Computers, Materials & Continua

Tech Science Press

Doi:10.32604/cmc.2025.062963

ARTICLE

CFH-Net: Transformer-Based Unstructured Road-Free Space Detection Network

Jingcheng Yang[1], **Lili Fan**[2] **and Hongmei Liu**[1,*]

[1]School of Science, Dalian Minzu University, Dalian, 116000, China
[2]School of Automation, Beijing Institute of Technology, Beijing, 100081, China

*Corresponding Author: Hongmei Liu. Email: liuhm@dlnu.edu.cn

Received: 31 December 2024; Accepted: 01 March 2025; Published: 19 May 2025

ABSTRACT: With the advancement of deep learning in the automotive domain, more and more researchers are focusing on autonomous driving. Among these tasks, free space detection is particularly crucial. Currently, many model-based approaches have achieved autonomous driving on well-structured urban roads, but these efforts primarily focus on urban road environments. In contrast, there are fewer deep learning methods specifically designed for off-road traversable area detection, and their effectiveness is not yet satisfactory. This is because detecting traversable areas in complex outdoor environments poses significant challenges, and current methods often rely on single-image inputs, which do not align with contemporary multimodal approaches. Therefore, in this study, we propose a CFH-Net model for off-road traversable area detection. This model employs a Transformer architecture to enhance its capability of capturing global information. For multimodal feature extraction and fusion, we integrate the CM-FRM module for feature extraction and introduce the novel FFX module for feature fusion, thereby improving the perception capability of autonomous vehicles on unstructured roads. To address upsampling, we propose a new convolution precorrection method to reduce model parameters and computational complexity while enhancing the model's ability to capture complex features. Finally, we conducted experiments on the ORFD off-road dataset and achieved outstanding results. The code is available at: https://github.com/qka1991/CFH-Net (accessed on 11 January 2025).

KEYWORDS: Freespace detection; semantic segmentation; terrain perception; unstructured environment

1 Introduction

The widespread deployment of Autonomous Vehicle Systems (AVS) faces the challenge of making safe, intelligent, and socially compatible decisions [1]. Achieving autonomous driving requires enabling vehicles to navigate without explicit maps or rules, demanding systems to make human-like decisions by understanding their environment [2]. A key aspect of this is improving the vehicle's perceptual ability to detect traversable areas, a topic that has received significant attention. However, AVs must operate not only in structured urban roads but also in unstructured environments such as off-road terrains. While urban traversable areas are often defined as obstacle-free, paved surfaces [3,4], off-road environments present additional challenges, such as unpredictable weather and obstacles like weeds and tree stumps. Particularly in structured road scenes and unstructured off-road terrains, distinct concepts exist, as depicted in Fig. 1a. For instance, in structured roads, free space primarily concentrates on urban roadways, while in rural off-road scenarios, the concept of traversable areas is relatively vague, presenting challenges in our work. Despite extensive research on structured urban roads, fewer studies address rural, mountain, or off-road terrains [5], and applying existing models to these environments often leads to suboptimal results. Contemporary models in traversable area detection utilize multimodal fusion techniques to enhance performance by combining data from different modalities. These approaches can be divided into two main categories: (1) single-network fusion, where features from RGB and other modalities are combined at the input stage [6,7], and (2) multi-network fusion, where independent networks process RGB and other modalities before combining the features for semantic prediction [8–10]. Given the complexities of off-road environments, this study employs the second approach to leverage cross-modal interactions, using RGB and radar point cloud data for enhanced detection. A key component of this approach is the CMR-FRM module for feature extraction and the novel FFX module for interactive feature fusion. Recent studies have also contributed valuable insights to this area of research [11,12]. The proposed CFH-Net model, designed for off-road free space detection, is evaluated on the ORFD dataset, achieving superior results. The CM-FRM module enhances feature extraction, and the FFX module improves multimodal fusion, boosting the vehicle's perception on unstructured roads. Furthermore, the encoder is redesigned with a convolutional precoding approach to reduce computational complexity and improve feature fusion. This results in a more efficient and effective model, advancing the field of off-road AVs.

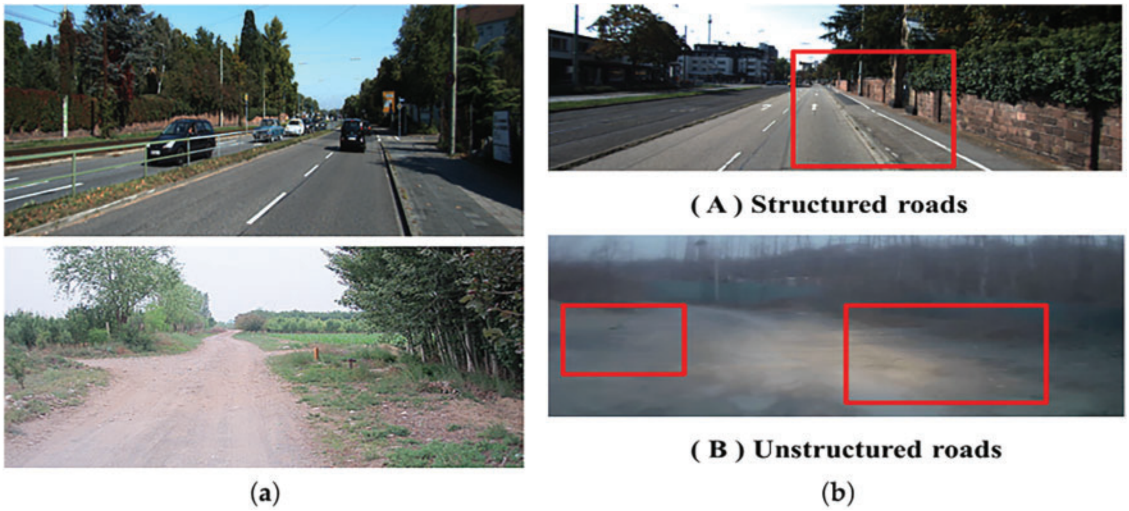

Figure 1: (a) Comparison between structured (urban [13]) and unstructured (off-road [14]) environments. (b) Red boxes highlight lane markings and vegetation interference, respectively

However, existing Transformer-based methods [15,16] remain limited in addressing off-road challenges such as ambiguous boundaries and sparse features. To bridge this gap, wepropose CFH-Net with three key innovations: (1) The CM-FRM module introduces dual-path channel-spatial rectification, dynamically suppressing noise in LiDAR and RGB modalities through cross-modal calibration, unlike the simplistic feature addition in OFF-Net [14]; (2) The FFX module employs efficient attention for global-context fusion, replacing traditional cross-attention to reduce computational overhead while enhancing distant-view perception; (3) In the upsampling process, a list of tensors is stacked and summed along the first dimension, replacing the traditional linear transformation approach. This design avoids unnecessary linear transformations and reshaping steps, thus reducing model complexity and minimizing the risk of overfitting. Moreover, direct feature fusion effectively preserves the original information, enhancing the model's expressive capacity.

In conclusion, the primary contributions of this paper are as follows:

- To address challenges such as decreased segmentation accuracy due to the diverse and complex environments in off-road traversable area detection, unclear segmentation in autonomous driving, and poor long-range road visibility, this paper proposes the CFH-Net model. This model aims to enhance the extraction of modal features from RGB and depth images, where integration is often insufficient. By applying the CFH-Net model, this study ultimately enhances the perception capabilities of autonomous vehicles in off-road environments.

- Regarding feature extraction, this paper utilizes the CM-FRM module instead of the baseline model's cross-attention mechanism to meet the more intricate feature extraction requirements in off-road scenarios. In the feature fusion aspect, the paper introduces the FFX module for feature fusion, aiming to enhance the fusion capability of environmental features in the second stage, thereby improving road detection performance in off-road environments. Lastly, the paper redesigns the decoder section using a hierarchical upsampling approach. This method employs multiple linear transformation layers to further intensify feature information while reducing parameter count and computational complexity.
- Considering the requirements of off-road free space detection tasks, the designed model is applied to the ORFD dataset and achieves outstanding results. The innovation presented not only enhances the baseline model but also attains optimal performance, thus making positive contributions to the research field of unstructured roadways.

2 Realted Work

2.1 Off-Road Environments and Adverse Conditions

Improving the perception capabilities of autonomous vehicles, particularly in off-road environments, is critical. However, models often underperform in these settings due to environmental challenges like vegetation, tree stumps, and sandy terrain, which hinder segmentation accuracy. Existing algorithms have not been sufficiently tested on unstructured road conditions, resulting in degraded performance when applied directly to such environments. These challenges include varying lighting, road uncertainty, and complex conditions such as heavy fog, storms, and low visibility, which degrade segmentation results. The definition of road boundaries is less clear under these conditions, as shown in Fig. 1b, making it difficult for models to perform effectively. Moreover, research has primarily focused on urban structured roads, with fewer studies addressing rural, mountainous, or highway environments [5]. For off-road free space detection, various deep learning architectures have been proposed, notably CNN-based networks like PSPNet [17], HRNetV2+OCR [18], and BiSeNetV2 [19]. While CNNs perform well in structured urban environments, their performance deteriorates in unstructured areas due to unclear boundaries and overlapping regions. Moreover, CNNs' limited receptive field reduces their effectiveness for traversable area detection tasks [14]. Transformer-based methods [15,16] show better performance but come with higher computational costs, prompting a shift toward more efficient Transformer-based models [20,21]. While these models excel on structured datasets like Cityscapes [22] and KITTI [23], they struggle with off-road free space detection due to ambiguous boundary definitions. Methods like Global Convolution Networks [24] struggle with off-road datasets where complex boundaries are less common, limiting their performance. Recently, the OFF-Net method [14] has addressed some of these issues but still struggles with road boundary problems, misinterpreting road elements (e.g., mistaking tree trunks for roads) and providing insufficiently smooth segmentation edges and coverage of distant scenes. In conclusion, while Transformer-based semantic segmentation models demonstrate excellent contextual awareness and computational efficiency, they have not effectively addressed challenges in off-road environments. To overcome these limitations, this paper adopts Zhang's Transformer architecture [14] and introduces the CFH-Net method, specifically designed to improve off-road traversable area detection under challenging conditions.

2.2 LiDAR and Camera Fusion Techniques

In recent years, multimodal fusion methods for autonomous driving perception have advanced significantly [25–27]. Perception is crucial for autonomous vehicles [28–30], and multimodal fusion has become a key approach to improving vehicle perception. Early single-modal techniques, such as those relying solely on cameras or LiDAR, suffer from limitations in off-road scenarios. Camera-based detectors often provide insufficient environmental information, especially in complex scenes or low-light conditions. Likewise, LiDAR systems face mechanical constraints, including varying resolutions and susceptibility to extreme weather [31,32]. However, fusing LiDARandcamera data has proven to enhance perception [32], leveraging the complementarity of the two modalities to address off-road challenges. Methods such as [6,7,26] demonstrate the potential of multimodal feature learning in shared networks, making fusion techniques essential for off-road free space detection tasks. The schematic diagram of our proposed model is illustrated in Fig. 2.

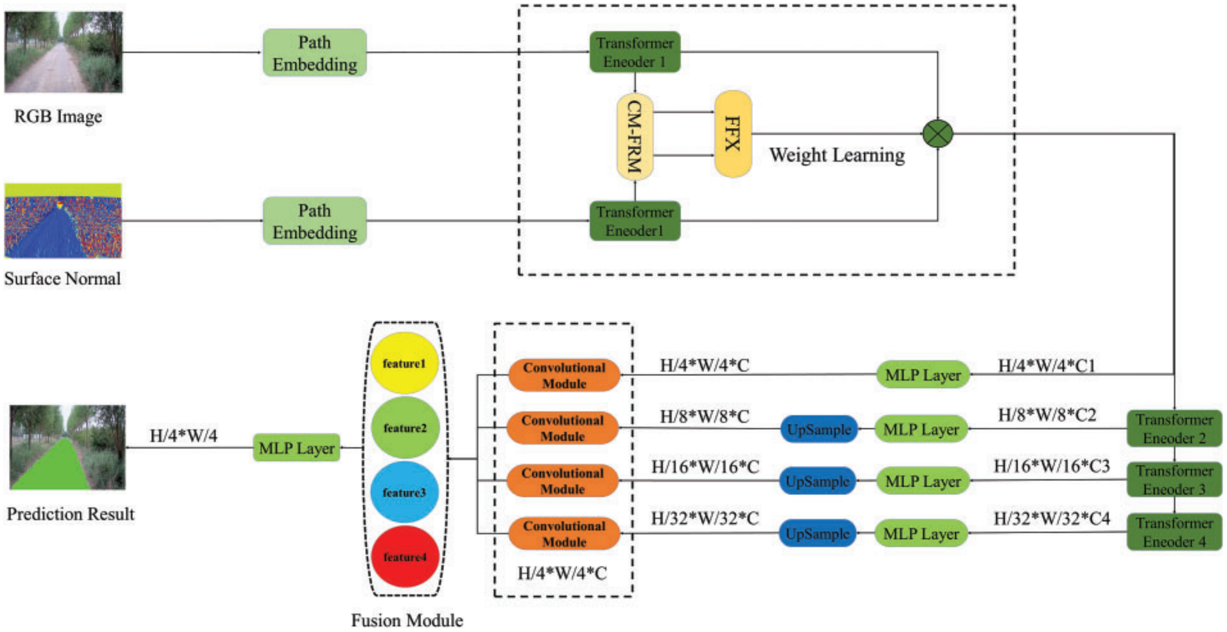

Figure 2: Off-road model diagram. The Transformer encoder extracts features from both the RGB image and surface normals, while the Transformer decoder predicts the free space output. CM-FRM and FFX are designed to fuse data from cameras and LiDAR, dynamically leveraging the advantages of each modality

Recent research on multimodal fusion for semantic segmentation has focused on integrating radar and visual features. Approaches like RGB-D Fusion Network (RDFNet) [33] and others have demonstrated success in fusing RGB and depth images to enhance semantic segmentation accuracy [34–36]. However, off-road environments present challenges such as boundary ambiguities and vegetation misclassifications. To address these, the SNE-RoadSeg model [37] integrates surface normal information with depth images to improve free-space detection. Surface normals, being consistent on the same road plane, are more easily identified than depth information [38], offering advantages for off-road tasks. This paper also incorporates surface normals as one modality in the fusion process to address these challenges. Upon evaluating the benchmark model, we found that its cross-attention component was overly simplistic, lacking in-depth consideration of the complex interrelationships between inputs. The direct addition of inputs followed by simple fully connected layers and a Sigmoid function inadequately captured the intricate dependencies, limiting the model's ability to learn effective features. This resulted in poor performance, particularly in misclassifying vegetation. To resolve this, we propose enhancements with the CM-FRM and FFX modules for more efficient feature extraction and fusion. Additionally, a new upsampling module is introduced to reduce model complexity and computational load, improving overall performance. These methods effectively address challenges such as poor road panoramic view performance and inaccuracies in model predictions, enhancing precision in free-space detection tasks.

3 Method

This study builds on the problem definition of Min et al. [14], proposing the CFH-Net network, which integrates camera and LiDAR data, including surface normal information from LiDARpoint clouds. Inspired by Fan et al. [39], surface normals are used as network inputs, aiding feature extraction and fusion with RGB images. Surface normals are calculated from dense depth images using Fan et al.'s method. Due to the large receptive field requirement for traversable region detection and the limited receptive fields of traditional CNNs, we adopted a Transformer-based architecture, optimizing feature extraction based on Chen et al.'s Transformer network. To address practical challenges, we introduce the CM-FRM module [40], along with the Feature Fusion Expert (FFX) module and an upsampling hierarchical method, enhancing feature extraction and model adaptability. Experimental results show significant improvements. Compared to methods like 2DPASS [41] and MSeg3D [42], our approach more effectively handles complex 3D terrain in off-road environments by leveraging surface normal features from LiDAR and combining them with a Transformer-based framework. While 2DPASS and MSeg3D perform well in structured environments, they may fail to capture 3D spatial information in off-road settings. Our network directly utilizes LiDAR geometric features (e.g., surface normals), enhancing multi-sensor data fusion and adapting to complex off-road terrains. Thus, our method excels in off-road traversable region detection, particularly for tasks involving large-scale, complex terrains.

3.1 Backbone Network

3.1.1 Encoder Network

To obtain multi-level features, the RGB image and surface normals [39] are first embedded into patches, with a resolution of H × W × 3 each. The output of patch embedding is hierarchical feature maps, with corresponding spatial resolutions of 1/4, 1/8, 1/16, and 1/32 of the input size. Utilizing a multi-head self-attention mechanism with Q, K, and V as heads, the interrelations between feature maps are computed. The mathematical formulation [14] is as follows:

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_{head}}}\right)V \tag{1}$$

This results in the final feature representation after multi-level feature processing and self-attention calculation. Next, we adopt the same method as [14] to use 3 × 3 convolutions to capture positional information of the Transformer encoder. The core formula for positional encoding [14] is as follows:

$$x_{out} = MLP\left(GELU\left(Conv_{3\times 3}\left(MLP\left(x_{in}\right)\right)\right)\right) + x_{in} \tag{2}$$

Here, x_{in} represents the features from the multi-head self-attention part, $GELU$ denotes the activation function proposed by Hendrycks [43], MLP refers to the fully connected neural network layer, and $Conv_{3\times 3}$ represents the 3 × 3 convolutional layer. As mentioned earlier, LiDAR point cloud data and RGB images offer complementary strengths in environmental perception. LiDAR provides precise geometric information but lacks detailed semantic context, while RGB images offer rich semantic understanding but may have accuracy limitations. These two types of data come with different noise characteristics. However, by using features from one modality to filter and calibrate the other, the effect of noise can be significantly reduced. To optimize the contributions from both modalities and improve detection performance, we developed a dynamic fusion module, as shown in Figs. 3 and 4.

Through our analysis, we believe that the cross-attention mechanism in the benchmark model is too simplistic for handling complex feature extraction and fusion. Therefore, we propose the Cross-Modal Feature Refinement Module (CM-FRM) [40], which refines features from parallel data streams at each stage of the extraction process. Additionally, we introduce the Feature Fusion and Exchange (FFX) module, which receives the refined features from the CM-FRM and performs the fusion process. The following sections provide a detailed explanation of how these modules function.

3.1.2 CM-FRM

In our research, we identified limitations in the cross-attention mechanism of the benchmark model, particularly in its ability to capture fine-grained features, especially along the edges of prediction maps and in distant regions. The original approach first combines RGB image features with surface normal features of LiDAR point clouds. These superimposed features are then passed through a Multi-Layer Perceptron (MLP) layer with a sigmoid activation function for cross-attention learning. While this cross-attention mechanism effectively learns the weights for each modality, it proves insufficient in refining fine-grained details. To address these issues, we introduced the Cross-Modal Feature Refinement Module (CM-FRM), which performs feature refinement across parallel streams at each stage of feature extraction, as depicted in Fig. 3 of the model diagram. This module enhances the refinement process and improves feature representation, particularly for fine-grained details. In the Channel-wise feature rectification part, to address noise and uncertainties in different modalities, CM-FRM processes features along two dimensions. This paper adopts the method proposed by Zhang et al. [40], which embeds the bimodal features $RGB_{in} \in R^{H \times W \times C}$ and $X_{in} \in R^{H \times W \times C}$ into two attention vectors along the spatial axis. Subsequently, global max-pooling and global average-pooling are applied along the channel dimension to RGB_{in} and X_{in} simultaneously to retain more information. The four resultant vectors are concatenated to form $Y \in R^{4C}$. Then, an MLP is applied followed by a sigmoid function to obtain $W^C \in R^{2\overline{C}}$, which is split into W^C_{RGB} and W^C_X:

$$W^C_{RGB}, W^C_X = split(\sigma(Fmlp(Y))) \tag{3}$$

Similarly to channel-level rectification, the formulation for spatial-level rectification is as follows:

$$RGB^S_{rec} = W^S_X * X_{in} \tag{4}$$

$$X^S_{rec} = W^S_{RGB} * RGB_{in} \tag{5}$$

where $*$ denotes spatial multiplication.

In the Spatial-wise feature rectification part, the rectified features for both RGB_{out} and X_{out} modalities are organized as follows:

$$RGB_{out} = RGB_{in} + \lambda_C RGB^C_{rec} + \lambda_S RGB^S_{rec} \tag{6}$$

$$X_{out} = X_{in} + \lambda_C X^C_{rec} + \lambda_S X^S_{rec} \tag{7}$$

where λ_C and λ_S are two hyperparameters, both set to 0.5 as default values. RGB_{out} and X_{out} are the rectified features after comprehensive calibration, which are then fed into the next stage for feature fusion.

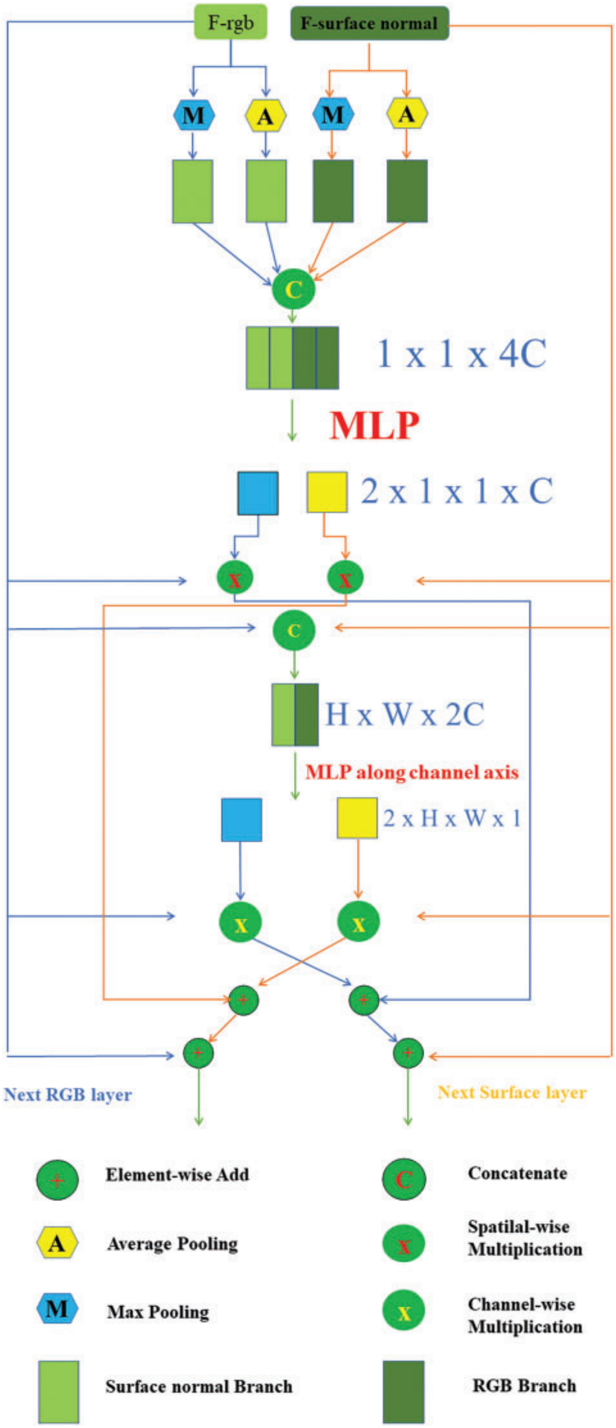

Figure 3: CM-FRM model diagram. CM-FRM with colored arrows as information flows of the two modalities

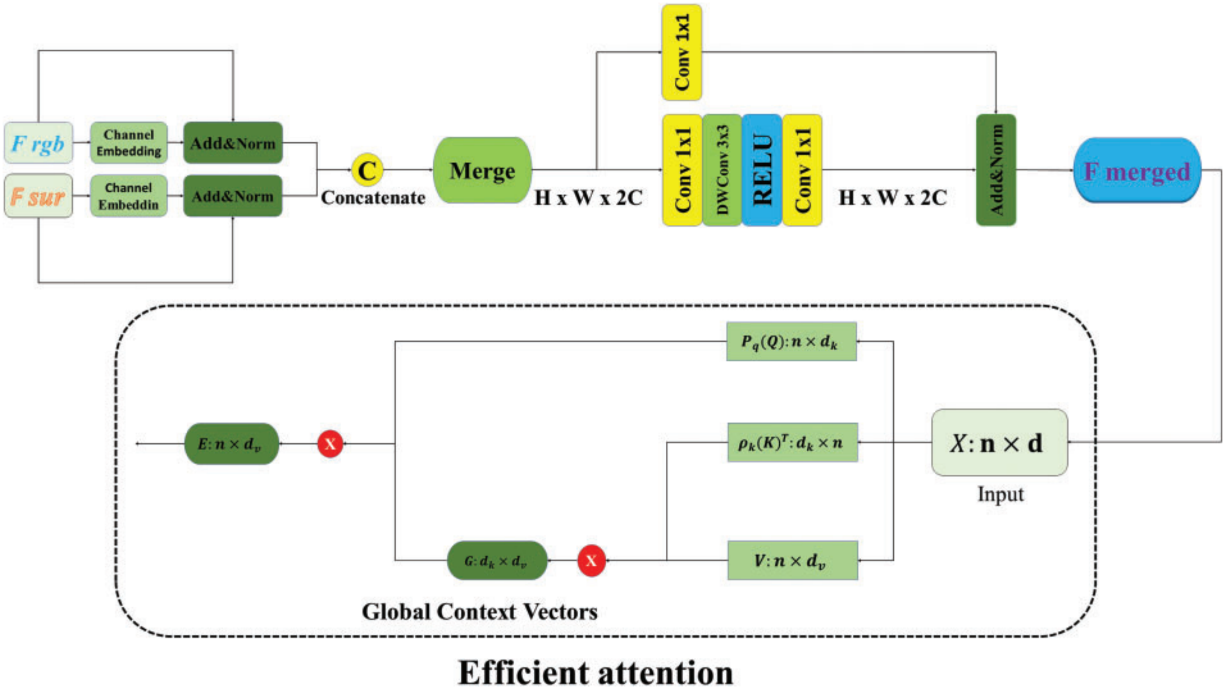

Figure 4: FFX model diagram. FFX is divided into two stages: information fusion and feature extraction

3.1.3 FFX

In the feature fusion part, we observed that although the original model's [40] feature fusion module conducted sufficient information exchange between modalities, the acquired feature information was not entirely suitable for off-road environments. Typically, cross-attention is widely used for feature extraction and fusion in text, but in off-road environments, due to challenges posed by adverse weather and complex scenes, the performance of cross-attention is not ideal, particularly evident in the edge regions and distant view acquisition of prediction maps. Inspired by Shen et al. [44] in efficient attention, we adopted a similar efficient attention mechanism to acquire both types of features. This mechanism is mathematically equivalent to dot-product attention but is faster and more memory-efficient in practice. Hence, this efficient attention mechanism performs better in perceiving complex environmental features and exhibits stronger capabilities in acquiring global feature information.

Summing up, we propose the FFX module for feature fusion. After obtaining feature maps at each layer, we perform lightweight and rapid fusion, using a simple channel embedding to merge features from both paths, achieved through a 1 × 1 convolutional layer. The model diagram is illustrated in Fig. 4. Furthermore, we believe that in this channel fusion process, it is essential to utilize information from surrounding regions for robust multimodal fusion.

Subsequently, this paper utilizes efficient attention [44] to acquire global fine-grained features. The feature of the efficient attention mechanism is as follows:

$$E(Q, K, V) = \rho_q(Q)(\rho_k(K)^T V) \tag{8}$$

Here, ρ_q and ρ_k are normalization functions for query and key features, respectively, achieving the same two normalization methods as in the point-wise attention [44].

$$Scaling\colon \rho_q(Y) = \rho_k(Y) = \frac{Y}{\sqrt{n}} \tag{9}$$

$$\begin{aligned} Softmax\colon \rho_q(Y) &= \sigma_{row}(Y) \\ \rho_k(Y) &= \sigma_{col}(Y) \end{aligned} \tag{10}$$

Here, $\sigma_{row}, \sigma_{col}$ apply softmax function along each row or column of the matrix Y. The efficient attention module is a specific implementation of computer vision data processing mechanisms. For input feature maps $X \in R^{h\times w\times d}$, this module flattens it into a matrix $X \in R^{hw\times d}$, applies an efficient attention mechanism to it, and reshapes the result into $h \times w \times d_v$. If $d_v \neq d$, further 1×1 convolution is applied to restore the dimension to d. Finally, the obtained features are added to the input features, forming a residual structure.

3.2 Decoder Network

To address the shortcomings of the original model in feature information acquisition and correspondingly reduce the parameter count and computational complexity of our model, we have made improvements. We have redesigned the Transformer decoder to effectively integrate local and global information. This decoder consists of MLP layers and components similar to hierarchical working principles. Firstly, features from the Transformer encoder are fed into MLP layers to aggregate channel information. During the feature concatenation process, we sequentially concatenate feature maps from different levels and use a fusion module to perform summation operations. This fusion module stacks tensors into a new tensor, where the dimensions of the new tensor correspond to the indices of the original tensor list, and then performs summation operations along the first dimension. This approach changes the linear transformation method traditionally performed before concatenation. This design simplifies the code logic, avoids additional linear transformation and reshape operations through feature map summation operations, making the process more concise and computationally efficient. Consequently, it reduces the parameter count, eliminates the need for additional linear transformation operations, reduces model complexity, and helps mitigate the risk of overfitting. Furthermore, this direct feature fusion method allows for more intuitive fusion of features from different levels, preserving the information of the original feature maps and enhancing the model's expressive power. Thus, this module design adopts a more concise fusion method, performing convolutional fusion on feature maps, and exhibits good performance. Finally, the features are fused to the same size (i.e., 1/4 of the in put size) to obtain the results of free space detection.

4 Experiments

4.1 Datasets

In this study, we selected the ORFD dataset [14] to evaluate the off-road free space detection task. The ORFD dataset was collected in off-road environments, which differ significantly from structured road environments due to factors such as terrain, vegetation, seasons, weather, and time. The dataset covers a variety of scenarios, as shown in Fig. 5, including different seasonal and weather conditions under various lighting conditions. It reflects the real-world characteristics of off-road roads, with the dataset's scale presented in Table 1. Data collection was carried out using the Pandora sensor fusion kit produced by SciTech, which consists of a 40-line mechanical LiDAR on the upper part and five cameras distributed at the lower part, including one color camera and four wide-angle black-and-white cameras. This diversity ensures that the dataset aligns with the real-world challenges of off-road environments.

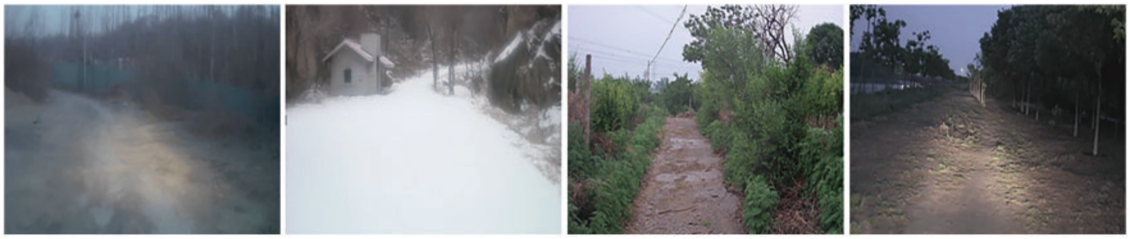

Figure 5: ORFD dataset demonstration: The image showcases RGB images from four different scenarios, including heavy fog, heavy snow, muddy roads after rain, and dimly lit conditions

Table 1: ORFD dataset

Split	Farmland	Woodland	Grassland	Countryside	Suny	Rainy	Foggy	Sonwy	Bright light	Daylight	Twilight	Darkness	Total	%
Train	2718	3180	361	2139	3803	2434	1136	1025	1019	4254	927	2192	8398	68.8
Val	356	302	0	587	356	302	0	587	0	356	302	587	1245	10.2
Test	1129	1064	0	1071	1071	405	720	359	361	710	359	1125	2555	20.9
Total	4203	4546	361	5230	5230	3141	1856	1971	1380	5320	1588	4510	12,198	100

The table presents a comprehensive overview of various scenarios from the ORFD dataset, highlighting the diversity and complexity of the environments used for evaluation.

4.2 Parameter Settings

In this experiment, we implemented the proposed method using PyTorch [45] and trained the model using the Stochastic Gradient Descent with Momentum (SGDM) optimizer [46]. The experimental environment was configured based on mmdetection. In the experimental setup, key hyperparameters include the use of the Adam optimizer with an initial learning rate of 0.00095. The momentum for SGD is set to 0.9, and weight decay is applied with a value of 0.0005. The maximum number of epochs is set to 30, with training beginning at epoch 1. The learning rate is adjusted using a lambda decay strategy, with decay occurring every 5 million iterations or every 25 epochs, and the decay factor is set to 0.9. The size of images used for training and testing was set to 1280 × 704. All experiments were conducted using one Nvidia RTX 3090 24 G GPU device. It took approximately two days to train the network for every 30 epochs.

4.3 Evaluation Metrics

Weusefivecommonmetricsfortheperformanceevaluationofoff-roadfreespacedetection: (1) Accuracy = $\frac{TP+TN}{TP+TN+FP+FN}$, (2) Precision = $\frac{TP}{TP+FP}$, (3) $Recall = \frac{TP}{TP+FN}$, (4) $F-score = \frac{2TP^2}{2TP^2+TP(FP+FN)}$, (5) $IOU = \frac{TP}{TP+FP+FN}$, where TP, TN, FP, FN represent the number of true positive, true negative, false positive, and false negative pixels, respectively.

4.4 Computational Efficiency

Compared to Off-Net, our model has an increased parameter count of approximately 10 M, but the actual inference speed remains largely unchanged. Moreover, we have reduced the parameter count by 165 M compared to SNE-RoadSeg (201.3 M), representing a significant improvement in efficiency. Although the computational cost of CFH-Net is slightly higher, its enhanced robustness in off-road scenarios—evidenced by a 5.2% improvement in F-score—demonstrates the value of this design choice.

After comparing the results with those reported in recent relevant literature, the outcomes of this study are as follows.

4.5 Results

The experimental results in this study, as illustrated in Fig. 6, include RGB images, surface normals, segmentation images, and label images. We conducted a detailed analysis of our model results, referencing recent research by Min et al. [14], Yan et al. [38], and Jeon et al. [47], as summarized in Table 2. Compared to models such as FuseNet and FtFoot, which take RGB and depth information as inputs, as well as SNE-RoadSeg and FSN-Swin models, along with OFF-Net model, which use RGB images and surface normals as inputs, our model exhibits superior performance. This suggests that surface normals are more suitable than depth for traversable area detection tasks. In the table, our model demonstrates a 3.5% improvement in F-score and a 5.9% improvement in Intersection over Union (IoU) compared to the baseline OFF-Net model. When compared to the latest FSN-Swin model, to the best of our knowledge, our approach achieves a 0.5% increase in both F-score and IoU. These results underscore the effectiveness of our method in capturing both local and global information, thereby enhancing the accuracy of traversable area detection.

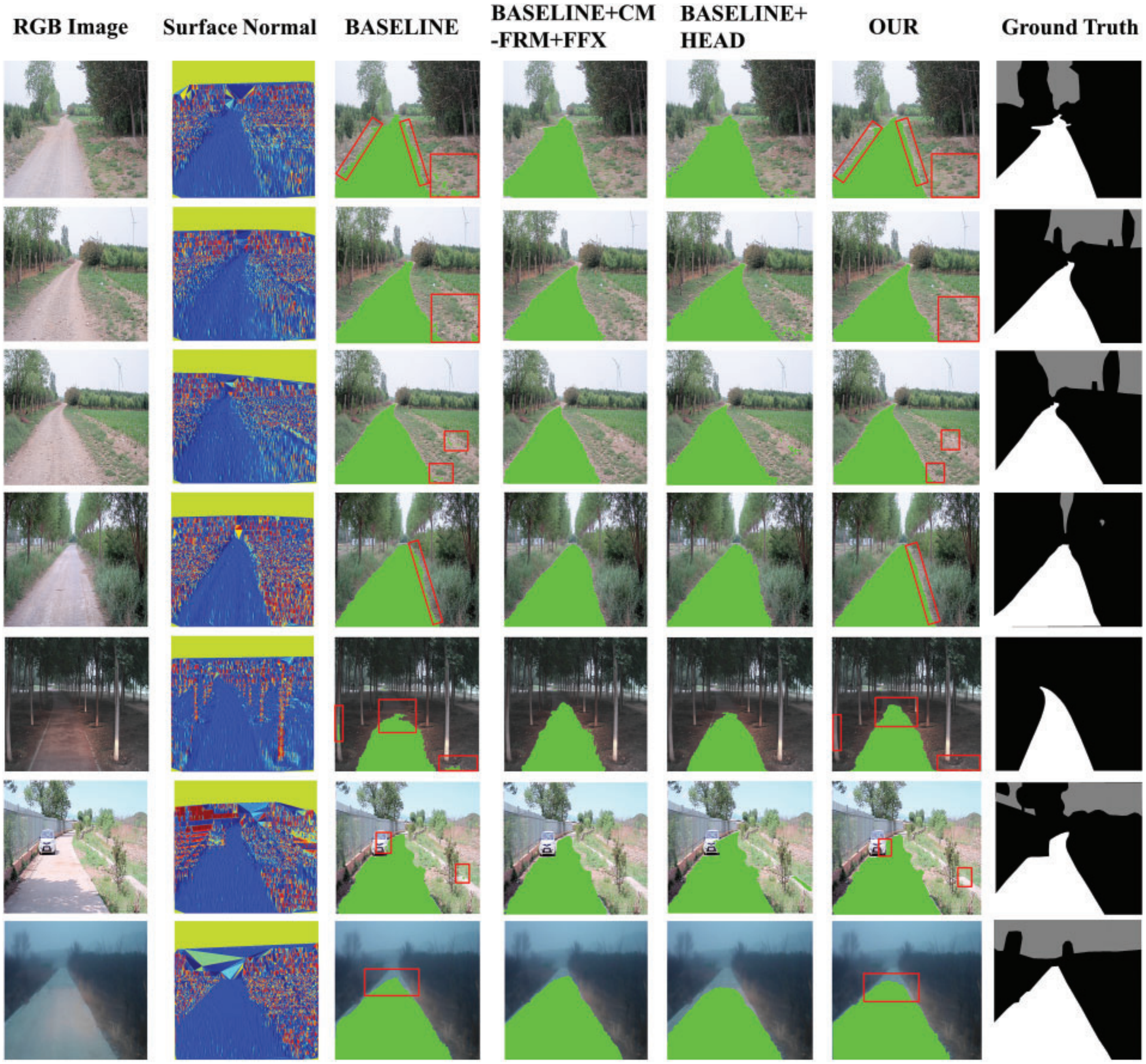

Figure 6: Experimental comparison figures

Table 2: Experimental results

Method	Modality	Accuracy	Precision	Recall	F-score	IOU
FtFoot [47] (2024)	RGB+Sparse depth	85.60%	76.80%	68.70%	72.50%	56.90%
FuseNet [36] (2017)	RGB+Sparse depth	87.40%	74.50%	85.20%	79.50%	66.00%
Test [37] (2020)	RGB+Surface normal	93.80%	86.70%	92.70%	89.60%	81.20%
Swin Transforme [48] (2021)	RGB	94.60%	84.20%	94.80%	89.20%	80.50%
OFF-Net [14] (2022)	RGB+Surface normal	94.50%	86.60%	94.30%	90.30%	82.30%
FSN-Swin [38] (2024)	RGB+Surface normal	96.40%	91.20%	95.80%	93.40%	87.70%
CFH-Net (our)	RGB+Surface normal	**96.50%**	**92.80%**	94.70%	**93.80%**	**88.20%**

From Fig. 6, it is evident that our CFH-Net model accurately estimates off-road free space. Specifically, in scenes unseen in the test set, the model demonstrates good generalization ability. Experimental results indicate that even in extreme rain or snow conditions, and amidst complex object scenarios or poor lighting conditions, the model maintains effective segmentation, successfully completing the semantic segmentation task of unstructured roads. Compared to the OFF-Net model, our model better meets the practical task requirements in terms of segmentation edges, distant views, and overall labeling. In the first and fourth comparisons, our model's edge predictions are clearer. In the first, second, third, fifth, and sixth comparisons, our model's predictions for identifying grass, vehicles, and trees are more accurate. In the fifth and seventh comparisons, our model's coverage of distant views is broader.

5 Quantitative Analysis

In Figs. 6 and 7, we conducted extensive ablation experiments. Through the presentation of data and comparative graphs, we observed a significant positive impact on the final prediction maps when the model adopted the new method.

(1) After incorporating the CM-FRM and FFX modules, as depicted in Fig. 7, both the F-score and IOU metrics exhibit a significant improvement trend. From the seven comparisons in Fig. 6, it is evident that our model outperforms the baseline model in all scenarios. Therefore, the proposed model in this study provides more accurate predictions at the boundary regions and more precise identification of components such as grass and trees.

(2) After adopting the decoder hierarchical approach, as illustrated in Fig. 7, both the F-score and IOU metrics demonstrate a noticeable upward trend, accompanied by a corresponding reduction in model parameters and computational complexity. With our decoder, the parameter count decreased from 38.711 to 38.514 M, reducing by nearly 200,000 parameters. From Fig. 6, it is evident that in the first, second, third, and sixth comparisons, employing our decoder results in better extension of the distant parts in the prediction maps. In the fifth comparison, the smoothness of our model's boundaries surpasses that of the baseline model. Thus, our model becomes smoother and more accurate in distant views and road edge regions.

(3) In the final comparison, where we simultaneously utilized all methods, the results indicate superior performance compared to individually adding single modules. This enhancement allows our model to better meet the requirements of practical tasks in segmentation edges, distant views, identification of grass, trees, obstacles, and overall labeling.

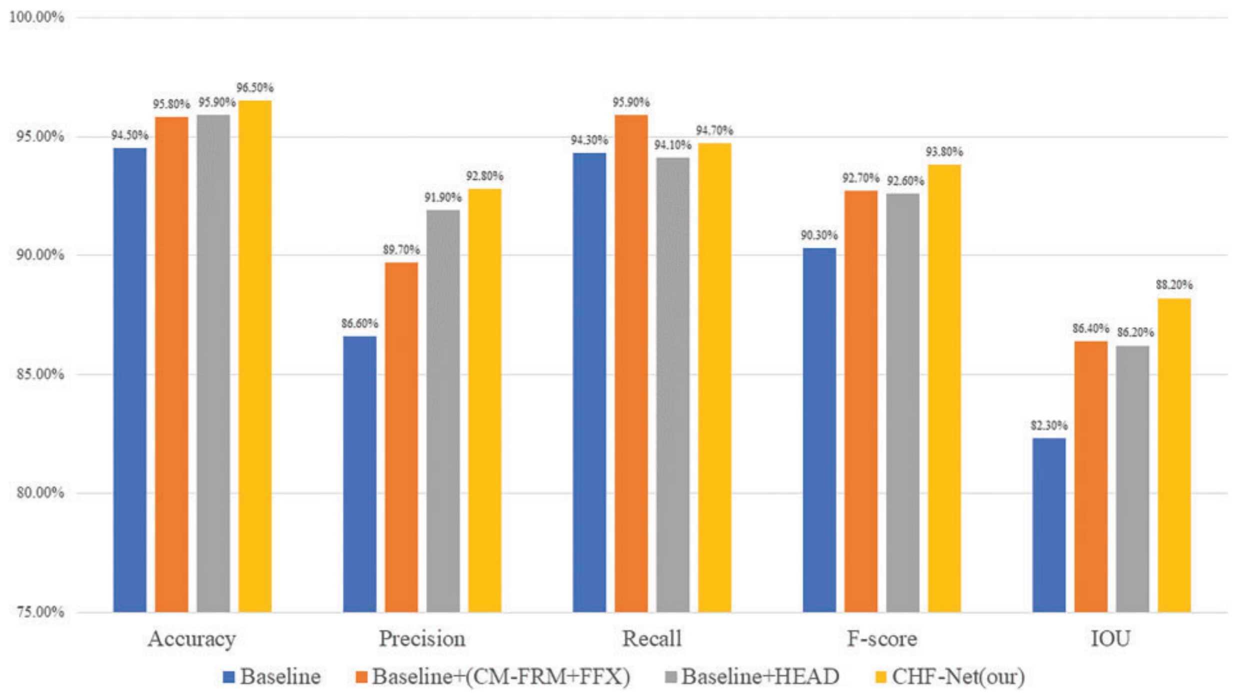

Figure 7: Ablation results

6 Fault Case

Due to the limited penetration of LiDAR, CFH-Net will recognize obstacles such as soil piles and roads as obstacles in heavy fog under low light conditions (Fig. 8). In addition, under low light conditions, farther road edges (<50 m) may shatter. Future work will integrate thermal imaging to address these issues

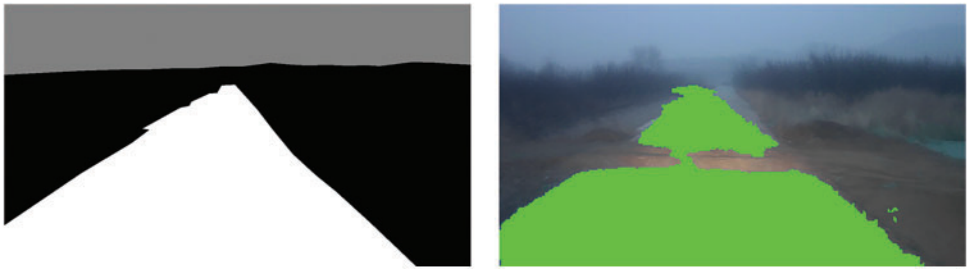

Figure 8: Comparison chart of prediction and GT labels under low light and heavy fog conditions

7 Conclusion

In this paper, we propose a novel model named CFH-Net for traversable area detection in off-road environments, utilizing a Transformer network architecture to capture contextual information. We reference the CM-FRM for feature extraction and introduce the innovative FFX module to fuse multimodal features dynamically, aggregating information from both cameras and LiDAR sensors to enhance accuracy. Additionally, we redesign the encoder using a convolutional hierarchical output approach, reinforcing the final feature fusion and reducing the impact on parameters and computational complexity. Ultimately, our model outperforms the baseline model in test results and surpasses other recent model algorithms. We utilize the ORFD dataset, which collects various off-road vehicle scenarios, aligning with the proposed task in this paper. This will aid future research in autonomous navigation in non-road environments and facilitate the development of robust models for unmanned vehicles in off-road environments. Moving forward, we plan to further explore various model algorithms on this dataset to advance research in unstructured road segmentation.

Acknowledgement: Not applicable.

Funding Statement: None.

Author Contributions: Study conception and design: Jingcheng Yang, Hongmei Liu; data collection: Jingcheng Yang; analysis and interpretation of results: Jingcheng Yang; draft manuscript preparation: Jingcheng Yang; review and editing: Lili Fan; funding acquisition: Hongmei Liu. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: Not applicable.

Ethics Approval: Not applicable.

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

References

1. Howard A, Seraji H. Real-time assessment of terrain traversability for autonomous rover navigation. In: Proceedings. 2000 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2000); 2000; Takamatsu, Japan. p. 58–63. doi:10.1109/IROS.2000.894582.
2. Badrinarayanan V, Kendall A, Cipolla R. Segnet: a deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans Pattern Anal Mach Intell. 2017;39(12):2481–95. doi:10.1109/TPAMI.2016.2644615.
3. Yao J, Ramalingam S, Taguchi Y, Miki Y, Urtasun R. Estimating drivable collision-free space from monocular video. In: 2015 IEEE Winter Conference on Applications of Computer Vision; 2015 Jan 5–9; Waikoloa, HI, USA: IEEE; 2015. p. 420–7. doi:10.1109/WACV.2015.62.
4. Tsutsui S, Kerola T, Saito S, Crandall DJ. Minimizing supervision for free-space segmentation. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW); 2018 Jun 18–22; Salt Lake City, UT, USA: IEEE; 2018. doi:10.1109/CVPRW.2018.00145.
5. Wigness M, Eum S, Rogers JG, Han D, Kwon H. A RUGD dataset for autonomous navigation and visual perception in unstructured outdoor environments. In: 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS); 2019 Nov 3–8; Macau, China: IEEE; 2019. p. 5000–7. doi:10.1109/iros40897.2019.8968283.
6. Cao J, Leng H, Lischinski D, Cohen-Or D, Tu C, Li Y. ShapeConv: shape-aware convolutional layer for indoor RGB-D semantic segmentation. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV); 2021 Oct 10–17; Montreal, QC, Canada: IEEE; 2021. p. 7068–77. doi:10.1109/ICCV48922.2021.00700.
7. Chen LZ, Lin Z, Wang Z, Yang YL, Cheng MM. Spatial information guided convolution for real-time RGBD semantic segmentation. IEEE Trans Image Process. 2021;30:2313–24. doi:10.1109/TIP.2021.3049332.
8. Chen X, Lin KY, Wang J, Wu W, Qian C, Li H, et al. Bi-directional cross-modality feature propagation with separation-and-aggregation gate for RGB-D semantic segmentation. In: Computer Vision—ECCV 2020. Cham: Springer International Publishing; 2020. p. 561–77. doi: 10.1007/978-3-030-58621-8_33.
9. Zhang Q, Zhao S, Luo Y, Zhang D, Huang N, Han J. ABMDRNet: adaptive-weighted bi-directional modality difference reduction network for RGB-T semantic segmentation. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2021 Jun 20–25; Nashville, TN, USA: IEEE; 2021. p. 2633–42. doi:10.1109/cvpr46437.2021.00266.
10. Deng F, Feng H, Liang M, Wang H, Yang Y, Gao Y, et al. FEANet: feature-enhanced attention network for RGB-thermal real-time semantic segmentation. In: 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS); 2021; Prague, Czech, Republic: IEEE. p. 4467–73. doi:10.1109/IROS51168.2021.9636084.

11. Xu J, Zhao L, Ren Y, Li Z, Abbas Z, Zhang L, et al. LightYOLO: lightweight model based on YOLOv8n for defect detection of ultrasonically welded wire terminations. Eng Sci Technol Int J. 2024;60:101896. doi:10.1016/j.jestch.2024.101896.
12. Abbas Z, Zhao L, Zeng J, Kao-walter S, Qi X. Bonding analysis of ultrasonic welded multi-wire joints with additional root gaps. Alex Eng J. 2025;116:20–34. doi:10.1016/j.aej.2024.12.082.
13. Fritsch J, Kühnl T, Geiger A. A new performance measure and evaluation benchmark for road detection algorithms. In: 16th International IEEE Conference on Intelligent Transportation Systems (ITSC 2013); 2013 Oct 6–9; The Hague, Netherlands: IEEE; 2013. p. 1693–700. doi:10.1109/ITSC.2013.6728473.
14. Min C, Jiang W, Zhao D, Xu J, Xiao L, Nie Y, et al. Orfd: a dataset and benchmark for off-road freespace detection. In: 2022 International Conference on Robotics and Automation (ICRA); 2022; Philadelphia, PA, USA: IEEE. p. 2532–8. doi:10.1109/ICRA46639.2022.9812139.
15. Zheng S, Lu J, Zhao H, Zhu X, Luo Z, Wang Y, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2021 Jun 20–25; Nashville, TN, USA: IEEE; 2021. p. 6877–86. doi:10.1109/cvpr46437.2021.00681.
16. Ranftl R, Bochkovskiy A, Koltun V. Vision transformers for dense prediction. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV); 2021 Oct 10–17; Montreal, QC, Canada: IEEE; 2021. p. 12159–68. doi:10.1109/ICCV48922.2021.01196.
17. Zhao H, Shi J, Qi X, Wang X, Jia J. Pyramid scene parsing network. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017 Jul 21–26; Honolulu, HI, USA: IEEE; 2017. p. 6230–9. doi:10.1109/CVPR.2017.660.
18. Wang J, Sun K, Cheng T, Jiang B, Deng C, Zhao Y, et al. Deep high-resolution representation learning for visual recognition. IEEE Trans Pattern Anal Mach Intell. 2021;43(10):3349–64. doi:10.1109/TPAMI.2020.2983686.
19. Yu C, Gao C, Wang J, Yu G, Shen C, Sang N. BiSeNet V2: bilateral network with guided aggregation for real-time semantic segmentation. Int J Comput Vis. 2021;129(11):3051–68. doi:10.1007/s11263-021-01515-2.
20. Xie E, Wang W, Yu Z, Anandkumar A, Alvarez JM, Luo P. SegFormer: simple and efficient design for semantic segmentation with transformers. In: Advances in neural information processing systems. Law St, San Diego, CA, USA: MIT Press; 2021.
21. Strudel R, Garcia R, Laptev I, Schmid C. Segmenter: transformer for semantic segmentation. arXiv:2105.05633. 2021.
22. Cordts M, Omran M, Ramos S, Rehfeld T, Enzweiler M, Benenson R, et al. The cityscapes dataset for semantic urban scene understanding. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016 Jun 27–30; Las Vegas, NV, USA: IEEE; 2016. p. 3213–23. doi:10.1109/CVPR.2016.350.
23. Geiger A, Lenz P, Stiller C, Urtasun R. Vision meets robotics: the KITTI dataset. Int J Robot Res. 2013;32(11):1231–7. doi:10.1177/0278364913491297.
24. Peng C, Zhang X, Yu G, Luo G, Sun J. Large kernel matters—improve semantic segmentation by global convolutional network. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017 Jul 21–26; Honolulu, HI, USA: IEEE; 2017. p. 1743–51. doi:10.1109/CVPR.2017.189.
25. 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. 2022;23(2):722–39. doi:10.1109/TITS.2020.3023541.
26. Wang C, Ma C, Zhu M, Yang X. PointAugmenting: cross-modal augmentation for 3D object detection. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2021 Jun 20–25; Nashville, TN, USA: IEEE; 2021. p. 11789–98. doi:10.1109/cvpr46437.2021.01162.
27. Wang Y, Mao Q, Zhu H, Deng J, Zhang Y, Ji J, et al. Multi-modal 3D object detection in autonomous driving: a survey. Int J Comput Vis. 2023;131:2122–52. doi:10.1007/s11263-023-01784-z.
28. Geiger A, Lenz P, Urtasun R. Are we ready for autonomous driving? The KITTI vision benchmark suite. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition; 2012 Jun 16–21; Providence, RI, USA: IEEE; 2012. p. 3354–61. doi:10.1109/CVPR.2012.6248074.

29. 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.
30. Sun P, Kretzschmar H, Dotiwalla X, Chouard A, Patnaik V, Tsui P, et al. Scalability in perception for autonomous driving: waymo open dataset. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2020 Jun 13–19; Seattle, WA, USA: IEEE; 2020. p. 2443–51. doi:10.1109/cvpr42600.2020.00252.
31. Xie L, Xu G, Cai D, He X. X-view: non-egocentric multi-view 3D object detector. IEEE Trans Image Process. 2023;32:1488–97. doi:10.1109/TIP.2023.3245337.
32. Bijelic M, Gruber T, Mannan F, Kraus F, Ritter W, Dietmayer K, et al. Seeing through fog without seeing fog: deep multimodal sensor fusion in unseen adverse weather. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2020 Jun 13–19; Seattle, WA, USA: IEEE; 2020. p. 11679–89. doi:10.1109/cvpr42600.2020.01170.
33. Yuan J, Zhou W, Luo T. DMFNet: deep multi-modal fusion network for RGB-D indoor scene segmentation. IEEE Access. 2019;7:169350–8. doi:10.1109/ACCESS.2019.2955101.
34. Wang J, Wang Z, Tao D, See S, Wang G. Learning common and specific features for RGB-D semantic segmentation with deconvolutional networks. In: European Conference on Computer Vision; 2016; Cham, Switzerland.
35. Lin D, Zhang R, Ji Y, Li P, Huang H. SCN: switchable context network for semantic segmentation of RGB-D images. IEEE Trans Cybern. 2020;50(3):1120–31. doi:10.1109/TCYB.2018.2885062.
36. Hazirbas C, Ma L, Domokos C, Cremers D. FuseNet: incorporating depth into semantic segmentation via *Fusi* on-based CNN architecture. In: Computer Vision—ACCV 2016: 13th Asian Conference on Computer Vision; 2016 Nov 20–24; Taipei, Taiwan. p. 213–28.
37. Fan R, Wang H, Cai P, Liu M. Sne-roadseg: incorporating surface normal information into semantic segmentation for accurate freespace detection. In: European Conference on Computer Vision; 2020; Cham, Switzerland: Springer.
38. Yan T, Wang Y, Lv H, Sun H, Zhang D, Yang Y. FSN-swin: a network for freespace detection in unstructured environments. IEEE Access. 2024;12:12308–22. doi:10.1109/ACCESS.2024.3354721.
39. Fan R, Wang H, Xue B, Huang H, Wang Y, Liu M, et al. Three-filters-to-normal: an accurate and ultrafast surface normal estimator. IEEE Robot Autom Lett. 2021;6(3):5405–12. doi:10.1109/LRA.2021.3067308.
40. Zhang J, Liu H, Yang K, Hu X, Liu R, Stiefelhagen R. CMX: cross-modal *Fusi* on for RGB-X semantic segmentation with transformers. IEEE Trans Intell Transp Syst. 2023;24(12):14679–94. doi:10.1109/TITS.2023.3300537.
41. Yan X, Gao J, Zheng C, Zheng C, Zhang R, Cui S, et al. 2DPASS: 2D priors assisted semantic segmentation on LiDAR point clouds. In: European Conference on Computer Vision—ECCV 2022; 2022; Cham, Switzerland. p. 677–95.
42. Li J, Dai H, Han H, Ding Y. MSeg3D: multi-modal 3D semantic segmentation for autonomous driving. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2023 Jun 17–24; Vancouver, BC, Canada: IEEE; 2023. p. 21694–704. doi:10.1109/CVPR52729.2023.02078.
43. Hendrycks D, Gimpel K. Gaussian error linear units (GELUs). arXiv:160608415. 2016.
44. Shen Z, Zhang M, Zhao H, Yi S, Li H. Efficient attention: attention with linear complexities. In: 2021 IEEE Winter Conference on Applications of Computer Vision (WACV); 2021 Jan 3–8; Waikoloa, HI, USA: IEEE; 2021. p. 3530–8. doi:10.1109/WACV48630.2021.00357.
45. Paszke A, Gross S, Massa F, Lerer A, Chintala S. PyTorch: an imperative style, high-performance deep learning library. In: Advances in neural information processing systems. Vancouver BC, Canada: MIT Press; 2019.
46. Polyak BT. Some methods of speeding up the convergence of iteration methods. USSR Comput Math Math Phys. 1964;4(5):1–17. doi:10.1016/0041-5553(64)90137-5.
47. Jeon Y, Son EI, Seo SW. Follow the footprints: self-supervised traversability estimation for off-road vehicle navigation based on geometric and visual cues. arXiv:240215363. 2024.
48. Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, et al. Swin transformer: hierarchical vision transformer using shifted windows. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV); 2021 Oct 10–17; Montreal, QC, Canada: IEEE; 2021. p. 9992–10002. doi:10.1109/ICCV48922.2021.00986.