Computers, Materials & Continua

Tech Science Press

Doi:10.32604/cmc.2025.064036

ARTICLE

Research on Vehicle Safety Based on Multi-Sensor Feature Fusion for Autonomous Driving Task

Yang Su[1,*], **Xianrang Shi**[1] **and Tinglun Song**[2]

[1]College of Energy and Power Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing, 210016, China
[2]Institute of Advanced Technology, Chery Automobile Co., Ltd., Wuhu, 241009, China

*Corresponding Author: Yang Su. Email: suyang@nuaa.edu.cn

Received: 02 February 2025; Accepted: 28 April 2025; Published: 19 May 2025

ABSTRACT: Ensuring that autonomous vehicles maintain high precision and rapid response capabilities in complex and dynamic driving environments is a critical challenge in the field of autonomous driving. This study aims to enhance the learning efficiency of multi-sensor feature fusion in autonomous driving tasks, thereby improving the safety and responsiveness of the system. To achieve this goal, we propose an innovative multi-sensor feature fusion model that integrates three distinct modalities: visual, radar, and lidar data. The model optimizes the feature fusion process through the introduction of two novel mechanisms: Sparse Channel Pooling (SCP) and Residual Triplet-Attention (RTA). Firstly, the SCP mechanism enables the model to adaptively filter out salient feature channels while eliminating the interference of redundant features. This enhances the model's emphasis on critical features essential for decision-making and strengthens its robustness to environmental variability. Secondly, the RTA mechanism addresses the issue of feature misalignment across different modalities by effectively aligning key cross-modal features. This alignment reduces the computational overhead associated with redundant features and enhances the overall efficiency of the system. Furthermore, this study incorporates a reinforcement learning module designed to optimize strategies within a continuous action space. By integrating this module with the feature fusion learning process, the entire system is capable of learning efficient driving strategies in an end-to-end manner within the CARLA autonomous driving simulator. Experimental results demonstrate that the proposed model significantly enhances the perception and decision-making accuracy of the autonomous driving system in complex traffic scenarios while maintaining real-time responsiveness. This work provides a novel perspective and technical pathway for the application of multi-sensor data fusion in autonomous driving.

KEYWORDS: Multi-sensor fusion; autonomous driving; feature selection; attention mechanism; reinforcement learning

1 Introduction

With the rapid development of autonomous driving technology, the demands for safety and reliability have been continuously increasing [1]. Traditional driving methods are fraught with limitations, such as driver fatigue and distraction, which can lead to traffic accidents. Early autonomous driving systems typically relied on single sensors (e.g., cameras or radar); however, their performance and reliability significantly deteriorated when confronted with complex and dynamic traffic environments, especially under adverse weather conditions. Currently, the integration of multiple sensors has become a mainstream trend. By combining various sensors, such as LiDAR, cameras, and millimeter-wave radar, vehicles can achieve a more comprehensive perception of their surrounding environment. For instance, the integration of

3D point cloud data from LiDAR and 2D images from cameras can significantly enhance the detection accuracy of pedestrians and obstacles. Moreover, advancements in deep learning technology have driven the optimization of object detection algorithms, making them more robust in complex scenarios [2]. In terms of decision-making, end-to-end autonomous driving technology is gradually maturing. For example, Tesla's FSD V12 system directly extracts information from raw sensor data using deep learning models, achieving seamless integration from perception to control, which greatly improves the efficiency and safety of autonomous driving. Furthermore, the application of reinforcement learning in autonomous driving has also made significant progress, optimizing the vehicle's decision-making capabilities in complex environments. Autonomous driving technology is progressively being implemented in various scenarios. For instance, Robotaxi and autonomous trucks have achieved significant results in testing and operations in specific areas. Companies such as Baidu and Didi have launched pilot projects for autonomous taxis in multiple cities in China. Meanwhile, autonomous trucks in the logistics sector have reduced operating costs through precise route planning and stable speed control [3].

Therefore, utilizing multi-sensor information for environmental perception and decision-making has become a key factor in enhancing the performance of autonomous driving systems [4]. Multi-sensor fusion integrates data from various types of sensors, providing more comprehensive and accurate environmental information, thereby strengthening the perception capabilities of autonomous driving systems [5,6]. Despite the progress made, the performance of autonomous driving systems in adverse weather conditions (such as rain, snow, and fog) and complex traffic scenarios remains unsatisfactory [7]. For instance, the performance of LiDAR deteriorates in rainy and foggy environments, while cameras exhibit insufficient sensitivity under low-light conditions, which limits the system's reliability. Although multi-sensor fusion has improved perception capabilities, it has also introduced complexities in data processing and increased computational burden [8]. Feature fusion learning often encounters issues of feature redundancy, which can lead to a decrease in system response speed. Furthermore, continuous action space algorithms based on reinforcement learning still face challenges concerning algorithm stability and generalization performance [9].

To address these challenges, this paper proposes a feature fusion learning and reinforcement learning method based on three modalities. We introduce Sparse Channel Pooling (SCP) and Residual Triplet-Attention (RTA) mechanisms to emphasize key features, reduce unnecessary information, and enhance the robustness and response speed of the system. Specifically, through the SCP mechanism, the model can adaptively filter out essential feature channels, minimizing redundant information and highlighting critical features. This process improves the performance of the autonomous driving system, particularly when processing large volumes of sensor data. Furthermore, the RTA mechanism facilitates effective alignment between features from different modalities, thereby enhancing the effectiveness of feature fusion. This cross-modal feature alignment not only strengthens the system's ability to integrate multi-sensor information but also mitigates the influence of irrelevant features by selectively utilizing different modal feature vectors.

In addition, we introduce a reinforcement learning module designed for learning and optimizing strategies in a continuous action space. This module is integrated with feature fusion learning to enable the entire system to learn efficient driving strategies in an end-to-end manner within the CARLA autonomous driving simulator. This approach aims to enhance the stability and generalization ability of autonomous driving systems.

The innovations of this study include:

(1) A feature fusion learning and reinforcement learning method based on three modalities is proposed to enhance the decision-making capabilities and response speed of the autonomous driving system. This approach effectively addresses the issues of cross-modal feature alignment and redundancy.

(2) SCP technology is introduced to improve the performance of the autonomous driving system by adaptively filtering important feature channels. This method reduces unnecessary features, highlights key elements, and alleviates the problems of feature redundancy and computational burden, significantly enhancing the model's robustness and real-time performance in complex environments.

(3) By sequentially utilizing feature vectors from different modalities, the influence of irrelevant features is diminished, thereby improving the system's ability to integrate multi-sensor information. This enhancement also boosts the effectiveness of reinforcement learning, enabling the system to perceive its environment and make decisions more efficiently and accurately. The combination of the reinforcement learning module and feature fusion learning facilitates end-to-end autonomous driving strategy learning, providing a novel perspective for the design of autonomous driving systems.

2 Related Work

2.1 Autonomous Driving Technology Based on Traditional Methods

In the early development of autonomous driving technology, most studies relied on traditional Computer Vision (CV) and machine learning methods. These approaches typically involved manual feature extraction and rule-based systems for tasks such as vehicle detection, lane recognition, and traffic sign recognition. Berkaya et al. [10] proposed a circular detection algorithm and an RGB-based color thresholding technique that leverages the color and shape characteristics of traffic signs. Nie et al. [11] developed a decision behavior model for free lane change execution. By comparing the lane change decision model based on Support Vector Machines (SVM) with the NAGLEA model, they demonstrated the superiority of the SVM model in predicting lane change behavior, achieving a success rate of 95%, which verifies the algorithm's effectiveness. On the other hand, Schubert et al. [12] utilized the vehicle's environmental perception data to create a Bayesian network model for developing lane change strategies. This model provides a decision-making framework based on probabilistic analysis by fusing surrounding information to facilitate lane changes. Despite achieving remarkable results during a specific period, traditional methods are limited by a series of drawbacks, including restricted generalization ability, insufficient real-time performance, poor robustness, and limited integration capabilities. These limitations become particularly evident when facing changing environmental conditions, real-time response requirements, and sensor noise.

2.2 Autonomous Driving Technology Based on Deep Learning

As deep learning has made significant strides in the fields of visual recognition and speech processing, it has also been effectively applied to tackle challenges in autonomous driving, thus overcoming some limitations of traditional methods [13]. Deep learning techniques, particularly Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN), are widely utilized in tasks related to vehicle environment perception, behavior prediction, and decision-making, resulting in substantial improvements in the performance of autonomous driving systems. For instance, Liu et al. [14] enhanced recognition capabilities by constructing a feature cone neural network based on bipartite attention, achieving effective recognition of small objects in images with a correctness rate of 93.3% on the TT100k dataset. Nacir et al. [15] proposed a transfer learning method to train the YOLOv5 model. After training YOLOv5 on various image datasets, the model's initial parameters were transferred to the German traffic sign dataset, thereby improving accuracy while maintaining detection speed. Additionally, Liu et al. [8] encode image features and radar features separately, which are then fused through a shared network, significantly enhancing both detection efficiency and accuracy. To tackle the challenges of speed and accuracy in autonomous driving target detection, Cao et al. [16] developed the MCS-YOLO algorithm. This algorithm incorporates a coordinate attention module into the backbone network to optimize feature aggregation, utilizes a multi-scale small target structure to enhance small target detection, and combines the Swin Transformer with CNN to improve context information processing. Experiments have validated the effectiveness and advantages of MCS-YOLO in enhancing detection performance.

However, existing systems still face limitations when dealing with complex scenarios and overcoming sensor noise. To enhance the quality of feature representation, model networks often adopt complex structures. However, these structures can be negatively impacted by factors such as slow updates of map data and environmental changes, resulting in insufficient accuracy and stability in positioning and navigation. In contrast, our proposed multi-modal feature fusion autonomous driving system utilizes RGB images, depth images, and point cloud data to adaptively filter important feature channels through SCP and RTA mechanisms. This approach enables effective cross-modal feature alignment, significantly reducing feature redundancy while ensuring that key features take precedence. Combined with reinforcement learning, our system's performance in environmental perception, positioning and navigation, control, and driving strategy learning has been further enhanced.

3 Method

The purpose of this study is to explore the interaction modes of agents in complex environments and to obtain data and key reward information through interactions with the environment. These data are processed by a multi-sensor feature fusion module to extract representative features. Subsequently, the reinforcement learning module utilizes these extracted features to learn the control strategy for autonomous driving, enabling the direct generation of actions for the autonomous vehicle [17]. This research method provides new insights and technical support for the development of intelligent transportation systems. The overall algorithm framework is illustrated in Fig. 1.

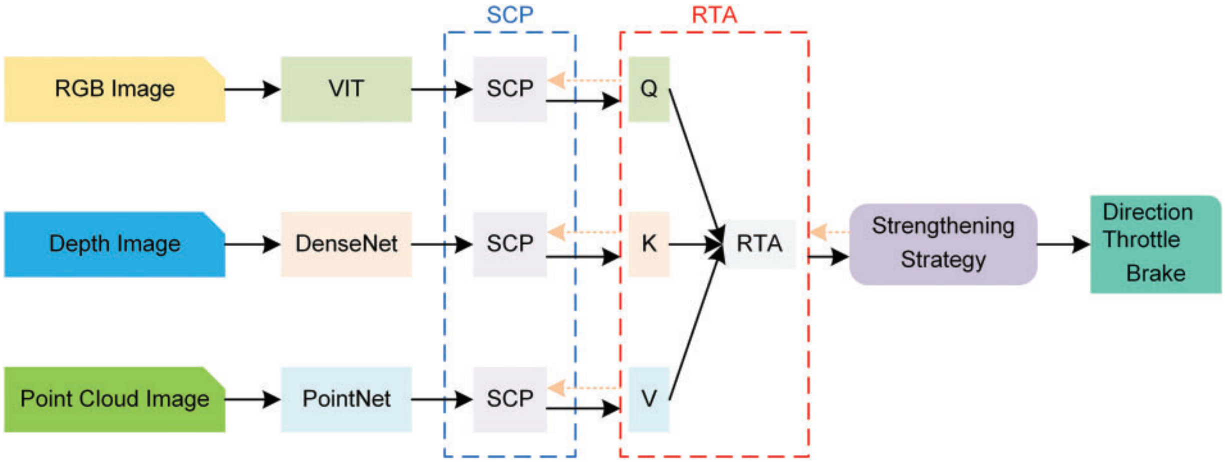

Figure 1: The overall algorithm framework diagram

3.1 Multi-Sensor Feature Fusion Module

This section addresses the design and implementation of the multi-sensor feature fusion module. The module comprises an RGB image feature extraction network based on Vision Transformer (ViT), a depth map feature extraction network based on Dense Convolutional Network (DenseNet), and a feature extraction network for LiDAR point cloud data based on PointNet++ [18,19]. By combining and coordinating these networks, the module achieves the extraction and fusion of features from different sensor data, facilitating subsequent reinforcement learning.

3.1.1 RGB Images

The module discussed in this article corresponds to the Encoder component of the Transformer, specifically referring to the ViT architecture [20]. The detailed network structure is presented in Fig. 2.

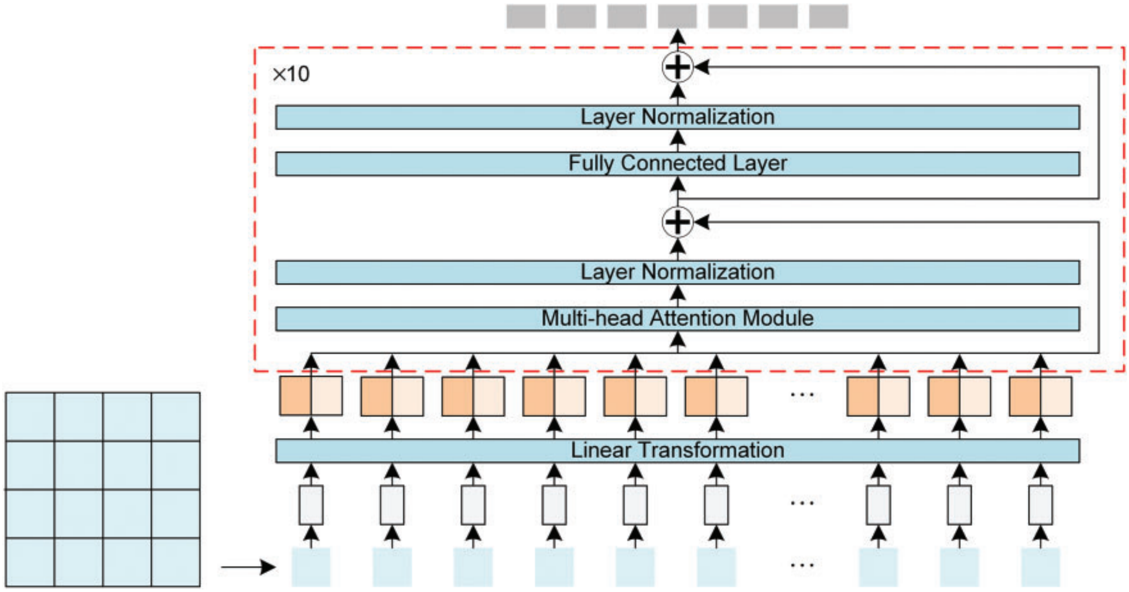

Figure 2: RGB image feature extraction algorithm based on ViT

In this study, the RGB input image, with a size of 320 × 320 pixels, is divided into 100 small blocks, each measuring 32 × 32 pixels. Each block is then flattened and transformed into a one-dimensional vector. Given the high dimensionality of these vectors, we apply a linear transformation layer to reduce their dimensions, similar to the word embedding process used in Natural Language Processing (NLP), where each transformed vector represents a token. To preserve the spatial relationships between image blocks, we add a position encoding to each token. Additionally, we introduce a class token to aggregate information from all image blocks, completing the preprocessing of the input.

The sequence of 100 tokens, after preprocessing, is passed to the decoder, which consists of 10 modules with the same structure but independent parameters. Each module allows the tokens in the sequence to exchange information with each other through the self-attention mechanism. Since each sub-image block contains different information, such as lane lines in some blocks and the sky in others, the attention mechanism learns to associate related image blocks effectively.

After the multi-head attention operation, the module performs normalization and establishes a residual connection with the original input. The fully connected layer then restores the token dimensions, normalizes the output once more, and creates another residual connection to form the final module output. This process is repeated across all 10 independent modules. The final output is the class-token vector, which can be obtained by taking the mean of all token dimensions or directly using the class-token vector learned by the network. This vector serves as the feature vector extracted from the depth map.

3.1.2 Depth Images

Since the VGG network [21] achieved remarkable results in the field of image recognition in 2014, CNNs have become a cornerstone in the field of CV. Building on the VGG architecture, researchers have proposed various innovative variants and derivative structures [22,23]. For instance, ResNet introduces residual connections, while mobile networks utilize depthwise separable convolutions to reduce model parameters. This study also applies deep CNNs for image feature extraction from depth maps, with the primary model structure based on DenseNet [24]. The specific network architecture is shown in Fig. 3.

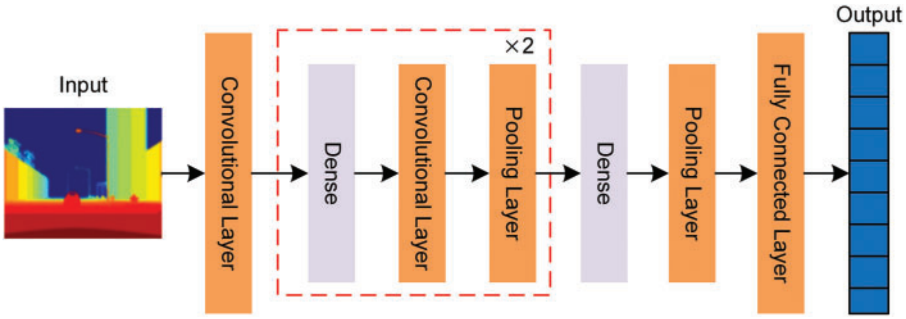

Figure 3: Depth image feature extraction algorithm based on DenseNet

The Residual Network (ResNet) assumes that if a deep network can learn identity mapping by adding several layers, its performance after training will not be worse than that of a shallower network. In contrast, DenseNet is built on the assumption that reusing features can enhance network performance [18]. The innovation of DenseNet lies in its direct connection approach, which differs from the ResNet's additive strategy. DenseNet enables direct transmission between input and output, bypassing nonlinear transformations.

In each Dense module, a dense connectivity strategy is employed, meaning that each layer receives the outputs of all previous layers as its input. The limited adaptability of a single Dense module to resolution variations poses significant challenges when handling multiple feature merges. As a result, DenseNet typically consists of multiple such modules to maintain both the functionality and flexibility of the network.

When processing the depth image, the input image is first down-sampled through a convolutional layer to reduce the complexity of subsequent processing. The image features are then sequentially passed through three Dense modules, each performing a dense connection operation to ensure effective circulation and reuse of features. The resulting feature map undergoes pooling, flattening, and fully connected layer processing. Through these steps, nonlinear dimensionality reduction is achieved, generating feature vectors that characterize the depth image's attributes.

3.1.3 Point Cloud Network

In three-dimensional spatial analysis, the point cloud consists of sets of three-dimensional points defined by $N\left(x, y, z\right)$ coordinates, valued for its compactness and high-fidelity representation of object details. The algorithm design must consider two key attributes: position invariance and rotation invariance. Position invariance means that the arrangement of points does not affect the content of the point cloud information. Even after reordering, the same object or scene is still represented. Rotation invariance implies that although all points undergo the same rotation and their coordinates change, the described object or scene remains unchanged.

PointNet achieves position invariance by mapping the points to high-dimensional space and applying maximum pooling [25]. Combined with the T-Net module, PointNet rotates the point cloud to ensure the data is processed under a unified reference frame, yielding a 1024-dimensional global feature vector to ensure rotation invariance [26]. However, PointNet may not fully capture the relationship between points, leading to the loss of local features.

In this study, the point cloud feature extraction network is primarily based on PointNet++, an improved version of PointNet. The network structure is shown in Fig. 4. By hierarchically processing the point cloud data, this network captures features at different scales, effectively overcoming the limitations of PointNet in capturing local details.

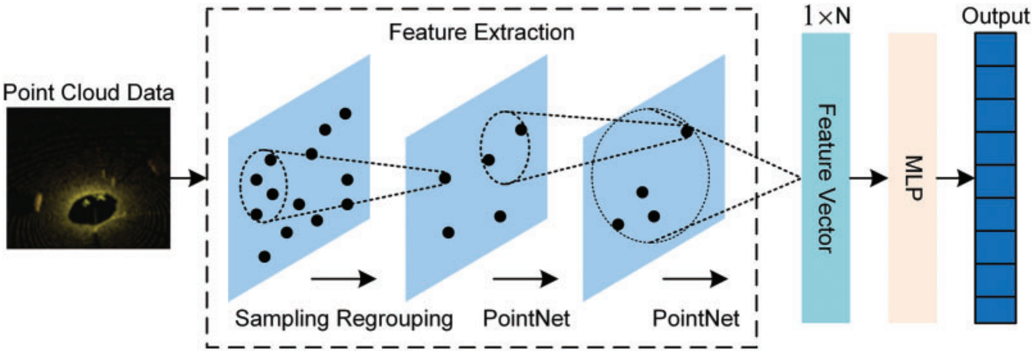

Figure 4: Feature extraction algorithm of point cloud data based on PointNet++

The introduction of PointNet++ aims to address the limitations of PointNet, with its primary enhancement being the ability to extract local features [19]. This method partitions the point cloud based on distance measurements, extracting features from each region individually, and progressively expands the range of feature extraction. This approach is similar to the layer-by-layer expansion of receptive fields in convolutional networks. By hierarchically extracting local features, they are eventually aggregated and passed to the fully connected layer for classification or other tasks. In the network architecture of this paper, the feature extraction part of PointNet++ is used to generate features from point cloud data. The input point cloud is first partitioned according to spatial distance, then sampled and reorganized, after which PointNet handles the feature extraction [26]. The final feature vector is transformed to the desired dimension using a multilayer perceptron (MLP).

A key detail when processing the input data is that the number of points, N, in the point cloud data, varies at different stages of the experiment, meaning the number of points is not fixed. If the number of points exceeds the network's requirements, random sampling is applied to reduce the count. If there are insufficient points, the required number is achieved through repeated sampling.

3.1.4 Sparse Channel Pooling Mechanism

In this section, inspired by graph pooling [27], we introduce the implementation of SCP technology, as shown in Fig. 5. SCP plays a crucial role in our model by reducing data redundancy and computational burden, selectively emphasizing the most informative feature channels.

Firstly, we use two 1 × 1 convolution kernels α and β to linearly transform the feature vector $f \in \mathfrak{R}^{C\times 1}$ output by the backbone network to obtain two different feature vectors f_{α} and f_{β}:

$$f_{\alpha} = a\left(f\right) \tag{1}$$

$$f_{\beta} = \beta\left(f\right) \tag{2}$$

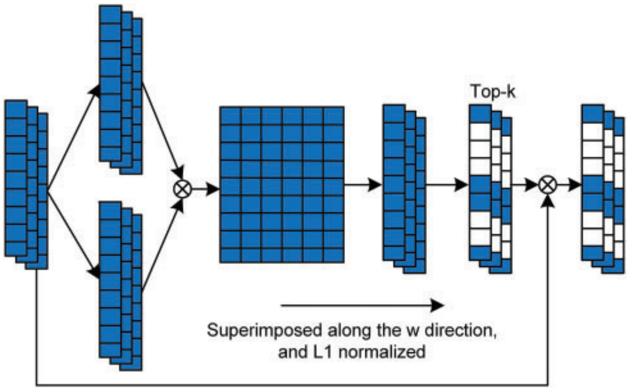

Figure 5: Sparse channel pooling structure

The purpose of these two convolution kernels is to map the original feature vectors to two different spaces so that the importance of each feature channel can be better evaluated in the next steps.

Next, we compute the outer product of eigenvectors f_α and f_β to obtain an importance matrix I:

$$I = f_a f_b^T \tag{3}$$

Each element I_{ij} of the importance matrix I represents the interactive importance between the i-th feature channel and the j-th feature channel.

Then, we sum up the importance matrix I to obtain a degree vector I_D, which represents the importance of each feature channel:

$$I_D = \sum_j I \tag{4}$$

In the degree vector I_D, we select the first k feature channels with the maximum value. These selected feature channels are considered to be the most important for model decision-making. We preserve these feature channels by:

$$f' = f \,\Theta\, I_D\,[:k] \tag{5}$$

here, Θ denotes the multiplication operation by elements, and $I_D\,[:k]$ denotes the first k largest elements in the degree vector I_D.

Through this method, we can adaptively select the feature channels that have the greatest impact on the model performance, reduce the computational burden of the model, and improve the decision accuracy of the model. This SCP technology provides an effective feature selection mechanism for multi-sensor data fusion, which helps to improve the performance of autonomous driving systems [28,29].

3.1.5 Residual Triplet-Attention Mechanism

Inspired by the self-attention mechanism [30], this paper introduces the RTA mechanism, as shown in Fig. 6. RTA is designed to perform deeper feature fusion using the features extracted by the backbone network (such as ViT, DenseNet, and PointNet++). By incorporating an attention mechanism, this approach highlights important feature channels across different modalities while suppressing irrelevant information.

This strategy enables the model to focus on critical information, reducing the computational burden caused by redundant features.

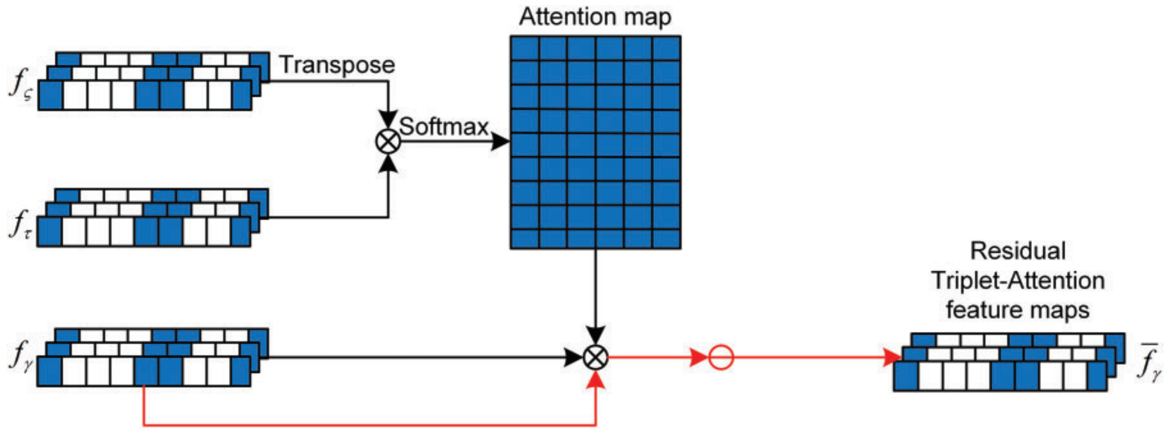

Figure 6: Residual Triplet-Attention structure diagram

Firstly, three different 1 × 1 convolution kernels ς, τ and γ are linearly transformed by the backbone network (ViT, DenseNet and PointNet++) and the SCP output $f' \in \mathfrak{R}^{Cx1}$ to obtain three different modal feature vectors f_ς, f_τ and f_γ. The expression is as follows:

$$f_\varsigma = \varsigma(f) \tag{6}$$

$$f_\tau = \tau(f) \tag{7}$$

$$f_\gamma = \gamma(f) \tag{8}$$

These feature vectors maintain the same feature dimension C. In order to align the key features between different modalities, RTA uses an attention mechanism, in which the feature vectors take turns to play the roles of Query, Key and Value. Specifically, when f_γ is a Query, the attention value A_γ can be calculated as follows:

$$A_\gamma = softmax\left(\frac{f_\varsigma \cdot f_\tau}{\sqrt{C}}\right) f_\gamma \tag{9}$$

Among them, $f_\varsigma \cdot f/\sqrt{C}$ is the attention map, which represents the correlation between different feature channels. The Softmax function ensures that the attention of each feature channel is between 0 and 1, thereby obtaining a weighted feature vector A_γ. In order to highlight the importance of key channels and reduce the computational burden of unnecessary features, we subtract the input Value feature f_γ from the attention value A_γ to obtain the sparse key feature $\overline{f_\gamma}$. The detailed formula is as follows:

$$\overline{f}_\gamma = A_\gamma - f_\gamma \tag{10}$$

This process only considers the case of f_γ as a Value feature. Similarly, we also need to calculate the case where f_ς and f_τ are used as Value features, respectively, to obtain $\overline{f_\varsigma}$ and $\overline{f_\tau}$. Finally, we concatenate the sparse key features $\overline{f_\gamma}$, $\overline{f_\varsigma}$ and $\overline{f_\tau}$ of these three different modalities as the input of reinforcement learning.

$$F = Concat(\overline{f}_\gamma, \overline{f}_\varsigma, \overline{f}_\tau) \tag{11}$$

In this way, RTA not only enhances the expression of important features, but also reduces the computational complexity by eliminating unimportant features, thereby improving the overall efficiency and performance of the model.

3.2 Reinforcement Learning Module

The randomness strategy adopted in this study is specifically implemented through the SAC algorithm, which has shown strong performance in the field of autonomous driving [31]. The innovation of the entire training framework lies in the combination of pre-training and offline training. This unique training approach begins by pre-training the model using a public dataset, which significantly enhances the model's ability to extract state information. This not only improves the model's understanding of the environment but also reduces the instability that may arise in the subsequent reinforcement learning process. Through such pre-training, the model is better equipped to capture key features in the driving environment, providing a solid foundation for subsequent decision-making [32].

In addition, the study also discusses the feature extraction method of multi-sensor fusion, which further enhances the system's generalization capability [33]. In autonomous driving scenarios, vehicles need to process data from various sensors, such as cameras, radars, and LiDAR. By fusing the data from these sensors, the model can acquire more comprehensive environmental information, enabling it to make more accurate and safer driving decisions. This multi-sensor fusion strategy allows the SAC algorithm to be applied to more complex and diverse driving scenarios, thereby expanding its potential for real-world applications.

In general, the application of the SAC algorithm in autonomous driving demonstrates its strong potential [34]. It not only addresses continuous action space problems but also enhances the model's performance and generalization ability through pre-training and multi-sensor fusion technology. These characteristics make SAC a highly promising research direction in the field of autonomous driving [35].

4 Research on Multi-Sensor Feature Fusion Learning for Autonomous Driving Tasks

4.1 Experimental Environment

In this experiment, the Windows 10 system is used as the platform, and the main configuration information is shown in Table 1.

Table 1: Experimental environment settings

Environment configuration	Parameter
IDE environment	Anaconda3-Windows-x86_64
GPU	NVIDIA RTX2080Ti (4)
Hard disk	1 T
CPU	Intel Xeon E5
Programming language	Python 3.10
Development framework	PyTorch 1.6.0, CUDA 10.1
Operation system	Windows 10 Pro (64-bit)
Operation system version	21H2 (19044.2846)

4.2 Model Pre-Training and Data Set Introduction

In this study, the pre-training experiment first designs an independent feature extraction network for different sensor data. The ViT is pre-trained on the CIFAR-100 dataset, which consists of 60,000 32 × 32 color images spanning 100 fine-grained categories, enabling ViT to learn rich visual features. To capture more complex visual semantic information, the DenseNet is pre-trained on the ImageNet dataset, containing over 14 million images across approximately 22,000 categories. Additionally, PointNet++ is pre-trained on the ModelNet dataset, which offers around 127,915 3D CAD models across 624 categories, specifically training the network to accurately extract and understand 3D geometric features. These pre-trained networks are then integrated with the reinforcement learning module, followed by end-to-end training in the CARLA simulator.

CARLA is an open-source autonomous driving research platform that offers a range of realistic urban environments, including diverse traffic conditions, weather scenarios, and road types. This allows us to test and optimize autonomous driving strategies in a safe and controlled setting. By simulating real-world driving scenarios in CARLA, our model learns to perform effective multi-sensor feature fusion under varying road conditions and make real-time decisions, thus establishing a solid foundation for the application of autonomous driving in the real world. Tables 2–4 are the pre-training parameter settings for each module.

Table 2: Hyperparameters of ViT network pre-training

Name	Value	Name	Value
Epochs	118	Activation function	ReLU
Learning rate	0.0001	Number of multi-head Self-attention	10
Dropout	0.2	Weight initialization method	Kaiming initialization
Optimizer	SGD	Image normalized mean	(0.485, 0.456, 0.406)
Mini batch size	128	Image normalization method	(0.229, 0.224, 0.225)

Table 3: Hyperparameters of PointNet++ network pre-training

Name	Value	Name	Value
Epochs	95	Mini batch size	48
Learning rate	0.0001	Image normalization method	Center point normalization
Number of points	4096	Weight initialization method	Kaiming initialization

Table 4: Reinforcement learning module training hyperparameters

Name	Value	Name	Value
Episode	100,000	Mini batch size	32
Learning rate	0.0001	Actor learning rate	0.0001
Return decay rate	0.99	Critic learning rate	0.0001

4.3 Evaluation Indicators

This study aims to comprehensively evaluate the fusion effect of multi-sensor information in autonomous driving operations. We used three main evaluation metrics: Top-1 Accuracy, Over Accuracy, and Mean Accuracy. Top-1 Accuracy is a standard metric for measuring the performance of models in single-label classification tasks. It represents the proportion of the most probable label predicted by the model compared to the true label across all test samples. This metric reflects the model's ability to correctly identify the most likely categories. Over Accuracy is a non-standard term, referring to cases where the model's performance exceeds the benchmark or expected performance level. This typically involves a comparison with a predefined threshold or another model. Mean Accuracy represents the average accuracy across all categories, providing a macro-level indicator of the model's overall classification ability. Through these comprehensive evaluation methods, we can accurately quantify the model's performance across different sensing tasks and datasets (including CIFAR-100, ImageNet, and ModelNet), ensuring a multi-angle and thorough evaluation.

4.4 Comparative Experiment

In this paper, to demonstrate the superiority of the proposed method, we compare it with the DLCE [11], TSingNet [14], BEVFusion [8], and MCS-YOLO [16] models. These experiments are conducted on three standard datasets: CIFAR-100, ImageNet, and ModelNet, to evaluate the effectiveness of the proposed approach. The detailed experimental results are presented in Table 5.

Table 5: Comparison of performance on ImageNet, CIFAR-100, and ModelNet datasets

Methods	ImageNet	CIFAR-100	ModelNet	
	Top-1 Acc (%)	Top-1 Acc (%)	Over Acc (%)	Mean Acc (%)
DLCE [11]	75.0	60.0	82.0	81.5
TSingNet [14]	78.0	63.0	84.0	83.0
BEVFusion [8]	76.5	62.5	83.5	82.7
MCS-YOLO [16]	80.0	66.0	85.0	84.5
Baseline	82.5	69.0	87.5	86.0
The proposed method	87.1	73.6	91.3	89.9

As shown in Table 5, the proposed Faster R-CNN++ network outperforms existing methods on the BDD-100K dataset. Specifically, on the ImageNet dataset, the proposed method achieves the highest Top-1 Accuracy of 87.1%, which is 4.6 percentage points higher than the Baseline and 7.1 percentage points higher than the latest multi-sensor fusion method. On the CIFAR-100 dataset, the proposed method also delivers the best performance, with a Top-1 Accuracy of 73.6%, surpassing the Baseline by 4.6 percentage points and the latest multi-sensor fusion method by 7.6 percentage points. For the ModelNet dataset, the proposed method achieves 91.3% in Over Accuracy and maintains high accuracy in Mean Accuracy at 89.9%.

The experimental results demonstrate that the proposed method significantly improves Top-1 Accuracy, Over Accuracy, and Mean Accuracy, highlighting its superiority in handling more complex and diverse image classification tasks. This improvement is attributed to SCP technology, which enhances the model's speed and generalization ability by making the feature map sparser and reducing unnecessary computations. The RTA mechanism leverages residual connections and attention to refine the initial features, effectively prioritizing important features and boosting the model's ability to recognize complex and fine-grained details. Additionally, this mechanism helps mitigate the issue of gradient vanishing, contributing to more stable training of deep networks.

4.5 Ablation Experiment

To verify the effectiveness of each component in the proposed method, we conducted ablation experiments. Specifically, we evaluated the impact of the RTA mechanism and SCP technology on the model's performance across three datasets: CIFAR-100, ImageNet, and ModelNet. The experimental setups are as follows: Baseline: a standard deep learning model without any attention mechanism or sparsity; +SCP: the introduction of SCP technology, without the attention mechanism; +RTA: the inclusion of the RTA mechanism, but without SCP; Proposed Method: the complete method that combines both SCP and RTA. The experimental results are presented in Table 6.

Table 6: Results of the ablation experiment

Methods	**ImageNet**	**CIFAR-100**	**ModelNet**	
	Top-1 Acc (%)	**Top-1 Acc (%)**	**Over Acc (%)**	**Mean Acc (%)**
Baseline	75.0	60.3	82.0	81.5
+SCP	78.5	63.6	84.5	83.3
+RTA	76.5	62.5	83.5	82.7
The proposed method	87.1	73.6	91.3	89.9

The experimental results show that the +SCP module improves performance by at least 3.3% in Top-1 Accuracy, 1.5% in Over Accuracy, and 1.8% in Mean Accuracy compared to the baseline model. This improvement is due to the sparsity of the +SCP module, which not only reduces feature map redundancy but also enhances computational efficiency. Additionally, this method helps improve the model's ability to generalize unknown data.

Furthermore, the +RTA module shows an improvement of at least 0.5% in Top-1 Accuracy, 0.5% in Over Accuracy, and 1.2% in Mean Accuracy compared to the baseline model. These results demonstrate that the RTA mechanism effectively boosts the model's recognition capability for complex and fine-grained features by emphasizing important features and suppressing irrelevant ones.

The proposed method, which combines both the SCP module and the RTA module, outperforms the baseline model by at least 11.9% in Top-1 Accuracy, 9.3% in Over Accuracy, and 8.4% in Mean Accuracy.

In summary, the results highlight that the RTA mechanism and SCP are two crucial components for improving algorithm performance. These components complement each other, collectively enhancing the model's feature extraction ability and generalization performance. Additionally, these findings support our hypothesis that the model's application performance across various fields can be significantly improved by enhancing the representation of important features through attention mechanisms and reducing computational complexity through sparsity.

4.6 Reinforcement Learning Training Results Analysis

4.6.1 Assess the Lane Keeping Effect

To evaluate the effectiveness of the proposed training strategy in lane-keeping performance, this study conducted a detailed test on both the training map and the non-training starting point of the test map.

The corresponding experimental results are presented in Table 7. As a comparison, a baseline case was established, using a random strategy network with only RGB images as input. This was compared and analyzed against the multi-sensor fusion method proposed in this study. In the table, the boxed numbers indicate the starting point of the training and the corresponding map, while the unlabeled data points represent the starting points of the test.

Table 7: Maximum bonus values for different scenarios and starting points

Methods	MAP	Start 1	Start 2	Start 3	Start 4
Baseline	Town 1	8.81	10.50	10.01	10.34
	Town 2	9.84	8.67	8.75	7.83
	Town 3	7.68	9.16	6.20	9.56
	Town 4	8.24	9.10	6.85	6.27
Random strategy	Town 1	11.8	11.2	10.6	7.7
	Town 2	10.8	9.17	7.15	10.3
	Town 3	7.81	10.89	10.0	8.46
	Town 4	9.93	10.00	10.3	10.01

Upon reviewing the experimental data, it can be observed that in this task, the multi-sensor fusion method proposed in this study does not show a significant improvement over the benchmark method, and the difference in generalization ability between the two is relatively small. This phenomenon can be attributed to the relative simplicity of the lane-keeping task. Since the agent primarily needs to focus on the lane lines, the RGB image information plays a crucial role, which explains the observed experimental results. Nevertheless, overall, the method proposed in this study still slightly outperforms the baseline method. This is mainly due to the contribution of other sensors, such as LiDAR, which provide additional information and offer extra functionalities like collision avoidance.

4.6.2 Evaluation of Vehicle Following Effect

To better evaluate the effectiveness of the proposed training method for vehicle tracking operations, this study selected various non-training starting points across a range of test environments, including training maps, for a series of experimental validations. These tests are designed to thoroughly assess the adaptability and robustness of the strategy in diverse scenarios, with the results presented in Table 8. Additionally, to benchmark the performance of the multi-sensor fusion method introduced in this study, a baseline case was chosen. This baseline was trained using a random strategy network that relied solely on RGB image input. In the result table, data points marked with boxes indicate the starting positions used in training, while unboxed data points represent the starting positions in the test phase.

A careful analysis of the data in the table reveals that the multi-sensor fusion method outperforms the baseline significantly in this demanding vehicle-following evaluation. This notable difference is also evident in the generalization tests, where the multi-sensor fusion method demonstrates superior performance. The experiments show that deep reinforcement learning algorithms, when based on multi-sensor data, can effectively enhance the lane tracking capabilities of autonomous vehicles and hold strong potential for practical applications. This improvement not only boosts the vehicle's environmental perception but also provides richer data for navigating complex traffic scenarios, thereby advancing intelligent driving systems toward a higher level of autonomy.

Table 8: The maximum reward value of the starting point in different scenarios

Methods	MAP	Start 1	Start 2	Start 3	Start 4
Baseline	Town 1	1.83	1.07	0.68	0.75
	Town 2	1.60	1.48	1.65	1.07
	Town 3	1.19	0.95	0.50	0.73
	Town 4	0.65	0.51	0.75	0.94
Random strategy	Town 1	2.08	1.03	2.04	0.98
	Town 2	1.78	1.50	2.23	1.45
	Town 3	2.15	1.01	1.97	2.20
	Town 4	1.72	1.11	1.23	1.40

5 Conclusion

This study presents an innovative learning model for multi-sensor feature fusion in the context of autonomous driving, with the goal of enhancing both the perception accuracy and response speed of autonomous driving systems in complex environments. By integrating vision, radar, and laser scanning data from three different modalities, our model optimizes the feature fusion process through the use of SCP and a residual ternary attention mechanism. Simulation results demonstrate that this approach effectively filters out irrelevant feature channels, reduces the impact of redundant signals, and enhances the model's robustness against external interference. The residual ternary attention mechanism facilitates the precise alignment of cross-modal features, alleviating the computational burden caused by unnecessary features and boosting the system's computational efficiency. These improvements lead to significant gains in the real-time responsiveness and decision-making accuracy of the autonomous driving system.

Additionally, by incorporating a reinforcement learning module and combining it with feature fusion learning, we trained an efficient end-to-end driving strategy in the CARLA autonomous driving simulator. The results demonstrate that this integrated approach not only improves the generalization ability and stability of the autonomous driving system but also enables more efficient and accurate environmental perception and decision-making.

In summary, the contributions of this study are as follows:

(1) A feature fusion framework for autonomous driving, based on three modalities, is proposed. This framework combines deep learning and reinforcement learning techniques to enhance the environmental adaptability and decision-making capabilities of autonomous vehicles.

(2) A novel SCP technique is introduced to reduce feature redundancy and emphasize task-relevant features, thereby improving the system's real-time performance and accuracy.

(3) A residual ternary attention mechanism is designed to enhance the integration efficiency of multi-sensor data, optimizing the feature fusion process through effective cross-modal feature alignment.

While the SCP technique and RTA mechanism proposed in this study have achieved notable success in reducing feature redundancy and improving feature fusion efficiency, challenges remain in terms of computational complexity and real-time performance when processing large-scale sensor data. Future work could focus on further optimizing these mechanisms, such as by incorporating more efficient feature selection algorithms or adopting lightweight network architectures to reduce computational load and enhance real-time responsiveness. Additionally, with the continuous advancement of sensor technologies—such as higher-resolution cameras, more precise LiDAR, and millimeter-wave radar—it will be crucial to explore how to better integrate data from these emerging sensors. This may involve developing new feature extraction and fusion methods that fully capitalize on the strengths of different sensor types, thereby further improving the perception accuracy of autonomous driving systems.

Acknowledgement: Not applicable.

Funding Statement: The authors received no specific funding for this study.

Author Contributions: Conceptualization, Yang Su and Xianrang Shi; methodology, Yang Su; software, Xianrang Shi; validation, Tinglun Song, Yang Su, and Xianrang Shi; formal analysis, Xianrang Shi; investigation, Yang Su; resources, Yang Su; data curation, Yang Su; writing—original draft preparation, Yang Su; writing—review and editing, Yang Su; visualization, Tinglun Song; supervision, Yang Su; project administration, Yang Su; funding acquisition, Xianrang Shi. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The data that support the findings of this study are available from the corresponding author upon reasonable request.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

1. Cai Y, Luan T, Gao H, Wang H, Chen L, Li Y, et al. YOLOv4-5D: an effective and efficient object detector for autonomous driving. IEEE Trans Instrum Meas. 2021;70:1–13. doi:10.1109/TIM.2021.3065438.
2. Li L, Zhao W, Wang C. Cooperative merging strategy considering stochastic driving style at on-ramps: a bayesian game approach. Automot Innov. 2024;7(2):312–34. doi:10.1007/s42154-023-00248-x.
3. Natan O, Miura J. Towards compact autonomous driving perception with balanced learning and multi-sensor fusion. IEEE Trans Intell Transp Syst. 2022;23(9):16249–66. doi:10.1109/TITS.2022.3149370.
4. Xiang C, Feng C, Xie X, Shi B, Lu H, Lv Y, et al. Multi-sensor fusion and cooperative perception for autonomous driving: a review. IEEE Intell Transp Syst Mag. 2023;15(5):36–58. doi:10.1109/MITS.2023.3283864.
5. Wang X, Li K, Chehri A. Multi-sensor fusion technology for 3D object detection in autonomous driving: a review. IEEE Trans Intell Transp Syst. 2023;25(2):1148–65. doi:10.1109/TITS.2023.3317372.
6. Li L, Zhao W, Wang C, Fotouhi A, Liu X. Nash double Q-based multi-agent deep reinforcement learning for interactive merging strategy in mixed traffic. Expert Syst Appl. 2024;237:121458. doi:10.1016/j.eswa.2023.121458.
7. Chen J, Li SE, Tomizuka M. Interpretable end-to-end urban autonomous driving with latent deep reinforcement learning. IEEE Trans Intell Transp Syst. 2021;23(6):5068–78. doi:10.1109/TITS.2020.3046646.
8. Liu Z, Tang H, Amini A, Yang X, Mao H, Rus DL, et al. Bevfusion: multi-task multi-sensor fusion with unified bird's-eye view representation. In: 2023 IEEE International Conference on Robotics and Automation (ICRA); 2023 May 29–Jun 2; London, UK. p. 2774–81. doi:10.1109/ICRA48891.2023.10160968.
9. Kiran BR, Sobh I, Talpaert V, Mannion P, Al Sallab AA, Yogamani S, et al. Deep reinforcement learning for autonomous driving: a survey. IEEE Trans Intell Transp Syst. 2021;23(6):4909–26. doi:10.1109/TITS.2021.3054625.
10. Berkaya SK, Gunduz H, Ozsen O, Akinlar C, Gunal S. On circular traffic sign detection and recognition. Expert Syst Appl. 2016;48:67–75. doi:10.1016/j.eswa.2015.11.018.
11. Nie J, Zhang J, Wan X, Ding W, Ran B. Modeling of decision-making behavior for discretionary lane-changing execution. In: 2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC); 2016 Nov 1–4; Brazil: Rio de Janeiro. p. 707–12. doi:10.1109/ITSC.2016.7795631.
12. Schubert R, Schulze K, Wanielik G. Situation assessment for automatic lane-change maneuvers. IEEE Trans Intell Transp Syst. 2010;11(3):607–16. doi:10.1109/TITS.2010.2049353.
13. Grigorescu S, Trasnea B, Cocias T, Macesanu G. A survey of deep learning techniques for autonomous driving. J Field Robot. 2020;37(3):362–86. doi:10.1002/rob.21918.

14. Liu Y, Peng J, Xue JH, Chen Y, Fu ZH. TSingNet: scale-aware and context-rich feature learning for traffic sign detection and recognition in the wild. Neurocomputing. 2021;447:10–22. doi:10.1016/j.neucom.2021.03.049.
15. Nacir O, Amna M, Imen W, Hamdi B. YOLO V5 for traffic sign recognition and detection using transfer learning. In: 2022 IEEE International Conference on Electrical Sciences and Technologies in Maghreb (CISTEM); 2022 Oct 26–28; Tunis, Tunisia. p. 1–4. doi:10.1109/CISTEM55808.2022.10044022.
16. Cao Y, Li C, Peng Y, Ru H. MCS-YOLO: a multiscale object detection method for autonomous driving road environment recognition. IEEE Access. 2023;11:22342–54. doi:10.1109/ACCESS.2023.3252021.
17. Kaelbling LP, Littman ML, Moore AW. Reinforcement learning: a survey. J Artif Intell Res. 1996;4:237–85. doi:10.1613/jair.301.
18. Liu P, Zhang C, Qi H, Wang G, Zheng H. Multi-attention DenseNet: a scattering medium imaging optimization framework for visual data pre-processing of autonomous driving systems. IEEE Trans Intell Transp Syst. 2022;23(12):25396–407. doi:10.1109/TITS.2022.3145815.
19. Qi CR, Yi L, Su H, Guibas LJ. PointNet++: deep hierarchical feature learning on point sets in a metric space. Adv Neural Inf Process Syst. 2017;30:1–10. doi:10.48550/arXiv.1706.02413.
20. Ando A, Gidaris S, Bursuc A, Puy G, Boulch A, Marlet R. Rangevit: towards vision transformers for 3D semantic segmentation in autonomous driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2023 Jun 17–24; Vancouver, BC, Canada. p. 5240–50. doi:10.48550/arXiv.2301.10222.
21. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556. 2014.
22. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2016 Jun 27–30; Las Vegas, NV, USA. p. 770–78. doi:10.1109/CVPR.2016.90.
23. Szegedy C, Ioffe S, Vanhoucke V, Alemi A. Inception-v4, inception-resnet and the impact of residual connections on learning. In: Proceedings of the AAAI Conference on Artificial Intelligence; 2017 Feb 4–9; San Francisco, CA, USA. doi:10.1609/aaai.v31i1.11231.
24. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2017 Jul 21–26; Honolulu, HI, USA. p. 4700–8. doi:10.48550/arXiv.1608.06993.
25. Li Y, Ma L, Zhong Z, Liu F, Chapman MA, Cao D, et al. Deep learning for lidar point clouds in autonomous driving: a review. IEEE Trans Neural Netw Learn Syst. 2020;32(8):3412–32. doi:10.1109/TNNLS.2020.3015992.
26. Cui Y, Chen R, Chu W, Chen L, Tian D, Li Y, et al. Deep learning for image and point cloud fusion in autonomous driving: a review. IEEE Trans Intell Transp Syst. 2021;23(2):722–39. doi:10.1109/TITS.2020.3023541.
27. Lee J, Lee I, Kang J. Self-attention graph pooling. In: International Conference on Machine Learning, LCML; 2019 Jun 9–15; Long Beach, CA, USA. p. 3734–43. doi:10.48550/arXiv.1904.08082.
28. Zhou H, An L, Zhu R. A grouping feature selection method based on feature interaction. Intell Data Anal. 2023;27(2):361–77. doi:10.3233/IDA-226551.
29. Wang W, Guo M, Han T, Ning S. A novel feature selection method considering feature interaction in neighborhood rough set. Intell Data Anal. 2023;27(2):345–59. doi:10.3233/IDA-216447.
30. Shaw P, Uszkoreit J, Vaswani A. Self-attention with relative position representations. arXiv:1803.02155. 2018.
31. Yang J, Zhang J, Xi M, Lei Y, Sun Y. A deep reinforcement learning algorithm suitable for autonomous vehicles: double bootstrapped soft-actor-critic-discrete. IEEE Trans Cogn Dev Syst. 2021;15(4):2041–52. doi:10.1109/TCDS.2021.3092715.
32. Liu H, Huang Z, Mo X, Lv C. Augmenting reinforcement learning with transformer-based scene representation learning for decision-making of autonomous driving. IEEE Trans Intell Veh. 2024;9(3):4405–21. doi:10.1109/TIV.2024.3372625.
33. Marsh B, Sadka AH, Bahai H. A critical review of deep learning-based multi-sensor fusion techniques. Sensors. 2022;22(23):9364. doi:10.3390/s22239364.

34. Gao M, Chang DE. Autonomous driving based on modified sac algorithm through imitation learning pretraining. In: 2021 21st International Conference on Control, Automation and Systems (ICCAS); 2021 Oct 12–15; Jeju, Republic of Korea. p. 1360–4. doi:10.23919/ICCAS52745.2021.9649939.
35. Wu K, Wang H, Esfahani MA, Yuan S. Learn to navigate autonomously through deep reinforcement learning. IEEE Trans Ind Electron. 2021;69(5):5342–52. doi:10.1109/TIE.2021.3078353.