Chinese Journal of Mechanical Engineering

ORIGINAL ARTICLE Open Access

Adaptive Multi-modal Fusion Instance Segmentation for CAEVs in Complex Conditions: Dataset, Framework and Verifications

Pai Peng, Keke Geng, Guodong Yin*, Yanbo Lu, Weichao Zhuang and Shuaipeng Liu

Abstract

Current works of environmental perception for connected autonomous electrified vehicles (CAEVs) mainly focus on the object detection task in good weather and illumination conditions, they often perform poorly in adverse scenarios and have a vague scene parsing ability. This paper aims to develop an end-to-end sharpening mixture of experts (SMoE) fusion framework to improve the robustness and accuracy of the perception systems for CAEVs in complex illumination and weather conditions. Three original contributions make our work distinctive from the existing relevant literature. The Complex KITTI dataset is introduced which consists of 7481 pairs of modified KITTI RGB images and the generated LiDAR dense depth maps, and this dataset is fine annotated in instance-level with the proposed semi-automatic annotation method. The SMoE fusion approach is devised to adaptively learn the robust kernels from complementary modalities. Comprehensive comparative experiments are implemented, and the results show that the proposed SMoE framework yield significant improvements over the other fusion techniques in adverse environmental conditions. This research proposes a SMoE fusion framework to improve the scene parsing ability of the perception systems for CAEVs in adverse conditions.

Keywords: Connected autonomous electrified vehicles, Multi-modal fusion, Semi-automatic annotation, Sharpening mixture of experts, Comparative experiments

1 Introduction

1.1 Motivations and Technical Challenges

Connected autonomous electrified vehicles (CAEVs) offer high potential to improve road safety, boost traffic efficiency and minimize carbon emissions [1], as well as reduce vehicle wear, transportation times and fuel consumption [2, 3]. Perception systems in CAEVs [4] are fundamental to decision making, route planning, obstacle avoidance and trajectory tracking [5], etc. As the most essential sensor of perception systems, the visual camera can provide detailed shape and texture information of the surroundings, which can be used to detect lane geometry, traffic signs and object class. In recent years, vision-based state-of-the-art deep neural network models [6–9] are prevalently used in scene understanding for CAEVs, demonstrating impressive performance in object detection as well as semantic and instance segmentation. However, the camera is vulnerable to lighting and weather conditions, examples range from low luminosity at night-time, extreme brightness disparity in sun glare to rainy or snowy weather, resulting in image degradation. Consequently, the performance of the deep neural network models decreases enormously or even fails which can potentially lead to catastrophic consequences.

To this end, perception systems [10] in CAEVs usually exploit the complementary and comprehensive information from multi-modal sensors like vision cameras, LiDARs and Radars to accurately perceive the surrounding traffic conditions. As active sensors, LiDARs offer accurate 3D information of the surroundings in the form of point cloud by emitting and receive laser beams. And

*Correspondence: ygd@seu.edu.cn
School of Mechanical Engineering, Southeast University, Nanjing, China

it is robust to extreme lighting conditions and is less influenced by adverse weather conditions. But due to the inherent sparse characteristics of LiDAR points, it cannot capture the fine texture and shape of objects. The combined information of camera and LiDAR is resilient to commonly observed perceptual variations, hence many deep learning based works [11–13] have been devoted to fusing features from camera and LiDAR and have shown promising results in environmental perception.

Besides, deep learning approaches are data-hungry, and their performance is strongly correlated with the amount of available training data. There are many published multi-modal datasets with fine annotated ground truth that can be used for fusing cameras and LiDARs, such as KITTI [14], ApolloScape [15], BLVD [16], nuScenes [17], H3D [18], etc. Yet, most of them do not encompass complex weather and illumination conditions, making deep learning models that are benchmarked on them perform poorly in adverse conditions. Moreover, common multi-modal datasets provide only bounding box annotations, which may contain noise from the background or other objects, while instance-level annotations provide a more detailed and natural parsing of a scene by precisely segmenting each object instance in the images. Further improvements in deep neural network models will only be possible when moving to the instance-level segmentation, which can be challenging as labeling data is extremely labor-intensive.

1.2 Original Contributions

This paper is aiming to improve the accuracy and robustness of perception systems for CAEVs in complex illumination and weather conditions using the combined information from the camera and LiDAR. As depicted in Figure 1, the KITTI dataset is used to train and benchmark our deep learning models for its diverse and rich data types, wide application for autonomous driving research. Due to the limited available annotated images (only 200 frames) of the KITTI instance segmentation benchmark, all 7481 training images from the KITTI object detection benchmark are annotated at instance-level by a semi-automatic annotation approach. Also, the corresponding 7481 training LiDAR point cloud frames are projected into the camera coordinate and then up-sampled as dense depth maps (DDM) according to the range data. Since the KITTI dataset is recorded only during daytime and on sunny days, the dataset is further modified by adjusting the brightness, adding Gaussian noise, motion blur, or Gaussian blur to simulate the adverse environmental conditions such as night, light rain and sun glare in the real environment.

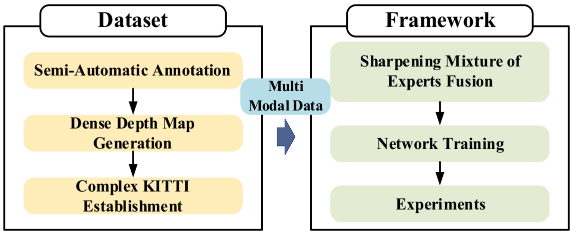
Dataset
Semi-Automatic Annotation
Dense Depth Map Generation
Complex KITTI Establishment
Multi Modal Data
Framework
Sharpening Mixture of Experts Fusion
Network Training
Experiments

Figure 1 Functional scheme of the adaptive multi-modal fusion instance segmentation pipeline

To combine RGB image and dense depth map more effectively for instance segmentation, inspired by recent multi-modal fusion models [12, 19], a sharpening mixture of experts (SMoE) fusion network is proposed based on the real-time instance segmentation network YOLACT [20] to automatically learn the contribution of each modality for instance segmentation in complex scenes. The proposed model mainly consists of three components: expert networks that extract high-level semantic information from each modality, the SMoE that adaptively weights and fuses features from expert networks and the prediction head that further learns complementary fused features to yield robust and accurate segmentation. The single modal approach and several different fusion architectures are compared on the modified KITTI dataset, and the results demonstrate the proposed SMoE fusion network can significantly improve the accuracy and robustness of instance segmentation in complex illumination and weather conditions.

In summary, this paper makes the following important contributions. First, a multi-modal dataset with fine annotated instance segmentation for complex conditions is established, which is based on the popular KITTI dataset. In total, 48220 instance masks are annotated. To the best of our knowledge, there exist no such fine and amount instance annotation efforts on the KITTI dataset. Second, the sharpening mixture of experts fusion network is proposed to learn the robust kernels from complementary modalities. Third, a comprehensive comparison of several fusion architectures, as well as single modal approach, on the multi-modal dataset is implemented. The proposed SMoE framework yields significant improvements over other fusion techniques in adverse environmental conditions.

1.3 Outline of the Paper

The rest of this paper is organized as follows. The related works are firstly reviewed in Section 2. Then the procedure of semi-automatic annotation, dense depth map generation and image modification from KITTI dataset are introduced in Section 3. In Section 4, the proposed SMoE fusion network is described in detail. Extensive experimental results and corresponding analyzes are reported in Section 5. Finally, conclusions are drawn in Section 6.

2 Related Works

2.1 Instance-Level Annotation

The performance of deep learning approaches is drastically affected by the amount and variety of training data, which requires large scales of high-quality annotations. However, manually labeling the ground truth instance mask typically requires 20–30 s for one object [21], with such an amount of images in nowadays dataset, it can be an extremely time-consuming and inefficient task. Several works attempt to simplify the very detailed per-pixel annotations with weakly labeled data. Xu et al. [22] and Lin et al. [23] use the scribble to annotate objects, while research in Refs. [24, 25] employs only a few points to label the object. However, the performance of these weakly supervised approaches can not compete with fully supervised ones.

Some people use the existing annotation information to produce instance segmentation. Chen et al. [26] exploit annotated 3D bounding boxes with information in the form of stereo, noisy point cloud, 3D CAD models as well as appearance models to perform accurate instance segmentation. Zhang et al. [27, 28] take advantage of Ref. [26] to further predict instance-level segmentation with depth ordering from different scales of image patches, then combine predictions into the final annotations using the Markov random field. But these methods only focus on cars which is far from enough in the autonomous driving scenario.

Significant efforts have been made to faster the instance-wise labeling procedure. Polygon-RNN [29, 30] uses a recurrent neural network to generate a polygon outlining the object instance based on the feature maps predicted by a CNN, and its performance is affected by the manually drawn bounding boxes. Fluid annotation [31] enables annotators to edit the full image annotation predicted by a neural network model to shorten the labeling time, yet, the output segmentation cannot serve as the ground truth for a benchmark for its imprecision. Similar to Ref. [31], Voigtlaender et al. [32] iterate automatically creating and manually correcting masks process until yielding ground-truth as precisely as possible, but the process could be time-consuming due to the continuous iteration.

2.2 Multi-Modal Fusion

2.2.1 Fusion Modalities

Some works focus on fusing the images from visual and thermal cameras. Zhou et al. [33] present a night-vision context enhancement algorithm by fusing these two modalities with a guided filter. Ha et al. [34] leverage the combined information to realize the semantic segmentation of street scenes for autonomous vehicles. Geng et al. [11] add the thermal camera to the visual perception system to boost the human and vehicle detection performance in low-observable conditions. In addition, the RGB images and depth images from the RGB-D camera are commonly employed to advance the ability of indoor perception for domestic robots [12, 35].

Fusing RGB image from a visual camera with a 3D LiDAR point cloud is the most common way in the literature. There are several fashions to represent LiDAR point clouds in the fusion procedure. Researches in Refs. [36, 37] utilize PointNet [38] to directly process the raw point clouds, together with RGB images feature extracted by Faster R-CNN [8], the two output streams are then fused to predict 3D object detection. Besides, 3D point clouds can be projected onto the 2D grid-based maps and processed by 2D networks. Bird's eye view map of LiDAR point cloud is widely applied to 3D environment perception [13, 39, 40] for it explicitly shows the objects' positions on the ground plane, offering easy access to localize the objects of interest.

Given the extrinsic and intrinsic matrix, the LiDAR point cloud can be projected onto the camera plane, and this sparse representation is then up-sampled as LiDAR dense depth map or dense reflectivity map. Asvadi et al. [41, 42] utilize these LiDAR modalities combined with the RGB images for vehicle detection. In this work, 3D point clouds are encoded with a dense depth map and fused with RGB images for instance segmentation.

2.2.2 Fusion Methods

Deep neural networks offer a wide range of choices to fuse the multi-modal features at different stages due to their hierarchical nature. Early fusion [43] directly stacks the channels of multi-modal inputs while late fusion [44] combines the network outputs of specific modality. Early fusion is characterized by high forward computation speed while late fusion has the advantage of high flexibility. Middle fusion neutralizes the characteristics of early fusion and late fusion: it combines the multi-modal feature maps at intermediate layers. Ref. [13] designs interactions among features of the intermediate layers and Ref. [34] employs a short-cut mechanism to realize middle fusion.

Typical fusion operations include addition, average mean and concatenation. Tong et al. [19] propose a novel sharpening fusion operation, and it strengthens the strong features and superimposes the weak features according to the calculated thresholds, effectively utilizing the characteristics of each modality. However, the above fusion operations ignore the varying contributions of each modality to the deep neural network. To this end, Guan et al. [45] devise an illumination-aware weighting mechanism to learn individual contributions from visual and thermal images. Moreover, the mixture of experts (MoE) approach is proposed in Refs. [12, 46, 47] to explicitly assign learned weights over the feature maps of each sensing modality extracted by its domain-specific network called the expert. In this paper, inspired by Refs. [12] and [19], an SMoE fusion network is devised to automatically learn weights from multi-modal features of expert networks and further use sharpening fusion operation to combine the weighted representations to yield robust instance segmentation in complex conditions.

3 Dataset

3.1 Semi-Automatic Annotation

Aiming at producing high-quality annotations with much less human effort, a semi-automatic annotation approach is proposed to generate instance masks for the KITTI dataset. The state-of-the-art object instance segmentation framework Mask-RCNN [9] is employed to automatically produce instance masks, then the masks are further revised and refined using manual polygon annotations. Unlike methods [29, 30] that require user input for each object, our labeling approach can automatically segment all objects, the annotators only need to focus on improving results for difficult cases. Two highly relevant road users to CAEVs, i.e., car and person, are considered in the annotation network.

The annotation network is trained on the 200 KITTI instance segmentation training images and the fine annotated Cityscapes instance segmentation dataset [48], for that the images in Cityscapes dataset are captured in Germany, which ensures consistency with the traffic environment of KITTI dataset. Furthermore, the names and instance classes of KITTI instance segmentation benchmark and Cityscapes dataset follow the same principle. Cityscapes dataset contains 2975 training, 500 validation and 1525 test images. Since the ground truth instances of the test set are not available, the annotation network is trained on the training set and validation set, as well as the KITTI instance segmentation training set for the first iteration. To reduce the number of redundant segments and misclassifications, the predicted segments are then sorted by detection score and non-maximum suppression (NMS). The trained model is used to produce segmentation masks on the first 1000 images of the KITTI dataset, then revise and refine the results by fixing wrong annotations manually. Then the manually revised annotations are used as additional training data. The automatic producing and manual revising process is iterated 8 times until all 7481 images have been fine annotated. To reveal the effectiveness of the semi-automatic annotation approach, the typical iteration results are shown on the 7000th image in the KITTI dataset in Figure 2. It can be seen that the mask quality of the people and the distant car (in the red box) becomes better with the increase of iterations.

In total, 48220 segmentation masks are annotated, including 7919 person masks and 40301 car masks, which makes the dataset viable for training and evaluating deep learning based techniques. To the best of our knowledge, there exist no such fine and amount instance annotation efforts on the KITTI dataset.

3.2 Dense Depth Map Generation

Given a 3D LiDAR point $(X\ Y\ Z)$, its coordinates $(u\ v)$ in the camera view can be yielded by:

$$z \cdot (u\, v\, 1)^{\mathrm{T}} = M_i \cdot M_e \cdot (X\ Y\ Z)^{\mathrm{T}}, \tag{1}$$

where M_i is the intrinsic matrix of the camera; M_e is the extrinsic matrix from LiDAR coordinate to camera coordinate; z represents the depth information of the object in camera coordinate.

As can be seen from Figure 3a, this sparse LiDAR representation is then up-sampled using the Delaunay Triangulation [42, 49], which generates a mesh from the sparse representation. Then the empty pixels are interpolated via the nearest neighbors. The generated dense depth map is further processed using the

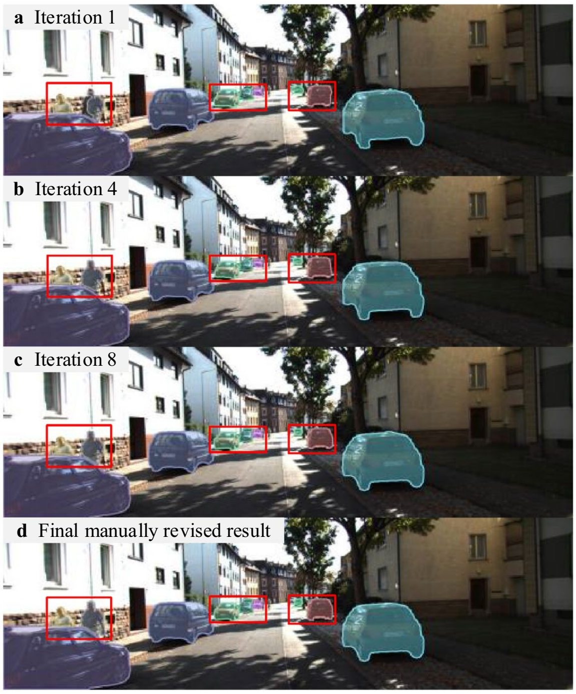

Figure 2 Iteration results of the proposed semi-automatic annotation procedure on the 7000th image of KITTI dataset

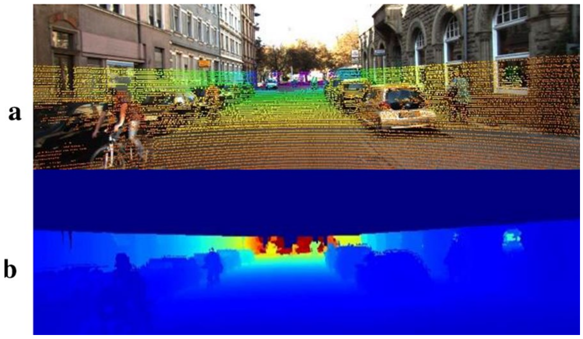

Figure 3 RGB image and generated dense depth map from KITTI dataset: **a** RGB image with superimposed sparse LiDAR data, **b** Generated dense depth map

jet colorization methodology. The DDM and corresponding RGB images of KITTI dataset are shown in Figure 3.

3.3 Complex KITTI Establishment

Since the well-known KITTI dataset is recorded only in optimal illumination and weather conditions, to train and evaluate the proposed multi-modal network, all 7481 RGB images in this dataset are modified to imitate the common complex environment, including night, light rain day, light rain night and sun glare.

As shown in Figure 4a, the camera image is prone to noise and motion blur due to the low illumination at night; this is simulated by reducing the brightness and adding Gaussian noise, as well as adding motion blur, to the images in KITTI dataset. The performance of LiDAR is less affected in light rain day or night, while the camera is much more disturbed in this condition. To get a better view, cameras in CAEVs are usually installed outside the driver's cab where the lens is vulnerable to raindrops; for cameras installed in the cab, the view field can be frequently blocked by the front windshield wiper, as depicted in Figure 4b–d. Therefore, the corresponding images are created from KITTI dataset using randomly generated Gaussian blur circles and black polygons to simulate the rainy weather. In sun glare, the camera images contain white spots, so this case is emulated using a randomly fitted white area, shown in Figure 4e. The above conditions are equally distributed over the modified KITTI dataset. Hereinafter, these modified images and corresponding LiDAR dense depth maps are entitled Complex KITTI dataset. According to the setting of Ref. [41], we divide the Complex KITTI dataset into a training set with 4489 image pairs, a validation set and a test set with 1496 image pairs, respectively.

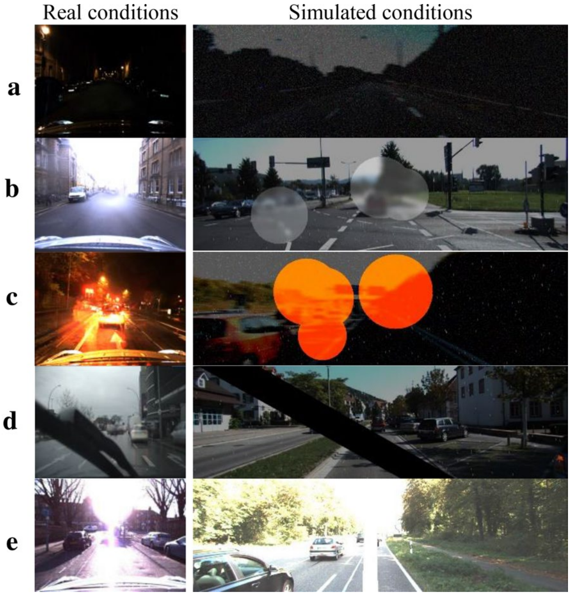

Figure 4 Real complex environment and corresponding simulated conditions on the KITTI dataset. **a** Night, **b** and **e** Light rain day, **c** Light rain night, **d** Glare. The real environment images **a**–**d** are from the Oxford RobotCar dataset [50], **e** is from EuroCity Persons dataset [51]

4 Fusion Method

4.1 Overall Architecture

A sharpening mixture of experts fusion network is proposed which builds upon the state-of-the-art real-time instance segmentation network YOLACT; the overall architecture is illustrated in Figure 5. It mainly consists of three parts: the frontend feature extraction networks that extract multi-scale high-level semantic information from each modality, the intermediate SMoE fusion network that adaptively weights and fuses features from expert networks and the final prediction head that

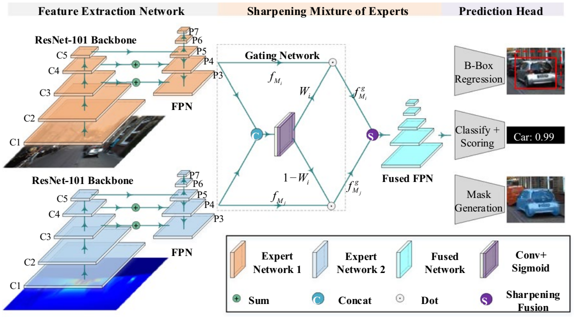

Figure 5 The overall architecture of the proposed SMoE model

learns complementary fused features to yield robust and accurate segmentation.

The ResNet-101 [52] and feature pyramid network (FPN) [53] are adopted as the frontend expert networks. ResNet-101 features a very deep convolutional network without facing degradation of the gradient and has achieved impressive performance in the ImageNet classification challenge, in our implementation, the last two fully connected layers of it are removed. Besides, the FPN is employed to deal with the significant object scale variation in the image.

Then the feature representations extracted from the FPN of each expert network are concatenated and used as the input of the gating network to learn the weights array assigning to the expert networks. The sharpening fusion layer strengthens the strong features and superimposes the weak features of the weighted output features of expert networks according to calculated thresholds. The prediction head layers are finally utilized to predict the instance segmentation.

4.2 Sharpening Mixture of Experts Fusion

4.2.1 Gating Network

The gating network is proposed to learn the contribution of individual modality M_i and M_j to the final instance segmentation, mapping the output feature representations of experts $f_{M_i} \in \mathrm{R}^{B\times C\times H\times W}$ and $f_{M_j} \in \mathrm{R}^{B\times C\times H\times W}$ to a probabilistic array $W \in \mathrm{R}^{B\times C\times H\times W}$, B, C, H and W denote the batch size, feature map channels, the height and the width of the training data. It is a compact network composed of three layers, i.e., a concatenation layer, a 1×1 convolution layer and a sigmoid layer. The output feature maps f_{M_i}, f_{M_j} of the expert networks are first concatenated together along channel C to an integrated probability map $P_{M_{ij}} \in \mathrm{R}^{B\times 2C\times H\times W}$, a 1×1 convolution layer with weights $Q \in \mathrm{R}^{B\times C\times 2C\times 1\times 1}$ is then employed to reduce the dimension of $P_{M_{ij}}$ and to weight the contribution of each modality. The output of the convolution layer $W \in \mathrm{R}^{B\times C\times H\times W}$ is normalized by the subsequent sigmoid layer, the above procedure can be written as:

$$W_i = \mathrm{sigmoid}\left(\sum \mathrm{concat}(f_{M_i}, f_{M_i}) \times Q\right), \tag{2}$$

The learned weights W_i are utilized to weight the contribution of RGB images and LiDAR dense depth images as follows:

$$\begin{cases} f^g_{M_i} = W_i * f_{M_i}, \\ f^g_{M_j} = (1 - W_i) * f_{M_j}. \end{cases} \tag{3}$$

where $*$ denotes Hadamard product.

4.2.2 Sharpening Fusion

The sharpening fusion network receives the weighted features $f^g_{M_i}$, $f^g_{M_j}$ from the gating network, then strengthens the strong features and superimposes the weak features according to the calculated thresholds, making full use of the complementary features to pursuit the boost in network performance.

First, the global threshold can be derived by

$$t = \mathrm{mean}\left(f^g_{M_i} + f^g_{M_j}\right). \tag{4}$$

Then the network calculates the larger value between the weighted features $f^g_{M_i}$ and $f^g_{M_j}$ in each spatial location, it is referred to as the pre-fusion features f_p. The strong features are defined as elements in the pre-fusion features that exceed the corresponding elements in the global threshold, otherwise, are the weak features. The strong features are enhanced via multiplying a gain factor a which is set to 2 in our implementation and the weak features are reinforced by adding the corresponding weighted features $f^g_{M_i}$ and $f^g_{M_j}$, the sharpening fusion process can be mathematically described as

$$f_{fused} = \begin{cases} a \cdot f_p, f_p > t, \\ f^g_{M_i} + f^g_{M_j}, f_p \le t. \end{cases} \tag{5}$$

4.3 Network Training

The proposed fusion network is implemented on the popular Pytorch framework. A two stage approach is utilized to train the model. In the first stage, the individual expert networks are trained to learn the semantic features of each modality in an end-to-end manner. The ResNet-101 backbone is pre-trained on ImageNet and the Xavier initialization [54] method is applied to the other layers. In the second training phase, the SMoE fusion network is fine-tuned with fixed weights of each expert network, which forces the gating network to focus on learning the complementary features extracted by the experts. The networks are trained for 80000 iterations with a batch size of 8 on one RTX 2080 Ti GPU by employing the ploy learning rate approach with an initial learning rate l_0 of 0.001, as follows

$$l_i = l_0 \cdot (1 - \frac{i}{i_{\max}})^{\gamma}, \tag{6}$$

where l_i is the current learning rate, i denotes the current iteration, $i_{\max}$ is the max iteration, γ is set to 0.9.

Stochastic gradient descent (SGD) is employed with a momentum of 0.9 and a weight decay of 5×10^4. The overall loss function L is a weighted sum of the classification loss L_{cls}; box regression loss L_{box} and mask loss L_{mask}, both L_{cls} and L_{box} are defined in the same way as in Ref. [55].

$$L = L_{cls}(c, c_{gt}) + \alpha L_{box}(b, b_{gt}) + \beta L_{mask}(m, m_{gt}). \tag{7}$$

The box regression loss L_{loc} is a Smooth L1 loss [6] between the predicted box b and the ground truth box b_{gt} parameters. The softmax cross entropy with c positive labels and 1 background label is used to train class prediction, selecting training examples using OHEM [56] with a 3:1 negative: positive ratio. To compute mask loss, the pixel-wise binary cross entropy between the predicted masks m and the ground truth masks m_{gt} is adopted. The data augmentations used in SSD are employed to train the network by randomly rotating, translating, skewing, scaling, vignetting, cropping, flipping color, modulating brightness and contrast, etc.

5 Experiments

This section presents a thorough comparison of the SMoE fusion framework using the RGB images and LiDAR dense depth images contained in the Complex KITTI dataset; different fusion architectures are compared and evaluated through instance segmentation tasks. The average precision (AP) at the different intersection over union (IoU) thresholds (AP, AP_{50}, AP_{75}) is employed as the metrics to evaluate the performance of instance segmentation, and only objects larger than 25 pixels in height are evaluated which follows the same principle with KITTI object detection benchmark.

5.1 Fusion Operation Comparison

To investigate the effectiveness of the proposed SMoE fusion operation, the network performance with serval different fusion operations, as well as the single modal scheme, are compared, and the performance of these models on the Complex KITTI dataset is shown in Table 1.

Table 1 Comparison of single modal and different fusion operations on the Complex KITTI dataset

Input	Approach	AP	AP_{50}	AP_{75}
RGB	Single modal	24.74	39.76	27.16
DDM	Single modal	24.39	41.42	27.27
RGB + DDM	Sum fusion [44]	25.22	43.09	27.85
RGB + DDM	Maximum fusion [57]	25.69	43.02	28.83
RGB + DDM	Sharpening fusion [19]	26.97	45.55	29.21
RGB + DDM	MoE + Sum fusion [12]	26.96	44.74	29.50
RGB + DDM	MoE + Maximum fusion	27.54	45.69	30.65
RGB + DDM	MoE + Sharpening fusion (ours)	29.76	49.36	32.07

The sum fusion in Ref. [44] directly adds the output features of expert networks, while the maximum fusion [57] chooses the maximum elements of the expert features at the same spatial position. The above two fusion operation and the sharpening fusion approach [19] all treat the expert networks of different modalities equally. As can be seen from Table 1, all fusion networks can effectively combine the complementary features of the RGB images and LiDAR dense depth map, achieving performance advancement than the single modality network. The sharpening fusion method outperforms the sum fusion and the maximum fusion approach, and even achieves close performance to the MoE+Sum fusion network, proving this fusion operation can fully utilize the complementary features of expert networks. The proposed SMoE approach yields the best performance for all fusion operations. This owes to the adaptive weighting mechanism of the gating network, making the framework can adequately assess the contribution of each expert network to the final segmentation result; meanwhile, the sharpening fusion operation is then used to learn the robust and accurate kernel from the weighted features.

5.2 Fusion Stage Comparison

To further verify the particular advantages of our SMoE fusion network, the performance of the fusion network at four different fusion stages, i.e., early fusion, middle fusion at backbone, middle fusion at FPN, late fusion, are compared, as depicted in Figure 6.

The RGB and DDM modalities are concatenated together in the early fusion method [43], the network

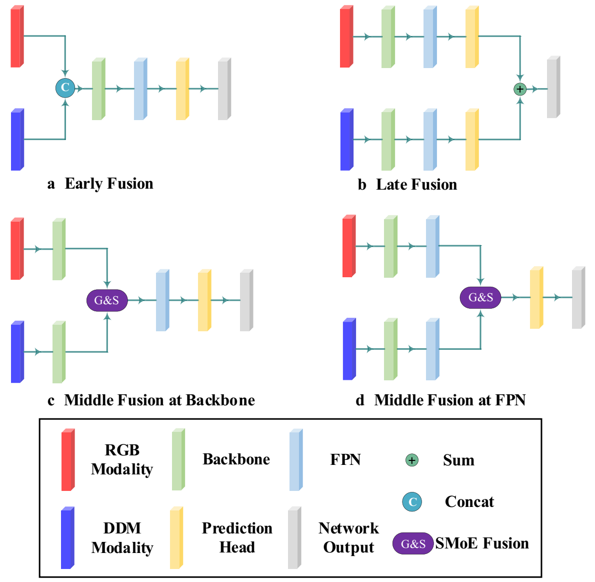

Figure 6 An illustration of different fusion stages

exploits the information of the raw data. Yet, this scheme does not perform better than the individual experts themselves as can be seen in Tables 1 and 2, this is primarily due to the inability to learn the joint RGB-D modality. The late fusion approach [44] directly integrates the predictions of each expert which means its performance is restricted by the expert networks, as it discards rich intermediate features which may bring benefits to the fusion network. The proposed SMoE fusion scheme is also implemented at backbone layers, i.e., middle fusion at backbone, and this type of fusion network has fewer parameters than the proposed model. As shown in Tables 1 2, middle fusion at backbone framework outperforms all other fusion techniques. However, it is inherently unable to learn the multi-scale feature representation of each expert from FPN layers, leading to a slight performance degradation compared with the proposed one.

Table 2 Comparison of different fusion stages on the complex KITTI dataset

Fusion Stage	AP	AP_{50}	AP_{75}
Early fusion [43]	16.69	30.30	17.96
Late fusion [44]	24.79	39.02	28.73
Middle fusion at backbone	28.21	46.86	30.75
Middle fusion at FPN (ours)	29.76	49.36	32.07

5.3 Qualitative Comparisons

To further evaluate the performance of the proposed SMoE fusion architecture, qualitative comparisons in instance segmentation are shown in Figure 7. In night scenes (a) and (d), the single RGB modality method fails to segment the objects in the shadows or under severe occlusion, whilst the DDM modality is less affected by illumination, motion blur or light rain, hence it outputs

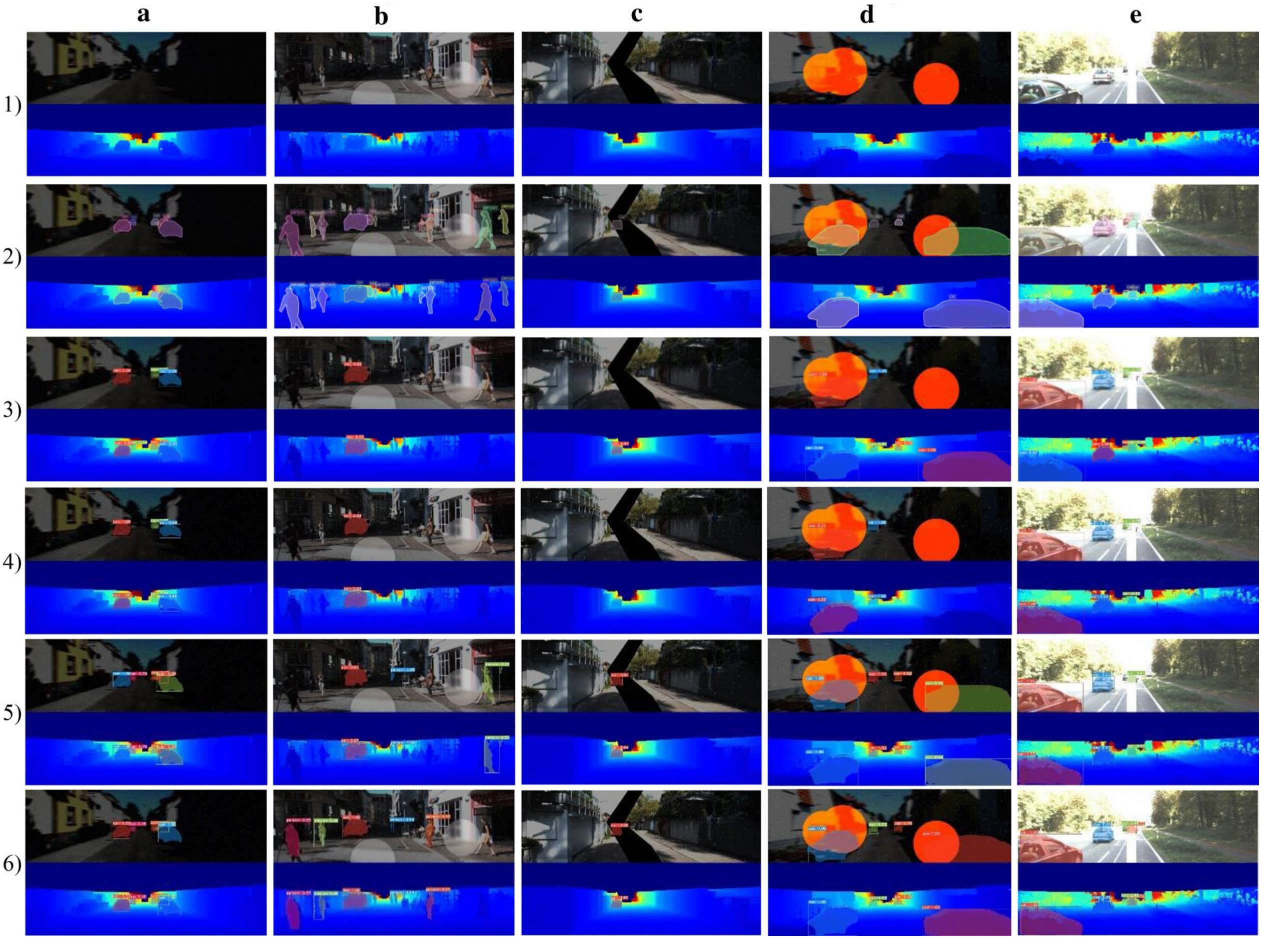

Figure 7 Qualitative comparison of segmentation in adverse conditions. **a** Night, **b** and **c** Light rain day, **d** Light rain night, **e** Glare; 1) Input images, 2) Ground truth, 3) Single modal, 4) Late fusion, 5) MoE+Maximum fusion, 6) SMoE fusion

better instance segmentation results than the single RGB modality approach. By effectively combining and learning the information of RGB and DDM modalities, the proposed SMoE fusion framework segments all instances correctly with high-quality masks. In light rain conditions (b), the camera is prone to raindrops which further scatter the light, leading to blurred images. The single modal approaches fail to segment persons in the image while the proposed SMoE fusion method segments the most instances, indicating our fusion method can learn robust information from the RGB and DDM modalities. In conditions (c) and (e), the distant cars are severely obscured by the front windshield wiper and glare, respectively, the single RGB modality method can not fix this issue. Comparatively, the proposed SMoE fusion network can accurately and robustly segment all instances using the integrated modalities. On the whole, the proposed SMoE fusion framework yields the best precision and mask quality in the above adverse conditions.

According to the above quantitative and qualitative results, the following conclusions can be drawn that the proposed sharpening mixture of experts fusion method can achieve robust and accurate performance in complex illumination and weather conditions, and is quite suitable for the perception system of CAEVs.

6 Conclusions

(1) The complex KITTI dataset is introduced which is annotated in instance-level by our proposed semi-automatic annotation procedure. This dataset consists of 7481 pairs of modified KITTI RGB images and the generated LiDAR dense depth maps.
(2) The SMoE fusion approach is proposed, it automatically learns an adaptive strategy for weighting the feature extraction networks and learning robust kernels from complementary modalities.
(3) In extensive experiments, the proposed SMoE framework outperforms instance segmentation using other fusion operation and stage techniques on the complex KITTI datasets, implying the robustness and accuracy of our approach in complex illumination and weather conditions.
(4) One direction to extend this work is to advance the performance of pedestrian detection of the proposed framework.

Acknowledgements
Not applicable.

Authors' contributions
GY provided fundamental ideas and all support conditions of this research. PP and KG were in charge of the trial and wrote the manuscript, they contributed equally to this work. YL and WZ assisted with building up the framework of the research. SL assisted in the experiments. All authors read and approved the final manuscript.

Authors' Information
Pai Peng, born in 1993, is currently a PhD candidate at *the School of Mechanical Engineering, Southeast University, Nanjing, China*. He received his M.S. degree from *Nanjing University of Science and Technology, China*, in 2019. His research interests include connected vehicles and deep learning.

Keke Geng, born in 1987, is an assistant professor at the *School of Mechanical Engineering, Southeast University, Nanjing, China*.

Guodong Yin, born in 1976, is currently a professor and a PhD candidate supervisor at the *School of Mechanical Engineering, Southeast University, Nanjing, China*.

Yanbo Lu, born in 1995, is currently a PhD candidate at *the School of Mechanical Engineering, Southeast University, Nanjing, China*.

Weichao Zhuang, born in 1990, is an assistant professor at *the School of Mechanical Engineering, Southeast University, Nanjing, China*.

Shuaipeng Liu, born in 1994, is currently a PhD candidate at *the School of Mechanical Engineering, Southeast University, Nanjing, China*.

Funding
Supported by National Natural Science Foundation of China (Grant Nos. 51975118, 52025121, 51975103, 51905095), National Natural Science Foundation of Jiangsu Province (Grant No. BK20180401).

Competing Interests
The authors declare no competing financial interests.

Received: 17 July 2020 Revised: 25 July 2021 Accepted: 6 August 2021
Published online: 30 August 2021

References
[1] Y Jiang, X Zhao, J Gong, et al. System design of self-driving in simplified urban environments. *Journal of Mechanical Engineering*, 2012, 48(20): 103–112. (in Chinese)
[2] J G Ibanez, S Zeadally, J Contreras-Castillo. Integration challenges of intelligent transportation systems with connected vehicle, cloud computing and internet of things technologies. *IEEE Wireless Communications*, 2015, 6(22): 122–128.
[3] X Tang, T Jia, X Hu, et al. Naturalistic data-driven predictive energy management for plug-in hybrid electric vehicles. *IEEE Transactions on Transportation Electrification*, 2021, 7(2): 497–508.
[4] F Rosique, P J Navarro, C Fernández, et al. A systematic review of perception system and simulators for autonomous vehicles research. *Sensors*, 2019, 19(3): 648.
[5] F Lin, Y Zhang, Y Zhao, et al. Trajectory tracking of autonomous vehicle with the fusion of dyc and longitudinal–lateral control. *Chinese Journal of Mechanical Engineering*, 2019, 32: 1–16.
[6] R Girshick. Fast R-CNN. *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, Santiago, Chile, December 7–13, 2015: 1440–1448.
[7] Y F Cai, H Wang, X Chen, et al. Vehicle detection based on visual saliency and deep sparse convolution hierarchical model. *Chinese Journal of Mechanical Engineering*, 2016, 29(4): 765–772.
[8] S Q Ren, K M He, R Girshick, et al. Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 2017, 39(6): 1137–1149.
[9] K M He, G Gkioxari, P Doll´ar, et al. Mask R-CNN. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 2020, 42(2): 386–397.
[10] L Hu, J Ou, J Huang, et al. A review of research on traffic conflicts based on intelligent vehicles. *IEEE Access*, 2020, 8: 24471–24483.
[11] K K Geng, W Zou, G D Yin, et al. Low-observable targets detection for autonomous vehicles based on dual-modal sensor fusion with deep learning approach. *Proceedings of the Institution of Mechanical Engineers Part D: Journal of Automobile Engineering*, 2019, 233(9): 2270–2283.

[12] O Mees, A Eitel, W Burgard. Choosing smartly: adaptive multimodal fusion for object detection in changing environments. *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Daejeon, South Korea, October 9–14, 2016: 151–156.

[13] X Chen, H Ma, J Wan, et al. Multi-view 3D object detection network for autonomous driving. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, USA, July 21-26, 2017: 6526–6534.

[14] A Geiger, P Lenz, R Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* Providence, Rhode Island, June 16–21, 2012: 3354–3361.

[15] P Wang, X Huang, X Cheng, et al. The apolloscape open dataset for autonomous driving and its application. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 2019, 42(10): 2702–2719.

[16] J Xue, J Fang, T Li, et al. Blvd: building a large-scale 5D semantics benchmark for autonomous driving. *International Conference on Robotics and Automation (ICRA)*, Montreal, Canada, May 20–24, 2019: 6685–6691.

[17] H Caesar, V Bankiti, A H Lang, et al. NuScenes: a multimodal dataset for autonomous driving. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*, Seattle, USA, June 16–18, 2020: 11621–11631.

[18] A Patil, S Malla, H Gang, et al. The h3d dataset for full-surround 3d multi-object detection and tracking in crowded urban scenes. *International Conference on Robotics and Automation (ICRA)*, Montreal, Canada, May 20–24, 2019: 9552–9557.

[19] J R Tong, L Mao, J Sun. Multimodal pedestrian detection algorithm based on fusion feature pyramids. *Computer Engineering and Applications*, 2019, 55(19): 214–222.

[20] D Bolya, C Zhou, F Xiao, et al. Yolact: real-time instance segmentation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, Long Beach, USA, June 16–20, 2019: 9157–9166.

[21] T Y Lin, M Maire, S Belongie, et al. Microsoft coco: common objects in context. *European Conference on Computer Vision(ECCV)*, Zürich, Switzerland, September 6–12, 2014: 740–755.

[22] J Xu, A G Schwing, R Urtasun. Learning to segment under various forms of weak supervision. *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, Santiago, Chile, December 7–13, 2015: 3781–3790.

[23] D Lin, J Dai, J Jia, et al. Scribblesup: scribble-supervised convolutional networks for semantic segmentation. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* Las Vegas, USA, June 26–July 1, 2016: 3159–3167.

[24] N Xu, B Price, S Cohen, et al. Deep interactive object selection. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* Las Vegas, USA, June 26–July 1, 2016: 373–381.

[25] A Bearman, O Russakovsky, V Ferrari, et al. What's the point: semantic segmentation with point supervision. *European conference on computer vision(ECCV)*, Amsterdam, Netherlands, October 8–16, 2016: 549–565.

[26] L C Chen, S Fidler, A L Yuille, et al. Beat the mturkers: automatic image labeling from weak 3d supervision. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Columbus, USA, June 24–27, 2014: 3198–3205.

[27] Z Zhang, A G Schwing, S Fidler, et al. Monocular object instance segmentation and depth ordering with CNNs. *IEEE International Conference on Computer Vision (ICCV)*, Santiago, Chile, December 7–13, 2015: 2614–2622.

[28] Z Zhang, S Fidler, R Urtasun. Instance-level segmentation for autonomous driving with deep densely connected MRFs. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* Las Vegas, USA, June 26–July 1, 2016: 669–677.

[29] L Castrejon, K Kundu, R Urtasun, et al. Annotating object instances with a polygon-RNN. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, USA, July 21–26, 2017: 4485–4493.

[30] D Acuna, H Ling, A Kar, et al. Efficient interactive annotation of segmentation datasets with polygon-RNN++. *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, Salt Lake City, USA, June 18–23, 2018: 859–868.

[31] M Andriluka, J R R Uijlings, V Ferrari. Fluid annotation: a human-machine collaboration interface for full image annotation. *Proceedings of the 26th ACM International Conference on Multimedia*, Seoul, South Korea, October 22–26, 2018: 1957–1966.

[32] P Voigtlaender, M Krause, A Osep, et al. Mots: multi-object tracking and segmentation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, Long Beach, USA, June 16–20, 2019: 7942–7951.

[33] Z Zhou, M Dong, X Xie, et al. Fusion of infrared and visible images for night-vision context enhancement. *Applied Optics*, 2016, 55(23): 6480–6490.

[34] Q Ha, K Watanabe, T Karasawa, et al. Mfnet: towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes. *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Vancouver, Canada, September 24–28, 2017: 5108–5115.

[35] A Valada, R Mohan, W Burgard. Self-supervised model adaptation for multimodal semantic segmentation. *International Journal of Computer Vision*, 2020, 128(5): 1239–1285.

[36] D Xu, D Anguelov, A Jain. Pointfusion: deep sensor fusion for 3D bounding box estimation. *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, Salt Lake City, USA, June 18–23, 2018: 244–253.

[37] K Shin, Y P Kwon, M Tomizuka. Roarnet: a robust 3D object detection based on region approximation refinement. *IEEE Intelligent Vehicles Symposium (IV)*, Paris, France, June 9–12, 2019: 2510–2515.

[38] C R Qi, H Su, K Mo, et al. Pointnet: deep learning on point sets for 3d classification and segmentation. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, USA, July 21–26, 2017: 77–85.

[39] J Ku, M Mozifian, J Lee, et al. Joint 3D proposal generation and object detection from view aggregation. *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Madrid, Spain, October 1–5, 2018: 1–8.

[40] Z Wang, W Zhan, M Tomizuka. Fusing bird's eye view lidar point cloud and front view camera image for 3D object detection. *IEEE Intelligent Vehicles Symposium (IV)*, Changshu, China, June 26–30, 2018: 1–6.

[41] A Asvadi, L Garrote, C Premebida, et al. Multimodal vehicle detection: fusing 3D-lidar and color camera data. *Pattern Recognition Letters*, 2017, 115: 20–29.

[42] A Asvadi, L Garrote, C Premebida, et al. Depthcn: vehicle detection using 3D-lidar and convent. *Proceedings of IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)*, Yokohama, Japan. October 16–19, 2017: 1–6.

[43] C Couprie, C Farabet, L Najman, et al. Indoor semantic segmentation using depth information. *International Conference on Learning Representations (ICLR)*, Scottsdale, USA, May 2–4, 2013: 1–8.

[44] J Long, E Shelhamer, T Darrell. Fully convolutional networks for semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2015, 39(4): 640–651.

[45] D Guan, Y Cao, J Yang, et al. Fusion of multispectral data through illumination-aware deep neural networks for pedestrian detection. *Information Fusion*, 2019, 50: 148–157.

[46] A Valada, J Vertens, A Dhall, et al. Adapnet: adaptive semantic segmentation in adverse environmental conditions. *IEEE International Conference on Robotics and Automation (ICRA)*, Singapore, May 29–June 3, 2017: 4644–4651.

[47] Y Cheng, R Cai, Z Li, et al. Locality-sensitive deconvolution networks with gated fusion for RGB-d indoor semantic segmentation. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, USA, July 21–26, 2017: 1475–1483.

[48] M Cordts, M Omran, S Ramos, et al. The cityscapes dataset for semantic urban scene understanding. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* Las Vegas, USA, June 26–July 1, 2016: 3213–3223.

[49] A Asvadi, L Garrote, C Premebida, et al. Real-time deep convnet-based vehicle detection using 3d-lidar reflection intensity data. *Robot 2017: Third Iberian Robotics Conference*, Seville, Spain, November 22–24, 2017: 475–486.

[50] W Maddern, G Pascoe, C Linegar, et al. 1 year, 1000 km: the oxford robotcar dataset. *The International Journal of Robotics Research*, 2017, 36(1): 3–15.

[51] M Braun, S Krebs, F Flohr, et al. Eurocity persons: a novel benchmark for person detection in traffic scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019, 41(8): 1844–1861.

[52] K M He, X Zhang, S Q Ren, et al. Deep residual learning for image recognition. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* Las Vegas, USA, June 26–July 1, 2016: 770–778.

[53] T Y Lin, P Doll´ar, R Girshick, et al. Feature pyramid networks for object detection. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, USA, July 21–26, 2017: 936–944.

[54] X Glorot, Y Bengio. Understanding the difficulty of training deep feedforward neural networks. *Journal of Machine Learning Research*, 2010, 9: 249–256.

[55] W Liu, D Anguelov, D Erhan, et al. Ssd: single shot multibox detector. *European Conference on Computer Vision(ECCV)*, Amsterdam, Netherlands, October 8–16, 2016: 21–37.

[56] A Shrivastava, A Gupta, R Girshick. Training region-based object detectors with online hard example mining. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* Las Vegas, USA, June 26–July 1, 2016: 761–769.

[57] O Prakash, A Kumar, A Khare. Pixel-level image fusion scheme based on steerable pyramid wavelet transform using absolute maximum selection fusion rule. *International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT)*, Kochi, India, December 3–5, 2014: 765–770.