

# ROD: RGB-Only Fast and Efficient Off-road Freespace Detection

Tong Sun<sup>1,2</sup>, Hongliang Ye<sup>3</sup>, Jilin Mei<sup>1,†</sup>, Liang Chen<sup>1</sup>, Fangzhou Zhao<sup>1</sup>, Leiqiang Zong<sup>4</sup>, Yu Hu<sup>1,5,†</sup>

**Abstract**—Off-road freespace detection is more challenging than on-road scenarios because of the blurred boundaries of traversable areas. Previous state-of-the-art (SOTA) methods employ multi-modal fusion of RGB images and LiDAR data. However, due to the significant increase in inference time when calculating surface normal maps from LiDAR data, multi-modal methods are not suitable for real-time applications, particularly in real-world scenarios where higher FPS is required compared to slow navigation. This paper presents a novel RGB-only approach for off-road freespace detection, named ROD, eliminating the reliance on LiDAR data and its computational demands. Specifically, we utilize a pre-trained Vision Transformer (ViT) to extract rich features from RGB images. Additionally, we design a lightweight yet efficient decoder, which together improve both precision and inference speed. ROD establishes a new SOTA on ORFD and RELIS-3D datasets, as well as an inference speed of 50 FPS, significantly outperforming prior models. Our code will be available at [https://github.com/STLIFE97/offroad\\_roadseg](https://github.com/STLIFE97/offroad_roadseg).

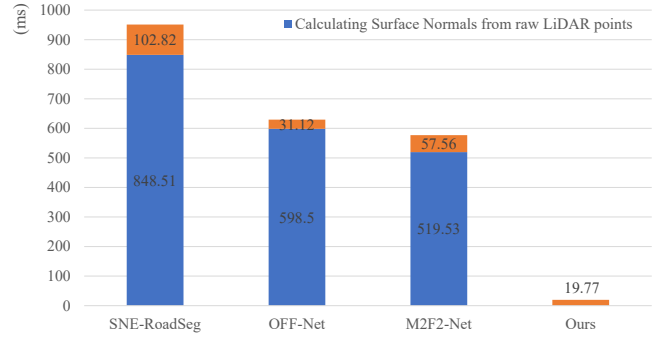
## I. INTRODUCTION

In recent years, autonomous driving is attracted growing attention and research. Freespace detection, a binary classification segmentation task, plays a fundamental role in autonomous driving systems by delineating navigable areas crucial for vehicle planning and control. The majority of research are mainly focused on urban on-road scenarios, characterized by well-defined features such as lanes and traffic signs [10][31]. In contrast, research on off-road scenarios receives less attention [37]. These off-road scenarios present a higher degree of complexity and diversity, with the freespace being less distinct [25]. Vehicles are required to traverse a variety of terrains, including grasslands, sandy areas, icy grounds, snowy regions, and muddy terrains. As illustrated in Fig. 1(a), the boundaries of traversable areas in off-road scenarios are blurred.

To cope with the above difficulties, previous SOTA methods predominantly rely on the multi-modal fusion of RGB imagery with LiDAR data [41][21]. While RGB images focus more on surface color, texture, and other visual information, LiDAR data focuses more on distance, depth, and position [34]. Multi-modal fusion can help these two modalities complement each



(a) RGB images of off-road scenes



(b) Inference Time

Fig. 1. (a) provides examples of RGB images in off-road scenes, where the boundaries of traversable areas are not clearly defined. (b) illustrates the inference time for SNE-RoadSeg [10], OFF-Net [25], the prior SOTA method M2F2-Net[41] and ours. Multi-modal fusion based methods do not meet real-time requirements due to the computational steps involved in generating surface normal maps.

other to achieve better performance, and thus multi-modal methods typically achieve higher accuracy [16]. Nonetheless, most methods for fusing LiDAR data require the calculation of surface normal maps [41], which is a time-consuming process. This huge computational overhead significantly affects the inference speed, making the algorithm less viable for real-time applications on vehicles. For high-speed autonomous systems such as off-road vehicles or drones operating in dynamic environments, achieving high frame rates is critical to ensure timely decision-making and collision avoidance. As shown in Fig. 1(b), the inference time for the M2F2-Net [41] is predominantly consumed by the generation of surface normal maps from LiDAR data, accounting for 90.02% of the total inference time. Such latency severely limits the practical deployment of LiDAR-based methods in scenarios requiring rapid response. Beyond computational constraints, LiDAR sensors also impose significant hardware costs and energy consumption.

To reduce reliance on surface normal map, this study proposes a novel approach that utilizes a pre-trained ViT model to extract features only from RGB images, thereby significantly enhancing the inference speed. With the development of ViT [9], the performance of large vision models [19][39]

†This work was supported by National Natural Science Foundation of China under Grant No.U23B2034, No.62203424, and No.62176250; and in part by the Innovation Program of Institute of Computing Technology, Chinese Academy of Sciences under Grant No. 2024000112.

<sup>1</sup>Research Center for Intelligent Computing Systems, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 100190, China.

<sup>2</sup>University of Chinese Academy of Sciences, Beijing, 100190, China.

<sup>3</sup>Astronomical Computing Research Center, Zhejiang Lab

<sup>4</sup>Beijing Special Vehicle Academy, Beijing, 100072, China.

<sup>5</sup>School of Computer Science and Technology, University of Chinese Academy of Sciences, Beijing, 100190, China.

†Correspondence: Jilin Mei, Yu Hu, {meijilin, huyu}@ict.ac.cn

improves as more data becomes available for pre-training. As shown in Fig. 2, the ViT encoder, which is employed for feature extraction, is kept frozen. Only the simple yet powerful decoder is trained. The RGB images are fed into the ViT encoder to generate an image embedding and to extract features from the latent layers of the transformer encoder blocks. These features are then used for subsequent fusion, and finally fed into the decoder for feature integration and prediction, culminating in the generation of the final prediction mask.

Our contribution can be summarized as follows:

- We conduct an investigation into the factors affecting inference speed in freespace detection models and propose a novel method, ROD. ROD only utilizes RGB data. This method surpasses prior multi-modal fusion methods in terms of both accuracy and inference speed.
- ROD integrates a pre-trained ViT model into the off-road freespace detection task, and design a powerful decoder that effectively merges image embeddings with latent features.
- The proposed ROD achieves performance on the ORFD and RELLIS-3D Dataset, with 98.3% F1\_score and 96.7% IoU on the ORFD dataset, and 97.1% F1\_score and 97.8% Accuracy on the RELLIS-3D dataset. Additionally, ROD achieves an inference speed of 50 FPS, comfortably meeting real-time requirements.

The remainder of this paper is organized as follows: Section 2 reviews related work, Section 3 details our model, Section 4 presents experiments, and Section 5 concludes the study.

## II. RELATED WORKS

### A. Freespace Detection

Freespace detection involves assigning a label to each pixel in traversable regions, which is categorized under semantic segmentation. The majority of research in this area focuses on on-road scenarios [10][31]. With the development of deep learning, freespace detection has expanded to include both single-modal and multi-modal approaches.

Single-modal methods utilize either RGB images or LiDAR data. CNNs are widely used for RGB data, as studies [11][22] demonstrate their effectiveness in freespace detection. [11] develops a driving scene generator to augment training data, while [22] introduces the RPP(Residual network with Pyramid Pooling) model, combining full convolutional networks with residual and pyramid pooling. Within LiDAR-based methods, the most approaches employ occupancy grids and scene flow to delineate obstacles. MotionNet [38] adopts a voxel-based representation, and PointMotionNet [32] opts for a point-based representation. Nevertheless, single-modal models that rely on either LiDAR or RGB images are limited by the data type they process. This limitation reduces the breadth of information available to the model, therefore, leading to lower accuracy compared to multi-modal systems [17].

The multi-modal methods for freespace detection fuse RGB images with LiDAR data. PETRv2 [23] and CVT [34] utilize multi-modal camera fusion to create Bird's Eye View

(BEV) images for road segmentation. BiFNet [20] introduces a bi-directional fusion network for integrating point cloud images with BEVs. RoadSeg [2] proposes a method that leverages LiDAR data for road segmentation. SNE-RoadSeg [10] enhances segmentation performance by fusing surface normal maps with RGB image features. Its successor, SNE-RoadSeg+ [31], further refines the accuracy of surface normal estimation and improves the network's performance and speed. Building on these advancements, M2F2-Net [41] and RoadFormer [21] achieve the SOTA performance in freespace detection.

However, the computation of surface normal maps from LiDAR data in multi-modal methods is time-consuming, resulting in an inability to meet real-time requirements. Therefore, this paper aims to find a method that utilizes only RGB images to improve inference speed and preserve precision.

### B. Applications of Vision Transformer

The Vision Transformer (ViT) [9] adapts the transformer, usually for NLP, to image classification. Its streamlined design allows scalability and, with sufficient pre-training data, can surpass CNNs in transferability across tasks [9]. This paves the way for the development of large-scale models. These models are successfully applied to an array of visual tasks, further substantiating ViT's prowess in feature extraction. For instance, ViT demonstrates exceptional performance in segmentation applications [3][5][8], image inpainting [42], image editing [12], object detection [19], image captioning [33], object tracking [6][40], and 3D object reconstruction [29], among others.

Therefore, this paper attempts to integrate pre-train ViT model with freespace detection. However, a challenge with ViT is their large model size, which presents significant challenges for real-time applications, recent developments offer solutions. Lightweight ViT models such as MobileSAM [43], FastSAM [44] are proposed, they increase the inference speed but decrease the accuracy. Subsequently EfficientSAM [39] introduces a method called SAM-leveraged masked image pretraining (SAMI), which results in a lightweight ViT encoder. With less than 5% parameter of the original SAM, the performance drop is minimal, making it feasible for application in autonomous driving.

Thus, this study adopts the pre-trained ViT encoder from EfficientSAM for freespace detection. Leveraging the ViT encoder's prowess in feature extraction, we are able to achieve high accuracy using only RGB images.

### C. Off-road Dataset

Compared to a large number of on-road datasets, off-road datasets supporting freespace detection are few. The RUGD dataset [37] is designed for semantic understanding in off-road scenarios, including mountain trails, streams, parks, villages and so on. The RELLIS-3D [18] dataset is derived from RUGD and includes unique terrain like puddles. In addition, RELLIS-3D includes 3D LiDAR annotations. [25] proposed the ORFD dataset, which is an off-road freespace

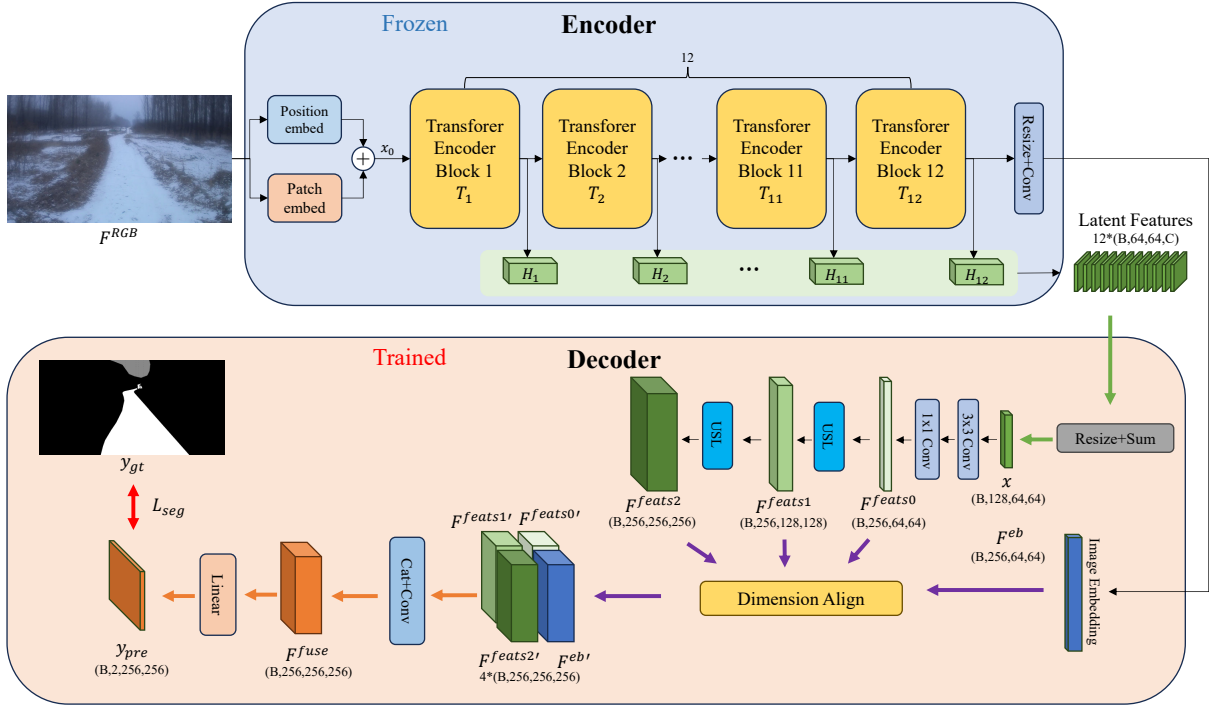


Fig. 2. The detailed architecture of the network proposed in this paper. (a) the frozen encoder part takes RGB data as input and processes it through 12 layers of transformer encoder block  $T_i$  to obtain the image embedding  $F^{eb}$  and latent features  $H_i$ . The number of channels  $C$  in  $H_i$  depends on the pre-trained model used, specifically,  $C=1280$  for ViT-H,  $C=1024$  for ViT-L,  $C=768$  for ViT-B,  $C=384$  for ViT-S and  $C=192$  for ViT-T. (b) a simple but powerful decoder is designed to fuse the  $H_i$  features and obtain features at different scales, then it outputs the prediction mask  $y_{pre}$ .

detection dataset. The dataset was collected under different scenarios (woodland, farmland, grassland, and countryside), different weather conditions (sunny, rainy, foggy and snowy), and different lighting conditions (bright light, daylight, dusk, and darkness). The proposed method will be tested on the RELLIS-3D [18] and ORFD [25] datasets.

### III. METHODOLOGY

#### A. Problem Definition

Freespace detection is a pixel-level classification problem, also known as semantic segmentation. The network  $f$  classifies each pixel or point to determine if it is passable. Given one RGB image frame  $F^{RGB}$  and its ground-truth mask  $y_{gt}$  with pixel-level annotations, the goal is to optimize the model parameters of  $f$  using the cross-entropy loss  $L_{seg}$ :

$$L_{seg} = - \sum_{H,W} y_{gt} \log(f(F^{RGB})) \quad (1)$$

where  $H$  and  $W$  represent the dimensions of the RGB image.

#### B. Overall Network Structure

The network architecture, as detailed in Fig. 2, is composed of two parts: the frozen encoder, which utilizes a pre-trained ViT-S model [39] to extract features from RGB images; and the trained decoder, which fuses these features to produce the prediction mask and updates its parameters using cross-entropy loss.

The RGB image  $F^{RGB}$  is utilized by the patch embed and position embed modules to generate positional and image

patch encodings. These encodings are summed and passed through a 12-layer transformer encoder block  $T_i$ , obtaining 12 features from the latent layers  $H_i$  and an image embedding  $F^{eb}$ . In the subsequent decode, the 12 feature maps  $H_i$  are merged to obtain the  $x$ . The channel dimension of  $x$  is expanded to create  $F^{feats0}$ , which is upsampled to produce  $F^{feats1}$ .  $F^{feats1}$  is then upsampled again to generate  $F^{feats2}$ . The features  $F^{feats0}$ ,  $F^{feats1}$ ,  $F^{feats2}$ , and  $F^{eb}$  are aligned and fused based on the dimensions of  $F^{feats2}$  to obtain  $F^{fuse}$ . Lastly,  $F^{fuse}$  is input to a linear layer to obtain the final prediction mask  $y_{pre}$ .

#### C. Architecture of the Encoder

The ViT encoder is utilized to extract high-performance features from only RGB images. A lightweight encoder, pre-trained by EfficientSAM [39], is employed. The encoder's details are shown in Fig. 2. The encoder processes the input  $F^{RGB}$ , and subsequently, it generates image embeddings  $F^{eb}$  and a set of latent features  $H_i$ .

Firstly, the RGB image  $F^{RGB}$  is processed by the Patch Embed module (*patch*), which segments it into a series of 16x16 pixel patches as embedded representations; meanwhile,  $F^{RGB}$  is fed to the Position Embed module (*pos*) to obtain positional embeddings and ensure that the positional embeddings match the dimensions of the input image. The resulting outputs are combined to form the initial feature vector  $x_0$ , as detailed by the following equation:

$$x_0 = pos(F^{RGB}) + patch(F^{RGB}) \quad (2)$$

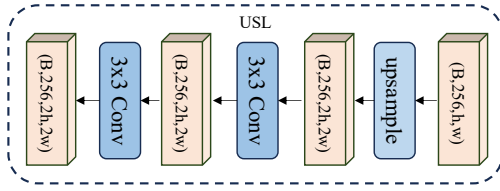


Fig. 3. The specific structure of the upsample layers (USL), where  $h$  and  $w$  represent the dimensions of the feature map.

The ViT encoder sets the depth of transformer encoder blocks  $T_i$  to 12 layers ( $i = 12$ ).  $x_0$  is fed into the first block  $T_1$ . Subsequently, each block  $T_i$  receives the output from the preceding block as its input, with the formula as follows:

$$H_i = \begin{cases} T_i(x_0) & \text{if } i = 1 \\ T_i(H_{i-1}) & \text{if } 1 < i \leq 12 \end{cases} \quad (3)$$

The ViT encoder processes data through 12 transformer blocks, resulting in 12 feature maps  $H_i$ . The last feature map  $H_{12}$  is input into *Resize+Conv* module, undergoes a dimension permute, and is passed through two convolutional layers (1x1 and 3x3 kernels) to produce the final image embedding  $F^{eb}$ , sized (B, 256, 64, 64).

#### D. Architecture of the Decoder

To balance precision and inference speed, we design a simple yet powerful seg decoder to predict segmentation masks for freespace detection. The architecture of the decoder is detailed in Fig. 2.

Firstly, we permute the dimensions of the 12 latent features  $H_i$  and downsampling them to (B, 128, 64, 64). Then summing these features to  $\sum_{i=1}^{12} H_i$ , and inputting the  $\sum_{i=1}^{12} H_i$  through two consecutive convolutional layers, each with a 3x3 kernel, then the outputs are merged back with  $\sum_{i=1}^{12} H_i$ , creating a residual connection, and obtaining the summed features  $x$ . The formula is as follows:

$$x = \text{Conv}(\text{Conv}(\sum_{i=1}^{12} H_i)) + \sum_{i=1}^{12} H_i \quad (4)$$

Fig. 3 illustrates the specific structures of the upsample layers (USL). The strategy consists of a sequence of two convolutional layers followed by a residual connection to retain finer details during the feature map's expansion.  $x$  is passed through a 1x1 convolution, which doubles the channel count from 128 to 256, obtaining the feature  $F^{feats0}$ . The features  $F^{feats0}$  are input into the upsample layers (USL) for bilinear upsample followed by two convolutional layers with a kernel size of 3x3, expanding the feature dimensions to 128x128 from 64x64. The output is merged with the upsampled  $F^{feats0}$  to produce  $F^{feats1}$ . This process is replicated on  $F^{feats1}$ , scaling the features from 128x128 to 256x256, to yield the feature  $F^{feats2}$ . The formula are as follows:

$$F^{feats0} = \text{Conv}(x) \quad (5)$$

$$F^{feats1} = \text{USL}(F^{feats0}) + \text{upsample}(F^{feats0}) \quad (6)$$

$$F^{feats2} = \text{USL}(F^{feats1}) + \text{upsample}(F^{feats1}) \quad (7)$$

Subsequently, the image embedding  $F^{eb}$  and the feature maps  $F^{feats0}$ ,  $F^{feats1}$ , and  $F^{feats2}$  are upsampled to match dimensions of (B, 256, 256, 256) through bilinear interpolation, resulting in  $F^{eb'}$ ,  $F^{feats0'}$ ,  $F^{feats1'}$ , and  $F^{feats2'}$ . The formula is as follows:

$$\begin{aligned} &F^{eb'}, F^{feats0'}, F^{feats1'}, F^{feats2'} \\ &= \text{interpolate}(F^{eb}, F^{feats0}, F^{feats1}, F^{feats2}) \end{aligned} \quad (8)$$

Ultimately, the feature maps  $F^{eb'}$ ,  $F^{feats0'}$ ,  $F^{feats1'}$ , and  $F^{feats2'}$  are concatenated to the dimensions (B, 4x256, 256, 256). This output is then processed by a 1x1 convolutional layer to downscale the channel number from 4x256 to 256, resulting  $F^{fuse}$ . Subsequently, a linear layer is applied to  $F^{fuse}$  to produce the final predicted mask  $y_{pre}$ .

## IV. EXPERIMENT

### A. Experimental Settings

Our method is evaluated on two datasets: the ORFD dataset [25], which provides 12,198 pairs of RGB images and LiDAR point clouds; and the RELIS-3D dataset [18], a large-scale dataset focusing on freespace detection in off-road scenarios. The ORFD dataset contains 8,398 training samples, 1,245 validation samples, and 2,555 test samples, with image resolutions of 1280x720. In contrast, RELIS-3D is a multi-modal dataset collected in off-road scenarios, comprising 13,556 LiDAR scans and 6,235 images. It includes 3,302 training samples, 983 validation samples, and 1,672 test samples, with images at a resolution of 1920x1200.

For testing on the ORFD dataset [25], this paper uses the evaluation metrics from the M2F2-Net [41], including IoU, Precision, Recall, Accuracy, F1\_score, and FPS.

For testing on the RELIS-3D dataset [18], this paper refers to the evaluation metrics from self-supervisions-only [28], utilizing Precision (Pre), Recall (Rec), Accuracy (Acc), and F1\_score as the assessment criteria.

This study employs the AdamW optimizer with its default parameters, initiating the learning rate at 1e-3. We adopt 'poly' decay policy for the learning rate, utilizing a decay power of 0.9. Batch size is set to 8. All experiments are conducted on a single NVIDIA RTX A6000 GPU.

### B. Comparisons on ORFD Dataset

Comparisons on the ORFD dataset are presented in TABLE I. We compare our approach with various methods, including U-Net [26], DeepLabV3Plus [4] (DLV3+-MNet and DLV3+-R101), FuseNet [15], MFNet [14], RTFNet [30], SNE-RoadSeg [10], OFF-Net [25], M2F2-Net [41], and RoadFormer [21]. Compared to methods that only use RGB [26][4], surpasses DLV3+-R101 with a 15.9% improvement in F1\_score and a 26.7% improvement in IoU. In methods

TABLE I  
COMPARISON TO OTHER METHODS ON THE TESTING SET OF ORFD

Method	Modality	Year	Accuracy	Precision	Recall	F1_score	IoU	Inference Time (ms)		FPS(total)
								CSN*	Model Inference	
U-Net [26]	RGB	2015	0.959	0.637	0.537	0.583	0.411	<b>0</b>	225.23	4.44
DLV3+-MNet [4]	RGB	2018	0.976	0.743	0.847	0.792	0.655	<b>0</b>	<b>14.77</b>	<b>67.70</b>
DLV3+-R101 [4]	RGB	2018	0.980	0.781	0.871	0.824	0.700	<b>0</b>	25.81	38.73
FuseNet [15]	RGB+LiDAR	2017	0.874	0.745	0.852	0.795	0.660	-	-	-
MFNet [14]	RGB+LiDAR	2017	-	0.896	0.903	0.899	0.817	-	-	-
RTFNet [30]	RGB+LiDAR	2019	-	0.842	0.967	0.900	0.818	-	-	-
SNE-RoadSeg [10]	RGB+LiDAR	2020	0.938	0.867	0.927	0.896	0.812	848.51	102.82	1.05
OFF-Net [25]	RGB+LiDAR	2022	0.945	0.866	0.943	0.903	0.823	598.50	31.12	1.58
M2F2-Net [41]	RGB+LiDAR	2023	0.981	0.973	0.955	0.964	0.931	519.53	57.56	1.73
RoadFormer [21]	RGB+LiDAR	2023	0.979	0.951	0.972	0.961	0.925	-	-	-
Ours	RGB	2024	<b>0.991</b>	<b>0.984</b>	<b>0.983</b>	<b>0.983</b>	<b>0.967</b>	<b>0</b>	19.77	50.56

\* Calculating Surface Normal(CSN) from raw LiDAR points.

— The necessary code related to this is not provided in related paper.

TABLE II  
COMPARISON TO OTHER METHODS ON THE TESTING SET OF RELIS-3D

Method	Modality	Year	Accuracy	Precision	Recall	F1_score
U-Net [26]	RGB	2015	0.977	0.809	0.856	0.832
DLV3+-MNet [4]	RGB	2018	0.966	0.570	0.856	0.832
DLV3+-R101 [4]	RGB	2018	0.966	0.586	0.428	0.495
Real-NVP [35]	RGB+LiDAR	2020	0.5625	0.5710	<b>0.9742</b>	0.7001
AE Based [27]	RGB+LiDAR	2022	0.7348	0.7079	0.9181	0.7437
Self-Supervisions Only [28]	RGB+LiDAR	2023	0.9036	0.9164	0.8508	0.8622
M2F2-Net [41]	RGB+LiDAR	2023	0.955	0.925	0.964	0.944
Ours	RGB	2024	<b>0.9786</b>	<b>0.9702</b>	0.9721	<b>0.9712</b>

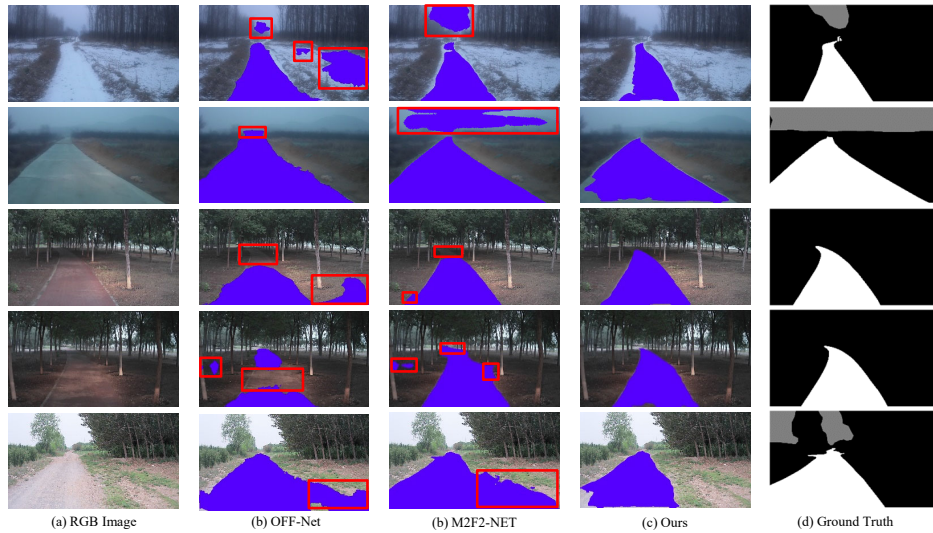


Fig. 4. Qualitative results of OFF-Net [25], M2F2-Net [41] and our method on ORFD dataset [25]. The red boxes are the area where OFF-Net and M2F2-Net predicts incorrectly but ours predicts correctly.

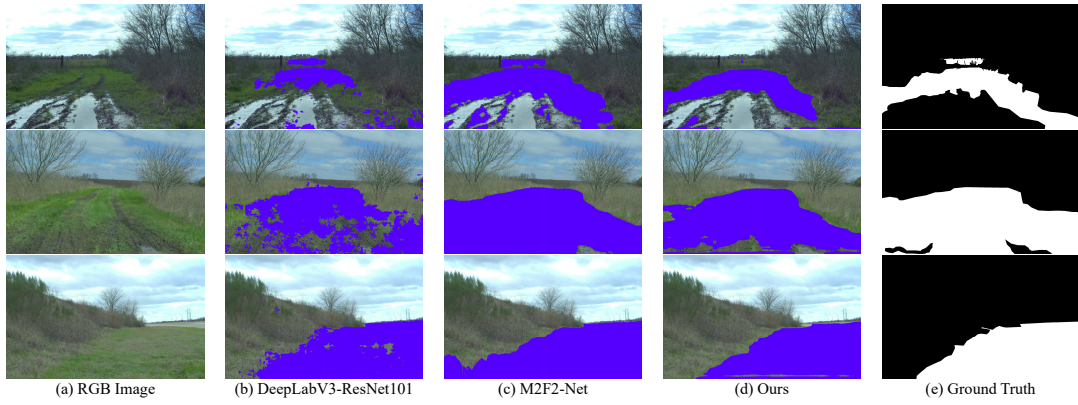


Fig. 5. Qualitative results of DLV3+-R101 [4], M2F2-Net [41] and our method on RELLIS-3D dataset [25].

TABLE III  
ABLATION STUDY ON ViT ENCODER SELECTION

ViT Encoder	Params (M)	IoU	F1_score	FPS
ViT-H	632	0.937	0.968	2.54
ViT-L	308	0.933	0.965	4.58
ViT-B	91	0.931	0.964	13.20
ViT-T	<b>9.8</b>	0.927	0.962	46.34
ViT-S	25	<b>0.967</b>	<b>0.983</b>	<b>50.56</b>

<sup>1</sup> ViT-H is the pre-trained model sam.vit.h.4b8939 in SAM [19];

<sup>2</sup> ViT-L is the pre-trained model sam.vit.l.0b3195 in SAM [19];

<sup>3</sup> ViT-B is the pre-trained model sam.vit.b.01ec64 in SAM [19];

<sup>4</sup> ViT-T is the pre-trained model efficient\_sam.vitt in EfficientSAM [39];

<sup>5</sup> ViT-S is the pre-trained model efficient\_sam.vits in EfficientSAM [39].

that fuse RGB with LiDAR, M2F2-Net and RoadFormer are the previous SOTA methods for off-road freespace detection. Our method, which utilizes only the RGB images, surpasses M2F2-Net with a 0.6% improvement in F1\_score and a 1.1% improvement in IoU. In terms of inference speed, our method, by avoiding the calculation of surface normal maps, is 25 times faster than M2F2-Net.

As shown in Fig. 4, we visualize the prediction results of our method and compare them with those of OFF-Net [25] and M2F2-Net [41] on the ORFD dataset. By comparing the predictions with Ground Truth, it becomes clear that our method outperforms OFF-Net and M2F2-Net. While all models excel in conditions with ample sunlight and clear weather. However, OFF-Net and M2F2-Net face several challenges in low light and adverse weather, including misidentifying the sky as ground, overlooking shaded passable roads, and incorrectly classifying roadside grass as part of the road.

### C. Comparisons on RELLIS-3D Dataset

Comparisons on the RELLIS-3D dataset are detailed in Table II. We compare our approach with various methods,

including U-Net [26], DeepLabV3Plus [4] (DLV3+-MNet and DLV3+-R101), Real-NVP [35], AE Based [27], Self-Supervisions Only [28] and M2F2-Net [41]. The results indicate that our method outperforms these methods, achieving an 2.2% improvement in Accuracy and a 2.7% improvement in the F1\_score.

As shown in Fig. 5, we visualize the prediction results of our method and compare them with those of DeepLabV3Plus-ResNet101 [4] (DLV3+-R101) and M2F2-Net [41] on the RELLIS-3D dataset. By comparing the predictions with Ground Truth, it becomes clear that our method outperforms DLV3+-R101 and M2F2-Net.

### D. Ablation Studies

To determine the most suitable Vision Transformer (ViT) as the encoder for our feature extraction module, we evaluate ViT-H, ViT-L and ViT-B from SAM [19], as well as ViT-T and ViT-S from EfficientSAM [39], on the ORFD dataset [25]. Our evaluations, as presented in Table III, indicates that ViT-H's performance, with a speed of only 3 FPS, could not meet the criteria for real-time applications. According to [39], instance segmentation models using ViT-T are 14% faster than those based on ViT-S. However, this paper only utilizes the encoder parts of ViT-T and ViT-S, both of which have 12 blocks and output 12 hidden features for subsequent feature fusion, thus their inference speeds are roughly comparable. Consequently, we choose ViT-S as the encoder for subsequent experimental comparisons.

## V. CONCLUSIONS

This paper introduces a fast and efficient method for off-road freespace detection. By employing a pre-trained ViT as the encoder, our approach effectively extracts features from RGB data alone, which enhances accuracy and significantly increases the model's inference speed by eliminating the need for surface normal map calculations. Additionally, we design a seg decoder that merges the encoder's latent features with image embeddings, ensuring greater detail retention and higher accuracy. Our method achieves SOTA performance on both the ORFD and RELLIS-3D datasets. Moreover, it performs inference at 50 FPS, significantly outperforming previous methods in terms of speed.

## REFERENCES

- [1] Jens Behley, Martin Garbade, Andres Milioto, Jan Quenzel, Sven Behnke, Cyrill Stachniss, and Jurgen Gall. Semantickitti: A dataset for semantic scene understanding of lidar sequences. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9297–9307, 2019.
- [2] Edouard Capellier, Franck Davoine, Véronique Cherfaoui, and You Li. Fusion of neural networks, for lidar-based evidential road mapping. *Journal of Field Robotics*, 38(5):727–758, 2021.
- [3] Jun Cen, Yizheng Wu, Kewei Wang, Xingyi Li, Jingkang Yang, Yixuan Pei, Lingdong Kong, Ziwei Liu, and Qifeng Chen. Sad: Segment any rgb. *arXiv preprint arXiv:2305.14207*, 2023.
- [4] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018.
- [5] Tianrun Chen, Lanyun Zhu, Chaotao Deng, Runlong Cao, Yan Wang, Shangzhan Zhang, Zejian Li, Lingyun Sun, Ying Zang, and Papa Mao. Sam-adapter: Adapting segment anything in underperformed scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3367–3375, 2023.
- [6] Yangming Cheng, Liulei Li, Yuanyou Xu, Xiaodi Li, Zongxin Yang, Wenguan Wang, and Yi Yang. Segment and track anything. *arXiv preprint arXiv:2305.06558*, 2023.
- [7] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016.
- [8] Ruining Deng, Can Cui, Quan Liu, Tianyuan Yao, Lucas W Remedios, Shunxing Bao, Bennett A Landman, Lee E Wheless, Lori A Coburn, Keith T Wilson, et al. Segment anything model (sam) for digital pathology: Assess zero-shot segmentation on whole slide imaging. *arXiv preprint arXiv:2304.04155*, 2023.
- [9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.
- [10] Rui Fan, Hengli Wang, Peide Cai, and Ming Liu. Sne-roadseg: Incorporating surface normal information into semantic segmentation for accurate freespace detection. In *European Conference on Computer Vision*, pages 340–356. Springer, 2020.
- [11] Rui Fan, Hengli Wang, Peide Cai, Jin Wu, Mohammad Junaid Bocus, Lei Qiao, and Ming Liu. Learning collision-free space detection from stereo images: Homography matrix brings better data augmentation. *IEEE/ASME Transactions on Mechatronics*, 27(1):225–233, 2021.
- [12] Shanghua Gao, Zhijie Lin, Xingyu Xie, Pan Zhou, Ming-Ming Cheng, and Shuicheng Yan. Editanything: Empowering unparalleled flexibility in image editing and generation. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 9414–9416, 2023.
- [13] Shuo Gu, Yigong Zhang, Jinhui Tang, Jian Yang, Jose M Alvarez, and Hui Kong. Integrating dense lidar-camera road detection maps by a multi-modal crf model. *IEEE Transactions on Vehicular Technology*, 68(12):11635–11645, 2019.
- [14] Qishen Ha, Kohei Watanabe, Takumi Karasawa, Yoshitaka Ushiku, and Tatsuya Harada. Mfnet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5108–5115. IEEE, 2017.
- [15] Caner Hazirbas, Lingni Ma, Csaba Domokos, and Daniel Cremers. Fusetnet: Incorporating depth into semantic segmentation via fusion-based cnn architecture. In *Computer Vision—ACCV 2016: 13th Asian Conference on Computer Vision, Taipei, Taiwan, November 20–24, 2016, Revised Selected Papers, Part I 13*, pages 213–228. Springer, 2017.
- [16] Keli Huang, Botian Shi, Xiang Li, Xin Li, Siyuan Huang, and Yikang Li. Multi-modal sensor fusion for auto driving perception: A survey. *arXiv preprint arXiv:2202.02703*, 2022.
- [17] Yu Huang, Chenzhuang Du, Zihui Xue, Xuanyao Chen, Hang Zhao, and Longbo Huang. What makes multi-modal learning better than single (provably). *Advances in Neural Information Processing Systems*, 34:10944–10956, 2021.
- [18] Peng Jiang, Philip Osteen, Maggie Wigness, and Srikanth Saripalli. Rellis-3d dataset: Data, benchmarks and analysis. In *2021 IEEE international conference on robotics and automation (ICRA)*, pages 1110–1116. IEEE, 2021.
- [19] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023.
- [20] Haoran Li, Yaran Chen, Qichao Zhang, and Dongbin Zhao. Bifnet: Bidirectional fusion network for road segmentation. *IEEE transactions on cybernetics*, 52(9):8617–8628, 2021.
- [21] Jiahang Li, Yikang Zhang, Peng Yun, Guangliang Zhou, Qijun Chen, and Rui Fan. Roadformer: Duplex transformer for rgb-normal semantic road scene parsing. *arXiv preprint arXiv:2309.10356*, 2023.
- [22] Xiaolong Liu and Zhidong Deng. Segmentation of drivable road using deep fully convolutional residual network with pyramid pooling. *Cognitive Computation*, 10:272–281, 2018.
- [23] Yingfei Liu, Junjie Yan, Fan Jia, Shuailin Li, Aqi Gao, Tiancai Wang, and Xiangyu Zhang. PetrV2: A unified framework for 3d perception from multi-camera images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3262–3272, 2023.
- [24] Kaustubh Mani, Swapnil Daga, Shubhika Garg, Sai Shankar Narasimhan, Madhava Krishna, and Krishna Murthy Jatavallabhula. Monolayout: Amodal scene layout from a single image. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1689–1697, 2020.
- [25] Chen Min, Weizhong Jiang, Dawei Zhao, Jiaolong Xu, Liang Xiao, Yiming Nie, and Bin Dai. Orfd: A dataset and benchmark for off-road freespace detection. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 2532–2538. IEEE, 2022.
- [26] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention—MICCAI 2015: 18th international conference, Munich, Germany, October 5–9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015.
- [27] Robin Schmid, Deegan Atha, Frederik Schöller, Sharmita Dey, Seyed Fakoorian, Kyohei Otsu, Barry Ridge, Marko Bjelonic, Lorenz Wellhausen, Marco Hutter, et al. Self-supervised traversability prediction by learning to reconstruct safe terrain. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 12419–12425. IEEE, 2022.
- [28] Junwon Seo, Sungdae Sim, and Inwook Shim. Learning off-road terrain traversability with self-supervisions only. *IEEE Robotics and Automation Letters*, 2023.
- [29] Qihong Shen, Xingyi Yang, and Xinchao Wang. Anything-3d: Towards single-view anything reconstruction in the wild. *arXiv preprint arXiv:2304.10261*, 2023.
- [30] Yuxiang Sun, Weixun Zuo, and Ming Liu. Rtfnet: Rgb-thermal fusion network for semantic segmentation of urban scenes. *IEEE Robotics and Automation Letters*, 4(3):2576–2583, 2019.
- [31] Hengli Wang, Rui Fan, Peide Cai, and Ming Liu. Sne-roadseg+: Rethinking depth-normal translation and deep supervision for freespace detection. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1140–1145. IEEE, 2021.
- [32] Jun Wang, Xiaolong Li, Alan Sullivan, Lynn Abbott, and Siheng Chen. Pointmotionnet: Point-wise motion learning for large-scale lidar point clouds sequences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4419–4428, 2022.
- [33] Teng Wang, Jinrui Zhang, Junjie Fei, Yixiao Ge, Hao Zheng, Yunlong Tang, Zhe Li, Mingqi Gao, Shanshan Zhao, Ying Shan, and Feng Zheng. Caption anything: Interactive image description with diverse multimodal controls. *arXiv preprint arXiv:2305.02677*, 2023.
- [34] Yuanbin Wang, Leyan Zhu, Shaofei Huang, Tianrui Hui, Xiaojie Li, Fei Wang, and Si Liu. Cross-modality domain adaptation for freespace detection: A simple yet effective baseline. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4031–4042, 2022.
- [35] Lorenz Wellhausen, René Ranftl, and Marco Hutter. Safe robot navigation via multi-modal anomaly detection. *IEEE Robotics and Automation Letters*, 5(2):1326–1333, 2020.
- [36] Li-Hua Wen and Kang-Hyun Jo. Deep learning-based perception systems for autonomous driving: A comprehensive survey. *Neurocomputing*, 489:255–270, 2022.
- [37] Maggie Wigness, Sungmin Eum, John G Rogers, David Han, and Heesung Kwon. 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)*, pages 5000–5007. IEEE, 2019.

- [38] Pengxiang Wu, Siheng Chen, and Dimitris N Metaxas. Motionnet: Joint perception and motion prediction for autonomous driving based on bird's eye view maps. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11385–11395, 2020.
- [39] Yunyang Xiong, Bala Varadarajan, Lemeng Wu, Xiaoyu Xiang, Fanyi Xiao, Chenchen Zhu, Xiaoliang Dai, Dilin Wang, Fei Sun, Forrest Iandola, et al. Efficientsam: Leveraged masked image pretraining for efficient segment anything. *arXiv preprint arXiv:2312.00863*, 2023.
- [40] Jinyu Yang, Mingqi Gao, Zhe Li, Shang Gao, Fangjing Wang, and Feng Zheng. Track anything: Segment anything meets videos. *arXiv preprint arXiv:2304.11968*, 2023.
- [41] Hongliang Ye, Jilin Mei, and Yu Hu. M2f2-net: Multi-modal feature fusion for unstructured off-road freespace detection. In *2023 IEEE Intelligent Vehicles Symposium (IV)*, pages 1–7. IEEE, 2023.
- [42] Tao Yu, Runseng Feng, Ruoyu Feng, Jinming Liu, Xin Jin, Wenjun Zeng, and Zhibo Chen. Inpaint anything: Segment anything meets image inpainting. *arXiv preprint arXiv:2304.06790*, 2023.
- [43] Chaoning Zhang, Dongshen Han, Yu Qiao, Jung Uk Kim, Sung-Ho Bae, Seungkyu Lee, and Choong Seon Hong. Faster segment anything: Towards lightweight sam for mobile applications. *arXiv preprint arXiv:2306.14289*, 2023.
- [44] Xu Zhao, Wenchao Ding, Yongqi An, Yinglong Du, Tao Yu, Min Li, Ming Tang, and Jinqiao Wang. Fast segment anything. *arXiv preprint arXiv:2306.12156*, 2023.
- [45] Brady Zhou and Philipp Krähenbühl. Cross-view transformers for real-time map-view semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13760–13769, 2022.
- [46] Farnoush Zohourian, Jan Siegemund, Mirko Meuter, and Josef Pauli. Efficient fine-grained road segmentation using superpixel-based cnn and crf models. *arXiv preprint arXiv:2207.02844*, 2022.