visually attractive region of a vehicle to classify models. However, an attention region in one image may be different in another image captured from a different shooting angle. In addition, considering the scale variation of vehicles in the captured images, most researchers [13]–[15] have conducted experiments on datasets with bounding box annotations. These frameworks generally remove the background from the images and retain only the vehicles to ensure that all the vehicles in the dataset have equivalent scales. This approach can alleviate the negative impacts of different scales of the vehicles on the final VMR performance; however, it does not fundamentally solve the problem. Overall, the existing VMR methods do not fully explore the pose information and multiscale information of vehicles in different images, corresponding to an inferior model performance.

In this paper, we leverage the vehicle pose information to solve the MV-VMR problem. A novel convolutional neural network (CNN) model known as the embedding pose CNN (EP-CNN) is proposed. EP-CNN is the first method capable of extracting pose information and embedding it into the classification network to realize MV-VMR. EP-CNN has the following characteristics.

1. EP-CNN is a two-stream network that includes a pose estimation subnetwork (PE-SubNet) and vehicle model classification subnetwork (VMC-SubNet). PE-SubNet extracts a vehicle's pose information, including the vehicle viewpoint and pose features. VMC-SubNet embeds the pose information into the classification.
2. Vehicle pose information is extracted as auxiliary information for MV-VMR and acquired based on the you only look once (YOLO) object detection model [16]. In this model, the anchor box dimension clustering strategy and loss functions are enhanced for viewpoint prediction and pose feature generation.
3. In VMC-SubNet, a new embedding block termed MultiSE is incorporated into the residual block to address the scale variation in vehicle models. Moreover, a novel fusing strategy is proposed to fuse the pose and classification features based on the hard attention mechanism, with attention masks generated based on the confidence information generated from PE-SubNet.

The remainder of the paper is organized as follows. Section II presents a brief review of related work on VMR. Section III presents the details of the proposed EP-CNN model. Section IV describes the experimental results and the analysis. Section V presents the concluding remarks.

II. Related Works

VMR is a fine-grained recognition problem that aims to identify the vehicle model. In this section, we summarize the existing fine-grained recognition methods, including general and specific methods focused on VMR.

A. General Fine-Grained Object Recognition

The challenge of fine-grained recognition mainly lies in the small interclass variances caused by highly similar subordinate categories and large intraclass variances in poses, viewpoints and occlusions. Many methods have been proposed to address these two challenges. We divide these methods into several categories, as they usually share certain common traits.

1) Methods Based on Constructing Label Structures: For the fine-grained recognition problem, a subcategory is often subordinate to a large category; thus, there must be a certain correlation between these subcategories. Therefore, several methods utilize the structural correlation between category labels and construct multilevel coarse-to-fine label structures to address the problem of large intraclass variances and small interclass variances in fine-grained recognition. For example, Zhang et al. [17] embedded label structures such as the hierarchy or shared attributes into the framework by generalizing the triplet loss. Moreover, Zhou et al. [18] detailed an approach to exploit the rich relationships through bipartite-graph labels.

2) Methods Based on Metric Learning: The purpose of metric learning is to cluster samples from the same class and push those from different classes apart via learning and embedding. Qian et al. [19] proposed a multistage metric learning framework that can be efficiently applied to tasks with large-scale high-dimensional data. Zhe et al. [20] reported a loss function based on the von Mises–Fisher distribution for metric learning to learn an embedded probability space on a hypersphere.

3) Methods based on Extracting Fine-Grained Features: These methods consider that the rigid and nonrigid body transformations of an object cause large intraclass variances. The approaches attempt to extract fine-grained features to capture subtle interclass variances. For example, Zhao et al. [21] proposed a diversified visual attention network to maximize the collection of discriminative fine-grained information.

Wei et al. [22] described a four-stream Mask-CNN model that segments the head, torso and object of birds using a fully convolutional network and simultaneously aggregates the selected object and part-level features. Lopez et al. [23] proposed a modular attention mechanism that is applied to the convolutional feature activations and can successfully find the most discriminative regions of the image.

4) Limitation of General Fine-Grained Object Recognition: Although the abovementioned methods are useful for general fine-grained object recognition, they cannot be applied directly to the MV-VMR problem to achieve satisfactory results. Several methods have not been trained using vehicle model datasets. Moreover, even the methods that have been trained on the vehicle model dataset do not consider the characteristics of vehicles and influence of the "multiview" problem, which limits their MV-VMR performance and leads to suboptimal results.

B. Fine-Grained VMR

Several methods have been proposed for fine-grained VMR. These methods leverage the distinct characteristics of vehicles, e.g., rigidity and distinguishable parts (headlights, grille, etc.). Vehicle recognition methods can be divided into the following three categories.

1) Part-Based Methods: Part-based methods detect discriminative parts, such as the headlamp and grille, to obtain expressive appearance features. Krause et al. [7] proposed an object representation model that detects important parts and describes fine-grained appearances. He et al. [8] proposed a framework in which cars are first detected using a part-based detector before VMR. Although most of these methods can achieve a high recognition accuracy, they are heavily dependent on artificial annotations such as part annotations and bounding boxes. Hence, these methods are labor-intensive for MV-VMR.

2) Methods Using Attention Mechanisms: To avoid the high costs associated with artificial annotation, many researchers have employed attention-mechanism methods in recent years. Zheng et al. [14] proposed a multiattention CNN (MA-CNN) method that learns part information through a multiattention network without any bounding box or part annotation. Yu et al. [24] proposed two attention mechanisms by modeling the human visual system to achieve VMR. Zhang et al. [25] applied a gradient-based attention module to extract the attention region and transform the training data into a new set for the experts. Ji et al. [26] used the attention transformer module to force the network to capture discriminative features in their proposed attention convolutional binary neural tree method. Sun et al. [27] introduced diversification blocks that function as an attention mechanism and mask out the salient features. In this manner, the network can search for subtle differences between similar-looking categories. Ding et al. [28] proposed a dynamic perception framework that achieves fine-grained recognition by adopting adaptive effective receptive fields and applying channel attention and spatial attention mechanisms. In contrast to the aforementioned methods, we combine the attention mechanism with the vehicle viewpoint to find discriminative features for MV-VMR according to the current vehicle viewpoint.

3) Methods Using Viewpoint Information: Although the vehicle viewpoint is of significance to solve the VMR problem, few studies of VMR have applied this information, probably because its importance is not fully recognized. To the best of our knowledge, the only research on incorporating the viewpoint information is [29], in which the BoxCars dataset is proposed. This dataset includes three-dimensional (3-D) bounding box information, which is difficult to obtain to solve for the MV-VMR problem. Due to the presence of the 3-D bounding box, the viewpoint can be easily extracted and encoded as three two-dimensional (2-D) vectors that express the front/rear, side and roof or encoded by rasterizing bounding boxes and passed to the net.

Moreover, viewpoint information has also been utilized in vehicle Re-ID, which is a task similar to VMR. For example, in [30], a viewpoint-aware network (VANet) was proposed. VANet has two metric learning branches for different viewpoint relationships, which creates two feature spaces for the learning metric under an S-view (similar view) and a D-view (different view) relationship.

However, the usefulness of the viewpoint is not fully explored in the context of the MV-VMR task. In this paper, we comprehensively explore the effects of the vehicle viewpoint and pose features.

4) Other Methods: Certain existing methods have been inspired by the habits of human observation. For example, Hu et al. [31] proposed a multitask CNN that localizes vehicles in the first stage and recognizes subclasses in the second stage. Several approaches utilize the vehicle 3-D information to facilitate fine-grained VMR. Sochor et al. [29], [32] collected a large fine-grained vehicle dataset with 3-D bounding boxes and leveraged 3-D information to enhance the results for the fine-grained recognition of vehicles. Other researchers focused on applying new strategies to their network structure. For example, Chen et al. [33] used multibranch CNNs with three scale images as input and a local loss module after each branch to simultaneously realize multiview vehicle type recognition (VTR) and fine-grained VTR, i.e. VMR. Tian et al. [10] proposed an iterative discrimination CNN (ID-CNN) approach that iteratively applies CNN for multiconvolutional region feature extraction. Chen et al. [34] introduced a destruction and construction learning (DCL) stream to automatically learn from discriminative regions.

5) Limitation of Fine-grained VMR: Although part-based and attention-based methods have laid the foundation for VMR, they cannot satisfactorily address the MV-VMR problem. Specifically, part-based methods require manually generated part annotations that are difficult to obtain, especially when multiview images are considered. Similarly, the attention-based methods focus only on discovering attention areas and extracting features in those areas. These approaches do not consider the multiview information, which is essential for MV-VMR, as highlighted in this work. In other words, the existing VMR methods cannot achieve satisfactory results for the MV-VMR problem, and there remains considerable scope for improvement.

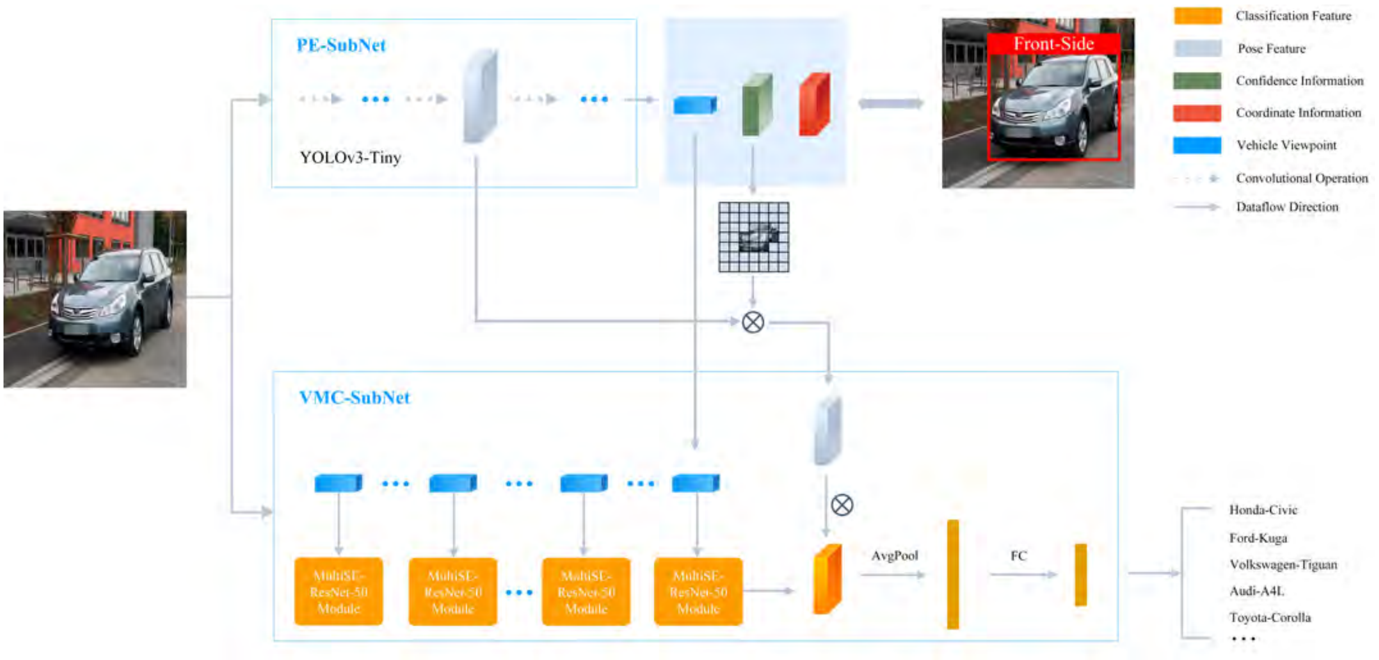


Fig. 2. Main structure of EP-CNN model.

C. Summary

MV-VMR is a branch of fine-grained VMR methods in which the adopted datasets contain vehicle images obtained from many angles rather than those limited to the front/back view. Therefore, MV-VMR focuses on the impact of different viewpoints on the task of VMR. The proposed method builds upon the idea of using visual attention to differentiate vehicles in attention-based methods. However, instead of adding an attention module, we introduce pose information to ensure that the network can focus on discriminative features from a certain viewpoint. Simultaneously, we adopt channel weights for use in VMC-SubNet, such that the network can select features that are conducive to MV-VMR.

III. Methodology

A. Concept

Vehicle viewpoint: Angle of the image obtained according to the direction of the vehicle. In our method, we categorize the vehicle images into five viewpoints (V = 5), i.e., front, rear, side, front-side and rear-side, which are sufficient for comprehensively describing the vehicle viewpoint.

Pose features: Features contained in the feature maps of PE-SubNet. Since PE-SubNet is used for pose estimation, the pose features can be obtained from feature maps of the network.

Pose information: The vehicle viewpoint and pose features help address the MV-VMR problem.

B. EP-CNN Model Structure

Because the vehicle pose information is important for solving MV-VMR, we design a novel CNN-based model named EP-CNN. The core idea of EP-CNN is to embed pose information into the pipeline of our deep learning model. As shown in Fig. 2, the EP-CNN is composed of two parts: the PE-SubNet and VMC-SubNet. The input image is propagated into two branch networks in parallel.

PE-SubNet is a multitask model based on the enhanced YOLOv3-tiny algorithm. The algorithm divides the original image into $N \times N$ grids, and each grid cell predicts the vehicle viewpoint, vehicle position coordinates, and possibility of it containing a vehicle, i.e., the confidence. Using these information, we can locate the vehicle by nonmaximal suppression. We consider the viewpoint predicted by the grid with the highest confidence as the vehicle viewpoint and the extracted features as the pose features.

VMC-SubNet is a traditional CNN used to extract fine-grained vehicle features. We propose an improved squeeze-and-excitation (SE) block [35], namely, a MultiSE block, to address the scale variation in vehicle models. The MultiSE block is added to the residual block and forms MultiSE-ResNet-50, which serves as the backbone of VMC-SubNet. The pose features and vehicle viewpoint obtained from PE-SubNet are incorporated in the VMC-SubNet to achieve superior classification results. By embedding pose information into the classification subnetwork, the MV-VMR problem can be addressed.

C. PE-SubNet

PE-SubNet estimates the vehicle viewpoint and generates pose features that represent the vehicle pose information. Due to the complex background information contained in the vehicle images, an object detection model must be used to accurately locate the vehicle, eliminate background interference, and obtain an accurate viewpoint. Two-stage object detection algorithms generally involve considerable training and testing periods. YOLOv3-tiny is an innovative one-stage detection algorithm that facilitates and accelerates object detection training and testing. This algorithm is a simplified version of YOLOv3 [16], in which many convolution layers of YOLOv3 are eliminated, and only two layers are retained for prediction with less than 1/7 of the parameters. We conduct several experiments to test the viewpoint prediction accuracy of YOLOv3-tiny. The results indicate that if a vehicle can be correctly detected, the viewpoint prediction accuracy is 99.8%. In other words, YOLOv3-tiny can effectively predict the vehicle viewpoint with minimal cost. Thus, we selected YOLOv3-tiny as the backbone of PE-SubNet.

To achieve the optimal result in our pose estimation task, we optimize YOLOv3-tiny to render it more suitable for our task. We improve YOLOv3-tiny in the following ways:

1) Anchor Box Dimension Clustering: We propose a new anchor box dimension clustering strategy, which is k-means++ running on the vehicle-related subset of the COCO (VR-COCO) dataset. This framework is different from k-means running on the COCO dataset, which is adopted in YOLOv3-tiny.

YOLOv3-tiny uses anchor boxes to predict the location of the object bounding boxes. Specifically, the algorithm runs k-means clustering on the bounding boxes of the COCO training set to automatically obtain reasonable prior anchor boxes. The clustering results of different k-means runs are different because the initial points are randomly selected, which affects the stability of the anchor box dimension clustering results. To solve this problem, we use k-means++ instead of k-means as the anchor box dimension clustering algorithm. Notably, k-means++ can significantly decrease the final error of the clustering results and calculation time. Although k-means++ requires additional time to calculate the initial points, the selected initial points enable the algorithm to rapidly converge during the iteration process; thus, k-means++ decreases the calculation time. The number of clustering centers must be specified in advance, and their positions must be verified. Therefore, many other widely used clustering methods, such as spectral clustering, MeanShift, affinity propagation, density-based spatial clustering of applications with noise (DBScan), and balanced iterative reducing and clustering using hierarchies (Birch), cannot be employed. K-means++ is a simple yet effective solution.

To ensure that the clustering results are consistent with the vehicle shape, we select the categories related to vehicles (i.e., car, bus and truck) in the COCO dataset as the training set. This dataset is known as the vehicle-related subset of COCO (VR-COCO). We run k-means and k-means++ clustering on the training set bounding boxes for various values of k and plot the average intersection over union (IOU) with the closest centroid, as shown in Fig. 3. For different numbers of centroids, the average IOU of k-means++ is generally higher than that of k-means. Furthermore, for k-means++, k = 6 yields the correct balance of recall and complexity. On the VR-COCO dataset, we can obtain the initial box dimensions of 6 clusters, which are (13×8), (25×18), (46×31), (80×56), (153×115) and (409×269), with an average IOU of 62.43%. Thus, k-means++ can be applied on the VR-COCO dataset to effectively match the vehicle shape and facilitate the learning process of the model.

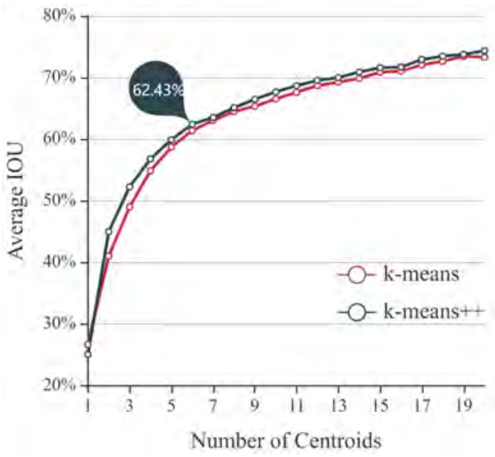


Fig. 3. Clustering box dimensions on the VR-COCO dataset.

2) Loss Function: YOLOv3-tiny is applied to estimate the vehicle pose information, but not for classification. Therefore, we add the viewpoint error to the loss function and eliminate the classification error. The adjusted loss function is as follows:

$$Loss = coordError + confError + viewError \tag{1}$$

The localization task of PE-SubNet is a simplified single class localization task. Therefore, we set each grid cell to predict only one bounding box. The coordinate error is calculated as:

$$\begin{aligned} coordError = \lambda_{coord} \sum_{i=1}^{N \times N} 1_i^{obj} \Big[(x - \hat{x})^2 + (y - \hat{y})^2 \\ + (w - \hat{w})^2 + (h - \hat{h})^2 \Big] \end{aligned} \tag{2}$$

where $1_i^{obj} = 1$ if grid cell i falls in the object and zero otherwise. We follow the setting of YOLOv3-tiny and set λ_{coord} as 5. Here, (x, y, w, h) represents the predicted object center coordinates, width and height, respectively, and $(\hat{x}, \hat{y}, \hat{w}, \hat{h})$ represents the ground-truth object center coordinates, width and height, respectively.

The confidence error indicates the probability that the grid cell belongs to the object and accuracy of the predicted bounding box. The detection error is calculated as:

$$\begin{aligned} confError = \sum_{i=1}^{N \times N} 1_i^{obj} [-\log(c)] \\ + \lambda_{\text{noobj}} \sum_{i=1}^{N \times N} \left(1 - 1_i^{obj}\right) [-\log(1 - c)] \end{aligned} \tag{3}$$

where c represents the predicted confidence. The confidence error may lead to class imbalance because most boxes do not contain any objects; thus, we follow the setting of YOLOv3-tiny and set λ_{noobj} as 0.5.

The viewpoint prediction task is a multiclass classification problem. The viewpoint is predicted only for the grid cell that

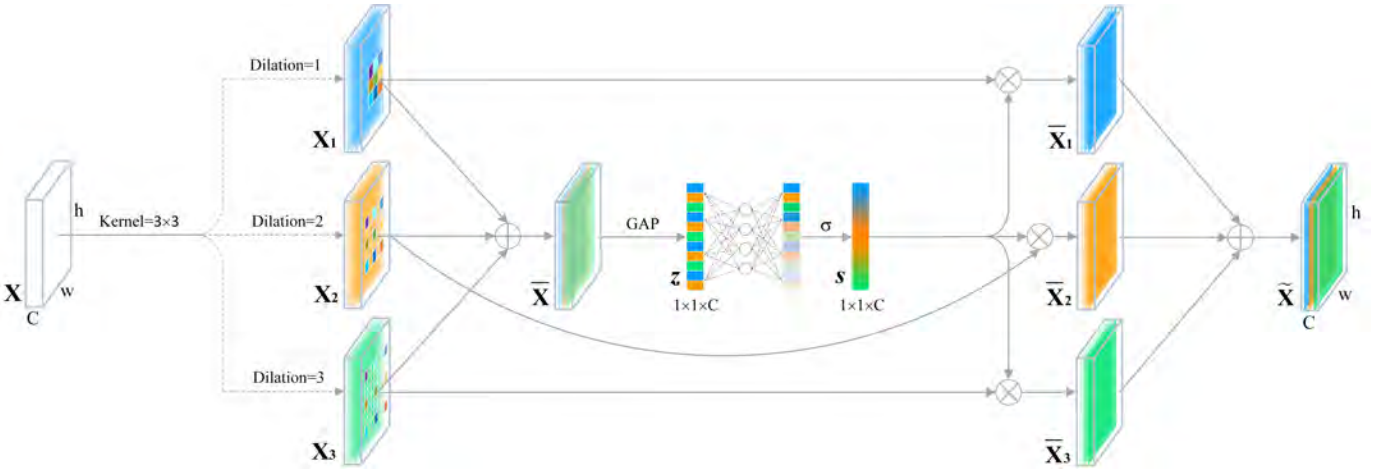


Fig. 4. Diagram of the proposed MultiSE block architecture.

falls in the object. The viewpoint prediction error is calculated as:

$$viewError = \sum_{i=0}^{N\times N} 1_i^{obj} \sum_{v\in \text{views}} \left[-\hat{p}_v \log\left(p_v\right)\right] \qquad (4)$$

where p_v represents the predicted class probability for viewpoint v, and $\hat{p}_v$ represents the ground-truth probability of the object.

The final output of PE-SubNet includes the vehicle's position (x, y, w, h) predicted by each grid cell, detection confidence of each grid cell, and viewpoint of the vehicle predicted by each grid cell. Therefore, the final output dimension of our network is $N\times N \times (1 + 4 + 5)$. We select the viewpoint predicted by the grid with the largest confidence as the vehicle viewpoint and choose the feature maps from YOLOv3-tiny as pose features with the same size as the classification features. PE-SubNet in the EP-CNN is pretrained to extract pose information when training VMC-SubNet.

D. VMC-SubNet

VMC-SubNet embeds the vehicle viewpoint and pose features extracted by PE-SubNet into its classification network architecture to enhance the useful features that flow through the network pipeline. We use ResNet-50 [36] as the backbone network. First, we propose an improved SE block termed the MultiSE block, which is proven to be more efficient than the SE block [35] for VMR. Second, we embed the vehicle viewpoint into the MultiSE block, which is proven to enhance the MV-VMR results. Thrid, through the fusion of the classification and pose features, the fine-grained features of the vehicle model, which are proven to be more distinguishable for MV-VMR, are enhanced.

1) MultiSE Block: The SE block, which explicitly models the interdependencies between channels via the SE operation, is proven to be efficient for object classification. Thus, the SE block can enhance significant features and suppress features that are not useful for the current task. In the MV-VMR task, vehicles of different scales exist in different images (Fig. 1(c)). Small-scale vehicles tend to have subtle features compared to complex background information, which is not conducive to recognition. In contrast, large-scale vehicles occupy nearly the whole image. Scale variation can affect the final recognition accuracy. Inspired by TridentNet [37], which has a multibranch structure to address the scale variation problem, we combine multibranch structures with the SE block and propose a new MultiSE block. As illustrated in Fig. 4, three parallel branches of dilated convolution with different dilation rates are adopted, and the SE block is implemented after dilated convolution on each branch. Scale-aware features are obtained by the fusion of feature maps from three branches.

For the feature map $\mathbf{X} \in \mathbb{R}^{H\times W\times C}$ obtained from the backbone network, we separately perform dilated 3×3 convolution operations with dilation rates of 1, 2, and 3 at the three branches, corresponding to receptive fields of 3×3, 5×5, and 7×7, respectively. Subsequently, we implement batch normalization and ReLU activation functions to generate three scale-aware feature maps $\mathbf{X}_i \in \mathbb{R}^{H\times W\times C}, i = 1, 2, 3$. To decrease the model complexity and obtain a more comprehensive feature map, we employ grouped convolution with 32 groups, as suggested in ResNeXt [38], for each branch.

The obtained feature maps $\mathbf{X}_1$, $\mathbf{X}_2$ and $\mathbf{X}_3$ are fused via elementwise summation:

$$\overline{\mathbf{X}} = \mathbf{X}_1 + \mathbf{X}_2 + \mathbf{X}_3 \qquad (5)$$

where $\overline{\mathbf{X}} \in \mathbb{R}^{H\times W\times C}$ aggregates multiscale information and has rich features. $\overline{\mathbf{X}}$ is input to the SE module and set as discussed in [35]. Specifically, after global average pooling (GAP), $\overline{\mathbf{X}}$ is squeezed to z ($z \in \mathbb{R}^{C}$). After an excitation operation, i.e., an operation involving two fully connected (FC) layers with the softmax activation σ, the weight value s (s $\in \mathbb{R}^{C}$), which represents the importance of different channels, is obtained.

Channel feature recalibration is performed based on the obtained weight value, which is denoted as:

$$\overline{\mathbf{X}}_{i,c} = F_{\text{scale}}\ (\mathbf{X}_{i,c}, s_c) = s_c \mathbf{X}_{i,c} \qquad (6)$$

where $\mathbf{X}_{i,c}$ denotes the c-th channel of the feature map $\mathbf{X}_i$, F_{scale} represents the channel feature recalibration operation, and $\overline{\mathbf{X}}_{i,c}$ denotes the c-th channel of the feature map $\overline{\mathbf{X}}_i$ after channel feature recalibration.

The feature maps obtained via channel feature recalibration at three scales are fused via elementwise summation:

$$\tilde{\mathbf{X}} = \overline{\mathbf{X}}_1 + \overline{\mathbf{X}}_2 + \overline{\mathbf{X}}_3 \tag{7}$$

where $\tilde{\mathbf{X}}$ is the output of the MultiSE block.

The performance of the MultiSE block is evaluated as described in Section IV.D.1.

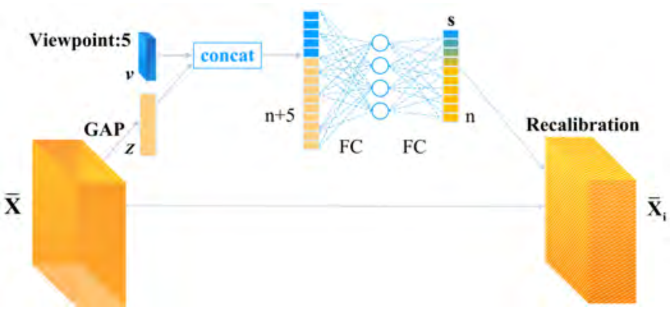


Fig. 5. Feature recalibration of different channels.

2) MultiSE Block with Embedded Vehicle Viewpoint: In the MultiSE and SE blocks, the "squeeze" operation is aimed at averaging the previous convolutional layer features according to each channel and yielding a vector with a length equal to the number of channels of the input feature map denoted as z, with length n. The "excitation" operation uses two FC layers with the squeezed vector as the input to obtain the weights of different channel features.

In the MV-VMR task, the vehicle images obtained from different angles correspond to significantly different appearances, and the features extracted from these images are considerably different for different viewpoints. Since a convolutional kernel usually focuses on only one feature, it is reasonable to choose convolutional features according to different viewpoints. We consider the vehicle viewpoint obtained from the image grid cell with the highest confidence in PE-SubNet as the predicted viewpoint. This entity is a 5-dimensional vector, with each dimension representing a direction. We consider the 5-dimensional vector v and squeezed feature vector z as the basis of generating channel weights. As shown in Fig. 5, the concatenated $n + 5$ vector as input and two FC layers are used for dimensionality-reduction mapping, i.e., the excitation operation generates the weights of n channels. The generation process of channel weights s can be expressed as:

$$s = \sigma\left(W_2 \delta\left(W_1 (z \oplus v)\right)\right) \tag{8}$$

where $\oplus$ represents concatenation, $W_1 \in \mathbb{R}^{\frac{C+5}{r} \times (C+5)}$, and $W_2 \in \mathbb{R}^{C \times \frac{C+5}{r}}$. δ and σ represent the ReLU and softmax activation functions, respectively. s represents the weight of each channel after the addition of the vehicle viewpoint.

3) Fusion of Pose and Classification Features: After adding the MultiSE block with viewpoint information to the residual block, the pose features P from PE-SubNet must be fused with the classification features F from VMC-SubNet. As shown in Fig. 2, we choose the feature maps from YOLOv3-tiny as the pose features, which have the same size as the classification features, to avoid losing useful information. In addition, we generate a hard attention mask M based on the confidence information $C \in \mathbb{R}^{N \times N}$ of the output of PE-SubNet. The confidence indicates the probability of containing a vehicle in this grid cell. The maximum value of confidence C is C_m, and the generation of mask M is formulated as:

$$M_{\mathrm{i},j} = \begin{cases} 1, & C_{i,j} \geq C_m - 0.1 \\ 0, & \text{otherwise} \end{cases} \tag{9}$$

where $C_{i,j}$ denotes the value of index i,j of the confidence matrix C, and $M_{i,j}$ denotes the value of the index i,j of the generation mask M.

Subsequently, as shown in Fig. 2, we perform the Hadamard product operation between the hard attention mask M and pose features P obtained from PE-SubNet, which can be considered as a hard spatial attention mechanism to eliminate the interference of image background information. The hard attention operation can be expressed as:

$$P' = M \odot P \tag{10}$$

In P', the grids with feature values of 0 are considered as the background, and the grids with non-zero feature values are the vehicle regions.

Next, the pose features after attention mechanism P' are fused with classification features F based on the Hadamard product operation:

$$F' = P' \odot F \tag{11}$$

We argue that the remaining regions are likely to contain the most discriminative and important features for recognition. Finally, we use average pooling and FC layers to obtain the classification result.

4) VMC-SubNet Loss function: In this section, we introduce the objective function of VMC-SubNet, which enables the model to focus on critical regions at different locations of the image. To train our model, the overall loss function consists of a cross-entropy loss (CE-Loss) and mutual channel loss (MC-Loss) [39]. The CE-Loss encourages the network to extract informative features focusing on the global discriminative regions of the image. The MC-Loss guides the model to highlight different local areas of the image and enables the network to learn features that are simultaneously discriminative and diverse.

IV. Experimental Results and Analysis

We conduct extensive experiments to evaluate the proposed neural network model (EP-CNN) on two fine-grained datasets. We aim to demonstrate that combining the MultiSE block with pose information, i.e., vehicle viewpoint and pose features, can significantly increase the recognition accuracy of multiview vehicle models.

Since the vehicle viewpoint is labeled on the CompCars web-nature dataset, we focus on comparing the results of state-of-the-art methods, ablation studies, and other analyses on the CompCars web-nature dataset. To prove that the proposed EP-CNN model can also be applied to datasets without viewpoint labels, we perform experiments on the Stanford Cars dataset and compare our method with state-of-the-art methods.

A. Dataset and Experimental Environment

Dataset: The CompCars web-nature dataset [40] contains 52,083 multiview vehicle images of 431 models, all of which contain the vehicle model, viewpoint, and bounding box label. We use 70% of the images as the training dataset and the other 30% as the testing dataset. The Stanford Cars dataset [41] contains 16,185 images of 196 models, all of which contain the vehicle model and bounding box label. We use 50% of the images as the training dataset and the other 50% as the testing dataset. We conduct experiments on the datasets without bounding boxes.

Experimental hardware environments: CPU: Intel Core i7-9700KF; Memory: 32 GB; Graphics Cards: Dual NVIDIA GeForce RTX 2080Ti; Video Memory: 11 GB.

B. Implementation Details

EP-CNN is implemented using the open-source PyTorch framework with CUDA version 10.0. To stabilize network the training, we implement the following training steps:

1. First, the original YOLOv3-tiny is pretrained on the VR-COCO dataset, and MultiSE-ResNet-50 is pretrained on ImageNet [42] with 1,000 classes of data. We resize all samples to 256×256/512×512 and crop five images of 224×224/448×448 from the center and four corners of the image. We perform a mirroring operation on the five images. Thus, for each sample, ten training images are obtained. Finally, we subtract the mean of the entire dataset for all input images.
2. We fine-tune PE-SubNet using the viewpoint and bounding box labels of the CompCars web-nature datasets. We do not augment training data in this case because the existence of coordinate labels is expected to complicate the augmentation process.
3. We use the finetuned PE-SubNet to generate the vehicle viewpoint, which is required by VMC-SubNet, and finetune VMC-SubNet using the training dataset. The training data are augmented, as in step 1. The entire network is comprehensively trained. We use the finetuned PE-SubNet to generate the vehicle viewpoint, which is needed by VMC-SubNet, and finetune VMC-SubNet using the training dataset. The training data are augmented, as in step 1. The entire network is comprehensively trained.

The hyperparameters and other settings in our experiments can be described as follows. The experimental optimization strategy is stochastic gradient descent, for which the initial learning rate is 0.01, and the batch size is 32. The learning rate is adjusted using a stepdown strategy such that the learning rate is decreased by a factor of 10 per 200,000 iterations, and the experiments are iterated 600,000 times. Moreover, we adopt the repeated augmentation (RA) [43] method in the training process. The proposed network EP-CNN does not require special initialization or part or bounding box annotations. Experiments are also verified using MindSpore.

TABLE I
COMPARISON OF RECOGNITION RESULTS ON THE COMPCARS WEB-NATURE DATASET.

Method	Size	#P(M)	GFLOPs	Anno.	Top-1
ResNet-50 [36]	224	25.5	4.1	×	94.0%
ResNeXt-50, 32×4d [38]	224	25.0	4.2	×	94.9%
SE-ResNet-50 [35]	224	26.9	4.1	×	95.1%
Res2Net-50,26w×4s [44]	224	24.5	4.3	×	95.2%
Res2NeXt-50,26w×4s [44]	224	23.5	4.2	×	95.4%
EfficientNet-B5 [45]	224	29.2	9.9	×	96.0%
FixResNet-50 [46]	224	25.5	4.1	×	97.0%
BoxCars [29]	-	-	-	✓	84.8%
FM-CNN [33]	227	-	-	×	91.0%
Location-Aware [31]	400	-	-	✓	94.3%
FR-ResNet [47]	224	⩾25.5	⩾4.7	×	95.3%
ABN [48] (ResNet-101)	323	62.5	16.0	×	97.1%
ID-CNN [10]	224	143.6	19.7	×	96.2%
ResNet152-CMP [15]	224	⩾59.0	⩾11.6	✓	97.0%
DCL [34] (ResNet-50)	448	25.5	17.5	×	97.5%
EP-CNN(ours)	224	28.0	4.7	×	**98.6%**
EP-CNN(ours)	448	28.0	18.9	×	**98.9%**

C. Comparison with State-of-the-Art Methods

1) Results on the CompCars Web-Nature Dataset: We evaluate the performance of our method on the CompCars web-nature dataset. The evaluation metrics include efficiency (i.e., network parameters and floating-point operations per second) and effectiveness (i.e., top-1 accuracy). The essential setting parameters are listed in Table I, including the input image resolution ("Size") and bounding box annotations ("Anno."), which significantly affect the VMR performance.

First, we compare our method with several classic state-of-the-art CNN models, especially the models extended based on ResNet-50, on the CompCars web-nature dataset. The results are summarized in Table I. In the table, "-" indicates that the information is not mentioned in the relevant paper, and the network names in the brackets indicate the backbone network used in the method.

As shown in Table I, EP-CNN achieves 4.6% gains in Top-1 accuracy versus the original ResNet-50 when the input size is 224×224, with only a slight increase in the model complexity. Compared with other models (i.e., ResNeXt-50, SE-ResNet-50, Res2Net-50, Res2NeXt-50 and FixResNet-50) that use ResNet-50 as the baseline, EP-CNN obtains superior and competitive results with a similar model complexity. Moreover, EP-CNN outperforms the advanced recognition network EfficientNet-B5 by 2.6%.

In addition, we compared EP-CNN with several state-of-the-art fine-grained VMR methods (i.e., BoxCars, FM-CNN, Location-Aware, FR-ResNet, ID-CNN and ResNet151-CMP) or fine-grained methods based on the CompCars dataset (i.e., ABN and DCL). We report the accuracies of these algorithms on CompCars from the original papers or codes. For an input image resolution of 224×224, the proposed EP-CNN achieves higher recognition accuracy than BoxCars, ID-CNN and ResNet152-CMP. Among these methods, BoxCars considers that the vehicle viewpoint information is useful for VMR, but the encoded vehicle viewpoints input to the network cannot efficiently enhance the final classification results.

When the input image resolution is 448×448, the proposed method outperforms the DCL. Although the input image resolutions of FM-CNN, Location-Aware and ABN are not regular, the proposed method achieves a highter performance. These results verify the efficiency and effectiveness of the proposed EP-CNN method.

2) Results on the Stanford Cars Dataset: The EP-CNN network can also be employed for datasets without the vehicle viewpoint information. For these datasets, we pretrain PE-SubNet using the CompCars web-nature dataset, fix the network parameters and fine-tune only VMC-SubNet based on these datasets.

Extensive experiments are conducted on the Stanford Cars dataset, which does not contain the vehicle viewpoint information. We use only the category label of the Stanford Cars dataset to fine-tune VMC-SubNet.

The experimental results are listed in Table II. Compared with the original ResNet-50 model, the EP-CNN achieves 5.7% gains in the Top-1 accuracy without a considerable increase in the model complexity. Compared with other models (i.e., ResNeXt-50, SE-ResNet-50, Res2Net-50, Res2NeXt-50 and FixResNet-50) that use ResNet-50 as the baseline, EP-CNN achieves superior and competitive results with a similar model complexity based on the Stanford Cars dataset. Moreover, EP-CNN outperforms the advanced recognition network EfficientNet-B5 by 3.0%. Compared with state-of-the-art fine-grained vehicle model classification methods (i.e., FR-ResNet, ID-CNN and ResNet152-CMP) or other fine-grained methods based on the Stanford Cars dataset in (i.e., FCAN, RA-CNN, MA-CNN, MAMC, DFL-CNN, TASN, AKEN, BiM-PMA, MOMN and BRAM), EP-CNN achieves competitive results. We report the accuracies of these algorithms on Stanford Cars from the original papers or codes.

TABLE II
COMPARISON OF THE RECOGNITION RESULTS ON THE STANFORD CARS DATASET WITHOUT BOUNDING BOXES.

Method	Size	#P(M)	GFLOPs	Anno.	Top-1
ResNet-50 [36]	224	25.5	4.1	×	87.8%
ResNeXt-50, 32×4d [38]	224	25.0	4.2	×	88.3%
SE-ResNet-50 [35]	224	26.9	4.1	×	88.2%
Res2Net-50,26w×4s [44]	224	24.5	4.3	×	89.8%
Res2NeXt-50,26w×4s [44]	224	23.5	4.2	×	89.1%
EfficientNet-B5 [45]	224	29.2	9.9	×	90.5%
FixResNet-50 [46]	224	25.5	4.1	×	91.2%
FR-ResNet [47] (ResNet-50)	224	⩾25.5	⩾4.1	×	90.6%
FR-ResNet (ResNet-50)	224	⩾25.5	⩾4.1	✓	93.1%
FCAN [49] (ResNet-50)	448	⩾25.5	⩾16.4	×	91.5%
FCAN (ResNet-50)	448	⩾25.5	⩾16.4	✓	93.1%
ID-CNN [10]	224	143.6	19.7	×	91.8%
RA-CNN [13] (VGG-19 [50])	448	265.9	117.7	✓	92.5%
MA-CNN [14] (VGG-19)	448	143.6	19.7	✓	92.8%
ResNet152-CMP [15]	224	⩾59.0	⩾11.6	✓	92.9%
MAMC [51] (ResNet-50)	448	434	192.8	×	92.8%
DFL-CNN [52] (ResNet-50)	448	29.8	20.5	×	93.1%
TASN [53] (ResNet-50)	448	40.7	25.6	×	93.8%
AKEN [54] (ResNet-50)	448	⩾25.5	⩾16.4	×	92.6%
BiM-PMA [55] (ResNet-50)	448	⩾25.5	⩾16.4	×	93.1%
MOMN [56] (ResNet-50)	448	32.1	20.7	×	93.2%
BRAM [57] (ResNet-50)	448	29.3	19.3	×	94.3%
EP-CNN(ours)	224	28.0	4.7	×	**93.5%**
EP-CNN(ours)	448	28.0	18.9	×	**94.6%**

D. Ablation Study

1) Ablation Study for EP-CNN Model: We analyze the importance of each strategy employed in EP-CNN via the ablation study. ResNet-50 is employed as the baseline of our evaluation. The ablation study results are listed in Table III. We analyze the results from the following aspects.

MultiSE block versus SE block: The MultiSE block is embedded in ResNet-50 in the same way as the SE block. As shown in Table III, the Top-1 accuracy of MultiSE-ResNet-50 is 2.0% higher than that of SE-ResNet-50. This phenomenon occurs because the MultiSE block can adaptively select more favorable features for VMR at multiple scales. After separately combining the vehicle viewpoint and pose features, the MultiSE block can achieve an accuracy that is 1.5% and 1.6% higher than that of the SE block, respectively. After combining the viewpoint and pose features, the performance of the MultiSE block is 1.3% higher than that

TABLE III
ABLATION STUDY FOR EP-CNN MODEL ON THE COMPCARS DATASET.

Baseline	SE	MultiSE	Pose information		Top-1
			Vehicle viewpoint	Pose features	Accuracy
ResNet-50					94.0%
	✓				95.1%
		✓			97.1%
	✓		✓		96.5%
		✓	✓		98.0%
	✓			✓	96.3%
		✓		✓	97.9%
	✓		✓	✓	97.3%
		✓	✓	✓	98.6%

of the SE block. Regardless of the combination of viewpoint and/or pose features, the performance of the MultiSE block is always higher than that of SE block. These findings prove the efficiency of the MultiSE block.

Effectiveness of pose information: As mentioned, the pose information includes the vehicle viewpoint and pose features. First, we combine the vehicle viewpoint with MultiSE-ResNet-50 and SE-ResNet-50. The Top-1 accuracies are increased by 0.9% and 1.4%. Second, we combine pose features with the two models. The Top-1 accuracies are separately enhanced by 0.8% and 1.2%. These results prove that the vehicle viewpoint and pose features are useful information for our MV-VMR task, and either of them can enhance the recognition accuracy.

We simultaneously combine the vehicle viewpoint and pose features, as listed in Table III. The Top-1 accuracies are higher than those obtained by combining either the viewpoint or pose features or neither the viewpoint nor pose features. This result proves that both vehicle viewpoint and pose features are complementary, i.e., the pose information is effective for the MV-VMR task.

TABLE IV
RESULTS OF DIFFERENT ANCHOR BOX DIMENSION CLUSTERING METHODS ON COMPCARS DATASET.

Anchor box dimension clustering settings	Detection accuracy (mAP) of PE-SubNet	Top-1 accuracy of EP-CNN
K-means(COCO)	97.4%	97.7%
K-means(VR-COCO)	98.7%	98.4%
K-means++(VR-COCO)	99.0%	98.6%

2) Ablation Study on Anchor Box Dimension Clustering: We perform experiments to prove the effectiveness of the anchor box dimension clustering methods. Table IV shows that in the configuration with k-means++ run on the VR-COCO dataset, the mAP of detecting the viewpoints in PE-SubNet is enhanced by 0.3% and 1.6%, and the final Top-1 recognition accuracies are increased by 0.2% and 0.9%, compared with k-means run on VR-COCO and COCO datasets, respectively. These experimental results prove the effectiveness of our anchor box dimension clustering strategy.

TABLE V
COMPARISON OF CLASSIFICATION ACCURACY WITH DIFFERENT LOSS FUNCTIONS FOR EP-CNN ON THE COMPCARS AND STANFORD CARS DATASETS.

Loss function	CompCars	Stanford Cars
Center Loss [58]	96.5%	92.3%
A-softmax Loss [59]	98.2%	93.4%
CE Loss	98.7%	94.1%
CE Loss + MC Loss [39]	98.9%	94.6%

3) Ablation Study on Loss: Table V presents the experimental results of using different classification loss functions in VMC-SubNet on the CompCars and Stanford Cars datasets. The methods use the same experimental settings with an image input size of 448×448. As shown in Table V, CE-Loss outperforms Center Loss and A-softmax Loss. The combination of CE-Loss and MC-Loss achieves the highest Top-1 accuracy, which is 0.2% and 0.5% higher than that achieved using only CE-Loss on the CompCars and Stanford Cars dataset, respectively.

TABLE VI
RESULTS OF DIFFERENT MULTISE BLOCKS FOR EP-CNN.

Method	Branch numbers	Dilation rate	#P	GFLOPs	Top-1 Accuracy (EP-CNN)
MultiSE block	2	1,2	26.3M	4.5	98.3%
	2	1,3	26.3M	4.5	97.8%
	3	1,2,3	28.0M	4.7	98.6%
	4	1,2,3,4	29.8M	4.9	98.5%

4) Evaluation of the Number of Branches for MultiSE Block: In this section, we explain why three branches are adopted in the MultiSE block based on the experimental results. If we only choose one branch, the MultiSE block is the same as the SE block, and the experimental results in Section IV.D.1 prove the effectiveness of MultiSE over the SE block.

Therefore, we consider the case of two or more branches. First, we separately choose two branches with dilation rates of 1, 2 and 1, 3. The results listed in Table VI prove that the corresponding Top-1 accuracies are 0.3% and 0.8% lower than those of the three branches, but the numbers of parameters and GFLOPs are slightly lower. Next, we set four branches with dilation rates of 1, 2, 3 and 4. As shown in Table

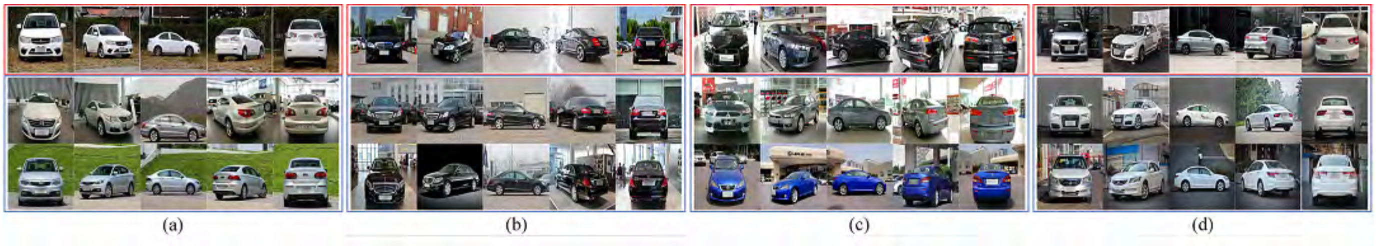


Fig. 6. Four vehicle models with the lowest area under the curve (AUC) values: (a) Class 319, (b) Class 330, (c) Class 345, and (d) Class 384.

VI, the Top-1 accuracy is slightly decreased and the number of parameters and GFLOPs increase, which means that a large number of branches cannot ensure a higher recognition accuracy. Moreover, the computation costs increase.

To achieve a balance between the recognition accuracy and computation cost, we choose three branches for the MultiSE block. In this case, the highest recognition accuracy can be achieved at only a slight computation cost overhead.

E. Failure Cases Analysis

1) Analysis from the Perspective of Dataset: To evaluate the performance of the EP-CNN method, we analyze the reasons for failed cases. We calculate the recognition accuracies of all classes and choose four classes of samples with the lowest recognition accuracy from the CompCars web-nature dataset.

By observing the datasets of these four classes, we can derive the following conclusions:

1. The training and test samples are not uniform. As shown in classes 330 and 384 in Fig. 6, the problem of uneven samples in the CompCars web-nature dataset is twofold. There exist considerably fewer samples in one class than in other classes. The training set in the CompCars web-nature dataset contains 36,456 vehicle images of 431 classes, with an average of 84 images for each vehicle model. However, only 30 and 17 images are present in classes 330 and 384, respectively, considerably fewer than the average value. Moreover, the intraclass distribution of images from different viewpoints is uneven. The images of classes 330 and 384 in the training set are focused on the side, front-side, and rear-side images. Consequently, the model has difficulity to recognize the front and rear vehicle images in the test set.
2. The appearances are only slightly different. As shown in Fig. 6, the vehicles in classes 319 and 345 appear similar to other vehicles of the same brand or other brands, leading to identification failure.

2) Analysis from the Perspective of Viewpoint: In Fig. 7, the rectangle represents the whole dataset. The two areas marked with different shades of blue represent success and failure cases with the viewpoint. The areas inside and outside the red ellipse represent the success and failure case without the viewpoint, respectively. Based on the statistical data of these areas, we can infer the area of A and B. A represents the cases in which using the viewpoint information causes the VMR to fail, accounting for 0.0186% of the testing dataset. B represents the cases in which using the viewpoint information causes the VMR to succeed, accounting for 1.651% of the testing dataset. The percentage of B is considerably larger than that of A, which indicates that even though the vehicle viewpoint information may cause the VMR to fail, the number of cases with success outweighs this risk.

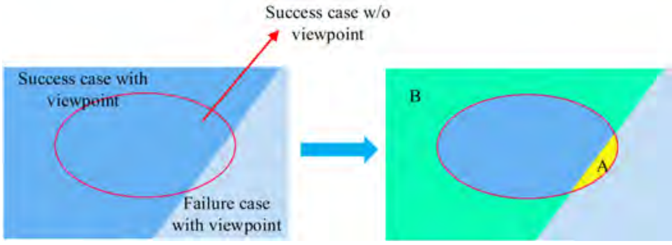

Success case w/o viewpoint
Success case with viewpoint
Failure case with viewpoint
B
A

Fig. 7. Statistics pertaining to VMR success/failure cases with/without viewpoint on the CompCars dataset.

F. Visual Analysis of Network Features

1) Visual Analysis based on the t-SNE Method: To analyze the feature extraction ability of the model, we decrease the dimensions of the features extracted from EP-CNN based on the CompCars dataset to two dimensions by using the t-SNE [60] [61] method and compare the visualization results with those of ResNet-50, SE-ResNet-50 and MultiSE-ResNet-50, as shown in Fig. 8. Moreover, we visualize the features from PE-SubNet. Each point in the figure represents a test sample, and points of the same color represent the same class. Generally, the last layer of CNN is used to map features to specific categories. Therefore, we choose the penultimate layer to extract features for visualization.

In Fig. 8(a), the sample points exhibit a clustering trend; however, substantial overlap exists between points of different classes. Thus, ResNet-50 learns useful features for fine-grained vehicle model classification, but different categories cannot be sufficiently distinguished.

As shown in Fig. 8(b), compared to the results shown in Fig. 8(a), all samples are effectively clustered, and the interclass variance is higher than that observed in Fig. 8(a). However, in the center region of the image, the level of overlap is higher than that in other regions but lower than that in Fig. 8(a). Thus, SE-ResNet-50 can extract more distinguishable features than ResNet-50. Fig. 8(c) shows the visualization of the features extracted from MultiSE-ResNet-50. As shown in Fig. 8(c), MultiSE-ResNet-50 has a superior feature extraction ability than SE-ResNet-50 and ResNet-50. Although various

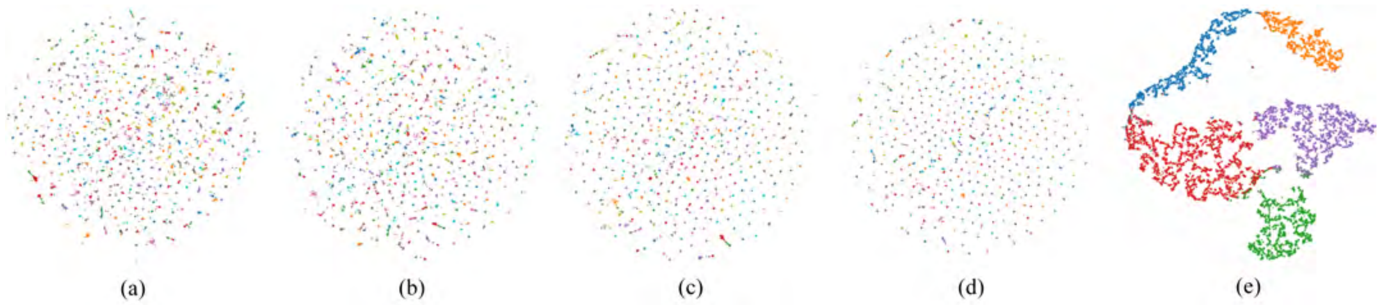


Fig. 8. Visualization of features after dimension reduction: (a) ResNet-50, (b) SE-ResNet-50, (c) MultiSE-ResNet-50, (d) EP-CNN and (e) PE-SubNet.

sample points are more clearly separated, and the same sample points are more tightly clustered, the interclass variance is not adequately large.

The features extracted by the EP-CNN model are shown in Fig. 8(d). Points of the same class are clustered compactly, and points of different classes are well distinguished. The degree of overlap is considerably lower than that of the features shown in Figs. 8(a), (b) and (c), which indicates that EP-CNN can extract more distinguishable features than the previously discussed models.

To verify the effectiveness of PE-SubNet, we visualize the features extracted from it. As shown in Fig. 8(e), five classes are obviously distinguished, which indicates that the vehicle viewpoint that is extracted from PE-SubNet can provide reliable information that guides the flow of the training process.

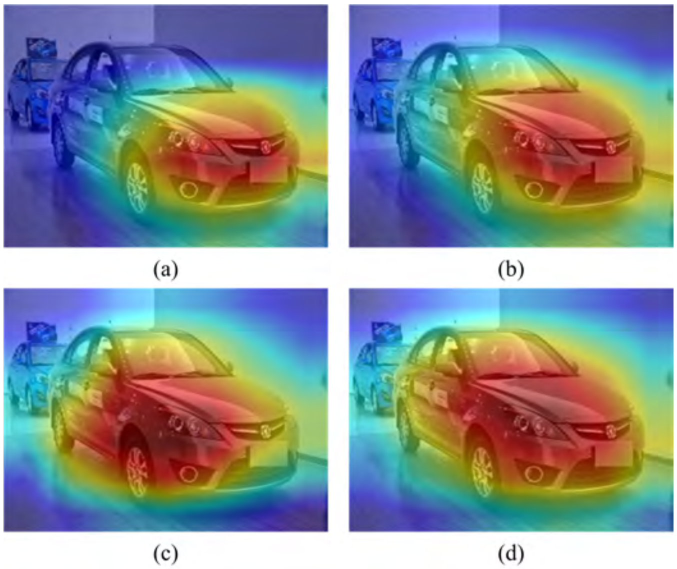


Fig. 9. Grad-CAM visualization results for different models. (a) ResNet-50. (b) SE-ResNet-50. (c) MultiSE-ResNet-50. (d) EP-CNN.

2) Visual Analysis based on the Class Activation Map (CAM) Method: To more intuitively compare the EP-CNN model and other networks in terms of the ability to extract vehicle features, we visualize the CAMs [62] using the gradient-weighted class activation mapping (Grad-CAM) [63] method for VMR. Grad-CAM is a visualization method that uses gradients to generate CAMs. This method is commonly employed to localize the discriminative regions for image classification. In CAMs, the discriminative regions of a vehicle used for classification are highlighted, and thus, CAMs function in a similar manner as attention maps.

As shown in Fig. 9, the ResNet-50 based Grad-CAM result covers only the front part of the vehicle (highlighted area) and focuses on the light and logo areas (shown in red). SE-ResNet-50 assigns a higher attention to the features in the hood section of the vehicle than ResNet-50. The highlighted area of MultiSE-ResNet-50 is larger than that of SE-ResNet-50, and the side-front area receives more attention. The EP-CNN has activation maps that tend to cover the key vehicle area, such as lights, and the vehicle logo and hoods receive considerable attention. These results prove that pose information enables the EP-CNN to extract more discriminative vehicle features.

V. Conclusion

This paper proposes a fine-grained VMR model, EP-CNN, is proposed. EP-CNN is composed of two subnetworks: one subnetwork is used for vehicle viewpoint estimation and pose feature extraction, and the other subnetwork is used for classification. During classification, multiview vehicle images are input, and the MultiSE block with pose information is added to the residual block. We use the MultiSE block to fully exploit the vehicle's multiscale characteristics and fuse the vehicle viewpoint and pose features into the classification network, thereby enhancing features that contribute more to the final classification and obtaining superior recognition results. The experimental results demonstrate that the proposed method can achieve a higher recognition accuracy than most classic CNN models and several state-of-the-art fine-grained vehicle model classification algorithms.

References

[1] C. N. E. Anagnostopoulos, I. E. Anagnostopoulos, V. Loumos, and E. Kayafas, "A license plate-recognition algorithm for intelligent transportation system applications," *IEEE Transactions on Intelligent Transportation Systems*, vol. 7, no. 3, pp. 377–392, 2006.

[2] S. Du, M. Ibrahim, M. Shehata, and W. Badawy, "Automatic license plate recognition (alpr): A state-of-the-art review," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 23, no. 2, pp. 311–325, 2013.

[3] Y. Yu, H. Li, J. Wang, H. Min, W. Jia, J. Yu, and C. Chen, "A multilayer pyramid network based on learning for vehicle logo recognition," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 5, pp. 3123–3134, 2021.

[4] Y. Yu, J. Wang, J. Lu, Y. Xie, and Z. Nie, "Vehicle logo recognition based on overlapping enhanced patterns of oriented edge magnitudes," *Computers & Electrical Engineering*, vol. 71, pp. 273–283, 2018.

[5] Z. Dong, Y. Wu, M. Pei, and Y. Jia, "Vehicle type classification using a semisupervised convolutional neural network," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 4, pp. 2247–2256, 2015.

[6] Y. Yu, Q. Jin, and C. Chen, "Ff-cmnet: A cnn-based model for fine-grained classification of car models based on feature fusion," in *2018 IEEE International Conference on Multimedia and Expo*, 2018, pp. 1–6.

[7] J. Krause, T. Gebru, J. Deng, L. Li, and L. Feifei, "Learning features and parts for fine-grained recognition," in *2014 22nd International Conference on Pattern Recognition*, 2014, pp. 26–33.

[8] H. He, Z. Shao, and J. Tan, "Recognition of car makes and models from a single traffic-camera image," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 6, pp. 3182–3192, 2015.

[9] 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, 2017.

[10] 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*, 2018, pp. 3279–3284.

[11] Y. Zhou, J. Yuan, and X. Tang, "A novel part-based model for fine-grained vehicle recognition," in *International Conference on Cloud Computing and Security*, 2018, pp. 647–658.

[12] S. Ghassemi, A. Fiandrotti, E. Caimotti, G. Francini, and E. Magli, "Vehicle joint make and model recognition with multiscale attention windows," *Signal Processing: Image Communication*, vol. 72, pp. 69–79, 2019.

[13] 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*, 2017, pp. 4438–4446.

[14] 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*, 2017, pp. 5209–5217.

[15] 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.

[16] A. Farhadi and J. Redmon, "Yolov3: An incremental improvement," in *Computer Vision and Pattern Recognition*, 2018, pp. 1804–2767.

[17] X. Zhang, F. Zhou, Y. Lin, and S. Zhang, "Embedding label structures for fine-grained feature representation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1114–1123.

[18] F. Zhou and Y. Lin, "Fine-grained image classification by exploring bipartite-graph labels," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1124–1133.

[19] Q. Qian, R. Jin, S. Zhu, and Y. Lin, "Fine-grained visual categorization via multi-stage metric learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3716–3724.

[20] X. Zhe, S. Chen, and H. Yan, "Directional statistics-based deep metric learning for image classification and retrieval," *Pattern Recognition*, vol. 93, pp. 113–123, 2019.

[21] 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.

[22] X. Wei, C. Xie, J. Wu, and C. Shen, "Mask-cnn: Localizing parts and selecting descriptors for fine-grained bird species categorization," *Pattern Recognition*, vol. 76, pp. 704–714, 2018.

[23] P. Rodríguez, D. Velazquez, G. Cucurull, J. M. Gonfaus, F. X. Roca, and J. Gonzàlez, "Pay attention to the activations: A modular attention mechanism for fine-grained image recognition," *IEEE Transactions on Multimedia*, vol. 22, no. 2, pp. 502–514, 2020.

[24] Y. Yu, L. Xu, W. Jia, W. Zhu, Y. Fu, and Q. Lu, "Cam: A fine-grained vehicle model recognition method based on visual attention model," *Image and Vision Computing*, vol. 104, p. 104027, 2020.

[25] L. Zhang, S. Huang, W. Liu, and D. Tao, "Learning a mixture of granularity-specific experts for fine-grained categorization," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 8331–8340.

[26] R. Ji, L. Wen, L. Zhang, D. Du, Y. Wu, C. Zhao, X. Liu, and F. 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. 10 468–10 477.

[27] G. Sun, H. Cholakkal, S. Khan, F. Khan, and L. Shao, "Fine-grained recognition: Accounting for subtle differences between similar classes," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 12 047–12 054.

[28] Y. Ding, Z. Han, Y. Zhou, Y. Zhu, J. Chen, Q. Ye, and J. Jiao, "Dynamic perception framework for fine-grained recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, 2021.

[29] J. Sochor, A. Herout, and J. Havel, "Boxcars: 3d boxes as cnn input for improved fine-grained vehicle recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3006–3015.

[30] R. Chu, Y. Sun, Y. Li, Z. Liu, C. Zhang, and Y. Wei, "Vehicle re-identification with viewpoint-aware metric learning," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 8282–8291.

[31] B. Hu, J. Lai, and C. Guo, "Location-aware fine-grained vehicle type recognition using multi-task deep networks," *Neurocomputing*, vol. 243, pp. 60–68, 2017.

[32] J. Sochor, J. Špaňhel, and A. Herout, "Boxcars: Improving fine-grained recognition of vehicles using 3-d bounding boxes in traffic surveillance," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 1, pp. 97–108, 2019.

[33] Z. Chen, C. Ying, C. Lin, S. Liu, and W. Li, "Multi-view vehicle type recognition with feedback-enhancement multi-branch cnns," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 9, pp. 2590–2599, 2019.

[34] Y. Chen, Y. Bai, W. Zhang, and T. 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.

[35] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7132–7141.

[36] 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*, 2016, pp. 770–778.

[37] Y. Li, Y. Chen, N. Wang, and Z. Zhang, "Scale-aware trident networks for object detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 6054–6063.

[38] 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*, 2017, pp. 1492–1500.

[39] D. Chang, Y. Ding, J. Xie, A. K. Bhunia, X. Li, Z. Ma, M. Wu, J. Guo, and Y.-Z. Song, "The devil is in the channels: Mutual-channel loss for fine-grained image classification," *IEEE Transactions on Image Processing*, vol. 29, pp. 4683–4695, 2020.

[40] 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*, 2015, pp. 3973–3981.

[41] J. Krause, M. Stark, J. Deng, and L. Feifei, "3d object representations for fine-grained categorization," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2013, pp. 554–561.

[42] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Association for Computing Machinery*, vol. 60, no. 6, pp. 84–90, 2017.

[43] H. Touvron, A. Vedaldi, M. Douze, and H. Jégou, "Fixing the train-test resolution discrepancy," *arXiv preprint arXiv:1906.06423*, 2020. [Online]. Available: https://arxiv.org/abs/1906.06423

[44] S. Gao, M. Cheng, K. Zhao, X. Zhang, M. Yang, and P. Torr, "Res2net: A new multi-scale backbone architecture," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 2, pp. 652–662, 2021.

[45] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proceedings of the 36th International Conference on Machine Learning*, vol. 97, 2019, pp. 6105–6114.

[46] H. Touvron, A. Vedaldi, M. Douze, and H. Jégou, "Fixing the train-test resolution discrepancy: Fixefficientnet," *arXiv preprint arXiv:2003.08237*, 2020. [Online]. Available: https://arxiv.org/abs/2003.08237

[47] Y. Yu, Y. Fu, C. Yang, and Q. Lu, "Fine-grained car model recognition based on fr-resnet," *Acta Automatica Sinica*, vol. 47, no. 5, pp. 1125–1136, 2021.

[48] H. Fukui, T. Hirakawa, T. Yamashita, and H. Fujiyoshi, "Attention branch network: Learning of attention mechanism for visual explanation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10 705–10 714.

[49] X. Liu, T. Xia, J. Wang, Y. Yang, F. Zhou, and Y. Lin, "Fully convolutional attention networks for fine-grained recognition," in *Proceedings*

of the 31st AAAI Conference on Artificial Intelligence, 2017, pp. 4190–4196.

[50] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *3rd International Conference on Learning Representations*, 2015. [Online]. Available: http://arxiv.org/abs/1409.1556

[51] M. Sun, Y. Yuan, F. Zhou, and E. Ding, "Multi-attention multi-class constraint for fine-grained image recognition," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 805–821.

[52] Y. Wang, V. I. Morariu, and L. S. Davis, "Learning a discriminative filter bank within a cnn for fine-grained recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4148–4157.

[53] H. Zheng, J. Fu, Z. Zha, and J. Luo, "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*, 2019, pp. 5012–5021.

[54] Y. Hu, Y. Yang, J. Zhang, X. Cao, and X. Zhen, "Attentional kernel encoding networks for fine-grained visual categorization," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 1, pp. 301–314, 2021.

[55] K. Song, X. Wei, X. Shu, R. Song, and J. Lu, "Bi-modal progressive mask attention for fine-grained recognition," *IEEE Transactions on Image Processing*, vol. 29, pp. 7006–7018, 2020.

[56] S. Min, H. Yao, H. Xie, Z.-J. Zha, and Y. Zhang, "Multi-objective matrix normalization for fine-grained visual recognition," *IEEE Transactions on Image Processing*, vol. 29, pp. 4996–5009, 2020.

[57] C. Liu, H. Xie, Z. Zha, L. Yu, Z. Chen, and Y. Zhang, "Bidirectional attention-recognition model for fine-grained object classification," *IEEE Transactions on Multimedia*, vol. 22, no. 7, pp. 1785–1795, 2020.

[58] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *European Conference on Computer Vision*, 2016, pp. 499–515.

[59] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "Sphereface: Deep hypersphere embedding for face recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 212–220.

[60] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[61] L. Van Der Maaten, "Accelerating t-sne using tree-based algorithms," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 3221–3245, 2014.

[62] 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*, 2016, pp. 2921–2929.

[63] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 618–626.

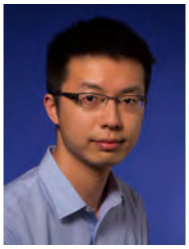
Ye Yu received her Ph.D. Degree in Computer Science and Technology from Hefei University of Technology (HFUT), Hefei, China, in May 2010. In June 2010, She joined the School of Computer and Information, HFUT, Hefei, China, where she is currently an Associate Professor. She was a Visiting Scholar with the University of North Texas, and the State University of New York at Buffalo. Her research interests include Computer Vision, Artificial Intelligence and Machine Learning, 3D Modeling and Virtual Reality. She is a member of the CCF (China Computer Federation) and CSIG (China Society of Image and Graphics).

Haitao Liu is pursuing his Master degree from School of Computer Science and Information, Hefei University of Technology, Hefei, China. His research interests include computer vision, artificial intelligence and machine learning.

Yuanzi Fu is pursuing his Master degree from School of Computer Science and Information, Hefei University of Technology, Hefei, China. He received the B.Sc. degree in Computer Science and Technology from Hefei University of Technology (HFUT), Hefei, China, in May 2018. His research interests include computer vision, artificial intelligence and machine learning.

Wei Jia received the B.Sc. degree in informatics from Central China Normal University, Wuhan, China, in 1998, the M.Sc. degree in computer science from Hefei University of Technology, Hefei, China, in 2004, and the Ph.D. degree in pattern recognition and intelligence system from University of Science and Technology of China, Hefei, China, in 2008. He has been a research associate professor in Hefei Institutes of Physical Science, Chinese Academy of Science from 2008 to 2016. He is currently an associate professor in Key Laboratory of Knowledge Engineering with Big Data, Ministry of Education, and in School of Computer Science and Information Engineering, Hefei University of Technology. His research interests include computer vision, biometrics, pattern recognition, image processing and machine learning.

Jun Yu received his Ph.D. degree in Pattern Recognition and Intelligence System from University of Science and Technology of China, Hefei, China, in 2010. He is currently an associate professor of the Department of Automation, University of Science and Technology of China. His research interests are multimedia computing, multi-modal information synthesis, perception and cognition. He has published more than 100 journal and conference papers, including TMM, TASLP, TCSVT, TCYB, TOMM, ACL, CVPR, MM, SIGGRAPH ASIA, VR, IJCAI. He has received two Best Paper Awards from IEEE ICME, IEEE FG, and has won 12 Grand Challenge Champion, 1st Runner-up and 2nd Runner-up Awards from premier conferences, such as ACM MM, IEEE ICME, IEEE FG. Email: harryjun@ustc.edu.cn.



Zhisheng Yan is currently an Assistant Professor in Department of Information Science and Technology, School of Computing, at George Mason University. He leads the Mason immErsive meDia computIng and Applications (MEDIA) Lab. Previously, he was an Assistant Professor in the Department of Computer Science at Georgia State University and a visiting researcher in the Department of Electrical Engineering at Stanford University. He received his Ph.D. degree in Computer Science and Engineering from University at Buffalo, The State University of New York. His research focuses on the systems and security issues of immersive computing systems, such as VR, AR, imaging, and video systems. His research has been recognized by several awards, including NSF CAREER Award, NSF CRII Award, ACM SIGMM Best PhD Thesis Award, University at Buffalo CSE Best Dissertation Award, ACM HotMobile'18 Best Demo Award, and IEEE HealthCom'14 Best Student Paper Runner-up.