SNE-RoadSegV2: Advancing Heterogeneous Feature Fusion and Fallibility Awareness for Freespace Detection

Yi Feng, Yu Ma, Stepan Andreev, Qijun Chen, *Senior Member, IEEE*, Alexander Dvorkovich, Ioannis Pitas, *Life Fellow, IEEE*, and Rui Fan, *Senior Member, IEEE*

***Abstract*— Feature-fusion networks with duplex encoders have proved to be an effective technique for solving the road freespace detection problem. However, despite the compelling results achieved by previous research efforts, the exploration of adequate and discriminative heterogeneous feature fusion, as well as the development of fallibility-aware loss functions, remains relatively scarce. This article makes several significant contributions to address these limitations: 1) it presents a novel heterogeneous feature fusion block (HF^{2}B), comprising a holistic attention module (HAM), a heterogeneous feature contrast descriptor (HFCD), and an affinity-weighted feature recalibrator (AWFR), enabling more in-depth exploitation of the inherent characteristics of the extracted features; 2) it incorporates both interscale and intrascale skip connections into the decoder architecture, while eliminating redundant ones, leading to both improved accuracy and computational efficiency; and 3) it introduces two fallibility-aware loss functions that separately focus on semantic-transition and depth-inconsistent regions, collectively contributing to greater supervision during model training. Our proposed SNE-RoadSegV2, which incorporates all these innovative components, demonstrates superior performance in comparison to all other free-space detection algorithms across multiple public datasets.**

***Index Terms*— Duplex encoders, fallibility-aware loss, feature fusion, freespace detection, holistic attention.**

Received 18 September 2024; revised 15 December 2024; accepted 22 January 2025. Date of publication 26 February 2025; date of current version 13 March 2025. This work was supported in part by the National Natural Science Foundation of China under Grant 62473288 and Grant 62233013; in part by the Science and Technology Commission of Shanghai Municipal under Grant 22511104500; in part by the Fundamental Research Funds for the Central Universities, NIO University Program (NIO UP); in part by Xiaomi Young Talents Program; and in part by the Research Leading to these Results has also Received Partial Funding from European Commission-European Union (under HORIZON EUROPE (HORIZON Research and Innovation Actions) (TEMA) HORIZON-CL4-2022-DATA-01-01) under Grant 101093003. The Associate Editor coordinating the review process was Dr. Gui-Bin Bian. *(Corresponding author: Rui Fan.)*

Yi Feng, Yu Ma, and Qijun Chen are with the College of Electronics and Information Engineering, Tongji University, Shanghai 201804, China (e-mail: fengyi@ieee.org; mayu2002@tongji.edu.cn; qjchen@tongji.edu.cn).

Stepan Andreev is with the Microwave Photonics Department, Telecommunications Center, Moscow Institute of Physics and Technology, Dolgoprudny, 141701 Moscow, Russia (e-mail: Andreev.sn@mipt.ru).

Alexander Dvorkovich is with the Multimedia Technology and Telecom Department, Telecommunications Center, Moscow Institute of Physics and Technology, Dolgoprudny, 141701 Moscow, Russia (e-mail: dvorkovich.av@mipt.ru).

Ioannis Pitas is with the Department of Informatics, University of Thessaloniki, 541 24 Thessaloniki, Greece (e-mail: pitas@csd.auth.gr).

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

Digital Object Identifier 10.1109/TIM.2025.3545498

I. Introduction

As a vital piece of the autonomous driving puzzle, reliable collision-free space (freespace for short) detection holds significant importance in autonomous driving systems, as it directly impacts a vehicle's ability to make informed decisions and ensures dependable navigation [1]. In recent years, freespace detection has attracted considerable attention in research, with ongoing efforts aimed at addressing corner cases in complex and dynamic environments. Nevertheless, regardless of whether the approach is explicit programming-based or data-driven, the utilization of 3-D information is growing in significance for freespace detection, primarily due to the valuable spatial geometry information it provides [2]. Feature-fusion networks with duplex encoders, designed to extract heterogeneous features from multiple data sources or modalities and fuse them to provide a more comprehensive understanding of the environment, have emerged as a viable solution to tackle this problem [3], [4], [5], [6].

The performance of a feature-fusion freespace detection network depends not only on the input data type but also on how these features are fused [7]. A current bottleneck lies in the simplistic and indiscriminate fusion of heterogeneous features, often causing conflicting feature representations and erroneous detection results [2]. For example, in the SNE-RoadSeg series [1], [8], their adopted feature fusion strategy essentially performs an element-wise summation between RGB and surface normal feature maps at each stage, while neglecting the inherent differences in feature characteristics and their respective reliability [9]. Furthermore, as the network goes deeper, such an asymmetric feature fusion strategy tends to diminish the proportion of RGB features in the decoder input. This, in turn, leads to unsatisfactory performance, particularly in areas such as pavements or other lanes, where surface normals closely resemble those of freespace, demanding a greater reliance on color or textural information.

In the decoder aspect, we observe two phenomena: 1) interscale skip connections provide an advantage in achieving more

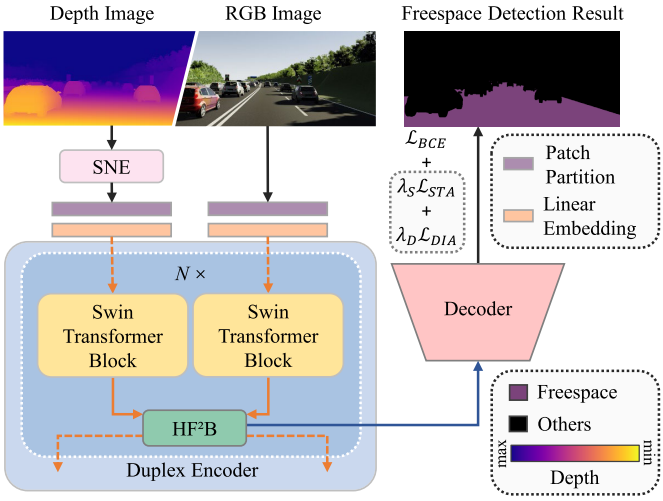

Fig. 1. Overview of our proposed SNE-RoadSegV2.

comprehensive feature decoding, primarily due to their ability to capture both fine-grained and coarse-grained details [10]. Unfortunately, however, they are not fully utilized and 2) the currently adopted intrascale skip connections appear to be excessively redundant for this task. Moreover, the utilization of pixel-wise binary cross-entropy (BCE) loss has been a common practice in freespace detection. Nevertheless, no prior endeavors have been undertaken to delve into the fallible cases, notably the misclassifications occurring near semantic transition and depth-inconsistent regions.

As shown in Fig. 1, to address the aforementioned limitations, we first dive deeper into the discriminative feature fusion strategies presented in recent universal semantic segmentation studies [1], [8]. Subsequently, we introduce a novel HF^{2}B to process the RGB and surface normal features, which are encoded using two independent Swin Transformer backbones [11]. Our technical contributions in this part are threefold: 1) a HAM to model the interdependencies between heterogeneous features across three dimensions (spatial, channel, and scale); 2) an HFCD to effectively underscore both the shared and unique characteristics in the holistically attentive features; and 3) an AWFR to jointly emphasize and suppress heterogeneous features before their input into the decoder. Additionally, we contribute to a lightweight, yet more effective decoder, which incorporates interscale skip connections while pruning redundant ones. Our decoder demonstrates greater feature decoding capabilities, while simultaneously reducing computational complexity. Finally, we design two new loss functions based on semantic annotations and depth data to provide deeper supervision during our model training process. This contribution also results in improved overall performance, particularly in error-prone areas. The effectiveness of each contribution is validated through extensive experiments conducted across public datasets. In a nutshell, our contributions are as follows.

1) We propose SNE-RoadSegV2, a novel feature-fusion freespace detection approach, achieving state-of-the-art (SoTA) performance across multiple public datasets.
2) We introduce HF^{2}B, consisting of an HAM, an HFCD, and an AWFR, for comprehensive heterogeneous feature description, recalibration, and fusion, resulting in more coherent feature representations.
3) We design a lightweight, yet more effective decoder, incorporating interscale skip connections while pruning redundant intrascale ones, demonstrating greater efficacy and computational efficiency.
4) We develop two novel fallibility-aware loss functions, which focus particularly on reducing misclassifications in semantic transition and depth-inconsistent regions, leading to improved overall performance.

This article is structured as follows. Section II presents an overview of the SoTA freespace detection and feature fusion methods. In Section III, we introduce the proposed SNE-RoadSegV2 framework and fallibility-aware loss functions. In Section IV, we present the experimental results across several public datasets. Section V provides a detailed discussion and concludes this article.

II. Related Work

A. Data-Driven Freespace Detection

While it is feasible to employ universal semantic segmentation networks [12], [13], [14], [15] for this task, it has been observed that task-specific approaches [1], [8], [16] consistently deliver superior performance. Early task-specific freespace detection approaches [17], [18], [19], [20] thoroughly rely on RGB images and were found to be highly sensitive to environmental factors, notably illumination and weather conditions [1]. Given the increased prevalence of range sensors, particularly LiDARs, feature-fusion networks [2], [21], [22] have emerged as a more robust choice in this domain. In terms of network architecture, these approaches are characterized by duplex-encoder architectures [1], [16], where each encoder extracts hierarchical features from a specific data source or modality. The extracted heterogeneous features are subsequently fused, enabling the network to gain a more comprehensive understanding of the environment [2]. As for the input data, the most commonly used spatial geometric information includes depth/disparity maps [23], [24], LiDAR point clouds [2], [21], and surface normal information [1], [8], [16]. Extensive experiments conducted in previous studies [1], [8], [16] have conclusively demonstrated that surface normals provide the most informative spatial geometric information for freespace detection, owing to their representation of road plane characteristics. Therefore, in this article, we adopt the pipeline introduced in [1], [8], and [16], which utilizes a duplex-encoder architecture to extract heterogeneous features from RGB images and surface normal information. However, it is important to note that our focus differs from those of these previous works. Our emphasis lies in designing heterogeneous feature fusion strategies, developing a lightweight, yet more effective decoder, and introducing task-specific loss functions.

B. Heterogeneous Feature Fusion

Heterogeneous feature fusion plays a pivotal role in various computer vision tasks, such as salient object detection [25],

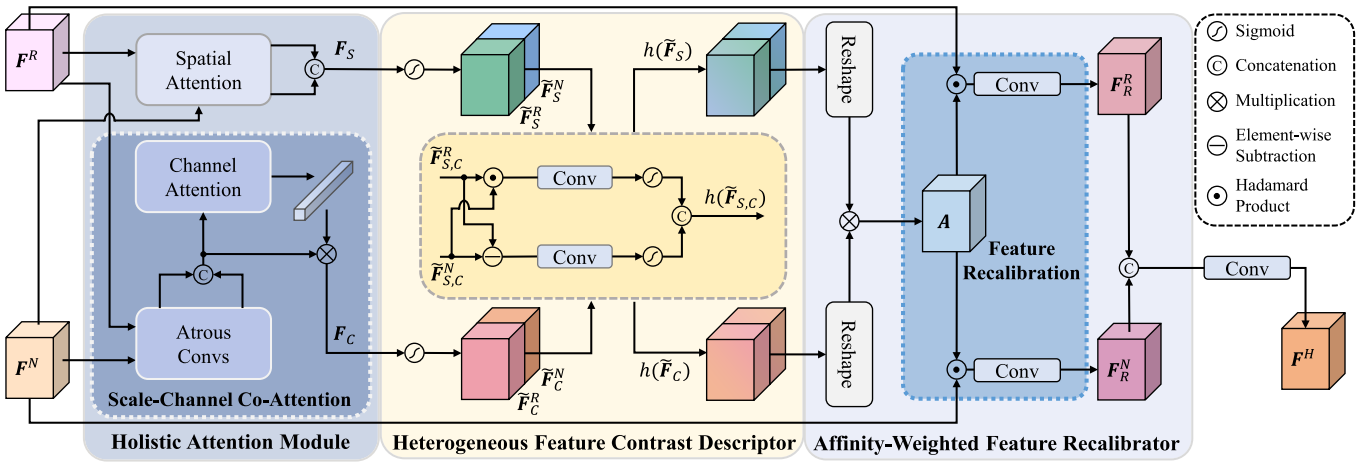

Fig. 2. Illustration of our proposed heterogeneous feature fusion block.

[26], [27] and scene parsing [28], [29], [30]. Although this topic may not have received extensive attention in freespace detection, it is worth noting that notable related works were proposed in the broader field of semantic segmentation. For instance, the cross feature module (CFM) was proposed in [25] for refining features at multiple levels while simultaneously suppressing background noise. Additionally, the separation-and-aggregation gate (SAGate) [26] is another general-purpose heterogeneous feature fusion method that incorporates complementary information through feature recalibration and aggregation to generate selective representations for segmentation. Moreover, Pang et al. [27] introduced the dynamic dilated pyramid module (DDPM), which generates adaptive kernels for efficient feature decoding. As these prior studies generally overlook the appropriate discrimination between the inherent differences of heterogeneous features, our primary focus in this article is directed toward addressing this aspect.

C. Attention Mechanisms

Attention mechanisms are vital components in modern deep learning models, allowing for effective concentration on specific elements of input data, ultimately leading to a more comprehensive understanding of the environment [31], [32], [33]. As a representative example, the squeeze-and-excitation network (SENet) [31] dynamically recalibrates channel-wise feature responses by explicitly modeling dependencies between channels, enabling the network to emphasize informative channels while suppressing less relevant ones. In addition to channel attention, the convolutional block attention module (CBAM) [34] introduces attention from another dimension—spatial. This lightweight and highly compatible module sequentially computes channel and spatial attention maps and multiplies them with the input feature maps to achieve more adaptive feature refinement. On the other hand, Swin Transformer [11] is a general-purpose Transformer backbone developed specifically for fundamental computer vision tasks. It is highly regarded for its hierarchical representation learning approach, which computes self-attention locally within nonoverlapping shifted windows. This innovative design contributes to its compelling performance in applications, including image classification, object detection, and semantic segmentation. In this article, we first extend CBAM to three dimensions: the original two-plus scale. Moreover, we use two Swin Transformers as the backbone networks in the SNE-RoadSegV2 duplex encoder, and comprehensive experiments in Section IV provide evidence of its superior performance compared to convolutional neural networks (CNNs).

III. Methodology

A. Architecture Overview

Fig. 1 provides readers with an overview of the SNE-RoadSegV2 architecture, consisting of three key elements: duplex feature embedding, heterogeneous feature fusion, and lightweight, yet effective feature decoding. A pair of input RGB image $\boldsymbol{I}^R$ and surface normal map $\boldsymbol{I}^N$, translated from a depth image $\boldsymbol{I}^D$ using a surface normal estimator (SNE) [35], are first tokenized into nonoverlapping patches and transformed into a high-dimensional feature space through a trainable linear projection in the patch embedding module [11]. The embedded features are subsequently fed into the Swin Transformer blocks [11] to produce hierarchical heterogeneous features $\mathcal{F}^R = \{\boldsymbol{F}_1^R, \ldots, \boldsymbol{F}_k^R\}$ and $\mathcal{F}^N = \{\boldsymbol{F}_1^N, \ldots, \boldsymbol{F}_k^N\}$. Each pair of heterogeneous features $\boldsymbol{F}_i^{R,N} \in \mathbb{R}^{C\times H\times W}$ undergoes a comprehensive fusion process through HF^{2}B. Finally, a lightweight, yet more effective decoder that incorporates both interscale and intrascale skip connections is designed to further boost the efficiency and accuracy of freespace detection. The proposed architecture is trained by minimizing a loss with fallibility awareness incorporated. Sections III-A–III-C will provide a detailed explanation of the HF^{2}B, decoder, and loss functions in sequence.

B. Heterogeneous Feature Fusion Block

The core problem of feature encoding for freespace detection revolves around the effective fusion of heterogeneous features extracted from various data sources. As detailed in Section I, heterogeneous features in the previous works were fused without appropriate discrimination between their

inherent differences [1], [8], [21], [23]. Our HF^2 block is specifically designed to overcome this limitation. As depicted in Fig. 2, meaningful heterogeneous features are first selectively emphasized and suppressed across three dimensions (spatial, channel, and scale) through an HAM. These features are further enhanced through an HFCD, which improves the representations of their shared and distinct heterogeneous data characteristics. The original heterogeneous features are ultimately weighted through an AWFR to emphasize the aspects that are important to both or either of the features.

1) Holistic Attention Module: Before contrasting and fusing heterogeneous features, it is imperative to emphasize or attenuate specific spatial regions and channels across multiple scales [15], [34]. Drawing inspiration from CBAM [34], we first apply spatial attention to both RGB and surface normal feature maps $\boldsymbol{F}^R$ and $\boldsymbol{F}^N$, resulting in spatially weighted feature maps $\boldsymbol{F}_S^R$ and $\boldsymbol{F}_S^N$, respectively, which are then concatenated to form $\boldsymbol{F}_S$

$$\boldsymbol{F}_S \triangleq \left[\boldsymbol{F}_S^R; \boldsymbol{F}_S^N\right] = \left[f_s\left(\boldsymbol{F}^R\right); f_s\left(\boldsymbol{F}^N\right)\right] \in \mathbb{R}^{2C\times H\times W} \quad (1)$$

where $f_s(\cdot)$ denotes the spatial attention operation, allowing the model to prioritize important regions of an image while suppressing less relevant areas.

As illustrated in Fig. 2, another branch of HAM incorporates multiscale contextual information along with channel attention. A series of atrous convolutional layers [15] are initially employed to generate features $\boldsymbol{F}_A$ with progressively expanded receptive fields

$$\boldsymbol{F}_A = \left[f_a\left(\boldsymbol{F}^R\right); f_a\left(\boldsymbol{F}^N\right)\right] \in \mathbb{R}^{2C\times H\times W} \quad (2)$$

where $f_a(\cdot)$ denotes the multiscale context aggregation operation. A channel attention operation f_c is subsequently applied to further model the interdependencies between heterogeneous features at both scale and channel levels, resulting in scale-channel attentive feature maps $\boldsymbol{F}_C$ as follows:

$$\boldsymbol{F}_C \triangleq \left[\boldsymbol{F}_C^R; \boldsymbol{F}_C^N\right] = f_c(\boldsymbol{F}_A) \in \mathbb{R}^{2C\times H\times W}. \quad (3)$$

In contrast to prior works [31], [34] which exclusively focus on channel attention and lack the consideration of multiple scales, our designed scale-channel co-attention mechanism provides a more comprehensive perspective on heterogeneous features.

2) Heterogeneous Feature Contrast Descriptor: After modeling the interdependencies of the heterogeneous features across three separate dimensions using (1)–(3), we further explore the way to contrast these features more comprehensively and effectively. Unlike relevant prior arts, for example, CFM [25], SAGate [26], and DDPM [27], which primarily emphasize feature commonality, our investigation delves deeper into both their shared and distinct characteristics. To this end, we first normalize the spatial-attentive and scale-channel co-attentive features $\boldsymbol{F}_S$ and $\boldsymbol{F}_C$ using a sigmoid function $\sigma(\cdot)$, yielding $\tilde{\boldsymbol{F}}_S = [\tilde{\boldsymbol{F}}_S^R; \tilde{\boldsymbol{F}}_S^N]$ and $\tilde{\boldsymbol{F}}_C = [\tilde{\boldsymbol{F}}_C^R; \tilde{\boldsymbol{F}}_C^N]$, respectively. Performing an element-wise product operation between $\tilde{\boldsymbol{F}}_{S,C}^R$ and $\tilde{\boldsymbol{F}}_{S,C}^N$ activates jointly emphasized parts between heterogeneous features, while performing an element-wise subtraction operation between $\tilde{\boldsymbol{F}}_{S,C}^R$ and $\tilde{\boldsymbol{F}}_{S,C}^N$ activates features that are important only in either the RGB image or the surface normal map. Upon such inspiration, we formulate our HFCD as follows:

$$h\left(\tilde{\boldsymbol{F}}_{S,C}\right) = \left[w_{h,1}\left(\tilde{\boldsymbol{F}}_{S,C}^R \odot \tilde{\boldsymbol{F}}_{S,C}^N\right); w_{h,2}\left(\tilde{\boldsymbol{F}}_{S,C}^R \ominus \tilde{\boldsymbol{F}}_{S,C}^N\right)\right] \quad (4)$$

where $\odot$ refers to the Hadamard product and $\ominus$ denotes element-wise subtraction. The operators $w_{h,i}$ represent a combination of convolutional, batchnorm, and sigmoid layers. As shown in Fig. 2, the described heterogeneous feature contrast $h\left(\tilde{\boldsymbol{F}}_{S,C}\right)$ is subsequently utilized to construct an affinity volume $\boldsymbol{A}$ for further feature recalibration.

3) Affinity-Weighted Feature Recalibrator: Constructing a volume that contains element-wise weights to jointly recalibrate (emphasize and de-emphasize) heterogeneous features is another significant contribution to our designed HF^{2}B. As $h(\tilde{\boldsymbol{F}}_S)$ and $h(\tilde{\boldsymbol{F}}_C)$ describe the contrasting aspects between the heterogeneous features at the spatial and scale-channel levels, respectively, we employ these two volumes to construct an affinity volume $\boldsymbol{A} \in \mathbb{R}^{C\times H\times W}$ as follows:

$$\boldsymbol{A} = \text{Reshape}\left(h(\tilde{\boldsymbol{F}}_S)\right) \text{Reshape}\left(h(\tilde{\boldsymbol{F}}_C)\right)^\top \quad (5)$$

which provides the original heterogeneous features $\boldsymbol{F}^R$ and $\boldsymbol{F}^N$ with element-wise weights between 0 and 1. Specifically, a higher affinity value indicates greater importance of that element in both types of feature maps, an intermediate affinity value indicates greater importance of an element in either of the feature maps, while a lower affinity value indicates a redundant element in both types of feature maps that should be neglected. Finally, $\boldsymbol{F}^R$ and $\boldsymbol{F}^N$ are weighted by $\boldsymbol{A}$ to form $\boldsymbol{F}_R^R$ and $\boldsymbol{F}_R^N$ for the next stage of feature encoding and then concatenated to generate the recalibrated heterogeneous features as follows:

$$\boldsymbol{F}^H = w_a\left(\left[\underbrace{w_r\left(\boldsymbol{F}^R \odot \boldsymbol{A}\right)}_{\boldsymbol{F}_R^R}; \underbrace{w_n\left(\boldsymbol{F}^N \odot \boldsymbol{A}\right)}_{\boldsymbol{F}_R^N}\right]\right) \in \mathbb{R}^{C\times H\times W} \quad (6)$$

where w_r, w_n, and w_a denote convolutional layers. Compared to other SoTA heterogeneous feature fusion strategies, our proposed HF^{2}B adaptively assigns weights to the original features, taking into account both the elements of agreement and disagreement in importance as determined by HFCD and AWFR. Such a way of seeking common ground while preserving differences enhances the comprehensiveness of feature fusion for freespace detection. The superior performance of HF^{2}B and the effectiveness of each component are demonstrated in Section IV-C.

C. Lightweight, Yet More Effective Decoder

Fig. 3 illustrates the decoder architectures of RoadSeg [1] (identical to UNet++ [36]), UNet3+ [10], and our proposed SNE-RoadSegV2. As claimed in [10], UNet++ fails to effectively utilize multiscale features, and UNet3+ was designed specifically to resolve this limitation. However, before designing our decoder, we must address the following question: Does UNet3+ consistently outperform UNet++ in the freespace detection task? Through an extensive series of experiments with both CNN and Transformer backbones, we regret to

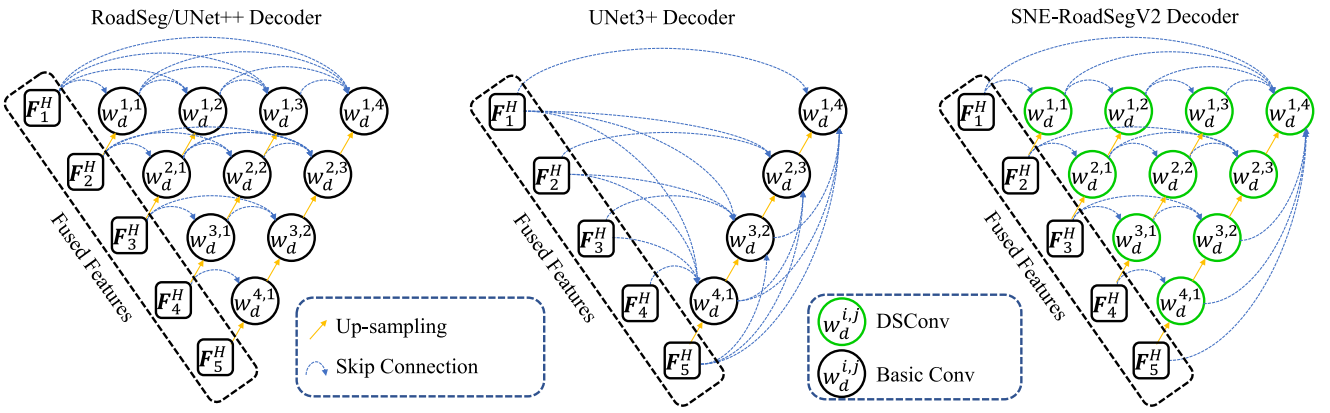

Fig. 3. Comparisons among decoder architectures of RoadSeg/UNet++, UNet3+, and our proposed SNE-RoadSegV2. $w_d^{i,j}$ denotes a basic convolutional layer (Basic Conv) or depth-wise separable convolutional layer (DSConv) with batchnorm and sigmoid layers.

report that the answer is negative. However, both the intrascale skip connections in UNet++ and the interscale skip connections in UNet3+ remain indispensable. Thus, we design the SNE-RoadSegV2 decoder, which combines the strengths of both UNet++ and UNet3+, through enormous experimental efforts. As shown in Fig. 3, we maintain skip connections only from a given node to its adjacent and final nodes at each stage. This modification reduces redundant information propagation without compromising decoder performance. Additionally, we adopt the interscale skip connections used in UNet3+ to capture both fine-grained and coarse-grained details. Moreover, we replace the basic convolutions in the decoder with depth-wise separable convolutions [37] to further reduce its computational complexity. Section IV-C quantitatively demonstrates that our SNE-RoadSegV2 decoder outperforms UNet++ and UNet3+ in terms of both efficiency and accuracy. Despite its superior performance, we consider the employed decoder as an experimental contribution.

D. Fallibility-Aware Loss Functions

In previous works, the pixel-wise BCE loss has been a commonly used primary loss function during supervised model training. However, such efforts have not considered the specific characteristics of real-world driving scenarios. Misclassifications are frequently observed near the transition regions between different semantic categories [25]. In addition, the depth data are not explicitly utilized to supervise model training. Hence, we propose a novel modification to the conventional BCE loss function $\mathcal{L}_{\text{BCE}}$ by introducing two weighting factors that prioritize these error-prone regions. The overall loss is formulated as follows:

$$\mathcal{L} = \mathcal{L}_{\text{BCE}} + \lambda_S \mathcal{L}_{\text{STA}} + \lambda_D \mathcal{L}_{\text{DIA}} \tag{7}$$

where λ_S and λ_D balance the semantics transition-aware loss $\mathcal{L}_{\text{STA}}$ and the depth inconsistency-aware loss $\mathcal{L}_{\text{DIA}}$, respectively. The ablation studies on their individual efficacy and the selection of hyperparameters λ_S and λ_D are provided in Section IV-C.

1) Semantics Transition-Aware Loss: A transition region between different semantic categories can be considered as a mixture of semantic labels. For each pixel $\boldsymbol{q}$ with a neighborhood system $\mathcal{N}_{\boldsymbol{q}} = \mathcal{F}_{\boldsymbol{q}} \cup \mathcal{B}_{\boldsymbol{q}}$, where $\mathcal{F}_{\boldsymbol{q}}$ and $\mathcal{B}_{\boldsymbol{q}}$ denote the foreground (freespace) and background (others) sets, respectively, we determine its likelihood $\omega_S(\boldsymbol{q}) \in [0, 1]$ of belonging to a semantics transition region using the following expression:

$$\omega_S(\boldsymbol{q}) = \cos\left(\pi \left| \frac{\sum_{\boldsymbol{p}} I_{\mathcal{F}_{\boldsymbol{q}}}(\boldsymbol{p})}{\sum_{\boldsymbol{p}} I_{\mathcal{N}_{\boldsymbol{q}}}(\boldsymbol{p})} - \frac{1}{2} \right| \right) \tag{8}$$

where $I(\cdot)$ is the indicator function and $\boldsymbol{p} \in \mathcal{N}_{\boldsymbol{q}}$. This likelihood approaches 0 when either $\mathcal{F}_{\boldsymbol{q}}$ or $\mathcal{B}_{\boldsymbol{q}}$ is close to being an empty set and approaches 1 in semantics transition regions. Substituting (8) into the standard BCE loss yields $\mathcal{L}_{\text{STA}}$ as follows:

$$\mathcal{L}_{\text{STA}} = -\sum_{\boldsymbol{q}} \omega_S(\boldsymbol{q}) \Big(y_{\boldsymbol{q}} \log p_{\boldsymbol{q}} + (1 - y_{\boldsymbol{q}}) \log(1 - p_{\boldsymbol{q}}) \Big) \tag{9}$$

where $y_{\boldsymbol{q}} \in \{0, 1\}$ denotes the ground-truth label of $\boldsymbol{q}$ (1 for freespace and 0 otherwise), while $p_{\boldsymbol{q}} \in [0, 1]$ indicates the probability that $\boldsymbol{q}$ belongs to the freespace category. Section IV provides details on the selection of $\mathcal{N}_{\boldsymbol{q}}$ radius.

2) Depth Inconsistency-Aware Loss: When depth images are available, it is also advantageous to leverage these data to improve network training via an adaptive loss function. However, prior research efforts have not explored this aspect.

Let $\tilde{\boldsymbol{Q}} \in \mathbb{R}^{3 \times N}$ be a matrix storing the homogeneous coordinates $\tilde{\boldsymbol{q}}$ of the predicted freespace pixels $\boldsymbol{q}$. Considering that the height between the camera and the ground plane remains theoretically constant in each image when disregarding the camera pitch angle, we aggregate the y-coordinates of these pixels to derive a theoretical camera height $\hat{y}$, using the following expression:

$$\hat{y} = \left[0, \frac{1}{N}, 0\right] \boldsymbol{K}^{-1} \tilde{\boldsymbol{Q}} \boldsymbol{z} \tag{10}$$

where $\boldsymbol{z} \in \mathbb{R}^N$ stores the depth values $\boldsymbol{I}^D(\boldsymbol{q})$. A weight $\omega_D(\boldsymbol{q}) \in [0, 1)$ measuring the depth inconsistency can then be yielded as follows:

$$\omega_D(\boldsymbol{q}) = 1 - \exp\left(-\left| \frac{\hat{y}}{[0, 1, 0]\boldsymbol{K}^{-1}\tilde{\boldsymbol{q}}} - \boldsymbol{I}^D(\boldsymbol{q}) \right| \right). \tag{11}$$

TABLE I

QUANTITATIVE COMPARISON AMONG SOTA FREESPACE DETECTION ALGORITHMS ON THE KITTI ROAD OFFICIAL BENCHMARK.[1] THE SYMBOL ↑ INDICATES THAT HIGHER VALUES CORRESPOND TO BETTER PERFORMANCE, WHILE ↓ IMPLIES THE OPPOSITE. "RGB": RGB IMAGES, "DISP": DISPARITY IMAGES, "DEPTH": DEPTH IMAGES, "PC": POINT CLOUDS, AND "NORMAL": SURFACE NORMAL MAPS

Method	Input Data	MaxF (%) ↑	AP (%) ↑	Pre (%) ↑	Rec (%) ↑	FPR (%) ↓	FNR (%) ↓	Rank ↓
Hadamard-FCN [39]	RGB	94.85	91.48	94.81	94.89	2.86	5.11	35
RBANet [24]	RGB	96.30	89.72	95.14	97.50	2.75	2.50	20
HA-DeepLabv3+ [40]	RGB + Disp	94.83	93.24	94.77	94.89	2.88	5.11	36
DFM-RTFNet [41]	RGB + Disp	96.78	94.05	96.62	96.93	1.87	3.07	15
USNet [23]	RGB + Depth	96.89	93.25	96.51	97.27	1.94	2.73	13
LRDNet+ [21]	RGB + LiDAR PC	96.95	92.22	96.88	97.02	1.72	2.98	12
PLB-RD [42]	RGB + LiDAR PC	97.42	**94.09**	97.30	97.54	1.49	2.46	8
PLARD [2]	RGB + LiDAR PC	97.03	94.03	97.19	96.88	1.54	3.12	11
BJN [43]	RGB + LiDAR PC	94.89	90.63	96.14	93.67	2.07	6.33	32
LidCamNet [30]	RGB + LiDAR PC	96.03	93.93	96.23	95.83	2.07	4.17	21
CLCFNet [22]	RGB + LiDAR PC	96.38	90.85	96.38	96.39	1.99	3.61	19
NIM-RTFNet [44]	RGB + Normal	96.02	94.01	96.43	95.62	1.95	4.38	22
SNE-RoadSeg [1]	RGB + Normal	96.75	94.07	96.90	96.61	1.70	3.39	16
SNE-RoadSeg+ [8]	RGB + Normal	97.50	93.98	97.41	97.58	1.43	2.42	6
RoadFormer [16]	RGB + Normal	97.50	93.85	97.16	**97.84**	1.57	**2.16**	6
RoadFormer+ [45]	RGB + Normal	**97.56**	93.74	97.43	97.69	1.42	2.31	3
SNE-RoadSegV2 (Ours)	RGB + Normal	97.55	93.98	**97.57**	97.53	**1.34**	2.47	5

It approaches 0 or 1 when the free-space depth is consistent or inconsistent, respectively. $\mathcal{L}_{\text{DIA}}$ is thus, formulated as follows:

$$\mathcal{L}_{\text{DIA}} = -\sum_{\boldsymbol{q}} \omega_D(\boldsymbol{q})\Big(y_{\boldsymbol{q}} \log p_{\boldsymbol{q}} + (1 - y_{\boldsymbol{q}}) \log(1 - p_{\boldsymbol{q}})\Big). \tag{12}$$

IV. EXPERIMENTS

We perform both qualitative and quantitative comparisons between SNE-RoadSegV2 and other SoTA freespace detection algorithms on the KITTI Road dataset [47] (medium-sized) and the Cityscapes dataset [46] (large-scale). Subsequently, we conduct extensive experiments to validate the effectiveness of our proposed encoder, decoder, and loss functions, as detailed in the ablation studies. To further demonstrate the superior performance of our network, we also provide additional experiments carried out on the vKITTI2 dataset [48] (large-scale, yet synthetic) and the KITTI Semantics dataset [49] (real-world, yet small-sized) in the supplement at https://mias.group/SNE-RoadSegV2 to validate the robustness of our free-space detection framework.

A. Datasets, Evaluation Metrics, and Implementation Details

The details on the four datasets are as follows.

1) *KITTI Road [47]:* This dataset provides real-world RGB-D data (image resolution: 1242 × 375 pixels) for the evaluation of data-fusion freespace detection algorithms. Following the study presented in [1], we split the dataset into three subsets: training (173 images), validation (58 images), and testing (58 images) to conduct ablation studies and hyperparameter selection experiments.

2) *Cityscapes [46]:* This dataset provides real-world stereo images (resolution: 2048 × 1.024 pixels), each manually annotated with 34 semantic classes. In our experiments, we preprocess the ground-truth annotations by categorizing them into two groups: freespace and others. As there is no depth ground truth available, we utilize a pretrained RAFT-Stereo [50] to generate depth images. We adhere to the official split of training, and validation sets, with 2975 and 500 images in each set.

Adhering to the experiments presented in [1], we quantify the model's performance using accuracy (Acc), precision (Pre), recall (Rec), $F1$-score (Fsc), and intersection over union (IoU) [38]. Additionally, when submitting the results obtained from the best-performing model to the KITTI Road benchmark, we also compute the maximum $F1$-measure (MaxF), average precision (AP), false-positive rate (FPR), and false negative rate (FNR) [38].

Our experiments are conducted using an Intel Core i7-12700k CPU and an NVIDIA RTX 4090 GPU. The Adam optimizer [51] with an initial learning rate of 0.001 is used to minimize the loss function. A multistep learning scheduler with a decay rate of 0.5 for every 20 epochs is also employed. Each model is trained for a total of 100 epochs, with early stopping mechanisms applied to the validation set to prevent over-fitting. Common data augmentation techniques, such as random flipping, rotation, cropping, and brightness adjustment, are also applied to enhance the model's robustness.

B. Comparison With SoTA Methods

The quantitative and qualitative experimental results on the KITTI Road dataset are presented in Table I and Fig. 4, respectively, while the quantitative and qualitative experimental results on the Cityscapes dataset are given in Table II and Fig. 5, respectively. These results suggest that our pro-

[1]Results are publicly available at cvlibs.net/datasets/kitti/eval_road.php.

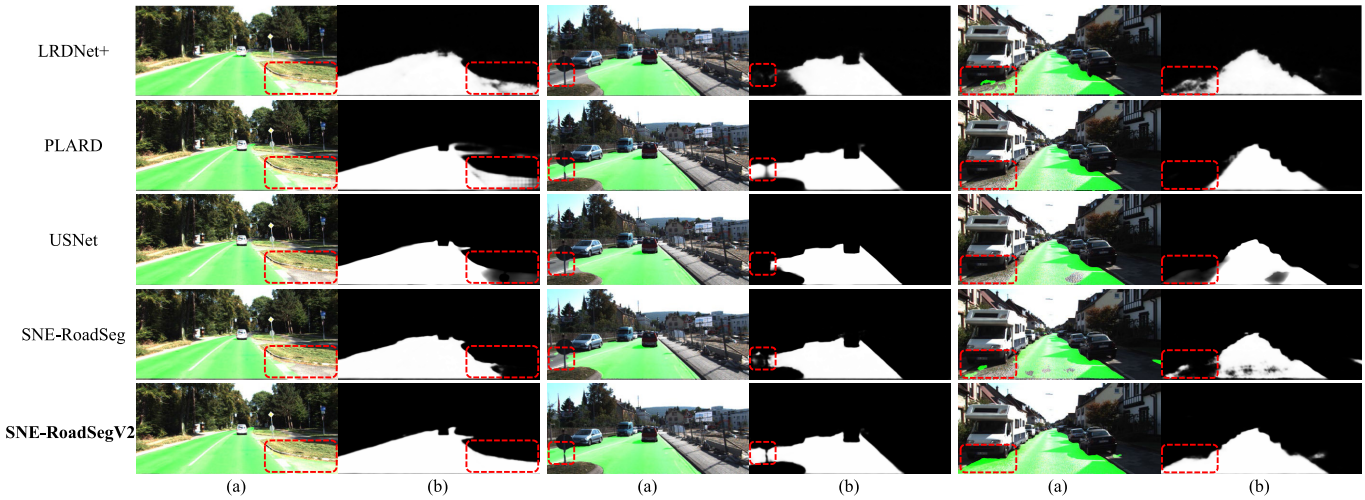

Fig. 4. Qualitative comparisons of SoTA freespace detection algorithms on the KITTI Road dataset [38]. The results of the compared algorithms are obtained using their officially published source codes and weights. (a) Freespace detection results. (b) Probability maps.

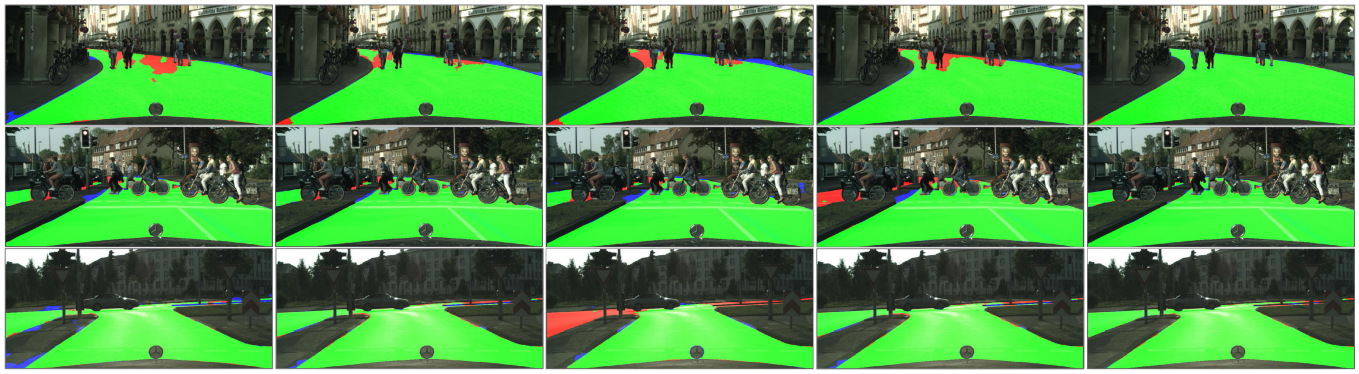

Fig. 5. Qualitative comparisons of SoTA freespace detection algorithms on the Cityscapes dataset [46]. The results are visualized with true-positive classifications in green, false-positive classifications in blue, and false-negative classifications in red.

TABLE II

COMPARISON AMONG SOTA FREESPACE DETECTION ALGORITHMS ON THE CITYSCAPES DATASET [46]

Method	Fsc (%) ↑	IoU (%) ↑	Acc (%) ↑
NIM-RTFNet [44]	92.02	85.22	96.07
RBANet [24]	93.81	88.34	96.50
USNet [23]	94.28	89.18	96.71
LRDNet+ [21]	94.71	89.95	97.02
PLARD [2]	95.28	90.99	97.15
SNE-RoadSeg [1]	96.49	93.22	97.68
SNE-RoadSegV2 (Ours)	**97.12**	**94.40**	**98.11**

posed SNE-RoadSegV2 demonstrates superior performance compared to all other SoTA free-space detection approaches, with an increase in MaxF by up to 2.72% on the KITTI dataset and an increase in IoU by 1.18% versus the second best on the Cityscapes dataset. The qualitative comparisons, with significantly improved regions highlighted by red dashed boxes, particularly near semantic transition and depth-inconsistent regions, also validate the effectiveness of our designed feature fusion block, decoder, and loss function. The ablation studies that validate the individual efficacy of these components are discussed in the next section.

TABLE III

ABLATION STUDY ON THE DESIGN OF HF^{2}B. "BASELINE": STANDARD FEATURE FUSION EMPLOYED IN SNE-ROADSEG, "SA": SPATIAL ATTENTION, "CA": CHANNEL ATTENTION, "AC": ATROUS CONVOLUTIONS

Baseline	HAM			HFCD	ACFR	Fsc (%) ↑
	SA	CA	AC			
✓						96.65
	✓			✓	✓	96.54
	✓	✓		✓	✓	97.11
	✓	✓	✓	✓	✓	**97.69**
	✓	✓	✓		✓	96.84
	✓	✓	✓	✓		97.08

C. Ablation Study

1) Encoder: We first explore the rationality of each component in HF^{2}B. As presented in Table III, we sequentially remove each component from HF^{2}B to quantify its impact on the overall performance. It is evident that each component in HF^{2}B contributes to an improvement in the overall performance, and the network achieves its peak performance when all components (HAM, HFCD, and AWFR) are integrated into HF^{2}B, which demonstrates the effectiveness of our design.

TABLE IV

COMPARISON BETWEEN OUR PROPOSED HF^{2}B AND OTHER SOTA HETEROGENEOUS FEATURE FUSION STRATEGIES

Strategy	KITTI Road Dataset		Cityscapes Dataset	
	Fsc (%)	IoU (%)	Fsc (%)	IoU (%)
CFM [25]	95.15	91.58	95.45	92.33
SAGate [26]	96.77	93.74	96.80	93.81
DDPM [27]	96.05	92.98	96.60	93.43
HF^{2}B (Ours)	**97.68**	**94.90**	**97.12**	**94.40**

TABLE V

COMPARISON OF DECODERS IN TERMS OF BOTH ACCURACY AND COMPUTATIONAL COMPLEXITY ON THE KITTI ROAD DATASET

Decoder	Fsc (%)	IoU (%)	Params (M)	FLOPS (G)
UNet++ [36]	96.27	93.90	13.62	78.44
UNet3+ [10]	95.64	93.41	14.70	164.63
Ours [10]	**97.58**	**94.50**	**6.71**	**60.33**

Additionally, we compare HF^{2}B with other SoTA heterogeneous feature fusion strategies. As shown in Table IV, HF^{2}B outperforms other compared methods on both datasets, with improvements of up to 1.26% in Fsc and 1.67% in IoU, respectively. These compelling results can be attributed to our novel contributions, particularly the exploitation of both shared and distinct characteristics of heterogeneous features in HFCD, and the effective feature recalibration based on the affinity volume constructed in AWFR.

2) Decoder: Quantitative comparisons of decoder performance among SNE-RoadSegV2, RoadSeg/UNet++, and UNet3+ are presented in Table V. These results demonstrate the SoTA performance of our decoder, with improvements in Fsc and IoU by up to 1.94% and 1.09%, respectively, while maintaining lower computational complexity, including a reduction in learnable parameters and FLOPS by up to 54.35% and 63.35%, respectively. These improvements can be attributed to the use of depth-wise separable convolution and the pruning of redundant skip connections.

3) Loss Function: Fig. 6 actually presents two experiments: 1) when $\lambda_S = \lambda_D = 0$ (only conventional BCE loss is used), the overall freespace detection performance on both datasets is the worst, demonstrating the effectiveness of our proposed fallibility-aware losses and 2) different ratios between λ_S and λ_D demonstrate that when $\lambda_S = 0.3$ and $\lambda_D = 0.1$ (the sum of these two weights is empirically set to 0.4, as discussed in [52], [53], and [54]), SNE-RoadSegV2 achieves the best performance on both datasets. While further hyperparameter tuning is possible, it is important to consider the risk of overfitting with limited data.

V. CONCLUSION AND FUTURE WORK

This article revisited the designs of heterogeneous feature fusion strategies, decoder architectures, and loss functions from prior research and introduced SNE-RoadSegV2, a novel, high-performing, SoTA freespace detection network. Breaking down our contributions further, our work contains six technical contributions: three novel components in the encoder, one decoder architecture, and two loss functions. The effectiveness of each contribution was validated through extensive experiments. Comprehensive comparisons with other SoTA algorithms unequivocally demonstrate the superiority of SNE-RoadSegV2. However, it can be observed that the MaxF scores of top-ranked algorithms have already achieved over 95% and are relatively close to each other, leaving little room for further performance improvement. This is due to the homogenized scenes and incorrect manual annotations in the dataset. Thus, our future work will focus on creating a more diverse, challenging autonomous driving dataset for fair and comprehensive algorithm evaluation.

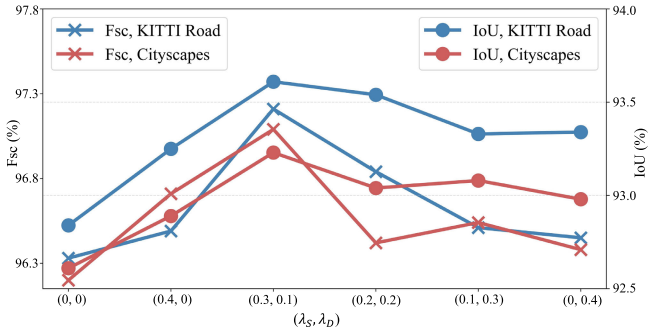

Fig. 6. Selection of hyperparameters λ_S and λ_D in (7).

ACKNOWLEDGMENT

Views and opinions expressed are, however, those of the authors only and do not necessarily reflect those of the European Union–European Commission. Neither the European Commission nor the European Union can be held responsible for them. It should be clarified that the collaboration among the authors is limited solely to this work and does not extend to any other projects.

REFERENCES

[1] 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.* Cham, Switzerland: Springer, 2020, pp. 340–356.

[2] Z. Chen, J. Zhang, and D. Tao, "Progressive LiDAR adaptation for road detection," *IEEE/CAA J. Autom. Sinica*, vol. 6, no. 3, pp. 693–702, May 2019.

[3] 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)*. Cham, Switzerland: Springer, 2017, pp. 213–228.

[4] Q. Ha, K. Watanabe, T. Karasawa, Y. Ushiku, and T. Harada, "MFNet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Sep. 2017, pp. 5108–5115.

[5] 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.

[6] J. Li, P. Yun, Q. Chen, and R. Fan, "HAPNet: Toward superior RGB-thermal scene parsing via hybrid, asymmetric, and progressive heterogeneous feature fusion," 2024, *arXiv:2404.03527*.

[7] X. Zhou, H. Wen, R. Shi, H. Yin, J. Zhang, and C. Yan, "FANet: Feature aggregation network for RGBD saliency detection," *Signal Process., Image Commun.*, vol. 102, Mar. 2022, Art. no. 116591.

[8] H. Wang, R. Fan, P. Cai, and M. Liu, "SNE-RoadSeg+: Rethinking depth-normal translation and deep supervision for freespace detection," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Sep. 2021, pp. 1140–1145.

[9] W. Zou, R. Long, Y. Zhang, M. Liao, Z. Zhou, and S. Tian, "Dual geometric perception for cross-domain road segmentation," *Displays*, vol. 76, Jan. 2023, Art. no. 102332.

[10] H. Huang et al., "UNet 3+: A full-scale connected UNet for medical image segmentation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 1055–1059.

[11] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10012–10022.

[12] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[13] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*. Cham, Switzerland: Springer, 2015, pp. 234–241.

[14] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder–decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.

[15] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, Atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2017.

[16] 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.

[17] M. J. AlvarezTheo, G. LeCunAntonio, and M. Lopez, "Road scene segmentation from a single image," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, 2012, pp. 376–389.

[18] L. Xiao, B. Dai, D. Liu, D. Zhao, and T. Wu, "Monocular road detection using structured random forest," *Int. J. Adv. Robot. Syst.*, vol. 13, no. 3, p. 101, May 2016.

[19] C.-A. Brust, S. Sickert, M. Simon, E. Rodner, and J. Denzler, "Convolutional patch networks with spatial prior for road detection and urban scene understanding," in *Proc. 10th Int. Conf. Comput. Vis. Theory Appl.*, 2015, pp. 510–517.

[20] D. Levi, N. Garnett, and E. Fetaya, "StixelNet: A deep convolutional network for obstacle detection and road segmentation," in *Proc. Brit. Mach. Vis. Conf.*, 2015, pp. 109.1–109.12.

[21] A. A. Khan, J. Shao, Y. Rao, L. She, and H. T. Shen, "LRDNet: Lightweight LiDAR aided cascaded feature pools for free road space detection," *IEEE Trans. Multimedia*, vol. 27, pp. 652–664, 2022.

[22] 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.

[23] 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.

[24] J.-Y. Sun, S.-W. Kim, S.-W. Lee, Y.-W. Kim, and S.-J. Ko, "Reverse and boundary attention network for road segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 876–885.

[25] J. Wei, S. Wang, and Q. Huang, "FNet: Fusion, feedback and focus for salient object detection," in *Proc. AAAI Conf. Artif. Intell.*, Feb. 2020, vol. 34, no. 7, pp. 12321–12328.

[26] X. Chen et al., "Bi-directional cross-modality feature propagation with separation-and-aggregation gate for RGB-D semantic segmentation," in *Proc. 16th Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, Aug. 2020, pp. 561–577.

[27] Y. Pang, L. Zhang, X. Zhao, and H. Lu, "Hierarchical dynamic filtering network for RGB-D salient object detection," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 235–252.

[28] S. Qiu, S. Anwar, and N. Barnes, "Semantic segmentation for real point cloud scenes via bilateral augmentation and adaptive fusion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 1757–1767.

[29] L. Sun, K. Yang, X. Hu, W. Hu, and K. Wang, "Real-time fusion network for RGB-D semantic segmentation incorporating unexpected obstacle detection for road-driving images," *IEEE Robot. Autom. Lett.*, vol. 5, no. 4, pp. 5558–5565, Apr. 2020.

[30] L. Caltagirone, M. Bellone, L. Svensson, and M. Wahde, "LiDAR–camera fusion for road detection using fully convolutional neural networks," *Robot. Auto. Syst.*, vol. 111, pp. 125–131, Jan. 2019.

[31] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.

[32] K. Xu et al., "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. 32nd Int. Conf. Mach. Learn. (ICML)*, vol. 37, 2015, pp. 2048–2057.

[33] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 30, Jun. 2017, pp. 5998–6008.

[34] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 3–19.

[35] Y. Feng, B. Xue, M. Liu, Q. Chen, and R. Fan, "D2NT: A high-performing depth-to-normal translator," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2023, pp. 12360–12366.

[36] Z. Zhou et al., "UNet++: Redesigning skip connections to exploit multiscale features in image segmentation," *IEEE Trans. Med. Imag.*, vol. 39, no. 6, pp. 1856–1867, Dec. 2019.

[37] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1251–1258.

[38] J. Fritsch, T. Kuhnl, 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.

[39] M. Oeljeklaus, *An Integrated Approach for Traffic Scene Understanding From Monocular Cameras*. Germany: VDI Verlag, 2021.

[40] 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.

[41] H. Wang, R. Fan, Y. Sun, and M. Liu, "Dynamic fusion module evolves drivable area and road anomaly detection: A benchmark and algorithms," *IEEE Trans. Cybern.*, vol. 52, no. 10, pp. 10750–10760, Mar. 2021.

[42] 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.

[43] B. Yu, D. Lee, J.-S. Lee, and S.-C. Kee, "Free space detection using camera-LiDAR fusion in a bird's eye view plane," *Sensors*, vol. 21, no. 22, p. 7623, Nov. 2021.

[44] H. Wang, R. Fan, Y. Sun, and M. Liu, "Applying surface normal information in drivable area and road anomaly detection for ground mobile robots," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2020, pp. 2706–2711.

[45] J. Huang et al., "RoadFormer+: Delivering RGB-X scene parsing through scale-aware information decoupling and advanced heterogeneous feature fusion," *IEEE Trans. Intell. Vehicles*, early access, Aug. 22, 2024, doi: 10.1109/TIV.2024.344825.

[46] M. Cordts et al., "The cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3213–3223.

[47] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 3354–3361.

[48] Y. Cabon, N. Murray, and M. Humenberger, "Virtual KITTI 2," 2020, *arXiv:2001.10773*.

[49] H. A. 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.

[50] L. Lipson, Z. Teed, and J. Deng, "RAFT-stereo: Multilevel recurrent field transforms for stereo matching," in *Proc. Int. Conf. 3D Vis. (3DV)*, 2021, pp. 218–227.

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

[52] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, "CCNet: Criss-cross attention for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 603–612.

[53] X. Li, Z. Zhong, J. Wu, Y. Yang, Z. Lin, and H. Liu, "Expectation-maximization attention networks for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9166–9175.

[54] Y. Yuan, X. Chen, and J. Wang, "Object-contextual representations for semantic segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, 2020, pp. 173–190.