2024 IEEE International Conference on Robotics and Automation (ICRA 2024)
May 13-17, 2024. Yokohama, Japan

SG-RoadSeg: End-to-End Collision-Free Space Detection Sharing Encoder Representations Jointly Learned via Unsupervised Deep Stereo

Zhiyuan Wu, Jiaqi Li, Yi Feng, Chengju Liu, Wei Ye, Qijun Chen, Rui Fan✉

Abstract— **Collision-free space detection is of utmost importance for autonomous robot perception and navigation. State-of-the-art (SoTA) approaches generally extract features from RGB images and an additional source or modality of 3-D information, such as depth or disparity images, using a pair of independent encoders. The extracted features are subsequently fused and decoded to yield semantic predictions of collision-free spaces. Such feature-fusion approaches become infeasible in scenarios, where the sensor for 3-D information acquisition is unavailable, or just when multi-sensor calibration falls short of the necessary precision. To overcome these limitations, this paper introduces a novel end-to-end collision-free space detection network, referred to as SG-RoadSeg, built upon our previous work SNE-RoadSeg. A key contribution of this paper is a strategy for sharing encoder representations that are co-learned through both semantic segmentation and unsupervised stereo matching tasks, enabling the features extracted from RGB images to contain both semantic and spatial geometric information. The unsupervised deep stereo serves as an auxiliary functionality, capable of generating accurate disparity maps that can be used by other perception tasks that require depth-related data. Comprehensive experimental results on the KITTI road and semantics datasets validate the effectiveness of our proposed architecture and encoder representation sharing strategy. SG-RoadSeg also demonstrates superior performance than other SoTA collision-free space detection approaches. Our source code, demo video, and supplement are publicly available at `mias.group/SG-RoadSeg`.**

I. Introduction

Today, the dream of summoning an autonomous car to your doorstep is not just possible, but increasingly commonplace [1]. The primary bottleneck hindering the advancement of autonomous driving technology lies in the domain of environmental perception [2]. Collision-free space detection is essential for autonomous robot perception and navigation, as it ensures the safety of passengers, pedestrians, and other drivers by allowing the vehicle to navigate reliably and efficiently through complicated and dynamic environments [3]. Regardless of whether the method is explicit programming-based or data-driven, 3-D information, particularly in the form of disparity or depth maps, is gaining prominence in the field of collision-free space detection due to the valuable spatial geometric information it provides [4].

Early collision-free space detection approaches were generally developed based on explicit programming, where the region of interest (RoI) was often assumed to be a planar surface represented by an explicit mathematical model [5]–[7]. Nonetheless, recent years have witnessed a significant transformation in this field, driven by state-of-the-art (SoTA) deep learning techniques [3]. This shift has led to significant advancements in the accuracy and robustness of collision-free space detection [1]. Among these algorithms, feature-fusion networks with a pair of independent encoders [1], [8]–[11], have demonstrated superior performance compared to single-modal networks [12]–[19]. These improvements can be attributed to the incorporation of additional sources or modalities of 3-D information, from which extracted features contain informative spatial geometric characteristics [3].

However, the demand for accurate 3-D information, *e.g.*, depth and surface normal maps, remains a significant limitation in feature-fusion methods. Fulfilling this requirement necessitates high-precision LiDAR-camera calibration [20]. When LiDAR is unavailable or LiDAR-camera calibration falls short of the necessary precision, obtaining accurate 3-D information for feature fusion becomes notably challenging. In our previous study [21], we moved away from using LiDAR and turned to stereo cameras to enhance data augmentation, resulting in improved performance in collision-free space detection. Nevertheless, our fusion approach remains constrained to the data level and does not extend to the more promising feature level. The latter requires accurate disparity maps, typically estimated using SoTA deep stereo networks.

Deep stereo networks [22]–[24] are generally trained via fully supervised learning with disparity ground truth obtained from LiDAR point clouds. While there have been extensive explorations into unsupervised learning strategies [25]–[27], introducing an additional network to independently learn stereo matching can result in substantial computational resource demands and deployment costs [28]. Moreover, the features extracted for stereo matching may differ greatly from those employed for collision-free space detection [29]. Therefore, the limitations of current methods are noteworthy when it comes to practical autonomous robot systems, underscoring the critical need for research focused on integrating stereo matching and collision-free space detection at the feature-sharing level.

To address the aforementioned challenges, this paper introduces **S**tereo-**G**uided **RoadSeg** (**SG-RoadSeg**), an end-to-end collision-free space detection network that shares

This research was supported by the National Key R&D Program of China under Grant 2020AAA0108100, the National Natural Science Foundation of China under Grants 62233013, 62173248, and 62176184, the Science and Technology Commission of Shanghai Municipal under Grant 22511104500, the Fundamental Research Funds for the Central Universities, and Xiaomi Young Talents Program. (✉ *Corresponding author: Rui Fan*).

The authors are with the College of Electronics & Information Engineering, Shanghai Research Institute for Intelligent Autonomous Systems, the State Key Laboratory of Intelligent Autonomous Systems, and Frontiers Science Center for Intelligent Autonomous Systems, Tongji University, Shanghai 201804, P. R. China. (e-mails: {gwu, lijiaqi220808}@tongji.edu.cn; fengyi@ieee.org; {liuchengju, yew, qjchen}@tongji.edu.cn; rui.fan@ieee.org)

encoder representations jointly learned by an unsupervised deep stereo network. SG-RoadSeg utilizes a weight-sharing hourglass network to extract feature maps, which are subsequently fused with the features extracted from disparity maps estimated by the unsupervised deep stereo network. The fused features are then decoded using densely-connected skip connections to perform collision-free space detection. Notably, stereo matching serves as an auxiliary functionality, and its output can be leveraged for other tasks that demand 3-D information. Specifically, the strategy to share encoder representations allows the feature maps extracted from RGB images to contain both semantic and spatial geometric information, greatly enhancing the network's ability to comprehend complex scenarios. To validate the effectiveness of our proposed SG-RoadSeg and the encoder representation sharing strategy, we conduct extensive experiments on the KITTI road [30] and semantics [31] datasets. Our qualitative and quantitative experimental results demonstrate that SG-RoadSeg outperforms all other SoTA single-modal and feature-fusion networks for collision-free space detection.

In summary, our novel contributions are as follows:

- SG-RoadSeg, a novel end-to-end collision-free space detection network guided by an auxiliary unsupervised deep stereo matching task.
- An encoder representation sharing strategy, enabling the features extracted from RGB images to contain both semantic and spatial geometric information.
- Extensive experiments on two public datasets to validate the feasibility of sharing encoder representations and to evaluate the performance of SG-RoadSeg.

II. Related Work

Collision-free space detection approaches can be divided into two groups: 1) conventional explicit programming-based [6] and 2) data-driven methods [1]. The former are generally based on the "flat road" assumption [32], which is, however, often violated in real-world scenarios [3]. Although previous studies [5]–[7] have made efforts to fit road disparity maps to minimize the impact of deviation from the "flat road" assumption, the achieved results are still far from satisfactory.

Recently, with the ongoing advancement of deep learning techniques, particularly convolutional neural networks (CNNs), the spotlight has shifted towards data-driven methods in the field of collision-free space detection [1]. Data-driven methods typically fall into two categories: single-modal and feature-fusion. Fully convolutional network (FCN) [12] marked a significant milestone in semantic segmentation. Building upon FCN, SegNet [13] introduces the encoder-decoder architecture and a pixel-wise classification layer, while U-Net improves upon it by adding skip connections to better preserve the spatial information of objects. Pyramid scene parsing network (PSPNet) [15], on the other hand, leverages a pyramid pooling module to gather contextual information for improved semantic segmentation performance. Moreover, DeepLabv3+ [16] employs atrous convolution and depth-wise separable convolution in both its atrous spatial pyramid pooling (ASPP) and decoder modules. This design choice enables the network to effectively capture multi-scale contextual information while simultaneously reducing computational complexity. As a result, it leads to significant enhancements in both the efficiency and accuracy of semantic segmentation.

While single-modal networks mentioned above have made great progress in semantic segmentation, their performance in collision-free space detection remains below expectations. To address this limitation, researchers have explored introducing another source or modality of vision data, such as disparity or depth images, to enhance scene understanding. FuseNet [8] was the pioneering work that introduced feature-fusion techniques into semantic segmentation. It utilizes two independent encoders to extract RGB and depth features, which are subsequently fused through element-wise summation. Building upon FuseNet, MFNet [9] extends the fusion strategy to include dilated convolutions. It introduces a "mini-inception" block for efficient feature concatenation, and a decoder with shortcuts to extract lower-level information. RTFNet [10] extends the use of RGB-thermal data for driving scene parsing. It introduces an "upception" block in its decoder, which transfers input through shortcuts to preserve more detailed information. In our previous work, SNE-RoadSeg [1], we designed a lightweight module to translate disparity or depth maps into surface normal maps, and fuse the features extracted from both RGB images and surface normal maps. Inspired by DenseNet [33], we introduced densely-connected skip connections for more flexible feature-fusion. These contributions collectively result in improved performance in collision-free space detection.

Although significant advancements in feature-fusion methods have brought collision-free space detection to a new level, there remains a notable drawback: the reliance on disparity or depth information. For example, in SNE-RoadSeg [1], the availability of depth maps for estimating surface normals is essential. In such circumstances, to obtain satisfactory collision-free detection results, we are compelled to derive disparity or depth information from LiDAR point clouds, which requires precise LiDAR-camera calibration. However, any deviation in the calibration process can indeed present challenges in obtaining optimal results for feature fusion [34]. Therefore, there is a pressing need to explore an end-to-end collision-free space detection method that does not rely on precise disparity or depth information. This forms the core focus of our proposed SG-RoadSeg, which not only resolves this challenge but also has the capability to generate additional disparity or depth information that can be used by other perception tasks requiring depth-related data.

III. Methodology

A. Architecture Overview

This section details SG-RoadSeg, an end-to-end collision-free space detection network sharing encoder representations jointly learned via an unsupervised deep stereo matching task. As depicted in Fig. 1, our proposed SG-RoadSeg 1) utilizes a weight-sharing hourglass network to extract features from RGB images, 2) fuses them with the features

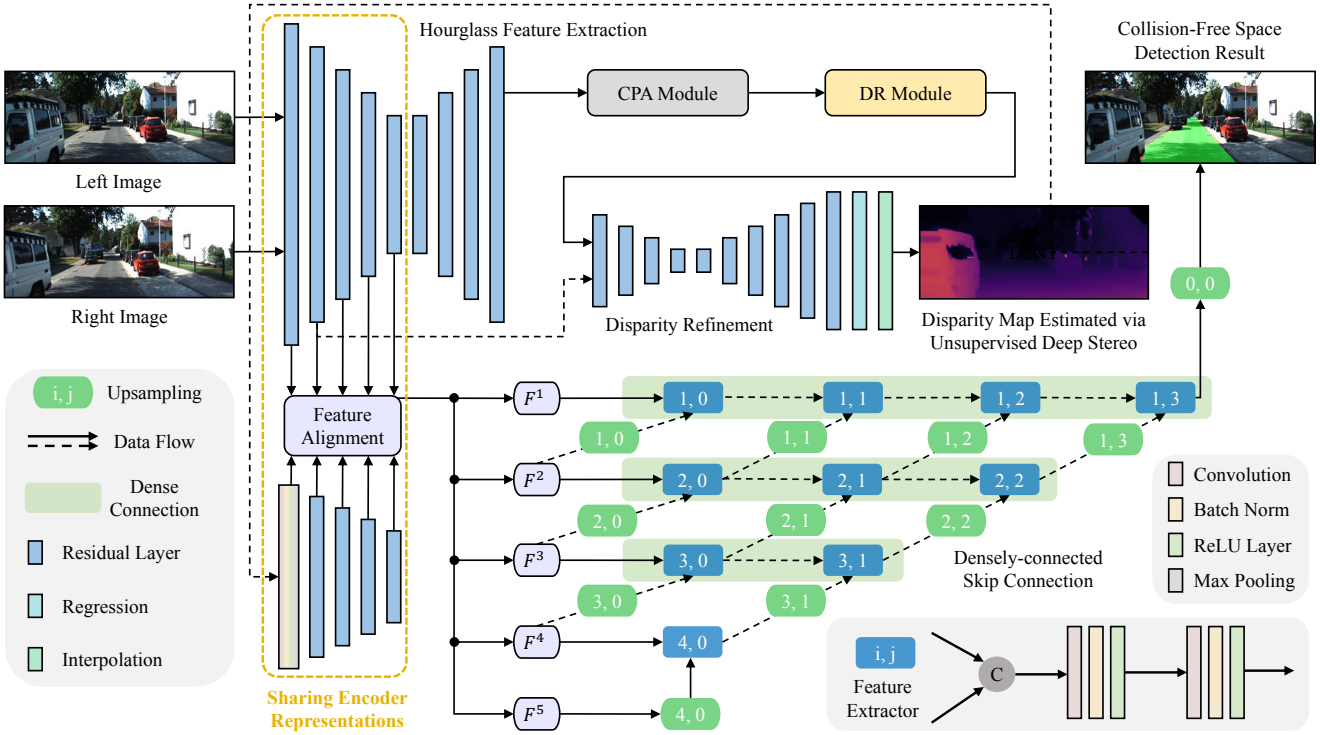

Fig. 1. An overview of our proposed SG-RoadSeg architecture.

extracted from disparity maps estimated using a deep stereo network (auxiliary functionality) trained via unsupervised learning, and 3) employs densely-connected skip connection to detect collision-free space.

B. Weight-Sharing Hourglass Cascaded Stereo Matching

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 first apply a weight-sharing hourglass network for feature extraction. The hourglass feature extraction network uses ResNet-152 [35] as its encoders, consisting of a series of residual blocks.

The extracted left and right feature maps are then fed into a cascaded parallax-attention (CPA) module [36] to regress matching costs, followed by a disparity regression (DR) module to generate initial disparities. Our CPA module is developed based on PASMNet [36], where the feature maps extracted from both left and right images are processed to update matching costs in a coarse-to-fine manner, from an initial cost volume $\boldsymbol{C}^0 \in \mathbb{R}^{\frac{H}{16}\times\frac{W}{16}\times\frac{W}{16}}$ whose elements are set to 0. The matching cost computation process in the m-th ($m > 0$) parallax-attention block $\boldsymbol{C}^m \in \mathbb{R}^{\frac{H}{2^{5-m}}\times\frac{W}{2^{5-m}}\times\frac{W}{2^{5-m}}}$ can be formulated as follows:

$$\boldsymbol{C}^{m+1} = \mathcal{U}(\boldsymbol{C}^m) + \sum_i\sum_j (\boldsymbol{\sigma}_Q^{m+1}\boldsymbol{\sigma}^i\boldsymbol{F}_l^i)^\top \boldsymbol{\sigma}_K^{m+1}\boldsymbol{\sigma}^j\boldsymbol{F}_r^j, \tag{1}$$

where $\mathcal{U}$ refers to an upsampling layer, $\boldsymbol{F}_l, \boldsymbol{F}_r \in \mathbb{R}^{\frac{H}{N}\times\frac{W}{N}\times C}$ represent the left and right feature maps, $\boldsymbol{\sigma}_Q, \boldsymbol{\sigma}_K \in \mathbb{R}^{C\times C}$ refer to 1×1 convolution kernels for query and key feature maps, and $\boldsymbol{\sigma}$ represents two layers of 3×3 convolution kernels. In the DR module, the initial disparity map $\boldsymbol{D} \in \mathbb{R}^{\frac{H}{4}\times\frac{W}{4}}$ is regressed from the last matching cost volume of the CPA module, which can be formulated as:

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

where $\boldsymbol{M} \in \mathbb{R}^{\frac{H}{4}\times\frac{W}{4}\times\frac{W}{4}}$ denotes the softmax parallax-attention map produced by the CPA module.

Subsequently, we utilize another hourglass network as the disparity refinement module, with ResNet-152 [35] as its encoder. As shown in Fig. 1, the second feature map in the hourglass network is concatenated with the initial disparity map $\boldsymbol{D}$, and then fed into the DR module, followed by a series of residual blocks to produce a residual disparity map $\boldsymbol{D}_r \in \mathbb{R}^{H\times W}$ and a confidence map $\boldsymbol{M}_c \in \mathbb{R}^{H\times W}$. The refined disparity map $\hat{\boldsymbol{D}}$ is calculated as follows:

$$\hat{\boldsymbol{D}} = (1-\boldsymbol{M}_c)\mathcal{U}(\boldsymbol{D}) + \boldsymbol{M}_c\boldsymbol{D}_r. \tag{3}$$

This unsupervised stereo matching process not only eliminates the dependence on disparity or depth data but also has the potential to produce additional disparity or depth information, which can be advantageous for other perception tasks that require depth-related data.

C. RoadSeg Sharing Encoder Representations

After obtaining a refined disparity map, we process an encoder representation sharing strategy, followed by a feature-fusion operation, formulated as:

$$\boldsymbol{F}_f^i = \mathcal{A}(\boldsymbol{F}_l^i) \oplus \mathcal{E}(\boldsymbol{F}_d^{i-1}), \tag{4}$$

where $\boldsymbol{F}_f^i \in \mathbb{R}^{\frac{H}{2^i}\times\frac{W}{2^i}\times C^i}$ represents fused feature maps, $\boldsymbol{F}_d^i \in \mathbb{R}^{\frac{H}{2^i}\times\frac{W}{2^i}\times C^i}$ represents disparity feature maps, and $\mathcal{E}$ refers to deep stereo encoder, inspired by [37]. Specifically,

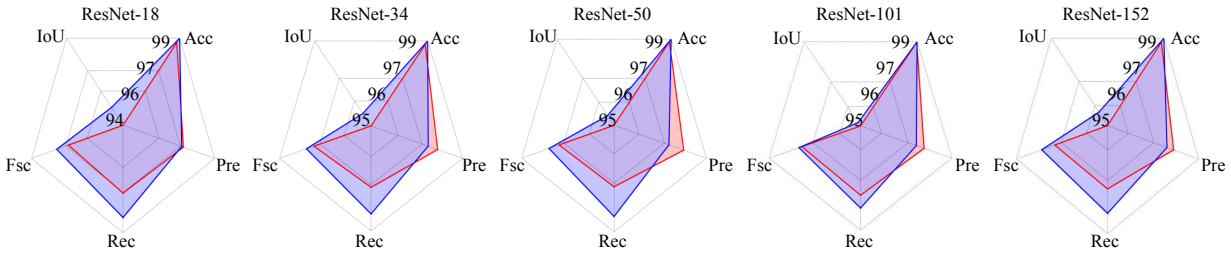

Fig. 2. Comparison (%) between SNE-RoadSeg and our proposed SG-RoadSeg with respect to different CNN backbones, where — presents the results achieved by SNE-RoadSeg and — presents the results achieved by our SG-RoadSeg.

the first block of the deep stereo encoder consists of a convolutional layer, a batch normalization layer, and a ReLU activation layer. Then, a max pooling layer and four residual layers (based on ResNet-152 [35]) are sequentially employed to gradually increase the number of feature map channels. Correspondingly, C^i represents the channels of the i-th feature map, and C^1-C^5 are 64, 256, 512, 1024, and 2048, respectively. $\boldsymbol{F}_l^i$ represents the left RGB feature maps in the encoding layers of the hourglass feature extraction network, $\mathcal{A}$ refers to a feature alignment operation to remap RGB feature maps into semantic feature space, and $\oplus$ refers to a feature-fusion operation. Specifically in this paper, we employ a 1×1 convolution layer, a batch normalization layer, and a ReLU activation layer for feature alignment, and addition for feature-fusion.

The encoder representation sharing strategy and feature-fusion operation enhance the capability of our feature encoders. The RGB feature maps contain not only semantic features but also spatial geometric information since they are used for both stereo matching and collision-free space detection tasks. Moreover, the fusion with the disparity feature maps $\boldsymbol{F}_d^i$ further enriches our network's understanding of spatial information.

Based on our previous work SNE-RoadSeg [1], we then utilize a densely-connected skip connection decoder, which consists of feature extractors and upsampling layers, to obtain the final semantic prediction of the collision-free space. Feature extractors are utilized to extract feature maps from the fused feature maps, and upsampling layers are utilized to increase the resolution while decreasing the feature map channels. We employ three convolutional layers in the feature extractor and upsampling layer, each with a 3×3 kernel size, a stride of 1, and a padding of 1.

D. Loss Function

SG-RoadSeg is trained by minimizing a pixel-wise cross-entropy loss $\mathcal{L}_d$. Concurrently, we train the auxiliary stereo matching network in an unsupervised way, by minimizing the loss function $\mathcal{L}_u$. The total loss is defined as follows:

$$\mathcal{L} = \mathcal{L}_d + \mathcal{L}_u, \tag{5}$$

where

$$\mathcal{L}_u = (\mathcal{L}_p + \lambda_s \mathcal{L}_s + \lambda_{pam} \mathcal{L}_{pam}). \tag{6}$$

λ_s and λ_{pam} are empirically set to 0.5 and 1, respectively. $\mathcal{L}_p$ represents the photometric loss function. Inspired by

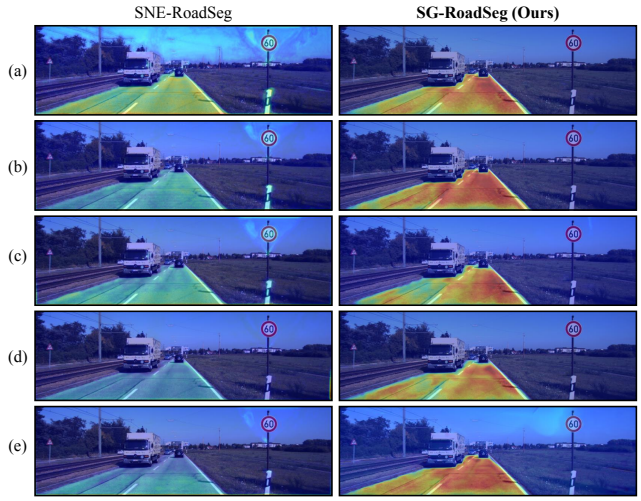

Fig. 3. Comparison between SNE-RoadSeg and our proposed SG-RoadSeg in terms of model decision-making explainability with respect to different CNN backbones. (a)-(e): ResNet-18, ResNet-34, ResNet-50, ResNet-101, and ResNet-152, respectively.

[38], [39], it consists of a mean absolute error (MAE) loss and a structural similarity index (SSIM) loss, which can be formulated as follows:

$$\begin{aligned}\mathcal{L}_p = \frac{1}{N}\sum_{\boldsymbol{p}} \frac{\alpha}{2}[1 - \mathcal{S}\left(\boldsymbol{I}^L(\boldsymbol{p}), \hat{\boldsymbol{I}}^L(\boldsymbol{p})\right)] \\ + (1-\alpha)\left\|\boldsymbol{I}^L(\boldsymbol{p}) - \hat{\boldsymbol{I}}^L(\boldsymbol{p})\right\|_1,\end{aligned} \tag{7}$$

where $\boldsymbol{p}$ represents a valid pixel, N represents the number of valid pixels, $\mathcal{S}$ represents an SSIM function, and α refers to a photometric parameter, which is set to 0.85. $\boldsymbol{I}^L$ represents the left RGB image, and $\hat{\boldsymbol{I}}^L$ represents the warped left image, which can be obtained using:

$$\hat{\boldsymbol{I}}^L = \mathcal{W}(\boldsymbol{I}^L, \hat{\boldsymbol{D}}), \tag{8}$$

where $\mathcal{W}$ refers to a warping operator. $\mathcal{L}_s$ represents the smoothness loss function, which is defined as follows:

$$\begin{aligned}\mathcal{L}_s = \frac{1}{N}\sum_{\boldsymbol{p}} (\left\|\nabla_x \hat{\boldsymbol{D}}(\boldsymbol{p})\right\|_1 e^{-\left\|\nabla_x \boldsymbol{I}^L(\boldsymbol{p})\right\|_1} \\ + \left\|\nabla_y \hat{\boldsymbol{D}}(\boldsymbol{p})\right\|_1 e^{-\left\|\nabla_y \boldsymbol{I}^L(\boldsymbol{p})\right\|_1}),\end{aligned} \tag{9}$$

where ∇_x and ∇_y represent the gradients in the x and y directions, respectively. $\mathcal{L}_{pam}$ refers to the PAM loss corresponding to the parallax-attention map in the CPA module, as introduced in [36] to achieve accurate and consistent stereo correspondences.

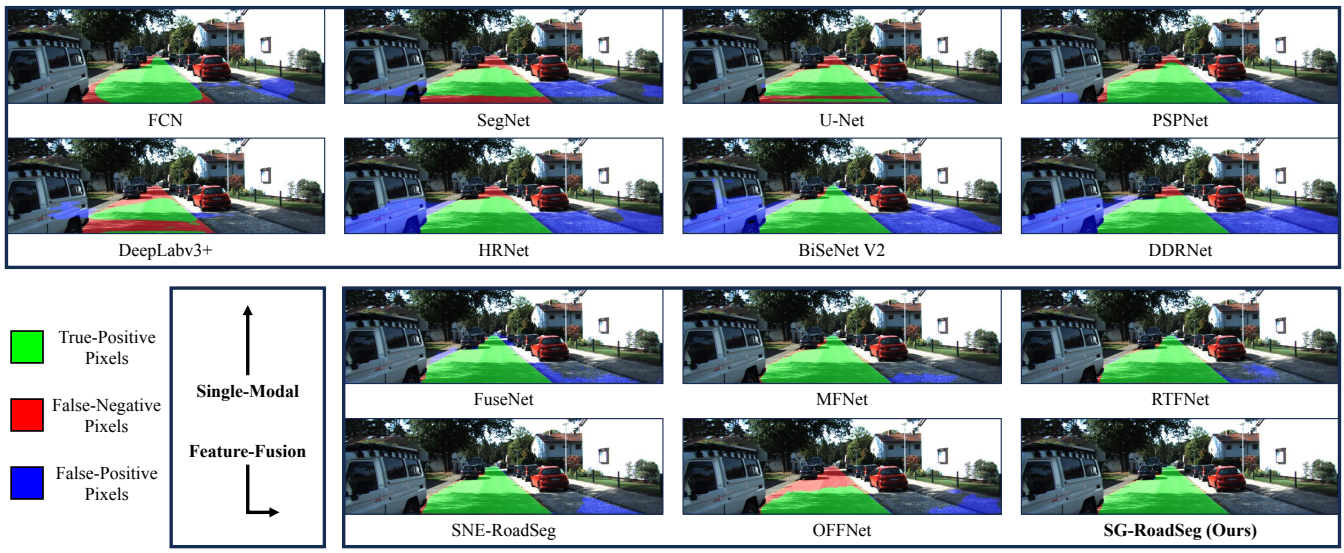

Fig. 4. Qualitative experimental results on the KITTI road [40] dataset.

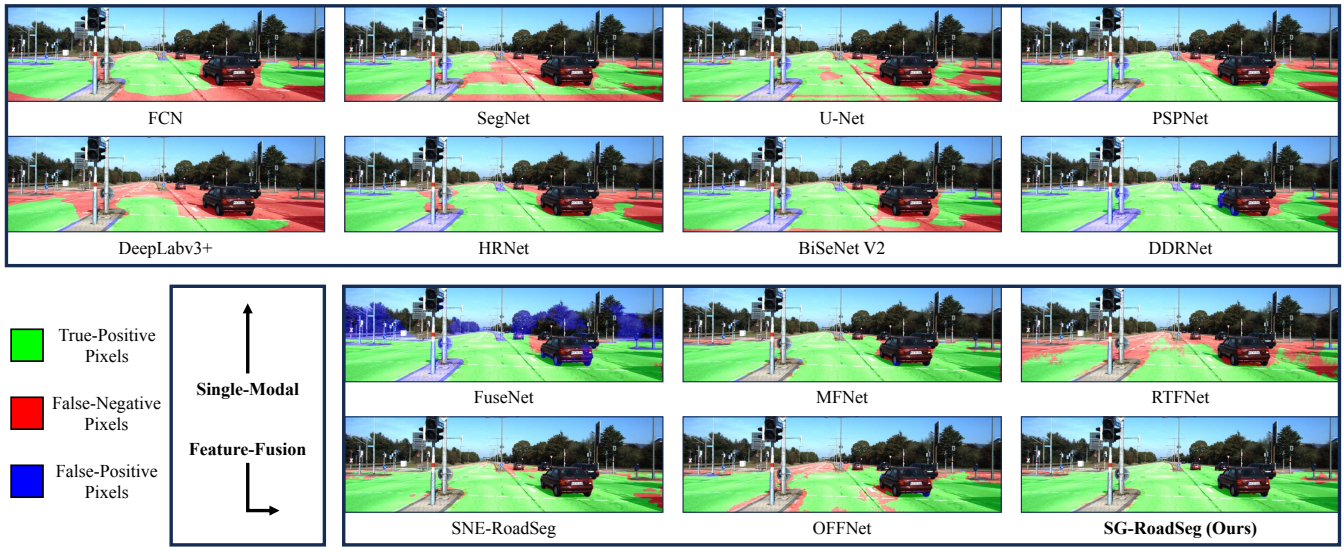

Fig. 5. Qualitative experimental results on the KITTI semantics [41] dataset.

IV. Experiments

A. Datasets and Experimental Setup

We carry out the experiments on the following two datasets to evaluate the performance of our proposed SG-RoadSeg for collision-free space detection:

- the KITTI road [40] dataset, containing 289 pairs of stereo images captured in real-world driving scenarios, as well as their pixel-level collision-free space detection ground-truth annotations.
- the KITTI semantics [41] dataset, containing 200 pairs of stereo images captured in real-world driving scenarios, as well as their semantic ground-truth annotations. To evaluate the performance of our proposed SG-RoadSeg, we extract pixels belonging to the 'road' class and consider them as the ground-truth annotations of the collision-free spaces.

Our experiments are conducted on an NVIDIA RTX 3090 GPU. For each dataset, we allocate 70% of images for training purposes, while the remaining data are used as the test set. The batch size is set to 2. We utilize the Adam [42] optimizer for modeling training. The initial learning rate is set to 1×10^{-3}. Training lasts for 200 epochs on each dataset. We utilize disparity maps generated by our SG-RoadSeg as inputs for other feature-fusion networks.

Five common metrics are used for the performance evaluation of collision-free space detection: (1) accuracy (Acc), (2) precision (Pre), (3) recall (Rec), (4) F-score (Fsc), and (5) intersection over union (IoU) [1].

B. Ablation Study

In this subsection, we conduct an ablation study to validate the superiority of our architecture that shares encoder representations. We compare our proposed structure to the baseline SNE-RoadSeg [1], using five different CNN backbones: ResNet-18, ResNet-34, ResNet-50, ResNet-101, and ResNet-152. The corresponding quantitative comparisons are given in Fig. 2, where SG-RoadSeg outperforms SNE-RoadSeg [1]

TABLE I

COMPARISONS OF SOTA COLLISION-FREE SPACE DETECTION NETWORKS ON THE KITTI ROAD [30] DATASET. THE BEST RESULTS ARE SHOWN IN BOLD FONT.

Networks	Acc (%) ↑	Pre (%) ↑	Rec (%) ↑	Fsc (%) ↑	IoU (%) ↑
FCN [12]	94.73	91.33	89.88	89.95	83.47
SegNet [13]	91.33	88.30	78.73	81.64	72.04
U-Net [14]	94.89	92.52	87.73	88.95	82.67
PSPNet [15]	94.31	88.09	94.79	90.48	83.80
DeepLabv3+ [16]	95.22	91.52	91.00	90.59	84.68
HRNet [17]	89.99	81.97	92.53	84.87	75.61
BiSeNet V2 [18]	84.28	77.26	90.08	79.04	67.84
DDRNet [19]	88.99	80.71	91.61	83.57	73.78
FuseNet [8]	94.06	87.13	96.08	90.49	83.36
MFNet [9]	98.04	96.79	96.06	96.13	93.23
RTFNet [10]	98.56	**97.65**	97.25	97.40	95.07
SNE-RoadSeg [1]	98.61	97.59	97.49	97.41	95.22
OFF-Net [11]	96.58	93.33	94.75	93.74	89.40
SG-RoadSeg (Ours)	**98.76**	97.52	**98.09**	**97.77**	**95.74**

TABLE II

COMPARISONS OF SOTA COLLISION-FREE SPACE DETECTION NETWORKS ON THE KITTI SEMANTICS [31] DATASET. THE BEST RESULTS ARE SHOWN IN BOLD FONT.

Networks	Acc (%) ↑	Pre (%) ↑	Rec (%) ↑	Fsc (%) ↑	IoU (%) ↑
FCN [12]	87.35	88.11	72.71	75.87	65.18
SegNet [13]	87.14	85.09	74.00	77.04	65.80
U-Net [14]	91.05	89.33	83.52	85.41	76.31
PSPNet [15]	94.55	92.37	91.36	91.48	85.35
DeepLabv3+ [16]	88.36	88.32	73.49	75.24	66.43
HRNet [17]	90.36	85.07	90.62	86.47	77.81
BiSeNet V2 [18]	89.99	84.83	87.44	85.47	76.21
DDRNet [19]	89.11	83.20	91.25	85.35	75.88
FuseNet [8]	54.82	65.48	69.92	53.31	37.05
MFNet [9]	96.44	94.36	95.06	94.57	90.22
RTFNet [10]	94.17	93.51	89.28	90.91	84.08
SNE-RoadSeg [1]	97.50	96.58	95.89	96.13	92.88
OFF-Net [11]	95.57	93.37	93.76	93.30	88.33
SG-RoadSeg (Ours)	**97.88**	**97.01**	**96.72**	**96.75**	**93.93**

for all five ResNet architectures (except in precision). Specifically, we observe an increase in IoU by 0.14-0.87% and an increase in F-score by 0.11-0.54%. Among the different CNN backbones, ResNet-152 achieves the best performance, consistent with its superior performance in image classification among the five ResNet architectures [35]. We also compare our proposed SG-RoadSeg with SNE-RoadSeg [1] in terms of model decision-making explainability. This is achieved by extracting feature maps before the final upsampling layer and computing the average over all channels to generate heat maps for visualization, as depicted in Fig. 3. These heat maps indicate that SG-RoadSeg pays more attention to the road areas, as they contain more spatial geometric information. This is attributed to the fact that RGB features before feature alignment operation are utilized not only for collision-free space detection but also for stereo matching, thereby providing both semantic and spatial geometric information, which enhances scenario understanding. Both Figs. 2 and 3 demonstrate that our proposed architecture significantly improves the distinguishability of road areas, and therefore, it improves the effectiveness of feature-fusion networks for collision-free space detection.

C. Performance Evaluation of Our Proposed SG-RoadSeg

In this subsection, we evaluate the effectiveness of our proposed SG-RoadSeg for collision-free space detection both qualitatively and quantitatively. We compare our model with SoTA networks on the KITTI road [30] and the KITTI semantics [31] datasets. The quantitative comparisons are given in Tables I and II, where our SG-RoadSeg achieves the best results on both datasets.

As shown in Figs. 4 and 5, feature-fusion networks achieve superior performance compared to single-modal networks. This improvement is attributed to the ability to obtain more features with the aid of disparity information, enabling them to leverage disparity maps to enhance scene understanding, especially in scenarios where RGB features alone are less informative. However, it is worth noting that feature-fusion methods may struggle with handling fine details in challenging scenarios, such as distinguishing roads from lanes, as indicated by the false-positive pixels in Fig. 4. This limitation may arise due to inaccuracies in the disparity information. On the other hand, SG-RoadSeg performs better in handling fine details even in the absence of disparity information. This is due to its capability to extract semantic features from RGB inputs based on spatial features and fuse them with disparity features through our encoder representation sharing strategy. In comparison to the baseline SNE-RoadSeg, SG-RoadSeg achieves an improvement of 0.36% in F-score and 0.52% in IoU on the KITTI road dataset and 0.62% in F-score and 1.05% in IoU on the KITTI semantics dataset. Furthermore, during the same training process, SG-RoadSeg also performs stereo matching as an auxiliary function, and its disparity outputs can be utilized in other feature-fusion networks, such as MFNet [9] and RTFNet [10]. This versatility enhances its potential applicability across various perception tasks.

V. CONCLUSION

In this paper, we presented two key technical contributions: (1) SG-RoadSeg, a novel end-to-end collision-free space detection network built upon our prior work SNE-RoadSeg, and (2) an innovative encoder representation sharing strategy to enrich the features extracted from RGB images. The stereo matching component embedded within SG-RoadSeg serves as an auxiliary functionality, capable of delivering accurate depth information without the need for disparity ground truth during unsupervised model training. Additionally, we also presented our contributions in the experimental evaluation aspect. Extensive experiments conducted on two widely used KITTI datasets demonstrate the feasibility of our proposed encoder representation sharing strategy and the superior performance of SG-RoadSeg in comparison to all other semantic segmentation networks. Our proposed architecture not only eliminates the requirement for other sensors to acquire 3-D information but also reduces the computational complexity of the overall architecture. This reduction in computational complexity makes it suitable for deployment on resource-limited autonomous robot systems.

REFERENCES

[1] R. Fan *et al.*, "SNE-RoadSeg: Incorporating surface normal information into semantic segmentation for accurate freespace detection," in *European Conference on Computer Vision (ECCV)*. Springer, 2020, pp. 340–356.

[2] M. Najibi *et al.*, "Motion inspired unsupervised perception and prediction in autonomous driving," in *European Conference on Computer Vision (ECCV)*. Springer, 2022, pp. 424–443.

[3] H. Wang *et al.*, "SNE-RoadSeg+: Rethinking depth-normal translation and deep supervision for freespace detection," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021, pp. 1140–1145.

[4] Y. Feng *et al.*, "Freespace optical flow modeling for automated driving," *IEEE/ASME Transactions on Mechatronics*, 2023, DOI: 10.1109/TMECH.2023.3300729.

[5] R. Fan and M. Liu, "Road damage detection based on unsupervised disparity map segmentation," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 11, pp. 4906–4911, 2020.

[6] A. Wedel *et al.*, "B-spline modeling of road surfaces with an application to free-space estimation," *IEEE Transactions on Intelligent Transportation Tystems*, vol. 10, no. 4, pp. 572–583, 2009.

[7] R. Fan *et al.*, "Pothole detection based on disparity transformation and road surface modeling," *IEEE Transactions on Image Processing*, vol. 29, pp. 897–908, 2020.

[8] C. Hazirbas *et al.*, "FuseNet: Incorporating depth into semantic segmentation via fusion-based CNN architecture," in *Proceedings of the Asian Conference on Computer Vision (ACCV)*. Springer, 2017, pp. 213–228.

[9] Q. Ha *et al.*, "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)*. IEEE, 2017, pp. 5108–5115.

[10] Y. Sun *et al.*, "RTFNet: RGB-thermal fusion network for semantic segmentation of urban scenes," *IEEE Robotics and Automation Letters*, vol. 4, no. 3, pp. 2576–2583, 2019.

[11] C. Min *et al.*, "ORFD: A dataset and benchmark for off-road freespace detection," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 2532–2538.

[12] J. Long *et al.*, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 3431–3440.

[13] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[14] O. Ronneberger *et al.*, "U-Net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Springer, 2015, pp. 234–241.

[15] H. Zhao *et al.*, "Pyramid scene parsing network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2881–2890.

[16] L.-C. Chen *et al.*, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 801–818.

[17] K. Sun *et al.*, "Deep high-resolution representation learning for human pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 5693–5703.

[18] C. Yu *et al.*, "BiSeNet V2: Bilateral network with guided aggregation for real-time semantic segmentation," *International Journal of Computer Vision*, vol. 129, pp. 3051–3068, 2021.

[19] H. Pan *et al.*, "Deep dual-resolution networks for real-time and accurate semantic segmentation of traffic scenes," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 3, pp. 3448–3460, 2022.

[20] A. Geiger *et al.*, "Are we ready for autonomous driving? the KITTI vision benchmark suite," in *2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2012, pp. 3354–3361.

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

[22] J.-R. Chang and Y.-S. Chen, "Pyramid stereo matching network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 5410–5418.

[23] L. Lipson *et al.*, "RAFT-Stereo: Multilevel recurrent field transforms for stereo matching," in *2021 International Conference on 3D Vision (3DV)*. IEEE, 2021, pp. 218–227.

[24] J. Li *et al.*, "Practical stereo matching via cascaded recurrent network with adaptive correlation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 16 263–16 272.

[25] T. Song *et al.*, "Unsupervised deep asymmetric stereo matching with spatially-adaptive self-similarity," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 13 672–13 680.

[26] N. Ying *et al.*, "Multi-directional broad learning system for the unsupervised stereo matching method," *Pattern Recognition*, vol. 142, p. 109648, 2023.

[27] S. Wang *et al.*, "Horizontal attention based generation module for unsupervised domain adaptive stereo matching," *IEEE Robotics and Automation Letters*, 2023, DOI: 10.1109/LRA.2023.3313009.

[28] P. L. Dovesi *et al.*, "Real-time semantic stereo matching," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 10 780–10 787.

[29] R. Fan *et al.*, *Autonomous Driving Perception*. Springer Nature, 2023.

[30] J. Fritsch *et al.*, "A new performance measure and evaluation benchmark for road detection algorithms," in *16th International IEEE Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2013, pp. 1693–1700.

[31] H. Alhaija *et al.*, "Augmented reality meets computer vision: Efficient data generation for urban driving scenes," *International Journal of Computer Vision*, 2018.

[32] A. B. Hillel *et al.*, "Recent progress in road and lane detection: A survey," *Machine Vision and Applications*, vol. 25, no. 3, pp. 727–745, 2014.

[33] G. Huang *et al.*, "Densely connected convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 4700–4708.

[34] H. Zhao *et al.*, "Dive deeper into rectifying homography for stereo camera online self-calibration," in *2024 International Conference on Robotics and Automation (ICRA)*, 2024, in press.

[35] K. He *et al.*, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[36] L. Wang *et al.*, "Parallax attention for unsupervised stereo correspondence learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 4, pp. 2108–2125, 2020.

[37] Z. Wu *et al.*, "S^3M-Net: Joint learning of semantic segmentation and stereo matching for autonomous driving," *IEEE Transactions on Intelligent Vehicles*, 2024, DOI: 10.1109/TIV.2024.3357056.

[38] C. Godard *et al.*, "Unsupervised monocular depth estimation with left-right consistency," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 270–279.

[39] Z. Yin and J. Shi, "GeoNet: Unsupervised learning of dense depth, optical flow and camera pose," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 1983–1992.

[40] J. Fritsch *et al.*, "A new performance measure and evaluation benchmark for road detection algorithms," in *16th International IEEE Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2013, pp. 1693–1700.

[41] H. Abu Alhaija *et al.*, "Augmented reality meets computer vision: Efficient data generation for urban driving scenes," *International Journal of Computer Vision*, vol. 126, pp. 961–972, 2018.

[42] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.