SG-RoadSeg+: End-to-End Freespace Detection Upgraded at Data, Feature, and Loss Levels

Ming-Ju Lee, Zhiyuan Wu, Chengju Liu, *Member, IEEE*, Wei Ye, Sergey Vityazev, *Senior Member, IEEE*, Qijun Chen, *Senior Member, IEEE*, and Rui Fan, *Senior Member, IEEE*

***Abstract*—Freespace detection is essential for autonomous robotic systems, particularly self-driving vehicles that depend on precise environmental perception for safe navigation. State-of-the-art methods typically use dual encoders to extract features from RGB images and other data sources, which are then fused and decoded for freespace prediction. This article presents SG-RoadSeg+, which incorporates the transformed disparity-azimuth–zenith (TAZ) map and a novel freespace disparity projection linearity (FDPL) loss function. The key advancement is the TAZ map, which integrates transformed disparity and surface normal maps to enrich spatial understanding, necessary for accurate freespace detection. In addition, the FDPL loss function optimizes the linear projection of the disparity during training, ensuring precise environmental measurements. Moreover, our data augmentation technique, using random perspective transformations, primarily enhances the system's generalization to real-world environments, enabling it to handle complex and dynamic scenarios more effectively. Our previous model, SG-RoadSeg, uses a deep stereo encoder, resulting in strong freespace detection performance. However, its reliance on a single estimated disparity map limits the available spatial information, potentially leading to inaccuracies. Comprehensive experiments on the KITTI Road and Semantics datasets show that SG-RoadSeg+ outperforms the previous version, with improvements of up to 0.73% in Intersection over Union (IoU) and 0.51% in F-score. Notably, applying the TAZ map to other feature-fusion models improves F-score by 3.27% and IoU by 6.44%, highlighting its effectiveness in boosting freespace detection accuracy.**

***Index Terms*—Environmental perception, freespace detection, navigation, robotic systems, self-driving vehicle.**

I. Introduction

TODAY, advancements in autonomous vehicle technology focus on overcoming critical challenges in environmental perception, such as real-time obstacle detection [1], [2], [3] and freespace recognition [4], which are essential for safe and efficient navigation in complex environments [5]. Despite advancements, environmental perception remains a key challenge for self-driving technology. Integrating data from multiple data sources, such as RGB images and disparity maps, has emerged as a crucial solution. This integration enhances environmental perception by providing valuable spatial and geometric information, thereby improving decision-making and navigation capabilities [6], [7], [8]. In particular, geometric data sources, such as surface normal maps and elevation maps, are widely applied in autonomous systems because they provide richer spatial and geometric context.

Accurate freespace detection is essential for autonomous robots to operate effectively in complex environments, ensuring the safety of passengers, pedestrians, and other road users [9]. Early approaches to freespace detection rely on explicit programming, often modeling the region of interest (RoI) as planar surfaces using mathematical representations [10], [11], [12]. However, recent advancements in deep learning have significantly improved this capability, boosting both accuracy and robustness [13]. Notably, feature-fusion networks with independent encoders [14], [15], [16] outperform single-modal networks [17], [18], [19], [20] by leveraging the complementary strengths of RGB images and geometric data. RGB images contribute rich texture details crucial for identifying fine scene features, particularly in dynamic conditions, while geometric data enhances spatial understanding through geometric features [21].

However, obtaining accurate geometric data sources, including depth or disparity maps, remains a significant challenge due to the reliance on precise LiDAR-camera calibration and the sensitivity of point clouds to errors such as crosstalk [22], [23]. Recent techniques, such as the LiDAR super-resolution method [24], have shown promise in enhancing point cloud quality without requiring additional sensors. Such high-quality point clouds are crucial for training deep stereo networks, which are generally trained via fully supervised learning with disparity ground truth obtained from LiDAR

Received 23 December 2024; revised 23 February 2025; accepted 29 May 2025. Date of publication 16 June 2025; date of current version 22 July 2025. This work was supported in part by the National Natural Science Foundation of China under Grant 62388101, Grant 62473288, Grant 62176184, Grant 62233013, and Grant 62333017; in part by the Fundamental Research Funds for the Central Universities; in part by the NIO University Program (NIO UP); and in part by the Xiaomi Young Talents Program. The Associate Editor coordinating the review process was Dr. Jim-Wei Wu. *(Corresponding author: Rui Fan.)*

Ming-Ju Lee, Zhiyuan Wu, Chengju Liu, and Qijun Chen are with the College of Electronics and Information Engineering, Tongji University, Shanghai 201804, China (e-mail: melolee@tongji.edu.cn; gwu@tongji.edu.cn; liuchengju@tongji.edu.cn; qjchen@tongji.edu.cn).

Wei Ye is with the College of Electronics and Information Engineering, Tongji University, Shanghai 201804, China, and also with the Shanghai Innovation Institute, Shanghai 200231, China (e-mail: yew@tongji.edu.cn).

Sergey Vityazev is with the Department of Basis of Radio Engineering and Telecommunication, Ryazan State Radio Engineering University, 390035 Ryazan, Russia (e-mail: vityazev.s.v@ieee.org).

Rui Fan is with the College of Electronics and Information Engineering, Shanghai Research Institute for Intelligent Autonomous Systems, Shanghai Institute of Intelligent Science and Technology, the State Key Laboratory of Intelligent Autonomous Systems, and the Frontiers Science Center for Intelligent Autonomous Systems, Tongji University, Shanghai 201804, China (e-mail: rfan@tongji.edu.cn).

This article has supplementary downloadable material available at https://doi.org/10.1109/TIM.2025.3579733, provided by the authors.

Digital Object Identifier 10.1109/TIM.2025.3579733

data [25], [26], [27], [28], [29], [30]. Nonetheless, the disparity between features required for stereo matching and those for freespace detection highlights the need for integrated approaches to bridge this gap [31]. Therefore, our previous work, SG-RoadSeg [32], proposes a state-of-the-art (SoTA) freespace detection network that effectively addresses these challenges. One of the unique contributions of SG-RoadSeg is its encoder representation sharing strategy, which enriches the features extracted from RGB images. This method allows for a more robust understanding of the scene by effectively combining information from multiple data sources. In addition, the embedded stereo matching component provides accurate depth information without requiring disparity ground truth during training, further enhancing the model's ability to handle complex environments. Despite its improvements, SG-RoadSeg relies on a single disparity map, limiting spatial and geometric information. In addition, the use of a single loss function may exacerbate these limitations.

To address these limitations, we present SG-RoadSeg+, an end-to-end freespace detection network with improvements at the data, feature, and loss levels. A key advancement in SG-RoadSeg+ is the development of the transformed disparity-azimuth–zenith (TAZ) map, which is specifically designed for freespace detection and integrates a broader range of informative geometric cues compared to the disparity map alone. A transformed disparity map refers to a processed disparity map that simulates a quasi-bird's-eye view, where pixels corresponding to freespace have similar values, while those representing road damages or anomalies differ significantly. In addition, the TAZ map incorporates azimuth and zenith components, which are derived from the surface normal map via spherical transformation. This offers a more robust representation of freespace, as the TAZ map exhibits greater similarity in freespace while embedding surface normal information. At the loss level, we introduce the freespace disparity projection linearity (FDPL) loss function, a novel constraint designed to better optimize the linear projection of the disparity during the training process. Finally, SG-RoadSeg+ benefits from a new data augmentation technique based on random perspective transformation, which dynamically strengthens the representation of freespace structures.

To validate the effectiveness of SG-RoadSeg+, we conduct extensive experiments on the KITTI Road [33] and Semantics [34] datasets. Our qualitative and quantitative experimental results demonstrate that SG-RoadSeg+ outperforms all other SoTA single-modal and feature-fusion networks for freespace detection, with the TAZ map emerging as the most informative geometric data source. In summary, our novel contributions are given as follows:

1) a new loss function devised to enhance the accuracy of freespace detection by optimizing the linear projection of freespace disparity;
2) RGB-TAZ feature fusion, initially generating a three-channel feature that exhibits great similarities in freespace and incorporates the surface normal as its geometric information, followed by fusion with the shared features from RGB images;
3) a novel data augmentation technique via random perspective transformation, introducing variations in perspective based on left-right image relationships, enhancing the diversity of training data.

The remainder of this article is organized as follows. Section II provides a review of related work. Section III introduces our proposed SG-RoadSeg+. Section IV presents experimental results and compares SG-RoadSeg+ with other SoTA single-modal and feature-fusion models. In Section V, we discuss the advantages and limitations of our method. Finally, we conclude this article and provide recommendations for future work in Section VI.

II. Related Work

Freespace detection approaches can be categorized into two groups: 1) conventional explicit programming-based methods [11] and 2) data-driven methods [9]. The former typically relies on the "flat road" assumption [35], which is often violated in real-world scenarios. Despite efforts in previous studies [10], [11], [12] to fit road disparity maps and minimize the impact of deviations from the "flat road" assumption, the results remain unsatisfactory.

The data-driven methods, including both single-modal and feature-fusion approaches, are increasingly adopted for improved detection performance. U-Net [17] is notable for its skip connections, which maintain high-resolution features crucial for freespace detection. However, its reliance on pixelwise classification limits its performance in dynamic or unstructured environments. PSPNet [18] improves contextual understanding by aggregating multiscale information. Nonetheless, its fixed-scale approach reduces adaptability in dynamic scenarios. DeepLabv3+ [19] uses atrous convolution to expand the receptive field, which is crucial for dense predictions. Still, its fixed receptive field size limits handling large-scale variations, which are essential for freespace detection.

Single-modal networks often struggle to handle complex scenes, particularly under low illumination or adverse weather conditions. To address these limitations, researchers have explored incorporating additional sources or modalities of vision data to improve scene understanding [36], [37], [38]. FuseNet [39] pioneered feature-fusion techniques in semantic segmentation by employing two independent encoders to extract RGB and depth features, which are then fused via elementwise summation. RTFNet [14] further extends the use of RGB-thermal data for driving scene parsing and introduces an "upception" block to preserve more detailed information. Inspired by DenseNet [40], SNE-RoadSeg [9] incorporates surface normal information for freespace detection and utilizes densely connected skip connections to improve feature decoding. These densely connected skip connections facilitate gradient flow and feature reuse, which are critical for deeper network architectures. In our previous work, SG-RoadSeg [32], we integrate a stereo matching network into the original framework for the guidance of geometric cues. Despite these advancements, such networks remain constrained by their reliance on basic geometric information, such as depth and disparity maps, which often fail to capture the complexities of real-world driving scenarios. SoTA feature-fusion approaches predominantly rely on geometric features derived from surface normals [41], [42], [43]. Although surface normals effectively characterize upward-facing planes, this assumption

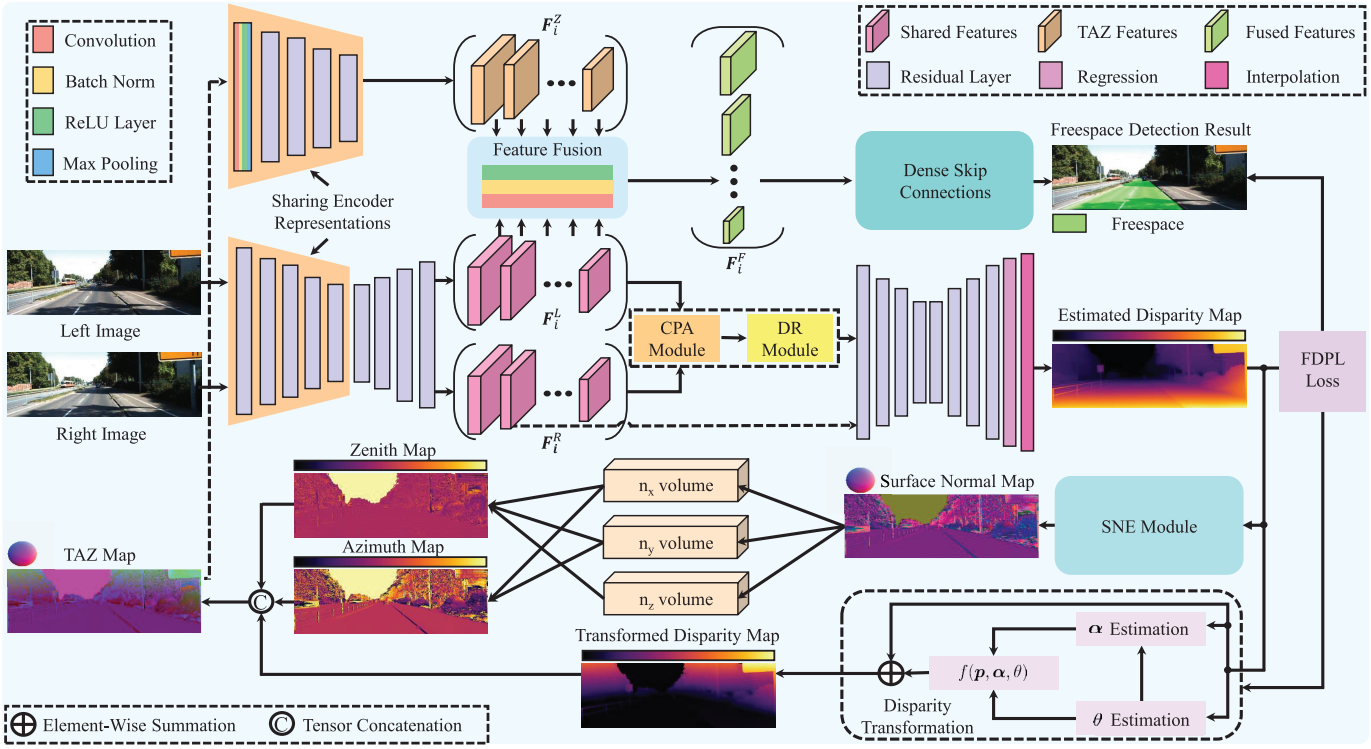

Fig. 1. Architecture of our proposed SG-RoadSeg+.

alone proves inadequate for accurate freespace detection. In complex outdoor environments, surfaces, such as the tops of vehicle roofs and sidewalks, can impede freespace detection. In this article, we introduce the TAZ map, which exhibits greater similarity in freespace while embedding surface normal information. This offers a more robust representation of freespace and integrates a broader range of informative geometric cues compared to disparity maps.

III. Methodology

A. Architecture Overview

In this article, SG-RoadSeg+ consists of four main components: 1) a weight-sharing hourglass network for extracting features from RGB images; 2) a deep stereo network trained via unsupervised learning to refine disparity maps and jointly optimize freespace detection and stereo matching using extracted RGB features; 3) RGB-TAZ feature fusion, which integrates geometric and semantic information by transforming disparity maps into TAZ maps and fusing them with encoded RGB features; and 4) densely connected skip connections to decode fused features and generate final freespace predictions.

Given a pair of stereo images $\boldsymbol{I}^L, \boldsymbol{I}^R \in \mathbb{R}^{H\times W\times 3}$, where H and W represent the height and width of the input images, we employ a weight-sharing hourglass network consisting of a series of residual blocks to extract features $\mathcal{F}^L = \{\boldsymbol{F}_1^L, \ldots, \boldsymbol{F}_n^L\}$ and $\mathcal{F}^R = \{\boldsymbol{F}_1^R, \ldots, \boldsymbol{F}_n^R\}$ from $\boldsymbol{I}^L$ and $\boldsymbol{I}^R$, respectively. Each feature map $\boldsymbol{F}_i^{L,R} \in \mathbb{R}^{(H/S_i)\times(W/S_i)\times C_i}$ represents the i-th level, and C_i and S_i denote the corresponding channel and stride numbers, respectively.

Using the features $\mathcal{F}^L$ and $\mathcal{F}^R$, we first employ a cascaded parallax-attention (CPA) module based on PASMNet [44], where $\mathcal{F}^L$ and $\mathcal{F}^R$ are processed to update matching costs in a coarse-to-fine manner, starting from an initial cost volume $\boldsymbol{V}_0 \in \mathbb{R}^{(H/16)\times(W/16)\times(W/16)}$ with elements set to zero. The cost matching process in the m-th ($m > 0$) parallax-attention block $\boldsymbol{V}_m \in \mathbb{R}^{(H/2^{5-m})\times(W/2^{5-m})\times(W/2^{5-m})}$ can be formulated as follows:

$$\boldsymbol{V}_{m+1} = \text{Upsample}(\boldsymbol{V}_m) + \sum_{i,j}\left(\boldsymbol{\sigma}_{m+1}^{Q}\boldsymbol{\sigma}_i\boldsymbol{F}_i^L\right)^{\top}\boldsymbol{\sigma}_{m+1}^{K}\boldsymbol{\sigma}_j\boldsymbol{F}_j^R \quad (1)$$

where $\boldsymbol{\sigma}_i^Q, \boldsymbol{\sigma}_i^K \in \mathbb{R}^{C_i\times C_i}$ refer to 1×1 convolution kernels for query and key feature maps, $\boldsymbol{\sigma}$ represents two layers of 3×3 convolution kernels, and $^{\top}$ denotes the transpose of the matrix. Through this coarse-to-fine strategy, disparities can be progressively refined, and better performance can be achieved. Then, we use the disparity regression (DR) module to generate the initial disparity map $\boldsymbol{D}^I \in \mathbb{R}^{(H/4)\times(W/4)}$, which can be formulated as follows:

$$\boldsymbol{D}^I = \sum_{k=0}^{\frac{W}{4}-1} k\boldsymbol{M}^P(:,:,k) \quad (2)$$

where $\boldsymbol{M}^P \in \mathbb{R}^{(H/4)\times(W/4)\times(W/4)}$ denotes the softmax parallax-attention map produced by the CPA module. Following this, another hourglass network is employed to refine disparities. As shown in Fig. 1, the second feature map in the hourglass network is concatenated with the initial disparity map $\boldsymbol{D}^I$ and then fed into the DR module, followed by a series of residual blocks to produce an auxiliary disparity map $\boldsymbol{D}^A \in \mathbb{R}^{H\times W}$ and a confidence map $\boldsymbol{M}^C \in \mathbb{R}^{H\times W}$. The refined disparity map $\boldsymbol{D}^R \in \mathbb{R}^{H\times W}$ is calculated as follows:

$$\boldsymbol{D}^R = \text{Upsample}(\boldsymbol{D}^I)(1-\boldsymbol{M}^C) + \boldsymbol{D}^A\boldsymbol{M}^C. \quad (3)$$

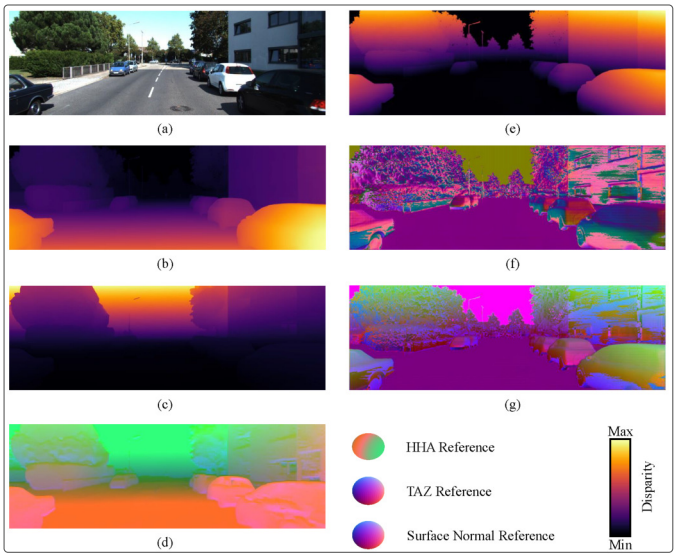

Fig. 2. Input data types. The shared color bar represents the value range for the disparity, elevation, and transformed disparity maps. (a) RGB image. (b) Disparity map. (c) Elevation map. (d) HHA map. (e) Transformed disparity map. (f) Surface normal map. (g) TAZ map.

This refinement process improves disparity estimation by leveraging structural information from the second feature map, effectively reducing errors in occluded and edge regions while maintaining high accuracy in well-initialized areas.

After obtaining a refined disparity map, the surface normal map, denoted as $\boldsymbol{N}^S \in \mathbb{R}^{H\times W\times 3}$, is generated using the SNE module [9]. To construct the transformed disparity map $\boldsymbol{D}^T \in \mathbb{R}^{H\times W}$, the following equation is applied [10]:

$$\boldsymbol{D}^T(\boldsymbol{p}) = \boldsymbol{D}^R(\boldsymbol{p}) - f(\boldsymbol{p}, \boldsymbol{\alpha}, \theta) \tag{4}$$

where f is a linear function representing the freespace disparities, $\boldsymbol{p} = (u; v)$ is a pixel in $\boldsymbol{D}^R$, $\boldsymbol{\alpha} = (\alpha_0; \alpha_1)$ stores the road profile model coefficients, and θ is the stereo rig roll angle. Next, the surface normal map $\boldsymbol{N}^S$ undergoes a spherical transformation to yield two maps: the azimuth map $\boldsymbol{N}^A \in \mathbb{R}^{H\times W}$ and the zenith map $\boldsymbol{N}^Z \in \mathbb{R}^{H\times W}$. These two maps, along with the transformed disparity map $\boldsymbol{D}^T$, are concatenated to form the TAZ map. Subsequently, we process the RGB-TAZ feature fusion operation under the encoder sharing strategy, where shared features are remapped into semantic feature space and fused with TAZ features. Finally, we employ the densely connected skip connection [9] to decode fused features and generate the semantic prediction of the freespace.

B. RGB-TAZ Feature Fusion

Combining RGB features with geometric features enhances freespace detection and improves the model's robustness by integrating both color information and spatial orientation, providing a more complete understanding of the environment. Freespace can be hypothesized as a ground plane where points exhibit similar values on both the transformed disparity map and the surface normal map [see Fig. 2(e) and (f)], making these maps highly informative for freespace detection [45]. Building on these ideas, we introduce the TAZ map, which highlights these similarities and preserves the orientation information from the surface normal map.

To enhance both interpretability and computational efficiency, we transform the surface normal map from Cartesian coordinates to Spherical coordinates, producing azimuth and zenith maps. Given the surface normal map $\boldsymbol{N}^S = (\boldsymbol{N}_x^S, \boldsymbol{N}_y^S, \boldsymbol{N}_z^S)$, the azimuth map $\boldsymbol{N}^A \in \mathbb{R}^{H\times W}$ and the zenith map $\boldsymbol{N}^Z \in \mathbb{R}^{H\times W}$ are computed as follows [9]:

$$\boldsymbol{N}^A(\boldsymbol{p}) = \arctan\left(\frac{\boldsymbol{N}_y^S(\boldsymbol{p})}{\boldsymbol{N}_x^S(\boldsymbol{p})}\right) \tag{5}$$

$$\boldsymbol{N}^Z(\boldsymbol{p}) = \arccos\left(\boldsymbol{N}_z^S(\boldsymbol{p})\right). \tag{6}$$

Next, we construct the TAZ map by concatenating the transformed disparity map $\boldsymbol{D}^T$, the azimuth map $\boldsymbol{N}^A$, and the zenith map $\boldsymbol{N}^Z$.

Subsequently, we perform RGB-TAZ feature fusion operation. Given the left RGB feature maps $\mathcal{F}^L = \{\boldsymbol{F}_1^L, \ldots, \boldsymbol{F}_n^L\}$ and TAZ feature maps $\mathcal{F}^Z = \{\boldsymbol{F}_1^Z, \ldots, \boldsymbol{F}_n^Z\}$, we obtain the fused feature sequence $\mathcal{F}^F = \{\boldsymbol{F}_1^F, \ldots, \boldsymbol{F}_n^F\}$ using our proposed RGB-TAZ feature fusion operation, which can be formulated as follows:

$$\boldsymbol{F}_i^F = \mathrm{ReLU}\left(\mathrm{Norm}\left(\underset{1\times 1}{\mathrm{Conv}}\left(\boldsymbol{F}_i^L\right)\right)\right) + f_{\mathrm{enc}}\left(\boldsymbol{F}_i^Z\right) \tag{7}$$

where $\boldsymbol{F}_i^{F,Z} \in \mathbb{R}^{(H/2^i)\times(W/2^i)\times C_i}$ represents the i-th feature map and C_1–C_5 are 64, 256, 512, 1024, and 2048, respectively. $f_{\mathrm{enc}}(\cdot)$ represents the feature encoding operation, starting with a convolutional layer, batch normalization, and ReLU activation, followed by a max pooling layer and four residual layers based on ResNet-152 [46]. ResNet-152 is selected for its superior performance over other ResNet architectures [32], with its structure dictating the channel sizes C_1–C_5. This design extracts and refines high-level spatial features from TAZ maps, ensuring effective fusion with RGB features.

C. Freespace Disparity Projection Linearity Loss

The road profile model coefficients $\boldsymbol{\alpha}$ in (4) can be obtained by minimizing as follows [10]:

$$E(\theta) = \|\boldsymbol{d} - \boldsymbol{T}(\theta)\boldsymbol{\alpha}\|_2^2 \tag{8}$$

where $\boldsymbol{T} = (\boldsymbol{1}_k, \boldsymbol{v}\cos\theta - \boldsymbol{u}\sin\theta)$, $\boldsymbol{d} = (d_1; \ldots; d_k)$ is a k-entry vector of disparities, $\boldsymbol{u} = (u_1; \ldots; u_k)$ is a k-entry vector of horizontal coordinates, and $\boldsymbol{v} = (v_1; \ldots; v_k)$ is a k-entry vector of vertical coordinates. The closed-form solution for (8) is given as follows [10]:

$$\hat{\boldsymbol{\alpha}} = \left(\boldsymbol{T}(\theta)^\top \boldsymbol{T}(\theta)\right)^{-1} \boldsymbol{T}(\theta)^\top \boldsymbol{d}. \tag{9}$$

Therefore, we can introduce an additional constraint via the FDPL loss $\mathcal{L}_e$, expressed as follows:

$$\mathcal{L}_e = h(E_p) = \lambda\left(1 - e^{-E_p}\right), \tag{10}$$

where

$$E_p = \frac{\boldsymbol{d}^\top\boldsymbol{d} - \boldsymbol{d}^\top\boldsymbol{T}(\theta)\left(\boldsymbol{T}(\theta)^\top\boldsymbol{T}(\theta)\right)^{-1}\boldsymbol{T}(\theta)^\top\boldsymbol{d}}{N}, \tag{11}$$

λ denotes the loss weight and N refers to the number of freespace pixels, determined by the freespace detection branch.

With the constraint of FDPL loss, the freespace disparity projection is refined and optimized, leading to improved accuracy in freespace detection.

To evaluate the effectiveness and robustness of the proposed FDPL loss, we compare it with three alternative loss functions that adopt different approaches to normalize E_p in (11) as follows:

$$\mathcal{L}_e^S = h_\sigma(E_p) = \lambda\left(\frac{1}{1+e^{-E_p}}\right) \tag{12}$$

$$\mathcal{L}_e^T = h_\tau(E_p) = \lambda\left(\frac{e^{E_p}-e^{-E_p}}{e^{E_p}+e^{-E_p}}\right) \tag{13}$$

$$\mathcal{L}_e^A = h_\alpha(E_p) = \lambda\left(\frac{\arctan(E_p)}{\pi}+\frac{1}{2}\right). \tag{14}$$

Specifically, we utilize the Sigmoid, Tanh, and Arctan functions, represented by $h_\sigma(E_p)$, $h_\tau(E_p)$, and $h_\alpha(E_p)$, respectively. These alternative loss functions were chosen due to their frequent adoption in related tasks, where nonlinear functions are often used to stabilize gradients and enhance optimization.

In addition to FDPL loss, we also employ a pixelwise cross-entropy loss $\mathcal{L}_c$ to supervise the semantic segmentation and train the stereo matching network in an unsupervised way by minimizing the loss function $\mathcal{L}_u$. The total loss can be written as follows:

$$\mathcal{L} = \mathcal{L}_e + \mathcal{L}_c + \mathcal{L}_u \tag{15}$$

where $\mathcal{L}_u$ is formulated as follows:

$$\mathcal{L}_u = \mathcal{L}_p + \lambda_s\mathcal{L}_s + \lambda_m\mathcal{L}_m \tag{16}$$

where $\mathcal{L}_p$ represents the photometric loss function [47], [48], $\mathcal{L}_s$ refers to the smoothness loss function, and $\mathcal{L}_m$ refers to the parallax-attention map loss [44]. λ_s and λ_m are empirically set to 0.5 and 1, respectively.

D. Data Augmentation via Random Perspective Transformation

Deep convolutional neural networks rely heavily on high-quality, labeled training data. While common augmentation methods such as reflections and rotations enhance data diversity, they may distort the geometric relationships in stereo rigs, compromising viewpoint integrity. To overcome this limitation, we propose a simple yet effective augmentation technique based on our previous study [49].

To correct the distortion caused by the stereo camera roll angle θ, we first apply an affine transformation to the left image $\boldsymbol{I}^L$ and its corresponding ground-truth image $\boldsymbol{I}^{\mathrm{GT}}$. With the corrected left image $\boldsymbol{I}^L$ and ground-truth image $\boldsymbol{I}^{\mathrm{GT}}$ (collectively denoted as $\boldsymbol{I}^C$), our random perspective transformation can be formulated as follows:

$$\boldsymbol{I}^G(\boldsymbol{p}_i) = \begin{cases} \boldsymbol{I}^C(\boldsymbol{p}_i - (d_i;0)), & W \geq u_i - d_i > 0 \\ 0, & \text{otherwise} \end{cases} \tag{17}$$

where

$$d_i = D\frac{(v_r - v_i)}{v_r}, \quad D \in [-T, T]. \tag{18}$$

$\boldsymbol{p}_i = (u_i; v_i)$ is a pixel in the generated image $\boldsymbol{I}^G$, v_r refers to the minimum vertical coordinate of the pixel in the freespace, and T is a preset threshold. By setting the reference row at v_r, our approach preserves the integrity of the freespace structure while introducing controlled variability through linear translation distance d_i. Specifically, the variable D, which varies within a specified range, creates diverse perspective transformations while maintaining alignment across rows.

IV. Experiments

A. Datasets and Experimental Setup

We utilize two public datasets to evaluate the performance of our SG-RoadSeg+ for freespace detection: the KITTI Road [33] dataset and the KITTI Semantics [34] dataset. The details are given as follows.

1) The KITTI Road [33] dataset is a real-world driving scenario dataset containing 289 pairs of stereo images, as well as their pixel-level freespace detection ground-truth annotations. It includes three categories of road scenes: urban unmarked (UU), urban marked (UM), and urban multiple marked (UMM). The resolution of KITTI Road images ranges from 370×1224 to 375×1242.
2) The KITTI Semantics [34] dataset contains 200 real-world images captured in a variety of urban driving scenarios. It provides ground-truth semantic annotations for 19 different classes. To evaluate the performance of SG-RoadSeg+, we extract pixels belonging to the "road" class and consider them the ground-truth annotation of the freespace.

Our experiments are conducted on an NVIDIA RTX 3090 GPU. For each dataset, we allocate 70% of images for training, while the remaining data are used as the test set. The batch size is set to 1. We utilize the Adam optimizer for model training. The initial learning rate is set to 1×10^{-3}. Training lasts for 200 epochs on each dataset.

Five common metrics are used for the performance evaluation of freespace detection: 1) accuracy (Acc); 2) precision (Pre); 3) recall (Rec); 4) F-score (Fsc); and 5) Intersection over Union (IoU) [9].

B. Ablation Studies

1) RGB-TAZ Feature Fusion: This study presents a detailed comparison of four widely recognized feature-fusion networks for freespace detection, along with our SG-RoadSeg+ model. Each network is evaluated using different input concatenations of RGB images with various feature maps, as shown in Table I. Results show that SG-RoadSeg+ consistently outperforms other feature-fusion networks across all input data types. In particular, SG-RoadSeg+ achieves the highest IoU and F-score when using the RGB+Z setup, highlighting the advantage of leveraging the TAZ map for feature fusion. Notably, the integration of our TAZ map significantly enhances the performance of other feature-fusion networks, yielding improvements of approximately 3.27% in F-score and 6.44% in IoU. In the Supplementary Material, we include a qualitative comparison that visually demonstrates the performance of SG-RoadSeg+ and other feature-fusion networks across different input data types.

TABLE I

PERFORMANCE COMPARISON AMONG FOUR FEATURE-FUSION NETWORKS AND OUR SG-ROADSEG+ WITH DIFFERENT SETUPS ON THE KITTI ROAD [33] DATASET. RGB + D, RGB + E, AND SO ON REPRESENT INPUT CONCATENATIONS OF RGB IMAGES AND FEATURE MAPS, INCLUDING DISPARITY MAPS (D), ELEVATION MAPS (E), HHA MAPS (H), TRANSFORMED DISPARITY MAPS (T), SURFACE NORMAL MAPS (N), AND TAZ MAPS (Z). THE BEST RESULTS ARE SHOWN IN BOLD FONT

Networks	RGB+D		RGB+E		RGB+H		RGB+T		RGB+N		RGB+Z	
	Fsc (%)	IoU (%)	Fsc (%)	IoU (%)	Fsc (%)	IoU (%)	Fsc (%)	IoU (%)	Fsc (%)	IoU (%)	Fsc (%)	IoU (%)
RTFNet	97.51	95.13	97.42	95.15	97.43	95.13	97.71	95.88	97.53	95.82	97.73	95.90
D-A CNN	97.54	95.20	97.23	95.10	97.26	95.01	97.64	95.62	97.56	95.61	97.70	95.72
MFNet	94.98	89.96	94.78	90.05	94.68	89.99	97.23	95.93	97.12	95.41	97.33	96.01
FuseNet	94.76	89.72	94.71	90.12	94.80	89.99	97.51	95.90	97.46	95.89	97.53	95.96
SG-RoadSeg+	**97.56**	**95.78**	**97.52**	**95.96**	**97.53**	**95.90**	**97.75**	**96.02**	**97.59**	**95.97**	**97.98**	**96.16**

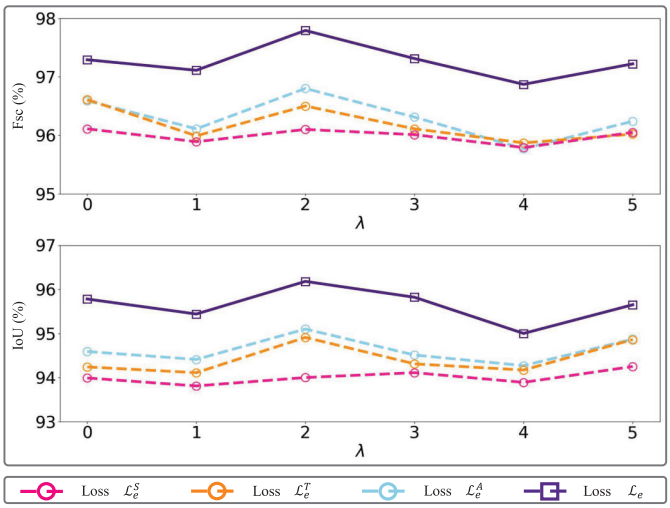

Fig. 3. Ablation study in terms of different FDPL loss functions and the selection of λ on the KITTI Road [33] dataset.

2) FDPL Loss Function: Fig. 3 illustrates the quantitative results of our freespace detection experiments, highlighting the impact of different loss functions and varying λ values from 0 to 5. It is evident that FDPL loss consistently improves the performance of SG-RoadSeg+ across all tested λ values. Specifically, SG-RoadSeg+ performs the best when utilizing the FDPL loss with $\lambda = 2$. However, it is important to note that smaller λ values are generally more effective for FDPL loss when training with limited data, as they help mitigate overfitting. In contrast, larger λ values can lead to marginal improvements in performance but are more likely to cause overfitting or diminishing returns beyond a certain threshold.

3) Component Contributions: In addition, we conduct an ablation study to isolate the individual contributions of the TAZ map, FDPL loss, and data augmentation, comparing their effects on SG-RoadSeg's baseline performance. Table II clearly shows the improvement brought by each component in SG-RoadSeg's performance. The incorporation of the TAZ map results in the largest improvement, with an increase of 0.53%, while the introduction of FDPL loss provides an additional boost of 0.23%. This may be due to the TAZ map's ability to more effectively represent the geometric structure of the environment, which plays a crucial role in improving freespace detection, while FDPL loss provides more subtle refinement in optimization. The observed performance boost from combining FDPL loss with RGB-TAZ feature fusion indicates the complementary roles of these components. RGB-TAZ feature fusion contributes to more accurate spatial representation, while FDPL loss refines the geometric details, leading to enhanced performance when the two are combined. Furthermore, while data augmentation yields a modest improvement of 0.24%, the model achieves its peak performance when all components are integrated, resulting in an increase in F-score by up to 0.99%. This demonstrates that each component plays a vital role, with TAZ map and FDPL loss enhancing spatial and geometric accuracy, while data augmentation ensures better generalization across varied input data.

TABLE II

ABLATION STUDY ANALYZING THE IMPACT OF INDIVIDUAL COMPONENTS: TAZ MAP (TAZ), FDPL LOSS (FDPL), AND DATA AUGMENTATION VIA RANDOM PERSPECTIVE TRANSFORMATION (DA) ON BASELINE'S PERFORMANCE

Baseline	TAZ	FDPL	DA	Fsc(%) ↑
✓				96.25
	✓			96.78
	✓	✓		97.01
			✓	96.49
	✓	✓	✓	97.24

4) Impact of ResNet Backbones: Finally, we evaluated the proposed architecture against the SG-RoadSeg [32], using five different CNN backbones: ResNet-18, ResNet-34, ResNet-50, ResNet-101, and ResNet-152. The quantitative results, presented in Fig. 4, demonstrate that SG-RoadSeg+ consistently surpasses the baseline SG-RoadSeg across all tested ResNet backbones. Among them, ResNet-152 achieves the highest performance, reflecting its advanced feature extraction capabilities and superior performance in semantic segmentation tasks. While ResNet-18 shows a relatively smaller improvement, this could be attributed to its limited capacity for extracting complex features, which restricts its ability to fully exploit the diverse visual information provided by the TAZ map. In contrast, the deeper ResNet models demonstrate more significant improvements, underscoring their ability to capture

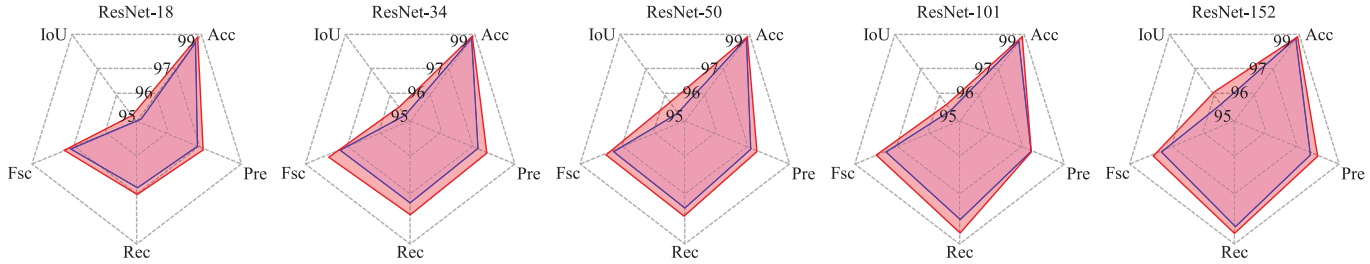

Fig. 4. Comparison (%) between SG-RoadSeg and SG-RoadSeg+ with respect to different CNN backbones, where —— presents the results achieved by SG-RoadSeg and —— presents the results achieved by SG-RoadSeg+.

Fig. 5. Qualitative comparison between SG-RoadSeg and SG-RoadSeg+ in complex scenarios. (a)–(c) Different example scenes under challenging conditions.

TABLE III

COMPARISONS OF SOTA FEATURE-FUSION NETWORKS ON THE KITTI ROAD [33] DATASET. THE BEST RESULTS ARE SHOWN IN BOLD FONT

Networks	Publication	Acc (%) ↑	Pre (%) ↑	Rec (%) ↑	Fsc (%) ↑	IoU (%) ↑
SNE-RoadSeg [9]	ECCV'20	98.61	97.59	97.49	97.41	95.22
CLCFNet [15]	ICRA'21	98.25	97.23	97.05	97.22	95.13
PLB-RD [50]	TCSVT'22	98.71	97.69	97.44	97.60	95.52
OFF-Net [16]	ICRA'22	96.58	93.33	94.75	93.74	89.40
US-Net [41]	ICRA'22	98.63	97.66	97.59	97.52	95.32
3MT-RoadSeg [42]	RA-L'23	98.50	96.40	96.15	96.27	93.15
M2F2-Net [43]	SIV'23	97.67	96.47	96.56	95.89	94.33
SG-RoadSeg [32]	ICRA'24	98.76	97.52	98.09	97.77	95.74
SG-RoadSeg+	TIM'25	98.89	97.68	98.26	97.98	96.16
DA-SG-RoadSeg+	TIM'25	**98.97**	**97.83**	**98.51**	**98.14**	**96.35**

TABLE IV

COMPARISONS OF SOTA FEATURE-FUSION NETWORKS ON THE KITTI SEMANTICS [34] DATASET. THE BEST RESULTS ARE SHOWN IN BOLD FONT

Networks	Publication	Acc (%) ↑	Pre (%) ↑	Rec (%) ↑	Fsc (%) ↑	IoU (%) ↑
SNE-RoadSeg [9]	ECCV'20	97.50	96.58	95.89	96.13	92.88
CLCFNet [15]	ICRA'21	97.47	96.11	95.46	95.92	92.69
PLB-RD [50]	TCSVT'22	97.98	96.84	96.12	96.37	93.01
OFF-Net [16]	ICRA'22	95.57	93.37	93.76	93.30	88.33
US-Net [41]	ICRA'22	97.55	96.67	95.98	96.22	92.97
3MT-RoadSeg [42]	RA-L'23	97.45	96.40	95.80	95.94	92.70
M2F2-Net [43]	SIV'23	96.89	95.25	95.63	95.11	92.35
SG-RoadSeg [32]	ICRA'24	97.88	97.01	96.72	96.75	93.93
SG-RoadSeg+	TIM'25	98.12	97.18	96.97	97.08	94.66
DA-SG-RoadSeg+	TIM'25	**98.30**	**97.34**	**97.16**	**97.32**	**95.17**

richer contextual and spatial details essential for accurate freespace detection.

C. Performance Evaluation

In this subsection, we evaluate the effectiveness of SG-RoadSeg+ for freespace detection both qualitatively and quantitatively. Qualitative results, as shown in Fig. 5, underscore SG-RoadSeg+'s superior ability to handle challenging scenarios over SG-RoadSeg, such as dynamic obstacles and cluttered environments. The improved performance of SG-RoadSeg+ can be attributed to the combination of RGB-TAZ feature fusion and FDPL loss. RGB-TAZ feature fusion enhances spatial representation by leveraging both semantic and geometric information, while FDPL loss refines geometric precision by optimizing disparity projection. This combination leads to a significant performance boost, particularly in handling complex scenarios with occlusion and dynamic motion. Furthermore, SG-RoadSeg+ strikes an effective balance between performance and computational efficiency, offering notable improvements in freespace detection while maintaining real-time feasibility. In comparison to SG-RoadSeg, SG-RoadSeg+ achieves an improvement of 0.21% in F-score and 0.42% in IoU on the KITTI Road dataset, and 0.33% in F-score and 0.73% in IoU on the KITTI Semantics dataset. Furthermore, it can be seen that the F-score and IoU of our SG-RoadSeg+, when trained on the generated dataset via random perspective transformation, are improved by 0.13% and 0.16% respectively on the KITTI Road dataset, and by 0.24% and 0.51% on the KITTI Semantics dataset. SG-RoadSeg+ shows a minimal increase in parameters (+0.007M) and a slight reduction in FPS (-0.53), indicating a small trade-off in speed for a significant improvement in performance. This characteristic enables future upgrades with faster feature-fusion networks, facilitating the development of models that provide both superior freespace detection and real-time performance. Quantitative comparisons are presented in Tables III and IV, demonstrating that SG-RoadSeg+ achieves the best results on both datasets. DA-SG-RoadSeg+ refers to SG-RoadSeg+ with data augmentation via random perspective transformation. While these feature-fusion methods demonstrate good performance, they face challenges in capturing fine details in complex scenarios, such as differentiating between freespace and sidewalks, as shown in Figs. 6 and 7. A major advantage of SG-RoadSeg+ lies in its use of the TAZ map, which demonstrates great similarity in freespace and incorporates geometric cues. This enables a clearer distinction between freespace and obstacles,

Fig. 6. Qualitative comparison between our proposed SG-RoadSeg+ and other SoTA feature-fusion networks on the KITTI Road [33] dataset. The true positive, false negative, and false positive pixels are shown in green, red, and blue, respectively.

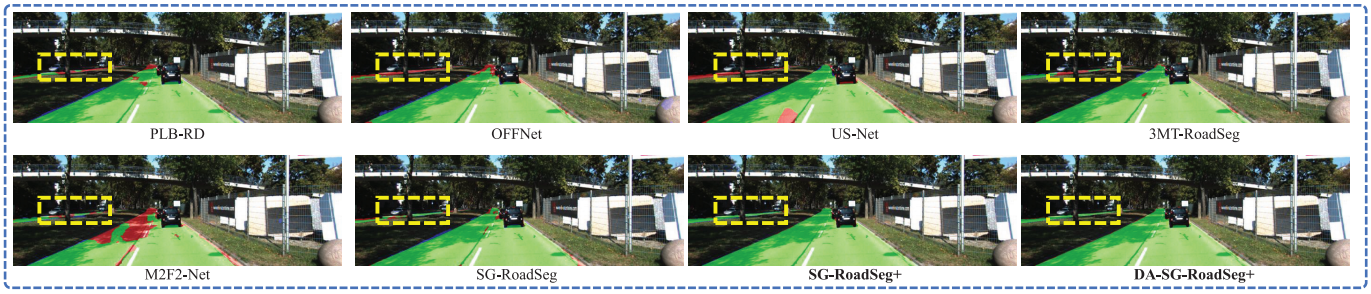

Fig. 7. Qualitative comparison between our proposed SG-RoadSeg+ and other SoTA feature-fusion networks on the KITTI Semantics [34] dataset. The true positive, false negative, and false positive pixels are shown in green, red, and blue, respectively.

addressing the difficulty of identifying subtle differences in challenging environments. For more detailed comparisons, we present quantitative and qualitative experiments on single-modal networks in the Supplementary Material.

V. Discussion

The experimental results shown in Section IV provide strong support for the claims made in Section I. First, the RGB-TAZ feature fusion and FDPL loss significantly enhance spatial representation and geometric precision, leading to notable performance improvements, especially in scenarios with occlusions and dynamic obstacles. Second, the incorporation of data augmentation techniques via random perspective transformation enables the generation of more training data in unstructured environments, further improving the model's robustness. Finally, SG-RoadSeg+ achieves an optimal balance between performance and computational efficiency, ensuring real-time feasibility while boosting freespace detection accuracy. We believe that our proposed SG-RoadSeg+ can be further improved after addressing the following limitations.

1) The accuracy of the TAZ map and the constraint effectiveness of FDPL loss may degrade in unstructured or highly cluttered environments. To overcome this, employing more advanced feature fusion techniques capable of fully leveraging both RGB and geometric information is essential.
2) SG-RoadSeg+ achieves a processing speed of 3.36 FPS for RGB input images with a resolution of 1024×320 pixels. Further optimizations in computational efficiency are necessary for real-time deployment in autonomous robotic systems. Therefore, investigating the integration of lightweight backbone architectures and leveraging hardware acceleration presents a promising avenue for future research to improve real-time processing efficiency.

VI. Conclusion and Future

In this article, we presented three key contributions. 1) FDPL loss, an additional constraint to optimize the projection of the freespace disparity; 2) TAZ map, demonstrating significant similarities in freespace and incorporating the surface normal as additional geometric information; and 3) data augmentation via random perspective transformation, greatly enhancing the diversity of our training dataset. Extensive experiments conducted on two KITTI datasets indicate that our SG-RoadSeg+ can produce more accurate and robust results.

In the future, we will investigate advanced feature fusion methods that jointly exploit RGB and geometric cues. We will also explore the use of lightweight network architectures combined with hardware acceleration for real-time deployment.

References

[1] H. Hu, T. Zhao, Q. Wang, F. Gao, L. He, and Z. Gao, "Monocular 3-D vehicle detection using a cascade network for autonomous driving," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–13, 2021.
[2] C. Chen, H. Qin, Y. Qin, and Y. Bai, "Real-time railway obstacle detection based on multitask perception learning," *IEEE Trans. Intell. Transp. Syst.*, vol. 26, no. 5, pp. 7142–7155, May 2025.
[3] H. Mu, G. Zhang, Z. Ma, M. Zhou, and Z. Cao, "Dynamic obstacle avoidance system based on rapid instance segmentation network," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 5, pp. 4578–4592, Nov. 2023.
[4] S. Li, Q. Yan, W. Shi, L. Wang, C. Liu, and Q. Chen, "HoloParser: Holistic visual parsing for real-time semantic segmentation in autonomous driving," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–15, 2023.
[5] R. Zhou, Y. Gao, Y. Wang, X. Xie, and X. Zhao, "A real-time scene parsing network for autonomous maritime transportation," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–14, 2023.
[6] Y. R. Wang, Y. Tian, J. Chen, K. Xu, and X. Ding, "A survey of visual SLAM in dynamic environment: The evolution from geometric to semantic approaches," *IEEE Trans. Instrum. Meas.*, vol. 73, pp. 1–21, 2024.

[7] S. Wen, S. Tao, X. Liu, A. Babiarz, and F. R. Yu, "CD-SLAM: A real-time stereo visual–inertial SLAM for complex dynamic environments with semantic and geometric information," *IEEE Trans. Instrum. Meas.*, vol. 73, pp. 1–8, 2024.

[8] J. J. Xie, J. Dou, L. Zhong, J. Di, and Y. Qin, "A dual-mode intensity and polarized imaging system for assisting autonomous driving," *IEEE Trans. Instrum. Meas.*, vol. 73, pp. 1–13, 2024.

[9] R. Fan, H. Wang, P. Cai, and M. Liu, "SNE-RoadSeg: Incorporating surface normal information into semantic segmentation for accurate freespace detection," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 340–356.

[10] R. Fan and M. Liu, "Road damage detection based on unsupervised disparity map segmentation," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 11, pp. 4906–4911, Nov. 2020.

[11] A. Wedel, H. Badino, C. Rabe, H. Loose, U. Franke, and D. Cremers, "B-spline modeling of road surfaces with an application to free-space estimation," *IEEE Trans. Intell. Transp. Syst.*, vol. 10, no. 4, pp. 572–583, Dec. 2009.

[12] R. Fan, U. Ozgunalp, B. Hosking, M. Liu, and I. Pitas, "Pothole detection based on disparity transformation and road surface modeling," *IEEE Trans. Image Process.*, vol. 29, pp. 897–908, 2020.

[13] Y. Feng et al., "SNE-RoadSegV2: Advancing heterogeneous feature fusion and fallibility awareness for freespace detection," *IEEE Trans. Instrum. Meas.*, vol. 74, pp. 1–9, 2025.

[14] Y. Sun, W. Zuo, and M. Liu, "RTFNet: RGB-thermal fusion network for semantic segmentation of urban scenes," *IEEE Robot. Autom. Lett.*, vol. 4, no. 3, pp. 2576–2583, Jul. 2019.

[15] S. Gu, J. Yang, and H. Kong, "A cascaded LiDAR-camera fusion network for road detection," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2021, pp. 13308–13314.

[16] C. Min et al., "Orfd: A dataset and benchmark for off-road freespace detection," in *Proc. Int. Conf. Robot. Autom. (ICRA)*, 2022, pp. 2532–2538.

[17] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. 18th Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, vol. 9351. Cham, Switzerland: Springer, 2015, pp. 234–241.

[18] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6230–6239.

[19] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder–decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Jan. 2018, pp. 833–851.

[20] H. Pan, Y. Hong, W. Sun, and Y. Jia, "Deep dual-resolution networks for real-time and accurate semantic segmentation of traffic scenes," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 3, pp. 3448–3460, Mar. 2023.

[21] M. Sun et al., "Attention-rectified and texture-enhanced cross-attention transformer feature fusion network for facial expression recognition," in *Proc. IEEE Trans. Ind. Informat.*, Mar. 2023, vol. 19, no. 12, pp. 11823–11832.

[22] Z. Yuan, X. Song, L. Bai, Z. Wang, and W. Ouyang, "Temporal-channel transformer for 3D LiDAR-based video object detection for autonomous driving," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 4, pp. 2068–2078, Apr. 2022.

[23] S. Cattini, D. Cassanelli, L. Ferrari, and L. Rovati, "A method for estimating object detection probability, lateral resolution, and errors in 3-D LiDARs," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–14, 2023.

[24] D. Tian, D. Zhao, D. Cheng, and J. Zhang, "LiDAR super-resolution based on segmentation and geometric analysis," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–17, 2022.

[25] G. Xu, X. Wang, X. Ding, and X. Yang, "Iterative geometry encoding volume for stereo matching," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 21919–21928.

[26] X. Wang, G. Xu, H. Jia, and X. Yang, "Selective-stereo: Adaptive frequency information selection for stereo matching," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 33, Jun. 2024, pp. 19701–19710.

[27] H. Zhao, H. Zhou, Y. Zhang, J. Chen, Y. Yang, and Y. Zhao, "High-frequency stereo matching network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 1327–1336.

[28] G. Xu, X. Wang, Z. Zhang, J. Cheng, C. Liao, and X. Yang, "IGEV++: Iterative multi-range geometry encoding volumes for stereo matching," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, May 12, 2025, doi: 10.1109/TPAMI.2025.3569218.

[29] P. Weinzaepfel et al., "CroCo v2: Improved cross-view completion pre-training for stereo matching and optical flow," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 17969–17980.

[30] Y. Wang, L. Wang, K. Li, Y. Zhang, D. Wu, and Y. Guo, "Cost volume aggregation in stereo matching revisited: A disparity classification perspective," *IEEE Trans. Image Process.*, vol. 33, pp. 6425–6438, 2024.

[31] H. Zhang, X. Ye, S. Chen, Z. Wang, H. Li, and W. Ouyang, "The farther the better: Balanced stereo matching via depth-based sampling and adaptive feature refinement," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 7, pp. 4613–4625, Jul. 2022.

[32] Z. Wu et al., "SG-RoadSeg: End-to-end collision-free space detection sharing encoder representations jointly learned via unsupervised deep stereo," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2024, pp. 11063–11069.

[33] J. Fritsch, T. Kühnl, and A. Geiger, "A new performance measure and evaluation benchmark for road detection algorithms," in *Proc. 16th Int. IEEE Conf. Intell. Transp. Syst. (ITSC)*, Oct. 2013, pp. 1693–1700.

[34] H. Abu Alhaija, S. K. Mustikovela, L. Mescheder, A. Geiger, and C. Rother, "Augmented reality meets computer vision: Efficient data generation for urban driving scenes," *Int. J. Comput. Vis.*, vol. 126, no. 9, pp. 961–972, Sep. 2018.

[35] A. Bar Hillel, R. Lerner, D. Levi, and G. Raz, "Recent progress in road and lane detection: A survey," *Mach. Vis. Appl.*, vol. 25, no. 3, pp. 727–745, 2014.

[36] J. Li, Y. Zhang, P. Yun, G. Zhou, Q. Chen, and R. Fan, "RoadFormer: Duplex transformer for RGB-normal semantic road scene parsing," *IEEE Trans. Intell. Vehicles*, vol. 9, no. 7, pp. 5163–5172, Jul. 2024.

[37] J. Huang, J. Li, S. Vityazev, A. Dvorkovich, and R. Fan, "DepthMatch: Semi-supervised RGB-D scene parsing through depth-guided regularization," *IEEE Signal Process. Lett.*, vol. 32, pp. 1–5, 2025.

[38] J. Huang et al., "RoadFormer+: Delivering RGB-X scene parsing through scale-aware information decoupling and advanced heterogeneous feature fusion," *IEEE Trans. Intell. Veh.*, early access, Aug. 22, 2024, doi: 10.1109/TIV.2024.3448251.

[39] C. Hazirbas, L. Ma, C. Domokos, and D. Cremers, "FuseNet: Incorporating depth into semantic segmentation via fusion-based CNN architecture," in *Proc. Asian Conf. Comput. Vis. (ACCV)*, 2017, pp. 213–228.

[40] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2261–2269.

[41] Y. Chang, F. Xue, F. Sheng, W. Liang, and A. Ming, "Fast road segmentation via uncertainty-aware symmetric network," in *Proc. Int. Conf. Robot. Autom. (ICRA)*, May 2022, pp. 11124–11130.

[42] E. Milli, Ö. Erkent, and A. E. Yılmaz, "Multi-modal multi-task (3MT) road segmentation," *IEEE Robot. Autom. Lett.*, vol. 8, no. 9, pp. 5408–5415, Sep. 2023.

[43] H. Ye, J. Mei, and Y. Hu, "M2F2-net: Multi-modal feature fusion for unstructured off-road freespace detection," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Jun. 2023, pp. 1–7.

[44] L. Wang et al., "Parallax attention for unsupervised stereo correspondence learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 4, pp. 2108–2125, Apr. 2020.

[45] R. Fan, H. Wang, Y. Wang, M. Liu, and I. Pitas, "Graph attention layer evolves semantic segmentation for road pothole detection: A benchmark and algorithms," *IEEE Trans. Image Process.*, vol. 30, pp. 8144–8154, 2021.

[46] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[47] C. Godard, O. M. Aodha, and G. J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 270–279.

[48] Z.-C. Yin and J.-P. Shi, "GeoNet: Unsupervised learning of dense depth, optical flow and camera pose," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 1983–1992.

[49] R. Fan et al., "Learning collision-free space detection from stereo images: Homography matrix brings better data augmentation," *IEEE/ASME Trans. Mechatronics*, vol. 27, no. 1, pp. 225–233, Feb. 2022.

[50] L. Sun, H. Zhang, and W. Yin, "Pseudo-LiDAR-based road detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 8, pp. 5386–5398, Aug. 2022.