ISSN 1330-3651 (Print), ISSN 1848-6339 (Online)

https://doi.org/10.17559/TV-20240219001336
Original scientific paper

Smart Vehicle Obstacle Detection by Camera and LiDAR Image Fusion

Qianying ZOU*, Fengyu LIU, Ruixin CHEN

Abstract: This paper proposes a novel obstacle detection method for autonomous vehicles that combines camera and LiDAR image fusion techniques. The proposed method employs the DeepLabV3+ algorithm with an attention mechanism for camera image segmentation and a centroid algorithm with scanning line bundle-based segmentation for LiDAR image processing. The processed images are then fused using the Local Non-Subsampled Shear Transform (LNSST) algorithm, which enhances the detail information and improves the recognition speed and accuracy. Experimental results demonstrate that the proposed method achieves superior performance in complex scenes, partially occluded objects, and long-range target detection compared to state-of-the-art algorithms. The proposed method significantly improves the environment perception capabilities of autonomous vehicles, contributing to safer and more efficient navigation in complex driving scenarios.

Keywords: attention mechanism; deep LabV3+ algorithm; high and low frequency subbands; obstacle detection

1 INTRODUCTION

With the advancement of artificial intelligence technology, autonomous driving technologies have become increasingly mature, covering key technologies such as vehicle control, path planning, and perceptual fusion. The core objective of these technologies is to enhance the safety of autonomous driving by improving the accuracy of obstacle recognition and reducing the response time of intelligent vehicle systems [1]. Currently, obstacle detection using visual and radar sensors is considered two mainstream technologies. Visual sensors, typically cameras, are capable of identifying the type of obstacles but face challenges in precisely locating them [2]. In contrast, LiDAR, a type of radar sensor, can accurately determine the position of obstacles but its accuracy in identifying the specific types of obstacles is relatively lower [3]. In order to improve the obstacle recognition accuracy and shorten the response time of the system for intelligent vehicles during motion, a single detection method appears to be insufficient. Therefore, the research pushes toward the use of multiple information fusion methods, aiming to combine the advantages of different sensors to obtain more comprehensive and accurate obstacle information. This approach not only improves the recognition efficiency, but also effectively reduces the response time of the system, which is crucial for improving the safety of self-driving vehicles.

In recent years, many important research results have been contributed by domestic and foreign scholars for intelligent vehicle obstacle recognition. Reference [4] introduces an algorithm that merges adaptive density clustering with multi-feature data association, enhancing the adaptability and tracking precision of dynamic obstacle clustering in LiDAR systems. The algorithm made progress in improving the accuracy and efficiency of obstacle detection, although the recognition accuracy in complex scenes still needs to be further optimized. Reference [5] proposes an improved corner-constrained LiDAR obstacle detection method, which solves the over-segmentation problem of the missing point cloud and accelerates the clustering speed, but there are problems such as incomplete segmentation and large impact on the detection effect. Reference [6] introduces an obstacle detection and tracking algorithm based on 3D LiDAR point cloud, which is effective for the identification and tracking of moving obstacles and can improve the GPS positioning accuracy. However, the algorithm has challenges in maintaining GPS accuracy during street driving of self-driving cars and needs further optimization. Reference [7] discusses a method where LiDAR enhances obstacle detection by projecting an enclosing box onto the machine vision image. However, this method faces challenges such as a high rate of false detections and reduced detection effectiveness in rainy conditions. Reference [8] developed an enhanced DBSCAN algorithm that conducts clustering and creates bounding boxes to identify obstacles, yet it still struggles with a high rate of false detections. Yoo et al. [9] proposed a 3D-CVF at SPA algorithm that significantly improves the single-modal performance by fusing camera and LiDAR features across views. Nonetheless, there is still room for improvement in the correction accuracy by accurately converting 2D camera features into spatial feature maps corresponding to LiDAR. Dou et al. [10] proposed a method SEG-VoxelNet based on RGB images and LiDAR point cloud information, which can accurately recognize vehicles with a detection rate exceeding the best performance at that time, but there is redundancy in the model. The CLOCs fusion algorithm of Pang et al. [11] is a low-complexity multimodal framework, which effectively improves the single-modal detection performance, but the long range detection is still challenging especially in complex environments and bad weather. The Sparse LiDAR and Stereo Fusion (SLS-Fusion) method by Mai N. A. M., Duthon P., Khoudour L., et al. [12] merges sparse LiDAR with stereo imagery to enhance depth estimation and 3D object detection. This integration offers precise environmental perception by utilizing both depth and visual details. However, the complexity of data processing leads to slower system response times, particularly with large datasets, impacting efficiency in real-time applications. Rukhovich et al. [13] developed the TheImvoxelNet algorithm converts 2D images into 3D voxel representations for object detection in 3D space and is suitable for both monocular and multi-view systems. Although suitable for general purpose 3D object detection, it is computationally expensive and needs further optimization for real-time applications. Palffy et al. [14] combined the fusion of visual images with 3D radar data of the target and proposed an obstacle recognition model based on CNN and SVM, which reduces the error obtained from the LIDAR scanning, but the LIDAR cost is high, and the practical applicability is not high. Reference [15] proposed a target detection system based on the fusion of LiDAR and camera information, which effectively solves the problem of low resolution of LiDAR and poor sensing distance of the camera, but it only meets the automatic formula racing. Reference [16] proposes an accurate obstacle recognition method by fusing laser 3D point cloud data with optical image grayscale information, and the algorithm can effectively solve the interference of image texture on obstacle recognition in the coarse obstacle avoidance processing by using a single grayscale information, but it has a large amount of computation and a high requirement for computing power. The improved Deeplab V3+ model in the Reference [17] improves accuracy and real-time performance by using MobileNet V2 instead of the Xception backbone network in the orchard scene segmentation and replacing the original ReLU with the ReLU6 activation function in the ASPP module. However, the recognition and segmentation effect is average for targets that are far away and have a small pixel percentage. Reference [18] proposed an improved semantic segmentation method for road scenes based on DeepLab v3+ network structure, combined with the attention mechanism to increase the weight of segmentation region, but there is still a lack of recognition accuracy in the face of complex scenes. Chen [19] and others' Painted-Point RCNN algorithm improves the accuracy and robustness of 3D object detection accuracy and robustness, nevertheless, its generalization ability under different lighting conditions and detection performance in variable environments still need to be further optimized.

Facing the challenges of inconspicuous image features and low accuracy of obstacle detection in complex scenes, especially the low detection rate under rainy conditions, this study proposes a novel obstacle detection method that combines camera image segmentation technology and LiDAR image point cloud processing technology.

Innovations of This Method:

1. Camera Image Processing: This method employs the DeepLabV3+ algorithm with the introduction of an attention mechanism for image segmentation. This innovative combination not only enhances the accuracy of obstacle recognition, especially in scenes with inconspicuous image features, but also ensures higher usability of the image as one of the input sources.

2. LiDAR Image Processing: The method introduces a centroid algorithm innovatively to reduce redundancy and complexity in the data during the rasterization process, thereby improving data processing speed.

3. LiDAR Data Processing: By adopting a scanning line bundle-based segmentation method, this approach effectively retains the original data characteristics, overcoming issues of data loss and deformation when projected onto a 2D plane. This results in processed LiDAR images becoming another high-value input source.

Fusion of Input Sources: Utilizing high and low frequency subband techniques to fuse the two input sources, this method not only enriches detail information but also significantly improves recognition speed and reduces the false detection rate.

2 RELATEDWORK

2.1 DeepLabV3+ Fusion Xception Modeling

DeepLab V3+ is an image semantic segmentation model that employs the Xception network as its backbone network to extract image features [20]. The Xception model, proposed by François Chollet [21], is constructed based on the depth-separable convolution technique, which was developed from the Inception v3 model [22]. The model is optimized by replacing the Inception module with a depth-separable convolution and incorporating ResNet-like jump connections. Although DeepLab V3+ performs well in semantic segmentation of images, it sometimes encounters difficulties in capturing and accurately segmenting the key information of an image. To address this problem, DeepLab V3+ introduces an attention mechanism to compensate for information that may be missed during the segmentation process in order to improve the accuracy of the segmentation results [23].

The DeepLab V3+ network optimizes the model performance by integrating two types of attention modules: in the encoding stage, a channel attention mechanism is used to strengthen the model's emphasis on different channel features; in the decoding stage, a spatial attention mechanism is employed to enhance the model's ability to process the spatial information of the image. This integration of dual attention mechanisms not only improves the model's ability in detail capture, but also further enhances the accuracy of the segmentation results.

Although the DeepLab V3+ model has improved in terms of enhancing segmentation accuracy, these improvements may not be sufficient to handle the integration of information between different modalities when fused with LiDAR data for obstacle detection in complex environments. In addition, the model may also face challenges in terms of real-time performance, especially when processing high-resolution images with a heavy computational burden, which may limit its usefulness in real-world applications of self-driving vehicles.

2.2 Point Cloud Voxelization Filtering

LiDAR generates highly accurate 3D point cloud maps by emitting laser pulses and receiving their reflected waves. These point cloud maps provide detailed information about the surrounding environment for intelligent vehicles [24]. However, due to uncertainties such as equipment accuracy, environmental variations, and operational errors, the generated point cloud data inevitably contain noise and outliers. If left untreated, these noisy and outlier points will seriously affect the accuracy and reliability of obstacle detection. To address this issue, this study adopts the voxel filtering method to preprocess the original point cloud.

The working principle of voxel filtering involves defining the size of the voxel to determine the length of each voxel in the *X*, *Y*, and *Z* axis directions. This size determines the granularity of the filtering. For each point in the point cloud, the grid coordinates of the voxel in which it is located are calculated as shown in Eq. (1).

$$
\begin{aligned}
X_\upsilon &= \left\lfloor \frac{x_i}{L_x} \right\rfloor \\
Y_\upsilon &= \left\lfloor \frac{y_i}{L_y} \right\rfloor \\
Z_\upsilon &= \left\lfloor \frac{z_i}{L_z} \right\rfloor
\end{aligned}
\tag{1}
$$

where the floor function $\lfloor \bullet \rfloor$ denotes the downward rounding operation. The point cloud data within each voxel raster will be replaced by the centroid of all the points in that raster. This processing not only preserves the overall contour and shape of the point cloud but also effectively filters out irrelevant noise and outliers. The expression is shown in Eq. (2).

$$C_u = \left(\frac{1}{N_u} \sum_{i=1}^{N_u} x_i ,\ \frac{1}{N_u} \sum_{i=1}^{N_u} y_i ,\ \frac{1}{N_u} \sum_{i=1}^{N_u} z_i \right) \tag{2}$$

where N_υ is the number of points within the voxel and (x_i, y_i, z_i) is the coordinate of the i th point within the voxel. Through these steps, all points in the original point cloud within the same voxel are reduced to a single centroid point, thereby reducing the complexity of the data while filtering out noisy and outlier points.

The effectiveness of voxel filtering directly depends on the size of the voxel: the larger the voxel, the more simplified the filtered point cloud is, but more details may be lost; the smaller the voxel, the more details are retained, but the computational load increases accordingly, and the ability to filter out noise weakens. Therefore, the voxel size needs to be adjusted according to the specific needs in practical applications.

2.3 LNSST Transform

In this study, the choice of image fusion technique is crucial because it directly affects the quality of fused images and the accuracy of obstacle detection [25]. The purpose of image fusion is to effectively combine different image information from cameras and LiDAR to obtain richer and more accurate environmental perception data. Among the many image fusion techniques, transform domain methods are especially critical [26], among which discrete wavelet transform is not widely used due to the lack of translation invariance and the presence of ringing phenomenon.

To overcome these limitations, Da et al. [27] proposed the Non-Downsampled Contour Wave Transform (NSCT), which can efficiently obtain the orientation information of an image and solve the problems of translation invariance and Gibbs phenomenon. Although NSCT performs well in image fusion, its computational efficiency is low, which is difficult to meet the demand for high timeliness. To address this problem, Guo et al. [28] further proposed a multi-scale and multi-orientation shear wave transform, which improves efficiency but still lacks translational invariance.

Non-Downsampled Shear Wave Transform (NSST) is a commonly used method in the field of image fusion, which not only maintains superior information capture and representation capabilities but also provides a significant improvement in computational efficiency compared to NSCT [29]. Its formula is shown in Eq. (3).

$$S(I) = \left\{ \left\langle I, \varphi_{a,s,t} \right\rangle : a \in A, s \in S, t \in \mathbb{Z}^2 \right\} \tag{3}$$

where $S(I)$ is the shear wave transform, depending on the scale parameter, shear parameter, A and S translation parameter and are the discrete sets of scale and shear respectively.

Based on NSST, there is a more suitable Local Non-Downsampled Shear Transform (LNSST) for image fusion. LNSST not only inherits all the advantages of NSST but also solves the problem of spectral aliasing and further optimizes key features such as multi-resolution, multi-directionality, and translation invariance [30]. When the dimension $n = 2$ is applied, the system function of LNSST is shown in Eq. (4).

$$
\begin{aligned}
& M_{AB}(\phi) = \{ \phi_{p,l,k}(k) = \mid \det A \mid^{p/2} \cdot \\
& \cdot \phi(B^l A^p x - k), (p,l) \in Z, k \in Z^2 \}
\end{aligned}
\tag{4}
$$

where $\phi \in L^2(R^2)$, A and B are the invertible matrices of 2×2 , $|\det B| = 1$, p are the scale parameters, l is the orientation parameter, and k denotes the spatial position.

3 METHOD

3.1 Algorithmic Flow

This study adopts a comprehensive methodology that aims to enhance the obstacle detection capability of intelligent vehicles in complex environments by combining image processing and point cloud processing techniques. The model is mainly divided into three core parts: first, an image segmentation model based on DeepLab V3+ and combining the channel attention mechanism; second, the preprocessing of LiDAR data and point cloud segmentation; and finally, the image fusion processing using the local non-subsampled shear transform (LNSST) algorithm. The following are the core ideas and implementation steps of these three parts:

1) In the process of constructing the image segmentation model, the study adopts a new approach: combining the DeepLab V3+ network model with the channel attention mechanism. Through this combination, the study introduces a channel attention module designed to significantly enhance the performance of DeepLab V3+. This improved model is able to effectively enhance the accuracy and processing efficiency of image segmentation tasks by precisely focusing on key feature channels in an image. This ability to focus on important features enables the model to better recognize and segment the target object when processing complex images, thus providing higher quality base data for further image analysis and applications.

2) In processing the LiDAR data, the study adopted a preprocessing step: point cloud rasterization preprocessing. This process uses a centroid algorithm to rasterize the point cloud data acquired by LIDAR scanning, aiming to lay the foundation for subsequent analysis. Next, the study applies a scanning harness-based segmentation algorithm specifically designed to reject invalid point cloud data on the ground. With this algorithm, it is possible to focus on the non-ground areas, depth-project these areas, and thus accurately segment the target obstacles. This processing step is directly related to the ability of the model to accurately recognize and localize obstacles, which is crucial for the navigation and safe operation of intelligent vehicles.

3) In the field of image fusion processing, the research utilizes the Local Non-Downsampled Shear Transform (LNSST) algorithm for exhaustive decomposition of processed camera images and point cloud images. This decomposition process produces two types of subbands: low frequency subbands and high frequency subbands. Each subband employs a specific fusion strategy to maximize its potential value. For the low-frequency subbands, the study adopted an energy attribute (EA)-based fusion method. This approach aims to maintain the overall information and contextual consistency of the fused image, ensuring that the underlying structure of the image is preserved. For the high frequency subbands, on the other hand, a modified Laplace energy sum (SML) is adopted for fusion, which helps to preserve and enhance the detailed features in the image and improve the clarity and richness of the image. Finally, the above fused subbands are processed by the inverse transformation of the LNSST algorithm, which will ultimately result in a fused image for obstacle detection that is both detailed and highly accurate. This image fuses data from different sources, which not only enriches the visual details, but also improves the accuracy of obstacle detection, providing more reliable environment sensing capability for intelligent vehicles.

3.2 Improved DeepLab V3+ Algorithm

The study was specifically optimized to address the limitations of the DeepLab V3+ algorithm for semantic segmentation of images. Although DeepLab V3+ performs well in synthesizing images, it sometimes falls short in capturing and accurately segmenting critical information in images [31]. To address this issue, the study introduces an attention mechanism that aims to compensate for important information that may be missed during segmentation, thus improving the accuracy of the segmentation results [32].

The optimization strategy studied consists of the integration of two key attention modules: in the encoding phase, a channel attention mechanism [33] is employed, which allows the model to focus on those feature channels that are more important for the segmentation task, while in the decoding process, a spatial attention mechanism [34] is introduced, by which the model is able to better understand the spatial relationships in the image, and thus, in complex scenes, to more accurately label and segment the target object.

Through this optimization of the DeepLab V3+ network model based on the attention mechanism, the study was able to significantly improve the performance of image segmentation in an obstacle detection system for smart vehicles. This not only enhances the model's ability to capture details in the image, but also improves the accuracy of recognizing and segmenting obstacles in complex environments, providing more reliable visual information for smart vehicles to support their safe and efficient navigation and operation. As shown in Fig. 1, the algorithmic flowchart of the DeepLab V3+ network model based on the attention mechanism is demonstrated, depicting in detail how the attention module is integrated into the whole segmentation process.

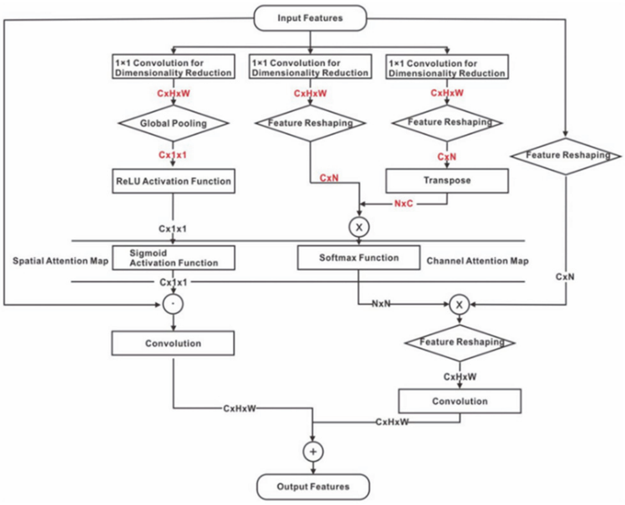

Figure 1 Flowchart of the algorithm of DeepLab V3+ incorporating the attention mechanism

Step 1 Channel Attention Module.

This module is designed to capture the dependencies between different channels in the feature map, thus enhancing the model's ability to recognize important information in the image.

1) Input feature acquisition. Assume that the input feature map obtained at the end of the Xception module processing is denoted as $F_1 \in \mathbb{R}^{C\times H\times W}$, where C is the number of channels, H and W are the height and width of the feature map, respectively; the feature map F_1 is downscaled by 1×1 convolution, and the convolution operation is shown in Eq. (5).

$$F' = Conv_{1\times1}\left(F_1\right) \tag{5}$$

where, F' is the feature map after dimensionality reduction, the size is still $C\times H\times W$ but the parameters and computation are reduced.

2) Global context information extraction. Global average pooling is performed on F' to extract the global context information for each channel as shown in Eq. (6).

$$Gp(D) = \frac{1}{H\times W}\sum_{h=1}^{H}\sum_{w=1}^{W} F'_{c,h,w} \tag{6}$$

where, $Gp(D)$ is the global pooling result for the first c channel, and $F'_{c,h,w}$ is the value of the feature map after dimensionality reduction in the c channel, height h, width w position.

3) Channel dependency modeling. Nonlinear features are extracted by ReLU and Sigmoid functions to model the dependencies between channels. The ReLU function is applied to increase the nonlinearity as shown in Eq. (7).

$$F'' = \mathrm{Re}LU(Gp(D)) \tag{7}$$

The channel importance is further refined using the Sigmoid function as shown in Eq. (8).

$$CA = \sigma(F'') \tag{8}$$

where CA denotes the channel attention map and σ is the Sigmoid function.

4) Weighted fusion. An element-by-element multiplication operation (dot product) is performed between the channel attention map CA and the original input feature F_1 to obtain the weighted channel attention feature H, as shown in Eq. (9).

$$H = CA \odot F_1 \tag{9}$$

where $\odot$ denotes the Hadamard product (element-by-element multiplication), and H is the final channel attention feature, which emphasizes the more important channels of the input features.

The channel attention module effectively identifies and enhances those channel features that are most critical to the task at hand, thereby improving the performance of the image segmentation model [35]. This approach not only enhances the model's ability to capture important information in the image, but also provides a powerful tool for deeper understanding of the image content.

Step 2 Spatial attention module.

1) Spatial feature acquisition. Let the deep feature maps extracted in the encoding part be $F_2 \in \mathbb{R}^{C\times H\times W}$, these feature maps contain the deeper visual information accessible in the network.

2) Feature downscaling and reshaping. The feature map F_2 is downscaled by a 1×1 convolution operation to generate two new feature maps B and C, which are still of the size $C\times H\times W$. The matrices B and C are reshaped into matrices $C\times N$ and $N\times C$, respectively, where $N = H\times W$, in order to perform the matrix multiplication, a step designed to prepare for the computation of the positional similarity.

3) Similarity matrix calculation. The product of two matrices is computed to obtain a similarity matrix $N\times N$ of S, representing the spatial relationship between any two points, as shown in Eq. (10).

$$S = B^T \cdot C \tag{10}$$

Each row of S is normalized by the Softmax function to ensure that the similarity scores sum to 1 as shown in Eq. (11).

$$S_{norm} = Soft\max(S) \tag{11}$$

where, S_{norm} is the spatial attention map processed by Softmax function, reflecting the positional similarity between any two points in the image.

4) Spatial feature generation. Multiply S_{norm} with the reshaped input feature F_2, where F_2 is first reshaped to $N\times C$ to match the S_{norm} dimension as shown in Eq. (12).

$$P = S_{norm} \cdot F_2' \tag{12}$$

where F'_2 is the reshaped form of the input feature F_2 and P is the weighted spatial attention feature, reflecting the importance of different spatial locations in the input feature.

The spatial attention module effectively models the interrelationships between spatial locations in the feature map. This detailed modeling of spatial relationships not only strengthens the model's understanding of the spatial information in the image, but also improves the accuracy of image segmentation and obstacle detection, providing more accurate and reliable visual perception capabilities for intelligent vehicles.

Step 3 Output Feature Generation.

After completing the processing of the channel attention module and the spatial attention module, the study obtains two weighted features: the channel attention feature H and the spatial attention feature P. These two features emphasize channel importance and spatial relevance in the input image, respectively. The final goal is to fuse these two features to generate an output feature that combines channel and spatial attention to be available for subsequent processing steps.

1) Re-morphing. Re-morph H and P to match the original input feature dimensions to ensure effective fusion.

2) Feature Fusion. In this study, this is done by weighted summation as shown in Eq. (13).

$$F_3 = \alpha \cdot H + \beta \cdot P \tag{13}$$

where α, β is the fusion weight, which is a pre-defined constant in this study, used to adjust the contribution of channel attention features and spatial attention features in the final output features.

3) The final output characteristics, as shown in Eq. (14).

$$F_3 = Activation(F_3) \tag{14}$$

where $Activation$ is the ReLU nonlinear activation function [36], which is used to increase the nonlinear capability of the model and to help in the learning of subsequent tasks.

The study was able to generate an output feature that synthesizes channel and spatial attention, a feature that effectively improves the accuracy and reliability of obstacle detection in smart vehicles. This not only enhances the model's ability to capture important information in images, but also provides a solid foundation for subsequent image understanding tasks.

In deciding to enhance the DeepLab V3+ algorithm and integrate attention mechanisms, the research comprehensively assessed its advantages over several other advanced 3D object detection and image segmentation technologies such as the 3D-CVF atSPA algorithm, the CLCOs algorithm, the ImvoxelNet algorithm, and the Painted-PointRCNN algorithm. Here are the primary reasons for the enhancements:

1. Comprehensive Performance Optimization: DeepLab V3+, as a mature image semantic segmentation framework, has demonstrated exceptional adaptability and effectiveness in diverse and dynamic real-world application scenarios. By integrating channel and spatial attention mechanisms, the algorithm exhibits greater accuracy and sensitivity in processing critical image information, especially in scenes with significant changes in lighting and perspective. In comparison, although the 3D-CVF at SPA and ImvoxelNet have potential in processing 3D data, their adaptability and effectiveness in the complex task of 2D image segmentation may not match the depth-learning-based DeepLab V3+ algorithm.

2. Precision in Feature Capture: DeepLab V3+ enhances focus on key feature channels in segmentation tasks through its channel attention module, significantly improving the model's ability to capture image details. This fine management of feature channels, while somewhat addressed by the CLCOs and Painted-PointRCNN algorithms, is more systematically and finely realized in DeepLab V3+, leading to superior performance. This capability is particularly important in applications requiring extreme accuracy, such as obstacle detection in intelligent vehicles.

3. Adaptability to Complex Environments: The incorporation of a spatial attention module has optimized the algorithm's understanding of image spatial relationships, enabling DeepLab V3+ to effectively perform image segmentation and object recognition in more complex environments. This feature, compared to the primarily 3D vision information processing-oriented 3D-CVF at SPA and ImvoxelNet algorithms, demonstrates DeepLab V3+'s strong adaptability across a wide range of 2D image processing scenarios.

3.3 LNSST Image Fusion Algorithm

The algorithm enhances the perception of the environment and the accuracy of target detection for intelligent vehicles, and its core process is as follows:

Step 1 Image Input and Decomposition

Assume that the input image I_1 is an RGB image and I_2 is a 2D depth map after the point cloud image is processed. The LNSST decomposition of the input image produces the low frequency subband L and the high frequency subband H, whose decomposition equations are shown in Eq. (15).

$$LNSST(I)=\{L,H\} \tag{15}$$

Step 2 Subband fusion strategy.

In the low-frequency subband fusion strategy, the study uses an energy attribute (EA)-based approach to maintain information and context consistency with the fusion equation shown in Eq. (16).

$$L_{fused}=fuse_{EA}(L_c,L_p) \tag{16}$$

where, L_c is the RGB image low-frequency sub-band, L_p is the 2D depth map low-frequency sub-band, and $fuse_{EA}$ is the fusion function based on the energy attribute, whose equations are shown in Eq. (17).

$$L_{fused}(x,y)=\frac{w_1(x,y)\cdot L_c(x,y)+w_2(x,y)\cdot L_p(x,y)}{w_1(x,y)+w_2(x,y)} \tag{17}$$

where $w_1(x,y)$ and $w_2(x,y)$ are weights calculated based on energy, corresponding to the two subbands at each location (x,y).

In the HF subband fusion strategy, the study uses the modified Laplace sum of energies (SML) approach to enhance the image details with the fusion equation shown in Eq. (18).

$$H_{fused}=fuse_{SML}(H_c,H_p) \tag{18}$$

where, H_c is the high frequency subband of RGB image, H_p is the high frequency subband of 2D depth map, and $fuse_{SML}$ is the fusion function of the modified Laplace energy sum, whose equation is shown in Eq. (19).

$$H_{fused}(x,y)=\begin{cases}H_c(x,y),\ if\ SML_c(x,y)>SML_p(x,y)\\ H_p,\ otherwise\end{cases} \tag{19}$$

where, $SML_c(x,y)$ and $SML_p(x,y)$ are the modified Laplace energy values of the RGB image and the 2D depth map at location (x,y), which are calculated as shown in Eq. (20).

$$SML_i(x,y)=\sum_{k=-N}^{N}\sum_{l=-N}^{N}\left|Laplace(H_i(x+k,y+l))\right| \tag{20}$$

where $Laplace(\cdot)$ is the Laplace operator used to compute the second derivative of the image to highlight the edge portions of the image; N is the size of the neighborhood considered in the computation of $SML_i(x,y)$, i.e., a region of $(2N+1)\times(2N+1)$ around the location (x,y); and H_i denotes the i th high-frequency subband.

Step 3 Fusion subband inversion.

The fused subbands are inverted by LNSST to obtain the final fused image as shown in Eq. (21).

$$I_{fused}=LNSST^{-1}(L_{fused},H_{fused}) \tag{21}$$

The LNSST algorithm produces high-quality fused images by fusing images from different sources - RGB images and 2D depth maps. This algorithm not only optimizes the visual effect, but also improves the accuracy of obstacle detection by reducing noise and clearly outlining objects. The algorithm fuses low-frequency subbands using an energy attribute-based approach, which helps to maintain the overall information and contextual consistency of the fused image, thus ensuring that the smart vehicle can accurately recognize and understand its surroundings. Meanwhile, the algorithm fuses high-frequency subbands using a modified Laplacian energy sum approach, which significantly enhances visual details, such as edges and textures, and thus improves the accuracy of target detection. This feature is particularly important when dealing with complex road environments. The LNSST image fusion algorithm is highly adaptable and can adjust the fusion strategy according to different application scenarios and detection needs, which increases the flexibility of the system.

In selecting the image fusion algorithm for enhancing environmental perception and target detection in smart vehicles, the research also compared it with several other advanced 3D object detection and image segmentation technologies, such as the 3D-CVF at SPA algorithm, the CLCOs algorithm, the ImvoxelNet algorithm, and the Painted-PointRCNN algorithm. Here are the main reasons for this choice:

1. Multisource Information Fusion: The LNSST image fusion algorithm particularly emphasizes extracting and merging information from RGB images and 2D depth maps, a feature less common in other algorithms. For example, while the 3D-CVF at SPA algorithm excels in fusing 2D images and 3D point cloud data, it primarily focuses on three-dimensional vision, which may not handle 2D depth images as comprehensively as the LNSST algorithm.

2. Maintaining Information and Contextual Consistency: LNSST employs a unique low-frequency subband fusion strategy based on an energy attribute method to effectively maintain the overall information and contextual consistency of images. This is a more precise method of information preservation compared to the general processing approach of the CLCOs algorithm, especially crucial in applications involving environmental perception and obstacle detection.

3. Enhancement of Details and Noise Reduction: The strategy of using a modified Laplacian energy sum for high-frequency subband fusion allows the LNSST algorithm to excel in enhancing image details, such as edges and textures. In contrast, although the ImvoxelNet and Painted-PointRCNN algorithms perform well in 3D object detection, they may not match the LNSST algorithm in the precision of enhancing details and noise handling in two-dimensional images.

4. Adaptability and Flexibility: The design of the LNSST image fusion algorithm allows for adjustment of the fusion strategy according to different application scenarios and detection needs. This flexibility is extremely important in practical applications of autonomous vehicles, which must adapt to variable environments. Other algorithms, while performing well under specific conditions, may lack the same adaptability and flexibility in adjustment.

4 EXPERIMENTAL ANALYSIS AND COMPARISON

4.1 Experimental Environment and Dataset

A high-performance Hewlett-Packard (HP) Z8 G4 graphics workstation was used as the host computer for the experiments. The computer is equipped with dual Intel Xeon 5218R CPUs, each with 20 cores running at 2.1 GHz, for a total of 40 cores. Graphics processing power is provided by an A4000 series 16 GB graphics card. System memory is 128 GB and the storage solution consists of a 512 GB solid state drive (SSD) and a 2TB mechanical hard disk. The operating system is linux, and this paper integrates data information from all relevant datasets in Tab. 1, totalling 2852534 images, 2000000 training set images, 427880 test set images, and 424654 validation set images.

Table 1 Data sets

Data set	Year of disclosure	Scene Distribution	Image source
KITTI	2011	Pedestrians and vehicles	KIT
Waymo	2020	Obstacles, traffic lights	google
nuScenes	2019	city street	Motional
Cityscapes	2015	City Scene	Daimler AGR&D
CADC	2020	Snow Driving	utoronto
TLR	2013	traffic signal	La Route Automastisee
Argoverse	2022	subdistrict	Argo AI
cityscapes	2021	City Scene	kaggle

In the experimental setup, to ensure that the model meets real-time performance requirements while achieving optimal performance, the research involved meticulous adjustment of several key hyperparameters. Firstly, considering the effective use of computational resources and the stability of model training, the batch size was set to 32. This value was determined through multiple experiments to ensure the best performance and resource efficiency given the current hardware configuration. Secondly, the initial learning rate of the model was set at 0.01, based on observations from preliminary experiments on model convergence speed and performance. A higher initial learning rate was used to accelerate early training, followed by gradually adjusting the learning rate decay strategy to ensure stability and precision during the model training process. In terms of optimizer selection, Stochastic Gradient Descent (SGD) was used, widely applied due to its efficiency and effectiveness in handling large datasets, and also helpful in preventing overfitting and enhancing the model's generalization capability. The training period was set to 50 epochs to fully utilize the data characteristics while keeping the training within an acceptable timeframe to meet real-time processing needs. Regarding input data processing, all images were resized to 640 × 640 pixels, standardizing the input data to enable the model to process and recognize image targets more efficiently and eliminate potential processing delays caused by inconsistent image sizes. Finally, to shorten the training time and enhance initial performance, pretrained weights were loaded at the beginning of the training. This strategy allows the model to demonstrate better performance early in the training, a key step in enhancing detection efficiency.

4.2 Evaluation Metrics ForFusion Algorithms

In order to comprehensively assess the performance of the fusion algorithm, this paper employs four key evaluation metrics, which collectively take into account the correctness of classification and the accuracy of localization. First, mean accuracy (mAP) is used to measure the overall performance of the model under different thresholds. Second, Precision (Precision) evaluates the model's ability to correctly identify the target, focusing on the accuracy of prediction. Again, Recall (Recall) measures the model's ability to recognize all relevant instances, focusing on coverage. Finally, the $F1$-score ($F1$-score) combines information from Precision and Recall to provide a comprehensive assessment of the overall balance of the model.

The average accuracy is calculated as shown in Eq. (22).

$$mAP = \frac{1}{n}\sum_{i=1}^{n} AP_i \tag{22}$$

where, n is the number of categories and AP_i is the average precision of the i th category, which is defined as shown in Eq. (23).

$$AP_i = \frac{1}{m}\sum_{j=1}^{m} P_{ij} R_{ij} \tag{23}$$

where, m is the number of detection results, P_{ij} is the precision of the th detection result, and R_{ij} is the recall of the j th detection result, which are defined as shown in Eq. (24) and Eq. (25), respectively.

$$P_{ij} = \frac{TP_{ij}}{TP_{ij} + FP_{ij}} \tag{24}$$

$$R_{ij} = \frac{TP_{ij}}{TP_{ij} + FN_{ij}} \tag{25}$$

where, TP_{ij} is the number of samples where the distance between the j th detection result and the true position is less than the threshold; FP_{ij} is the number of samples where the distance between the j th detection result and the true position is greater than the threshold (the classifier mistook some really negative samples as positive samples); FN_{ij} is the number of samples where the distance between the j th detection result and the true position is greater than the threshold (positive samples missed by the classifier).

The $F1$-score represents the reconciled mean of precision and recall, and its formula is shown in Eq. (26).

$$F1_{ij} = 2\frac{P_{ij} R_{ij}}{P_{ij} + R_{ij}} \tag{26}$$

4.3 Obstacle Detection Comparison Test

In order to clearly demonstrate the performance advantages of the obstacle detection algorithms constructed in this study for smart vehicle obstacle detection, a series of experiments were designed in this paper to compare and analyze the algorithms quantitatively and qualitatively. Specifically, several representative algorithms are selected for comparison, including 3D-CVF at SPA algorithm [37], CLCOs algorithm [38], ImvoxelNet algorithm [39] and Painted-PointRCNN algorithm [40]. By comparing these algorithms with the algorithms proposed in this paper, the effectiveness of each algorithm in obstacle detection is thoroughly evaluated. The purpose of such comparative experiments is to validate the effectiveness of this paper's algorithms in the obstacle detection task and to demonstrate their performance advantages over other algorithms.

4.3.1 Comparison of Different Algorithms on Evaluation Indicators

In order to objectively evaluate the detection performance of the algorithms proposed in this paper, four core metrics are used for quantitative analysis: mean accuracy (mAP), precision (Precision), recall (Recall), and F1-score. Together, these metrics measure the algorithms' accuracy and efficiency in obstacle detection. Further, in order to evaluate the model performance more comprehensively, the study calculates the average of the performance of these models on the smart vehicle obstacle detection image dataset, which is used as the final evaluation criterion. Through this approach, this paper is able to provide a comprehensive performance evaluation to ensure the comprehensiveness and reliability of the results. The related experimental results are shown in Tab. 2.

Table 2 Comparison of evaluation indexes of different algorithms for obstacle detection

Comparison Algorithm	Average precision	Comparison Algorithm	Average precision	Comparison Algorithm
3D-CVF atSPA algorithm	0.85	0.82	0.92	0.85
CLCOs algorithm	0.82	0.80	0.80	0.82
ImvoxelNet algorithm	0.88	0.85	0.85	0.87
Painted-PointRCNN algorithm	0.89	0.9	0.88	0.91
model of this paper	0.92	0.92	0.94	0.93

As shown in Tab. 2, the mean Average Precision (mAP) of the model in this paper is 0.92, which is 3.4% better than the best algorithm, Painted-PointRCNN. This advantage is due to the model's more effective fusion of RGB and LiDAR data, enhancing its ability to recognize obstacles in complex environments, especially under various occlusion conditions. Specifically, the model utilizes advanced feature extraction and fusion techniques to extract features from images and point clouds across multiple scales. During the fusion process, an adaptive weighting strategy is applied, allowing data from different modalities to complement each other, thereby enhancing the identification capabilities for target objects. These improvements allow the model to perform exceptionally well in complex urban traffic environments, improving its ability to recognize vehicles, pedestrians, and other road users.

In terms of precision, the model achieved a score of 0.92, which is 2.2% better than the Painted-PointRCNN algorithm. The improvement in precision indicates that the model is more effective in reducing false positives. The model employs advanced feature extraction technologies, including multi-layer convolutional neural networks and attention mechanisms, which can more precisely capture the characteristics of target objects and reduce false detections. Additionally, during the data preprocessing stage, the model applies more refined noise filtering and data enhancement, allowing it to maintain high detection accuracy under different environmental lighting and weather conditions.

In terms of recall, the model achieved a score of 0.94, which is 2.2% better than the 3D-CVF at SPA algorithm. This advantage is attributed to the model's stronger detection capabilities, particularly in handling occlusions and distant targets, ensuring that more true targets are detected. During training, the model used a large amount of data containing various occlusions and distant targets and employed dynamic data enhancement techniques, such as random occlusion and magnification of distant targets, enhancing the model's robustness in real scenarios.

In terms of the $F1$-score, the model achieved a score of 0.93, which is 2.2% better than the Painted-PointRCNN algorithm. The F1-score, a combined metric of precision and recall, reflects the overall superiority of the model in terms of accuracy and completeness. By maintaining high precision and improving recall, the model benefits from its innovative feature fusion strategy and efficient training methods. Furthermore, during the inference process, the model uses post-processing techniques based on graph neural networks, which effectively reduce false positives and false negatives, thereby enhancing overall detection performance.

The main reasons for these performance differences lie in the model's use of more advanced feature extraction and fusion technologies. It extracts features from images and point clouds at multiple scales and applies an adaptive weighting strategy during the fusion process, allowing data from different modalities to complement each other, thereby enhancing the ability to recognize target objects. The fine-tuning of noise filtering and data enhancement techniques during the data preprocessing and enhancement stages has improved the model's robustness and generalization capability across various environments. Moreover, the use of dynamic data enhancement techniques, including extensive training data containing various occlusions and long-distance targets, has bolstered the model's ability to detect in complex environments.

However, the current experimental setup has certain limitations. The experiments are primarily based on public datasets for training and testing. Although these datasets cover a variety of scenarios, the complexity and diversity of autonomous driving scenes in the real world are greater, and there may be some special cases that are not covered. The model's generalization capability needs further verification under different cities, times, and weather conditions. In practical applications, autonomous driving systems need to achieve real-time performance with limited computing resources, which poses higher demands on the model's computational complexity and optimization. To address these challenges, model compression and optimization techniques such as model pruning, quantization, and knowledge distillation could be considered to reduce computational complexity. Real-time performance can be enhanced through distributed computing and hardware acceleration (such as GPUs or TPUs). Furthermore, optimizing the inference process of the algorithm, by reducing unnecessary computational steps, can meet the real-time requirements of practical applications.

4.3.2 Comparison of Different Algorithms in Reasoning Time and Efficiency

Inference time refers to the time required for the model to make predictions on new data after training is completed; a lower inference time indicates that the model is able to process the data and provide the output in a shorter period of time. In this study inference time is measured in milliseconds (ms) as the inference speed of a single frame. In terms of operational efficiency, the number of frames processed per second (FPS) and the computational complexity (GFLOPS) were used as measures. With the same hardware and single-frame reasoning time, a higher number of frames processed per second usually implies a more efficient model, but the level of GFLOPS does not directly determine the efficiency of the model, but rather the balance between GFLOPS and FPS needs to be considered to ensure that the model achieves a reasonable balance between computational intensity and processing speed.

Table 3 Experimental results of different algorithms in terms of inference time and model efficiency

Comparison Algorithm	Reasoning time / ms	FPS	computational complexity (GFLOPS)
3D-CVF atSPA algorithm	120.7	64.8	217.78
CLCOs algorithm	116.2	53.3	186.13
ImvoxelNet algorithm	171.9	47.9	233.51
Painted-PointRCNN algorithm	162.5	58.6	220.71
model of this paper	93.7	62.1	196.98

The related experimental results are shown in Tab. 3. As demonstrated in Tab. 3, the model presented in this paper exhibits outstanding inference time performance, with an impressive 93.7 milliseconds, substantially faster than other algorithms. It is 78.2 milliseconds faster than the slowest ImvoxelNet algorithm and 22.5 milliseconds quicker than the fastest CLCOs algorithm, marking a performance improvement of approximately 45.5% over the ImvoxelNet and about 19.3% over the CLCOs. This indicates that the model is more efficient in processing data and delivering outputs. The model adopts a more efficient algorithmic architecture and optimized computational process, thus reducing the demand for computing resources while ensuring accuracy.

In terms of frames per second (FPS), the model achieves 62.1 FPS, slightly below the best-performing 3D-CVF at SPA algorithm at 64.8 FPS, suggesting that the model's processing speed is comparable to the best-performing algorithms. Although it is slightly lower than the 3D-CVF at SPA, the model's FPS remains higher than other compared algorithms, such as the CLCOs algorithm at 53.3 FPS and the Painted-Point RCNN algorithm at 58.6 FPS, demonstrating the model's superior real-time processing capabilities.

Regarding computational complexity, the model's 196.98 GFLOPS is lower than the 217.78 GFLOPS of the 3D-CVF at SPA algorithm, indicating that the model maintains high performance while being more computationally efficient. This balance allows the model to perform well in terms of inference speed, frames processed per second, and computational complexity, overall outperforming other compared algorithms.

The primary reasons for these performance differences are the model's adoption of a more efficient algorithmic architecture and an optimized computational process. Specifically, during the design phase, the model incorporates multi-layer convolutional neural networks and attention mechanisms, which enhance the efficiency of feature extraction. It also employs an adaptive weighting strategy, reducing redundant computations during the feature fusion process. These improvements not only enhance the detection performance of the model but also significantly reduce its computational complexity.

4.3.3 Comparison of the Effectiveness of Different Algorithms in Obstacle Detection

In this study, three main categories of cars, pedestrians and cyclists were selected for obstacle detection experiments in the integrated dataset [41] in order to evaluate the performance of different algorithms. In order to qualitatively analyze the detection effectiveness of the algorithms, the article shows a visual comparison of the detection results through Fig. 2, Fig. 3, Fig. 4, Fig. 5 and Fig. 6. These images present scenarios of varying complexity: car and person mixing, car and cyclist mixing at an intersection, car and cyclist mixing on a roadway, and a combined car, person, and cyclist mixing scenario, thus demonstrating the detection challenges from simple to difficult.

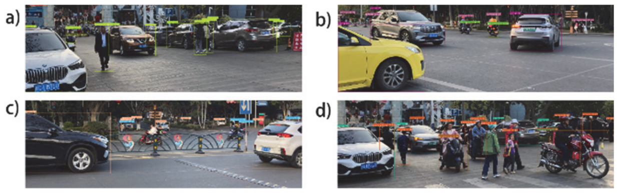

Figure 2 3D-CVF at SPA algorithm obstacle detection results

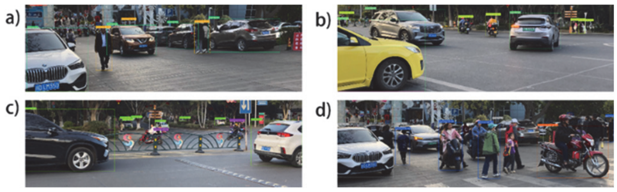

Figure 3 Obstacle detection results of CLCOs algorithm

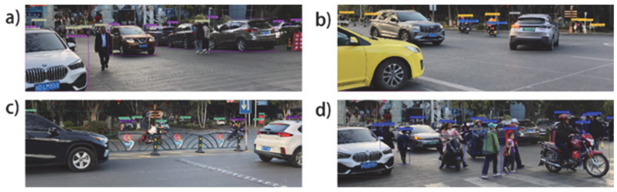

Figure 4 Obstacle detection results of ImvoxelNet algorithm

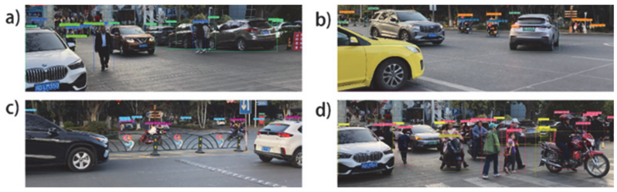

Figure 5 Painted-PointRCNN algorithm obstacle detection results

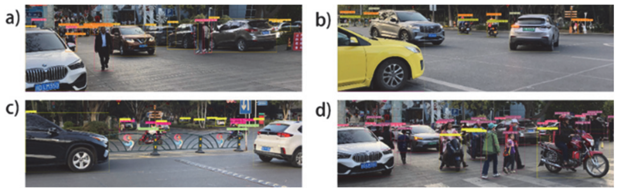

Figure 5 Obstacle detection results of this paper's algorithm

Through comprehensive analysis, as shown in Fig. 6, the algorithm in this paper demonstrates superior performance in complex scenarios, including occlusion and long-distance target detection, clearly outperforming other algorithms. This algorithm employs advanced data fusion techniques and deep learning methodologies, specifically enhanced feature extraction and target recognition mechanisms, which are optimally tailored for complex scenes, occlusions, and distant targets. Therefore, it surpasses other algorithms in detection accuracy and robustness, showing its strong performance in detection tasks.

As indicated in Fig. 3, the CLCOs algorithm exhibits the weakest performance in practical applications, particularly in detecting partially occluded and distant targets. Its deficiencies in fusing multimodal data result in poor adaptability to complex environments.

In contrast, as Fig. 2 and Fig. 4 reveal, while the 3D-CVF at SPA and ImvoxelNet algorithms perform better than CLCOs in some respects, they still have room for improvement in handling partially occluded targets. The 3D-CVF at SPA algorithm, although good at feature extraction, does not ideally manage occluded targets. Meanwhile, the ImvoxelNet algorithm performs well in detecting distant targets but could improve in detecting occluded targets.

As shown in Fig. 5, the Painted-PointRCNN algorithm performs better in detecting occluded targets but still needs to enhance its ability to detect long-distance targets. While the Painted-PointRCNN algorithm utilizes richer feature information to handle occluded targets, its feature extraction and fusion techniques have not yet achieved optimal results for long-distance targets.

Overall, the algorithm discussed in this paper exhibits the best comprehensive performance, primarily due to innovations and optimizations in feature extraction and fusion technologies. These improvements not only enhance the detection's accuracy and robustness but also increase the stability of the algorithm in processing complex scenes and distant targets. Comparative analysis highlights the significant superiority of this algorithm in detection tasks, demonstrating its potential and value in practical applications.

4.4 Obstacle Detection Ablation Experiment

The purpose of this ablation experiment is to evaluate the baseline model, DeepLab V3+ combined with the attention mechanism, using only LiDAR data inputs, and the effectiveness of the algorithms in this paper in the task of smart vehicle obstacle detection. The contribution of each component to the overall model performance is understood by systematically removing different parts of the model. Where the baseline model is the original DeepLab V3+ algorithm without any additional improvements or modifications, this model will be used as a benchmark for performance comparison. DeepLab V3 + incorporates the attention mechanism, i.e., an improved version of the attention mechanism integrated into the DeepLab V3+ model. LiDAR-only data input, i.e., in this experimental setup, the model will use only LiDAR data and not data from the camera. A comparison of the above models in terms of evaluation metrics is shown in Tab. 4.

Table 4 Comparative experimental results of different models in obstacle detection evaluation indexes

Comparison Algorithm	Average precision (mAP)	Accurate (Precision)	recall rate (Recall)	*F*1-score
baseline model	0.68	0.66	0.71	0.64
DeepLab V3 +				
Combining attention mechanisms	0.76	0.78	0.80	0.73
LiDAR data input only	0.62	0.61	0.67	0.64

As can be seen from Tab. 4, in terms of average accuracy, the high mAP of the model in this paper is due to the combined use of image and LiDAR data, as well as advanced fusion techniques and attention mechanisms. This multimodal fusion strategy improves the model's ability to recognize obstacles in complex environments, especially under different weather conditions. TheDeepLab V3 + combined with the attention mechanism model improves the accuracy of obstacle recognition by focusing on key features and regions in the image, which reduces the interference of the background noise but is still far less effective than the multisensor usage scenario. The lower mAP of the baseline model and the LiDAR-only data input model stems from the lack of fine-grained processing of complex scenes and the lack of sufficient contextual information to support accurate obstacle recognition. The comparison in terms of accuracy also highlights the advantages of this paper's model because of its structural optimization and algorithmic tuning, especially in terms of accurate edge detection and target localization, as well as effective noise suppression, which contribute to the reduction of false positives. DeepLab V3 + combined with Attention Mechanism. The model improves the sensitivity to the important features through the Attention Mechanism, which results in the reduction of false positives while maintaining a high level of accuracy, but the precision is lower than this paper's model because the model also needs to deal with more complex background information. The results of the recall rate show that the model in this paper is able to detect most of the actually existing obstacles, which is attributed to its strong feature extraction capability and effective fusion of different data sources, which enables the model to accurately recognize obstacles even in occluded and complex scenes. The low recall of other models is due to the fact that in some scenes, the model fails to recognize all relevant obstacles, especially targets that are difficult to distinguish in the image. The F1-score integrates precision and recall, providing a measure of the overall performance of the model. The high F1-score of the model in this paper reflects an excellent balance between precision and recall, which is due to the model's use of a multi-level, multi-angle processing strategy that effectively balances the need for detection speed and accuracy. DeepLab V3 + combined with the Attention Mechanism model, although it has made progress in improving recall and precision, may be due to an increase in the complexity of the model, which leads to a decrease in some cases of Processing efficiency decreases, which affects the F1-score. On the whole, the model in this paper performs better in all the indexes, thanks to its advanced algorithm design, efficient data fusion strategy, and the application of the attention mechanism, which together enhance the model's ability to recognize obstacles, especially in complex and changing environmental conditions. The lack of performance of the other models, on the other hand, stems from the limitation of data processing capability, the simplification of the model structure, or the lack of effective feature extraction and fusion mechanisms.

Table 5 Experimental results of different models in terms of inference time vs. model efficiency

Comparison Algorithm	Reasoning time / ms	FPS	computational complexity (GFLOPS)
baseline model	90.2	58.7	215.23
DeepLab V3 +Combining attention mechanisms	100.2	54.6	223.49
LiDAR data input only	83.2	68.3	189.11
model of this paper	93.7	62.1	196.98

As can be seen in Tab. 5, the LIDAR-only input model performs best in inference time, which indicates that the model is able to complete the prediction faster when processing LIDAR-only data due to the fact that LIDAR data has a simpler structure and is more straightforward to process compared to image data. The reason why the DeepLab V3 + model combined with the attentional mechanism has the longest inference time is because of the increased computational burden on the model due to the introduction of the attentional mechanism, although this helps to improve the model's recognition accuracy. The reasoning time of DeepLab V3 + combined with attention mechanism is the longest because the introduction of the attention mechanism increases the computational burden of the model, although it helps to improve the recognition accuracy of the model. The inference time of the model in this paper is 93.7ms, which is slightly lower than the inference time of the LIDAR-only data input, which indicates that the model maintains a good performance while keeping the inference efficiency within a good range. In terms of FPS, the LiDAR data input only model performs the highest, further demonstrating that the model is able to make predictions at a higher frequency when focusing on processing a single piece of radar data, which is more important for applications that require real-time or near-real-time processing. The DeepLab V3 + combined with Attention Mechanism model has the lowest FPS, which is in line with its longer inference time, and reflects the effect of the increased model complexity on the processing speed. The model in this paper performs better with 62.1 FPS, which maintains a high processing rate despite the introduction of a complex processing mechanism, indicating the effectiveness of the optimization strategy. In terms of computational complexity, the DeepLab V3 + combined with attention mechanism model is high, indicating that the model requires more computational resources while making high accuracy predictions. The LiDAR data input only model is the lowest, and the model structure is more efficient due to the relative simplicity of the LiDAR data. The model in this paper is slightly lower compared to the baseline model, indicating that the model in this paper focuses on optimizing computational efficiency while achieving high performance. In summary, the LiDAR data input only model performs optimally in terms of speed, but may have limited recognition ability in complex scenes. The DeepLab V3 + combined with the attention mechanism model sacrifices some of its efficiency in order to improve prediction accuracy. The model in this paper, on the other hand, ensures higher prediction accuracy while also taking into account inference speed and computational efficiency, demonstrating a balanced performance.

6 REFERENCES

[1] Huang, Z., Wu, J., & Lv, C. (2023). Efficient deep reinforcement learning with imitative expert priors for autonomous driving. *IEEE Transactions on Neural Networks and Learning Systems, 34*(10), 7391-7403. https://doi.org/10.1109/TNNLS.2022.3142822

[2] Janai, J., Güney, F., Behl, A., & Geiger, A. (2020). Computer vision for autonomous vehicles: Problems, datasets and state of the art. *Foundations and Trends® in Computer Graphics and Vision, 12*(1-3), 1-308. https://doi.org/10.1561/0600000079

[3] Adiuku, N., Avdelidis, N. P., Tang, G., & Plastropoulos, A. (2024). Improved hybrid model for obstacle detection and avoidance in robot operating system framework (rapidly exploring random tree and dynamic windows approach). *Sensors (Basel, Switzerland), 24*(7). https://doi.org/10.3390/s24072262

[4] Adnan, M., Slavic, G., Martin Gomez, D., Marcenaro, L., & Regazzoni, C. (2023). Systematic and comprehensive review of clustering and Multi-Target Tracking techniques for LiDAR point clouds in autonomous driving applications. *Sensors (Basel, Switzerland), 23*(13), 6119. https://doi.org/10.3390/s23136119

[5] Wen, L., Peng, Y., & Lin, M. (2024). Multi-Modal Contrastive Learning for LiDAR Point Cloud Rail-Obstacle Detection in Complex Weather. *Electronics, 13*(1). https://doi.org/10.3390/electronics13010220

[6] Wang, N., Shi, C., Guo, R., Lu, H., Zheng, Z., & Chen, X. (2023). InsMOS: Instance-aware moving object segmentation in LiDAR data. *In arXiv [cs.CV]*. https://doi.org/10.1109/IROS55552.2023.10342277

[7] Xie, G., Zhang, J., Tang, J., Zhao, H., Sun, N., & Hu, M. (2021). Obstacle detection based on depth fusion of lidar and radar in challenging conditions. *The Industrial Robot, 48*(6), 792-802. https://doi.org/10.1108/ir-12-2020-0271

[8] Zhang, C. Y., Chen, Z. H., & Han, L. (2021). Obstacle detection of Lidar based on Improved DBSCAN algorithm. Laser & Optoelectronics Progress, 58(24).

[9] Yoo, J. H., Kim, Y., Kim, J., & Choi, J. W. (2020a). 3D-CVF: Generating joint camera and LiDAR features using cross-view spatial feature fusion for 3D object detection. *Computer Vision ECCV 2020*, *Springer International Publishing,* 720-736. https://doi.org/10.1007/978-3-030-58583-9_43

[10] Dou, J., Xue, J., & Fang, J. (2019). SEG-VoxelNet for 3D Vehicle Detection from RGB and LiDAR Data. *International Conference on Robotics and Automation (ICRA)*. https://doi.org/10.1109/ICRA.2019.8793492

[11] Pang, S., Morris, D., & Radha, H. (2022). Fast-CLOCs: Fast camera-LiDAR object candidates fusion for 3D object detection. *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. https://doi.org/10.1109/WACV51458.2022.00380

[12] Mai, N. A. M., Duthon, P., Khoudour, L., Crouzil, A., & Velastin, S. A. (2021). Sparse LiDAR and stereo fusion (SLS-fusion) for depth estimation and 3D object detection. *11th International Conference of Pattern Recognition Systems (ICPRS 2021)*. https://doi.org/10.1049/icp.2021.1442

[13] Rukhovich, D., Vorontsova, A., & Konushin, A. (2022). ImVoxelNet: Image to voxels projection for monocular and multi-view general-purpose 3D object detection. *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. https://doi.org/10.1109/WACV51458.2022.00133

[14] Yao, S., Guan, R., Huang, X., Li, Z., Sha, X., Yue, Y., Lim, E. G., Seo, H., Man, K. L., Zhu, X., & Yue, Y. (2024). Radar-camera fusion for object detection and semantic segmentation in autonomous driving: A comprehensive review. *IEEE Transactions on Intelligent Vehicles, 9*(1), 2094-2128. https://doi.org/10.1109/tiv.2023.3307157

[15] Liu, Z., Cai, Y., Wang, H., Chen, L., Gao, H., Jia, Y., & Li, Y. (2022). Robust target recognition and tracking of self-driving cars with radar and camera information fusion under severe weather conditions. *IEEE Transactions on Intelligent Transportation Systems: A Publication of the IEEE Intelligent Transportation Systems Council, 23*(7), 6640-6653. https://doi.org/10.1109/tits.2021.3059674

[16] Jisen, W. (2021). A study on target recognition algorithm based on 3D point cloud and feature fusion. *IEEE 4th International Conference on Automation, Electronics and Electrical Engineering (AUTEEE)*. https://doi.org/10.1109/AUTEEE52864.2021.9668653

[17] Li, H., Zhang, J., & Wang, J. (2023). Extracting Citrus in Southern China (Guangxi Region) Based on the Improved DeepLabV3+ Network. *Remote Sensing, 2023*. https://doi.org/10.3390/rs15235614

[18] Rongrong, L. & Dongzhi, H. (2021). Semantic segmentation method of road scene based on Deeplabv3+ and attention mechanism. *IEEE 4th Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC).*

[19] Chen, W., Tian, W., Xie, X., & Stork, W. (2022). RGB image- and lidar-based 3D object detection under multiple lighting scenarios. *Automotive Innovation, 5*(3), 251-259. https://doi.org/10.1007/s42154-022-00176-2

[20] Wang, Z., Wang, J., Yang, K., Wang, L., Su, F., & Chen, X. (2022). Semantic segmentation of high-resolution remote sensing images based on a class feature attention mechanism fused with Deeplabv3+. *Computers & Geosciences, 158*(104969). https://doi.org/10.1016/j.cageo.2021.104969

[21] Shaheed, K., Mao, A., Qureshi, I., Kumar, M., Hussain, S., Ullah, I., & Zhang, X. (2022). DS-CNN: A pre-trained Xception model based on depth-wise separable convolutional neural network for finger vein recognition. *Expert Systems with Applications, 191*(116288). https://doi.org/10.1016/j.eswa.2021.116288

[22] Shaheed, K., Mao, A., Qureshi, I., Kumar, M., Hussain, S., Ullah, I., & Zhang, X. (2022). DS-CNN: A pre-trained Xception model based on depth-wise separable convolutional neural network for finger vein recognition. *Expert Systems with Applications, 191*(116288). https://doi.org/10.1016/j.eswa.2021.116288

[23] Cai, C., Tan, J., Zhang, P., Ye, Y., & Zhang, J. (2022). Determining strawberries' varying maturity levels by utilizing image segmentation methods of improved DeepLabV3+. *Agronomy (Basel, Switzerland), 12*(8), 1875. https://doi.org/10.3390/agronomy12081875

[24] Sakib, S. (2022). LiDAR Technology-An Overview. *IUP Journal of Electrical & Electronics Engineering, 2022*(1).

[25] Ma, W., Wang, K., Li, J., Yang, S. X., Li, J., Song, L., & Li, Q. (2023). Infrared and visible image fusion technology and application: A review. *Sensors (Basel, Switzerland), 23*(2), 599. https://doi.org/10.3390/s23020599

[26] Tang, W., He, F., Liu, Y., & Duan, Y. (2022). MATR: Multimodal medical image fusion via multiscale adaptive Transformer. *IEEE Transactions on Image Processing: A Publication of the IEEE Signal Processing Society, 31*, 5134-5149. https://doi.org/10.1109/TIP.2022.3193288

[27] Cunha, D., Zhou, A. L., & Do, J. (2006). The nonsubsampled contourlet transform: theory, design, and applications. *IEEE Transactions on Image Processing, 15*, 3089-3101. https://doi.org/10.1109/TIP.2006.877507

[28] Guo, K. & Labate, D. (2007). Optimally sparse multidimensional representation using shearlets. *SIAM Journal on Mathematical Analysis, 39*(1), 298-318. https://doi.org/10.1137/060649781

[29] Zhao, X., Chen, J., Said, A., Seregin, V., Egilmez, H. E., & Karczewicz, M. (2016). NSST: Non-separable secondary transforms for next generation video coding. *2016 Picture Coding Symposium (PCS).* https://doi.org/10.1109/PCS.2016.7906344

[30] Cheng, B., Jin, L., & Li, G. (2018). Infrared and visual image fusion using LNSST and an adaptive dual-channel PCNN with triple-linking strength. *Neurocomputing, 310*, 135-147. https://doi.org/10.1016/j.neucom.2018.05.028

[31] Yang, L. & Cai, H. (2024). Unsupervised video object segmentation for enhanced SLAM-based localization in dynamic construction environments. *Automation in Construction, 158*. https://doi.org/10.1016/j.autcon.2023.105235

[32] Khan, S. & Khan, A. (2022). FFireNet: Deep learning based forest fire classification and detection in smart cities. *Symmetry, 14*(10). https://doi.org/10.3390/sym14102155

[33] Du, Y., Zuo, X., Liu, S., Cheng, D., Li, J., Sun, M., Zhao, X., Ding, H., & Hu, Y. (2023). Segmentation of pancreatic tumors based on multi-scale convolution and channel attention mechanism in the encoder-decoder scheme. *Medical Physics, 50*(12), 7764-7778. https://doi.org/10.1002/mp.16561

[34] Su, E., Cai, S., Xie, L., Li, H., & Schultz, T. (2022). STAnet: A spatiotemporal attention network for decoding auditory spatial attention from EEG. *IEEE Transactions on Bio-Medical Engineering, 69*(7), 2233-2242. https://doi.org/10.1109/TBME.2022.3140246

[35] Cheng, J., Tian, S., Yu, L., Gao, C., Kang, X., Ma, X., Wu, W., Liu, S., & Lu, H. (2022). ResGANet: Residual group attention network for medical image classification and segmentation. *Medical Image Analysis, 76*(102313). https://doi.org/10.1016/j.media.2021.102313

[36] Varshney, M. & Singh, P. (2021). Optimizing nonlinear activation function for convolutional neural networks. *Signal, Image and Video Processing, 15*(6), 1323-1330. https://doi.org/10.1007/s11760-021-01863-z

[37] Sheng, H., Cai, S., Liu, Y., Deng, B., Huang, J., Hua, X.-S., & Zhao, M.-J. (2021). Improving 3D object detection with channel-wise Transformer. *In arXiv [cs.CV].* https://doi.org/10.1109/ICCV48922.2021.00274

[38] Müller, M. M., Said, R. S., Jelezko, F., Calarco, T., & Montangero, S. (2022). One decade of quantum optimal control in the chopped random basis. Reports on Progress in Physics. *Physical Society (Great Britain), 85*(7). https://doi.org/10.1088/1361-6633/ac723c

[39] Rukhovich, D., Vorontsova, A., & Konushin, A. (2022). ImVoxelNet: Image to voxels projection for monocular and multi-view general-purpose 3D object detection. *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV).* https://doi.org/10.1109/WACV51458.2022.00133

[40] Mahmoud, A. & Waslander, S. L. (2021). Sequential fusion via bounding box and motion Point Painting for 3D objection detection. *18th Conference on Robots and Vision (CRV).* https://doi.org/10.1109/CRV52889.2021.00013

[41] Atitallah, A. B., Said, Y., & Atitallah, M. A. B. (2023). Embedded implementation of an obstacle detection system for blind and visually impaired persons' assistance navigation. *Computers and Electrical Engineering, 108*. https://doi.org/10.1016/j.compeleceng.2023.108714

Contact information:

Qianying ZOU, PhD, Professor
(Corresponding author)
Geely University of China,
No. 123, Section 2, Chengjian Avenue, East New District, Chengdu, Sichuan Province
E-mail: zqy_bb@163.com

Fengyu LIU, Teaching Assistant
Geely University of China,
No. 123, Section 2, Chengjian Avenue, East New District, Chengdu, Sichuan Province
E-mail: i@liufengyu.cn

Chen RUIXIN, Student
Geely University of China,
No. 123, Section 2, Chengjian Avenue, East New District, Chengdu, Sichuan Province
E-mail: 1870984969@qq.com