IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

TerrFlat: Physics-Driven Geometry Representation for Structure-Aware Freespace Detection

Jingwei Yang[1], Liuyi Wang[1], Mengjiao Shen[1], Jiayuan Du[1], Chengju Liu[1], Qijun Chen[1]

Abstract— **Freespace detection in autonomous driving is limited by the lack of explicit geometric modeling, hindering generalization across complex terrains. Existing approaches are predominantly data-driven and neglect the physical structure of drivable surfaces. We propose Terrain Flat (TerrFlat), a physics-driven geometric representation that models road surfaces along three interpretable dimensions: lateral smoothness, longitudinal consistency, and vertical deviation. TerrFlat is constructed through geometric reasoning and projected into pixel-aligned maps via a differentiable projection, ensuring geometric–visual consistency. Building on this representation, we introduce a symmetric feature fusion module (SFFM) to integrate TerrFlat with visual features through bidirectional recalibration, improving semantic discrimination and boundary localization. Together, TerrFlat and SFFM form TerrFlat-Seg, a unified framework for physics-aware freespace perception. Experiments on KITTI-Road, Semantic-KITTI, and ORFD datasets demonstrate consistent improvements over existing baselines. Real-world validation on an automated guided vehicle platform further confirms the robustness of our approach.**

I. Introduction

The rapid development of embodied intelligence has accelerated the deployment of autonomous systems across various industries, including self-driving vehicles, augmented reality, smart homes, and logistics delivery [1]–[4]. For intelligent agents, reliable decision-making depends on accurate perception of the surrounding environment. Freespace detection, the task of identifying navigable areas in a scene, is therefore critical for safe and robust autonomous navigation [5].

Traditional methods rely on 2D visual cues from RGB images or 3D spatial cues from depth or LiDAR. RGB-based networks extract color and texture for semantic segmentation [6], while depth-based approaches use geometric priors to model freespace as planar or continuous surfaces for improved spatial reasoning. Although RGB–depth fusion enhances performance, existing methods still struggle in complex scenes with uneven terrain or ambiguous appearances [7].

Semantic segmentation dominates freespace detection for its pixel-level understanding [8], yet RGB-based models remain sensitive to adverse conditions [9]. While additional modalities such as LiDAR [10], radar [11], and thermal imaging [12] have been introduced, depth is often treated as an auxiliary cue, neglecting the physical properties of drivable surfaces [13]. This leads to implicit constraint learning, increased training difficulty, and limited generalization.

Existing 3D representations are generally designed for generic scene understanding rather than freespace-specific reasoning. Depth is often abstracted as point clouds or disparity maps without explicitly encoding the physical structure of navigable surfaces [14], resulting in suboptimal integration with semantic cues and reduced interpretability. Likewise, multi-modal fusion commonly relies on naive operations such as feature concatenation or summation [15], which fail to exploit the complementary nature of visual and geometric information. Prior work has explored geometric cues such as surface normals and raw depth to aid freespace detection [16], but there remains a lack of dedicated, physically interpretable 3D representations tailored for this task.

These limitations motivate a central question: *How can we design physics-driven 3D representations that explicitly encode road geometry and seamlessly integrate with RGB perception to achieve reliable freespace detection?*

To address these challenges, this work focuses on designing a physics-driven 3D geometric representation that clearly encodes interpretable road surface properties from depth information, and developing a structure-aware fusion framework to integrate these physical cues with RGB semantics for robust freespace detection. The main contributions are as follows:

- We propose a physics-driven geometric representation, called Terrain Flat (TerrFlat), which comprehensively encodes road geometry across three interpretable physical dimensions: lateral smoothness, longitudinal consistency, and vertical deviation, thereby enabling explicit geometric reasoning for freespace detection.
- We design a symmetric feature fusion module (SFFM) that integrates TerrFlat cues into visual perception through bidirectional recalibration, enhancing semantic awareness and boundary precision.

Together, TerrFlat and SFFM form the TerrFlat-Seg framework, a unified and physics-aware system for robust freespace detection across diverse environments.

II. Related Work

A. Geometric Representations for Freespace Perception

Recent research has explored various 3D representations to enhance scene understanding and improve freespace detection [17]. By leveraging spatial cues from depth, disparity, or

Manuscript received: July 12, 2025; Revised: November 1, 2025; Accepted: January 8, 2026. This paper was recommended for publication by Editor Hyungpil Moon upon evaluation of the Associate Editor and Reviewers comments. This paper is supported by the National Natural Science Foundation of China under Grants (62333017, 62173248, 62233013, 624B2105). (*Corresponding author: Qijun Chen; Chengju Liu*)

[1]The authors are with the College of Electronics & Information Engineering, Robotics and Artificial Intelligence Laboratory (RAIL), Tongji University, Shanghai 201804, China. (e-mails: {jw_yang, wly, 1910680, dujiayuan, liuchengju, qjchen}@tongji.edu.cn)

Digital Object Identifier (DOI): see top of this page.

reconstructed geometry, these methods extend beyond RGB-based perception to offer richer geometric understanding. For instance, elevation maps have been fused with RGB images to better segment uneven terrains, yet they generalize poorly in complex outdoor settings [18]. Similarly, HHA encoding, which captures disparity, height, and angle to gravity, has proven effective in indoor applications but remains limited in handling large-scale outdoor scenes [19].

Disparity-based features have also been employed to detect surface anomalies such as potholes, though they often rely on idealized assumptions like camera–road parallelism that rarely hold in practice [20]. Other works combining depth and surface normals have improved segmentation accuracy, but remain sensitive to noise, occlusion, and geometric variation in driving scenes [21].

Our previous study demonstrated that surface normals provide critical cues for identifying flat, traversable surfaces [16]. However, not all planar regions are drivable, and most 3D representations lack task-specific or physically grounded design. These challenges highlight the need for interpretable geometric representations that explicitly encode road-related physical characteristics for accurate freespace detection.

B. Multi-modal Semantic Segmentation for Freespace Detection

Given the limitations of RGB-only perception, many approaches adopt multi-modal fusion techniques that explicitly incorporate depth, infrared, thermal, or surface normal information [22]. Early methods such as FuseNet [7] and MFNet [23] use encoder–decoder structures to fuse RGB and additional modalities through element-wise addition or concatenation. While effective to some extent, these early fusion strategies often overlook the distinct semantic, physical, and contextual nature of each modality.

Recent approaches have introduced advanced fusion strategies. SNE-RoadSeg [15] and SNE-RoadSegV2 [24] incorporate surface normals into RGB networks and introduce attention-guided heterogeneous fusion, enhancing discriminative learning under complex conditions. However, they rely on static surface normal representations and lack explicit modeling of terrain continuity. USNet [25] uses uncertainty-aware fusion for better efficiency–accuracy trade-offs, while CMX [26] designs a unified fusion block to capture long-range cross-modal dependencies. Evi-RoadSeg [27] applies evidence-based late fusion but underutilizes geometric complementarity. RoadFormer [28] and RoadFormer+ [29] adopt Transformer-based fusion to jointly model RGB and geometric cues, improving multi-modal adaptation through hierarchical attention. Nevertheless, these methods remain purely data-driven, lacking physical constraints or geometric interpretability, which limits robustness on non-planar terrains.

Although multi-modal methods have advanced autonomous driving perception, they struggle to model the physical and geometric properties of drivable surfaces. Few explicitly encode structured road geometry or integrate it meaningfully, motivating physics-driven, task-specific multi-modal representations for more accurate and robust freespace detection in diverse environments.

III. Geometric Preliminaries

To incorporate physically grounded geometric cues, we first estimate surface normals in the surface normal generation module (Fig. 2) from depth images, supporting reasoning over flatness essential for freespace detection.

A 3D point is denoted as $\boldsymbol{q} = [x, y, z]^\top$, with neighboring points $\boldsymbol{P} = (\boldsymbol{p}_1, ..., \boldsymbol{p}_k)$. For each pixel, the surface normal $\boldsymbol{n} = [n_x, n_y, n_z]^\top$ is computed from the inverse depth $1/z$, where horizontal components are obtained via image gradients along u and v directions:

$$n_x = f_x \frac{\partial 1/z}{\partial u}, \quad n_y = f_y \frac{\partial 1/z}{\partial v}, \tag{1}$$

with f_x and f_y denoting camera focal lengths. The vertical component is computed by aggregating local derivatives:

$$n_z = -\Phi \left\{ \frac{\Delta x_i n_x + \Delta y_i n_y}{\Delta z_i} \right\}, \tag{2}$$

where $[\Delta x_i, \Delta y_i, \Delta z_i]^\top = \boldsymbol{p}_i - \boldsymbol{q}$ and $\Phi\{\cdot\}$ denotes a central tendency measurement. This estimation ensures geometric continuity in flat road segments and suppresses unreliable gradients near object boundaries. Surface normals are further optimized using curvature-based and correlation coefficient-based discontinuity discrimination, providing the foundation for the TerrFlat representation.

IV. Methodology

A. Overview

We propose TerrFlat-Seg (Fig. 1), a physics-driven geometric–visual fusion framework for freespace detection, built upon the encoder–decoder architecture of SNE-RoadSeg [15]. Detailed architectural designs of the underlying encoder–decoder backbone can be found in [15], while this section focuses on the extensions introduced in TerrFlat-Seg. The original framework employs surface normal cues for geometric reasoning, TerrFlat-Seg introduces a physics-driven TerrFlat representation that encodes road surface geometry along three interpretable dimensions and projects it into pixel-aligned maps via differentiable operations.

To integrate these geometric priors with visual features, we replace the original fusion strategy with SFFM at the encoder stage, enabling bidirectional recalibration through channel-wise semantic enhancement and spatial-wise boundary refinement. Together, TerrFlat and SFFM form a closed-loop geometric–visual pipeline, in which explicit geometric structure guides visual perception and the fused representations reciprocally reinforce geometric reasoning.

B. TerrFlat Geometric Representation

TerrFlat encodes the 3D geometric structure of the scene along three physically interpretable dimensions: lateral smoothness, longitudinal consistency, and vertical deviation. As shown in Fig. 2, each component captures a

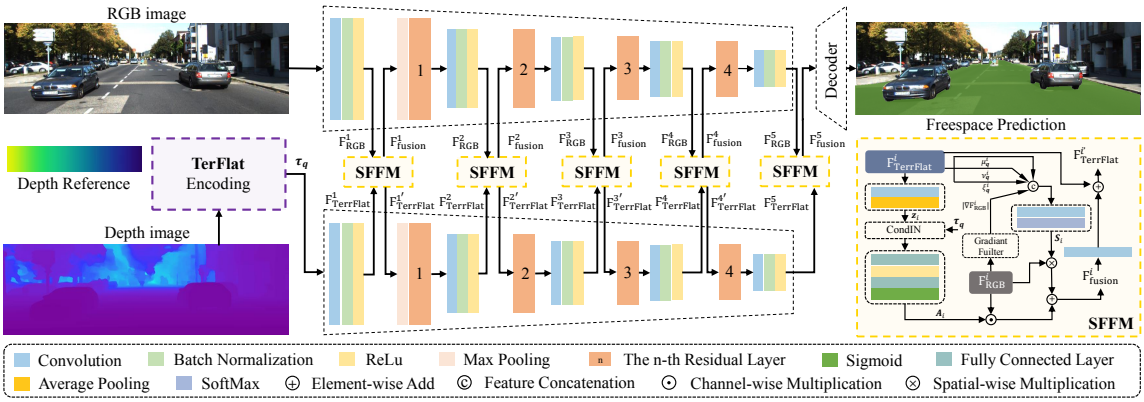

Fig. 1. Overall framework of TerrFlat-Seg, which integrates the physics-driven TerrFlat encoding and the SFFM for robust geometric–visual feature fusion in freespace detection.

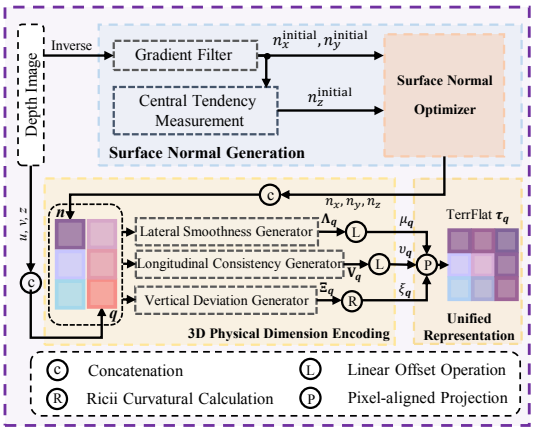

Fig. 2. Illustration of the TerrFlat encoding process. The encoding characterizes local planar geometry by projecting each point onto its normal-defined tangent plane, producing a physically meaningful representation that enhances drivable region perception.

complementary aspect of road geometry: lateral smoothness measures horizontal surface continuity via curvature tensors, longitudinal consistency models driving-direction deviations using normal torsion, and vertical deviation quantifies elevation changes through Riemannian geodesic distance and local curvature. These cues are combined into a unified vector representation $\boldsymbol{\tau_q}$, which is spatially aligned with RGB images via differentiable operations, enabling direct geometric–visual fusion. The following subsections provide detailed definitions and formulations for each component.

1) Lateral Smoothness: To characterize lateral surface smoothness, we define a curvature tensor that couples horizontal residual deviations with local second-order surface variations. The residual term

$$u_{\boldsymbol{q}} - \frac{\boldsymbol{n_q} \cdot \boldsymbol{q}}{||\boldsymbol{n_q}||^2} n_x \tag{3}$$

measures the lateral continuity of the surface, while the Hessian matrix $\nabla^2 u_{\boldsymbol{q}}$ captures curvature distribution (its trace reflects average curvature intensity). The lateral tensor is

$$\boldsymbol{\Lambda_q} = (u_{\boldsymbol{q}} - \frac{\boldsymbol{n_q} \cdot \boldsymbol{q}}{||\boldsymbol{n_q}||^2} n_x) \cdot (\nabla^2 u_{\boldsymbol{q}} - \frac{1}{2}\mathrm{tr}(\nabla^2 u_{\boldsymbol{q}}) \boldsymbol{I}_2) \cdot \boldsymbol{e}_u \boldsymbol{e}_u^\top + \gamma \cdot \det(\nabla^2 u_{\boldsymbol{q}}) \cdot \boldsymbol{I}_2, \tag{4}$$

where $\boldsymbol{I}_2$ is the 2×2 identity matrix, $\boldsymbol{e}_u$ is the lateral unit vector, and $\gamma = \exp(-\lambda ||\boldsymbol{n_q} \times \boldsymbol{g}||^2)$ enforces alignment with gravity ($\lambda > 0$, $\boldsymbol{g} = (0, 0, -1)^\top$). This tensor elevates the scalar offset to a second-order geometric characterization, capturing both local curvature and orientation relative to gravity.

2) Longitudinal Consistency:

$$v_{\boldsymbol{q}} - \frac{\boldsymbol{n_q} \cdot \boldsymbol{q}}{||\boldsymbol{n_q}||^2} n_y \tag{5}$$

while geodesic torsion

$$\chi_{\boldsymbol{g}} = -\boldsymbol{b_q} \cdot \frac{\partial \boldsymbol{n_q}}{\partial v} \tag{6}$$

captures the rotation of surface normals along the driving direction ($\boldsymbol{b_q}$ is the binormal vector from the Frenet-Serret frame). The longitudinal tensor is

$$\mathbf{V_q} = (v_{\boldsymbol{q}} - \frac{\boldsymbol{n_q} \cdot \boldsymbol{q}}{||\boldsymbol{n_q}||^2} n_y) \cdot \boldsymbol{I}_3 + \alpha \cdot (\chi_{\boldsymbol{g}} \cdot (\boldsymbol{n_q} \times \frac{\partial \boldsymbol{n_q}}{\partial v}) + \beta \cdot \nabla(\boldsymbol{n_q} \cdot \boldsymbol{g})), \tag{7}$$

where $\boldsymbol{I}_3$ is the 3×3 identity matrix, and $\beta > 0$ is an adaptive gravity-alignment coefficient controlling the sensitivity of the sigmoid function in

$$\alpha = \frac{1}{1 + \exp(-\beta(\boldsymbol{n_q} \cdot \boldsymbol{g} + 1))}. \tag{8}$$

This formulation adaptively weights longitudinal torsion contributions according to gravity alignment, thereby reducing slope misjudgment on slanted surfaces.

3) Vertical Deviation: For non-planar terrain, we introduce a Riemannian field [30] using surface parameterization $\boldsymbol{p} : U \subset \mathbb{R}^2 \to \mathbb{R}^3$. The inverse exponential map $\exp_{\boldsymbol{q}_0}^{-1}(\boldsymbol{q})$

projects a point onto the tangent space at $\boldsymbol{q}_0$, and the Riemannian norm

$$||\exp_{\boldsymbol{q}_0}^{-1}(\boldsymbol{q})||_{\boldsymbol{\varphi}} = \sqrt{\boldsymbol{\varphi}_{ij} \exp_{\boldsymbol{q}_0}^{-1}(\boldsymbol{q})^i \exp_{\boldsymbol{q}_0}^{-1}(\boldsymbol{q})^j} \tag{9}$$

measures geodesic distance under the metric tensor

$$\boldsymbol{\varphi}_{ij} = \boldsymbol{n}_{\boldsymbol{q}} \cdot \frac{\partial \boldsymbol{p}}{\partial x^i} \frac{\partial \boldsymbol{p}}{\partial x^j}, \tag{10}$$

where $\partial \boldsymbol{p}/\partial x^i$ are surface tangent vectors. The final vertical deviation vector is

$$\boldsymbol{\Xi}_{\boldsymbol{q}} = ||\exp_{\boldsymbol{q}_0}^{-1}(\boldsymbol{q})||_{\boldsymbol{\varphi}} \cdot \frac{\boldsymbol{n}_{\boldsymbol{q}}}{||\boldsymbol{n}_{\boldsymbol{q}}||} + \epsilon \cdot \text{Ricci}(\boldsymbol{q}) \cdot \boldsymbol{n}_{\boldsymbol{q}}, \tag{11}$$

where $\epsilon > 0$ is a curvature coupling coefficient and $\text{Ricci}(\boldsymbol{q})$ models global manifold distortion.

4) TerrFlat Representation: We integrate the three terrain attributes into a unified vector form $\boldsymbol{\tau}_{\boldsymbol{q}}$, which simultaneously captures lateral smoothness $\mu_{\boldsymbol{q}}$, longitudinal consistency $\nu_{\boldsymbol{q}}$, and vertical deviation $\xi_{\boldsymbol{q}}$. Each component is derived through geometric projection:

$$\begin{aligned} \boldsymbol{\tau}_{\boldsymbol{q}} &= \begin{bmatrix} \mu_{\boldsymbol{q}} \\ \nu_{\boldsymbol{q}} \\ \xi_{\boldsymbol{q}} \end{bmatrix} = \begin{bmatrix} \text{Re}(\boldsymbol{\Lambda}_{\boldsymbol{q}}) \\ \text{Re}(\mathbf{V}_{\boldsymbol{q}}) \\ ||\boldsymbol{\Xi}_{\boldsymbol{q}}||_{\boldsymbol{\varphi}} \end{bmatrix} \\ &= \begin{bmatrix} (u_{\boldsymbol{q}} - \frac{\boldsymbol{n}_{\boldsymbol{q}} \cdot \boldsymbol{q}}{||\boldsymbol{n}_{\boldsymbol{q}}||^2} n_x) \cdot \text{tr}(\boldsymbol{\Lambda}_{\boldsymbol{q}}) + \int_0^1 \nabla_s^2 n_x ds \\ (v_{\boldsymbol{q}} - \frac{\boldsymbol{n}_{\boldsymbol{q}} \cdot \boldsymbol{q}}{||\boldsymbol{n}_{\boldsymbol{q}}||^2} n_y) + \chi_{\boldsymbol{g}} \cdot ||\boldsymbol{n}_{\boldsymbol{q}} \times \frac{\partial \boldsymbol{n}_{\boldsymbol{q}}}{\partial v}|| \\ ||\exp_{\boldsymbol{q}_0}^{-1}(\boldsymbol{q})||_{\boldsymbol{\varphi}} + \epsilon \cdot \text{Ricci}(\boldsymbol{q}) \end{bmatrix}, \end{aligned} \tag{12}$$

where $\text{Re}(\cdot)$ extracts the linear offset term from the corresponding tensor field. $\mu_{\boldsymbol{q}}$ quantifies lateral smoothness by combining the horizontal offset residual with the trace of the curvature tensor $\boldsymbol{\Lambda}_{\boldsymbol{q}}$, which reflects the mean curvature along the lateral direction. The integral term $\int_0^1 \nabla_s^2 n_x ds$ describes the accumulated change in the surface normal curvature along a local arc-length s, capturing higher-order curvature variations that cannot be represented by the trace term alone. This ensures that lateral irregularities, such as road bumps or undulations, are smoothly encoded. $\nu_{\boldsymbol{q}}$ is achieved from the longitudinal tensor $\mathbf{V}_{\boldsymbol{q}}$ with the magnitude of normal rotation. It adaptively weights the torsion contributions according to gravity alignment, thus mitigating slope misjudgment on inclined or curved terrains. $\xi_{\boldsymbol{q}}$ captures vertical deviation from the vertical tensor $\boldsymbol{\Xi}_{\boldsymbol{q}}$ through a combination of the Riemannian geodesic distance $||\exp_{\boldsymbol{q}_0}^{-1}(\boldsymbol{q})||_{\boldsymbol{\varphi}}$ and the Ricci curvature term, reflecting both local elevation displacement and global manifold distortion. This vectorized representation aligns spatially with RGB images, enabling direct geometric-visual fusion.

C. Symmetric Feature Fusion Module

From the hierarchical TerrFlat features presented in Sect. IV-B.4, we extract the i-th level geometric features $\boldsymbol{F}^i_{\text{TerrFlat}} \in \mathbb{R}^{C \times H \times W}$ and the corresponding RGB features $\boldsymbol{F}^i_{\text{RGB}} \in \mathbb{R}^{C \times H \times W}$ $(i \in \{1, 2, ..., 5\})$. $\boldsymbol{F}^i_{\text{TerrFlat}}$ explicitly encodes three physically interpretable geometric cues for the i-th layer: lateral smoothness $\mu^i_{\boldsymbol{q}}$, longitudinal consistency $\nu^i_{\boldsymbol{q}}$, and vertical deviation $\xi^i_{\boldsymbol{q}}$, all spatially aligned with $\boldsymbol{F}^i_{\text{RGB}}$ to ensure effective fusion.

To overcome the limitations of traditional pixel-wise fusion, we propose a plug-and-play SFFM that employs bidirectional channel–spatial recalibration. Unlike conventional element-wise fusion, SFFM adaptively reweights features by embedding TerrFlat's geometric constraints $\boldsymbol{\tau}_{\boldsymbol{q}}$, enabling the network to prioritize structurally consistent and road-relevant cues.

1) Channel-wise Geometric Recalibration: To model global geometric importance, we first apply global average pooling to

$$\boldsymbol{z}_i = \text{GlobalAvgPool}(\boldsymbol{F}^i_{\text{TerrFlat}}) \in \mathbb{R}^{C \times 1 \times 1}. \tag{13}$$

The globally precomputed TerrFlat $\boldsymbol{\tau}_{\boldsymbol{q}}$ is used to normalize $\boldsymbol{z}_i$ through conditional instance normalization (CondIN), producing adaptive scaling $\zeta(\boldsymbol{\tau}_{\boldsymbol{q}})$ and shifting $\varrho(\boldsymbol{\tau}_{\boldsymbol{q}})$ parameters:

$$\text{CondIN}(\boldsymbol{z}_i, \boldsymbol{\tau}_{\boldsymbol{q}}) = \zeta(\boldsymbol{\tau}_{\boldsymbol{q}}) \cdot \hat{\boldsymbol{z}}_i + \varrho(\boldsymbol{\tau}_{\boldsymbol{q}}), \tag{14}$$

where $\hat{\boldsymbol{z}}_i$ denotes the instance-normalized feature. This process embeds TerrFlat-specific priors into the channel distribution, allowing subsequent layers to prioritize geometrically plausible road surfaces.

The normalized feature is then sequentially fed into a two-layer fully connected (FC) network with bottleneck ratio r, using rectified linear unit (ReLU) activation $\delta(\cdot)$ and sigmoid activation $\sigma(\cdot)$:

$$\boldsymbol{A}_i = \sigma(\boldsymbol{W}_2 \cdot \delta(\boldsymbol{W}_1 \cdot \text{CondIN}(\boldsymbol{z}_i, \boldsymbol{\tau}_{\boldsymbol{q}}))), \tag{15}$$

where the input is the concatenation of the channel feature $\boldsymbol{z}_i \in \mathbb{R}^C$ and the geometric vector $\boldsymbol{\tau}_{\boldsymbol{q}} \in \mathbb{R}^3$, giving a combined dimension of $C+3$, and the FC layers have weights $\boldsymbol{W}_1 \in \mathbb{R}^{C/r \times (C+3)}$ and $\boldsymbol{W}_2 \in \mathbb{R}^{C \times C/r}$. The first FC layer integrates geometric cues with channel features while compressing to C/r channels, and the second restores the original channel dimension. The output $\boldsymbol{A}_i \in \mathbb{R}^{C \times 1 \times 1}$ serves as a channel-wise attention map highlighting geometrically informative regions.

2) Spatial-wise TerrFlat-guided Refinement: In parallel, we design a spatial attention module that integrates local terrain cues from TerrFlat and appearance gradients to suppress inconsistent regions. For the i-th layer, three geometric components from $\boldsymbol{F}^i_{\text{TerrFlat}}$ ($\xi^i_{\boldsymbol{q}}$, $\mu^i_{\boldsymbol{q}}$, and $\nu^i_{\boldsymbol{q}}$) are concatenated with the visual gradient $|\nabla \boldsymbol{F}^i_{\text{RGB}}|$:

$$\boldsymbol{S}_i = \text{Softmax}(f_{\text{dwconv}}([\xi^i_{\boldsymbol{q}}, \mu^i_{\boldsymbol{q}}, \nu^i_{\boldsymbol{q}}, |\nabla \boldsymbol{F}^i_{\text{RGB}}|])) \in \mathbb{R}^{1 \times H \times W}, \tag{16}$$

where f_{dwconv} denotes a 3×3 depth-wise separable convolution. This map highlights regions where geometric (*e.g.*, smooth curvature) and visual (*e.g.*, consistent texture) structures align.

3) Bidirectional Fusion: SFFM performs bidirectional refinement to reinforce both modalities.

Geometric-to-Visual Refinement: The attention maps $\boldsymbol{A}_i$ and $\boldsymbol{S}_i$ are applied to recalibrate visual features:

$$\boldsymbol{F}^i_{\text{fusion}} = \boldsymbol{A}_i \odot \boldsymbol{F}^i_{\text{RGB}} + \boldsymbol{S}_i \otimes \boldsymbol{F}^i_{\text{RGB}}, \tag{17}$$

where $\odot$ and $\otimes$ denote channel-wise and spatial-wise multiplication, respectively.

Visual-to-Geometric Enhancement: To refine geometric features using visual context, we apply a learnable 1×1 convolution $f_{\text{atten}}(\cdot)$ with residual connection:

$$\boldsymbol{F}^{i'}_{\text{TerrFlat}} = \boldsymbol{F}^{i}_{\text{TerrFlat}} + f_{\text{atten}}(\boldsymbol{F}^{i}_{\text{fusion}}), \tag{18}$$

which can be naturally interpreted as explicitly solving the corresponding minimization problem

$$\min_{\boldsymbol{\theta}} ||\boldsymbol{F}^{i'}_{\text{TerrFlat}} - (\boldsymbol{F}^{i}_{\text{TerrFlat}} + f_{\text{atten}}(\boldsymbol{F}^{i}_{\text{fusion}}; \boldsymbol{\theta}))||_2^2. \tag{19}$$

where $\boldsymbol{\theta}$ denotes the learnable parameters.

4) Optimization and Training Strategy: All FC and convolutional layers are initialized with Kaiming normal initialization $W \sim \mathcal{N}(0, \sqrt{2/n_{\text{in}}})$, where n_{in} is the number of input units. The network is trained using a hybrid loss $\mathcal{L} = \mathcal{L}_{\text{CE}} + \eta\mathcal{L}_{\text{Dice}}$, where $\mathcal{L}_{\text{CE}}$ is cross-entropy loss, $\mathcal{L}_{\text{Dice}}$ is Dice loss, and η balances their contributions.

To ensure geometric consistency during optimization, gradients of the terrain descriptor $\boldsymbol{\tau_q}$ are propagated through differentiable geometric projections:

$$\frac{\partial \boldsymbol{\tau_q}}{\partial \boldsymbol{\theta}} = \frac{\partial(\boldsymbol{q} - \boldsymbol{Pq})}{\partial \boldsymbol{\theta}} = \frac{\partial \boldsymbol{q}}{\partial \boldsymbol{\theta}} - \frac{\partial(\boldsymbol{Pq})}{\partial \boldsymbol{\theta}}, \tag{20}$$

where $\boldsymbol{Pq}$ denotes the projection of $\boldsymbol{q}$ onto the reference surface. This ensures end-to-end optimization of geometric–visual fusion under physically grounded constraints.

V. Experiments

A. Datasets

We evaluate our method on three datasets: KITTI-Road, Semantic-KITTI, and ORFD. The KITTI-Road dataset [39] includes 289 RGB–depth pairs ($1,242 \times 375$). Since the official test set lacks annotations, we randomly split the training set into 70% for training and 30% for validation, and report all results on the validation subset. Semantic-KITTI [40] provides 200 annotated RGB images at the same resolution, and we follow the same 7:3 split because test labels are unavailable.

The ORFD dataset [41] contains 12,198 RGB–LiDAR pairs ($1,280 \times 720$), divided into 8,398 for training, 1,245 for validation, and 2,555 for testing. It offers pixel-level annotations for freespace, undrivable, and untraversable regions which are merged into a single non-freespace class for evaluation. Covering diverse environments (*e.g.*, woodland and farmland) and varying weather and illumination conditions (rain, fog, snow, daylight, twilight, darkness), ORFD enables reliable assessment under complex real-world scenarios.

B. Implementation Details and Evaluation Metrics

All experiments are implemented in PyTorch 1.12.1 on an NVIDIA L40 GPU. We optimize the network with stochastic gradient descent (SGD) [42] with momentum (0.9) and weight decay (5×10^{-4}), starting at a learning rate of 0.01, which is reduced by 0.1 every 50 epochs. Hyperparameters are empirically tuned based on road surface geometry.

During the TerrFlat encoding process, the gravity alignment coefficient $\lambda = 0.1$ emphasizes near-horizontal regions, the torsion sensitivity $\beta = 5.0$ controls the strength of longitudinal encoding, and the curvature coupling coefficient ϵ regulates the interaction between geodesic and Ricci-based components for vertical offset encoding. The hybrid loss function employs $\eta = 0.5$ to balance geometric and semantic supervision.

We use two standard metrics to evaluate freespace detection: F-score (Fsc) and Intersection over Union (IoU). They are defined as Precision $= n_{\text{TP}}/(n_{\text{TP}} + n_{\text{FP}})$, Recall $= n_{\text{TP}}/(n_{\text{TP}}+n_{\text{FN}})$, Fsc $= (2\times\text{Precision}\times\text{Recall})/(\text{Precision}+\text{Recall})$, IoU $= n_{\text{TP}}/(n_{\text{TP}} + n_{\text{FP}} + n_{\text{FN}})$, where n_{TP}, n_{FP}, and n_{FN} denote the numbers of true positive, false positive, and false negative pixels, respectively. Fsc and IoU are adopted as the primary evaluation metrics because they jointly balance precision and recall while directly quantifying the overlap between predicted and ground-truth freespace regions, providing a reliable measure of detection accuracy and spatial consistency.

C. Experimental Results

1) Overall Performance Evaluation: Table I presents comprehensive quantitative results across all datasets. Single-modal RGB methods are included for reference, while multi-modal approaches are grouped according to whether SFFM ablation is applicable. Highlighted values indicate the best performance within each category.

Across all datasets, TerrFlat consistently enhances freespace detection across diverse backbones by effectively encoding physically interpretable surface geometry. When combined with SFFM in TerrFlat-Seg (SNE-RoadSeg+TerrFlat+SFFM), geometric and visual cues interact more effectively, leading to smoother boundaries and higher overall accuracy. Minor performance drops in some configurations (*e.g.*, MFNet with depth, FuseNet with transformed disparity) are due to dataset bias or feature redundancy and remain within 1%, leaving the overall trend of consistent improvement unchanged.

Networks with tightly coupled fusion designs, such as RoadFormer+, cannot directly incorporate SFFM. It is worth noting that RoadFormer+ achieves higher peak performance by leveraging explicit global context modeling, at the cost of increased computational and memory complexity (Table II). In contrast, TerrFlat-Seg integrates physics-driven geometric priors and achieves competitive performance with significantly lower inference-time cost. Across backbones and datasets, the consistent performance gains obtained by incorporating TerrFlat demonstrate the general applicability of the proposed geometric representation.

Overall, the experimental results indicate that TerrFlat provides physically interpretable geometric cues, while SFFM effectively refines geometric–visual feature alignment. Their combination in TerrFlat-Seg consistently improves freespace detection across diverse architectures and scenarios.

2) Ablation Studies: We perform ablation studies to isolate the contributions of TerrFlat and SFFM using the re-

TABLE I

COMPREHENSIVE PERFORMANCE ACROSS KITTI-ROAD, SEMANTIC-KITTI, AND ORFD (ALL METRICS IN %). SINGLE-MODAL RGB METHODS ARE INCLUDED AS REFERENCE. MULTI-MODAL METHODS ARE GROUPED BY SFFM APPLICABILITY. Δ_{Fsc} AND Δ_{IoU} INDICATE PERFORMANCE GAINS FROM SFFM. BOLD VALUES INDICATE THE OVERALL BEST PERFORMANCE AMONG ALL METHODS. UNDERLINED VALUES INDICATE THE BEST PERFORMANCE AMONG METHODS WHERE SFFM IS APPLICABLE. NOTE THAT ROADFORMER+ DOES NOT INCORPORATE SFFM AND IS INCLUDED FOR COMPLETENESS.

Method	Geometric Input	KITTI-Road				Semantic-KITTI				ORFD			
		w/o SFFM		w/ SFFM		w/o SFFM		w/ SFFM		w/o SFFM		w/ SFFM	
		Fsc↑	IoU↑	Δ_{Fsc}	Δ_{IoU}	Fsc↑	IoU↑	Δ_{Fsc}	Δ_{IoU}	Fsc↑	IoU↑	Δ_{Fsc}	Δ_{IoU}
FCN [31]	-	81.2	68.4	-	-	77.2	62.9	-	-	72.4	56.8	-	-
SegNet [32]	-	88.8	79.9	-	-	67.0	50.3	-	-	87.6	77.9	-	-
U-Net [33]	-	91.8	84.8	-	-	89.6	81.1	-	-	85.1	74.0	-	-
PSPNet [6]	-	76.8	62.3	-	-	73.6	58.2	-	-	67.2	50.5	-	-
Deeplabv3+ [34]	-	94.5	89.5	-	-	94.4	89.4	-	-	87.3	77.5	-	-
HRNet [35]	-	77.7	63.5	-	-	73.2	57.7	-	-	68.5	52.1	-	-
Mask2Former [36]	-	95.7	87.8	-	-	93.4	84.1	-	-	87.9	81.0	-	-
ViT-Adapter [37]	-	94.4	89.4	-	-	94.9	90.3	-	-	91.7	88.9	-	-
ViT-CoMer [38]	-	95.5	91.4	-	-	95.6	91.5	-	-	91.9	89.1	-	-
RTFNet [22]	Depth	95.3	91.0	-	-	98.0	96.0	-	-	94.7	89.9	-	-
CMX [26]	Depth	96.8	93.9	-	-	97.3	94.7	-	-	95.0	90.4	-	-
USNet [25]	Depth	96.4	93.1	-	-	-	-	-	-	95.0	90.5	-	-
Evi-RoadSeg [27]	Surface Normal	96.6	94.2	-	-	-	-	-	-	94.5	89.5	-	-
RoadFormer [28]	Depth	97.4	94.9	-	-	98.8	97.7	-	-	87.6	78.0	-	-
	Transformed Disparity	97.3	94.8	-	-	98.4	96.9	-	-	88.1	78.7	-	-
	HHA	97.1	94.4	-	-	98.9	97.9	-	-	84.2	72.7	-	-
	Surface Normal	97.4	94.9	-	-	98.9	97.7	-	-	89.2	80.4	-	-
	TerrFlat (Ours)	**97.5**	95.1	-	-	98.9	97.8	-	-	89.3	80.7	-	-
RoadFormer+ [29]	Depth	97.4	94.9	-	-	99.2	98.5	-	-	94.9	90.2	-	-
	Transformed Disparity	97.4	95.0	-	-	99.4	**98.7**	-	-	94.3	90.9	-	-
	HHA	94.7	90.0	-	-	99.2	98.5	-	-	94.5	90.4	-	-
	Surface Normal	**97.5**	95.1	-	-	99.3	98.5	-	-	95.0	90.5	-	-
	TerrFlat (Ours)	**97.5**	**95.2**	-	-	**99.5**	**98.7**	-	-	**96.5**	**93.2**	-	-
MFNet [23]	Depth	92.7	86.3	-0.4	-0.5	98.2	96.4	+0.6	+0.3	91.9	85.1	-0.5	-0.9
	Transformed Disparity	95.9	92.1	+0.4	+0.8	98.4	96.5	+0.3	+0.0	91.5	84.3	+0.6	+1.1
	HHA	96.0	92.2	+0.2	+0.4	98.6	97.5	+0.1	+0.1	90.4	82.5	+1.0	+1.7
	Surface Normal	96.3	92.8	+0.3	+0.7	98.8	97.6	+0.1	+0.2	92.1	85.3	+0.0	+0.1
	TerrFlat (Ours)	96.6	93.3	+0.1	+0.3	98.9	97.8	+0.1	+0.3	93.7	88.2	+0.2	+0.3
FuseNet [7]	Depth	82.1	69.7	+7.3	+11.2	98.6	97.2	+0.1	+0.2	91.1	83.7	+0.3	+0.2
	Transformed Disparity	91.9	85.0	+0.7	+1.2	98.7	97.3	-0.5	-0.2	91.9	85.0	-0.4	-0.7
	HHA	86.0	75.4	+1.8	+2.9	98.8	97.6	+0.0	+0.3	91.5	84.4	+0.2	+0.3
	Surface Normal	90.9	83.3	+0.5	+0.8	98.7	97.4	+0.1	+0.1	91.1	83.7	+0.7	+1.2
	TerrFlat (Ours)	92.5	86.0	+1.7	+2.9	98.9	97.8	+0.0	+0.3	92.7	86.4	+0.4	+0.6
SNE-RoadSeg [15]	Depth	97.0	94.0	+0.2	+0.4	98.0	96.1	+0.4	+0.7	95.0	90.5	+0.2	+0.1
	Transformed Disparity	96.9	94.0	+0.2	+0.4	97.8	95.6	+0.6	+1.2	94.1	88.8	+0.4	+0.2
	HHA	96.7	93.6	+0.1	+0.2	98.4	96.9	+0.3	+0.6	93.9	88.5	+0.7	+0.7
	Surface Normal	96.9	94.0	+0.1	+0.1	98.3	96.7	+0.5	+0.9	94.6	89.8	+0.2	+0.4
	TerrFlat (Ours)	97.0	94.2	+0.2	+0.4	98.5	97.0	+0.4	+0.8	95.3	91.0	+0.5	+0.4

TABLE II

COMPUTATIONAL COMPLEXITY COMPARISONS.

Method	FLOPs (G) ↓	Activations (M) ↓
RoadFormer+	416.6	991.6
TerrFlat-Seg (Ours)	**166.9**	**351.8**

sults in Table I. In each multi-modal network, conventional geometric inputs (depth, transformed disparity, HHA, and surface normals) are replaced with TerrFlat to assess its independent effect. Across all backbones and datasets, TerrFlat consistently improves both Fsc and IoU, indicating that it provides richer and more stable geometric cues. On the challenging ORFD dataset with RoadFormer+, TerrFlat raises Fsc by 1.5% and IoU by 2.7% over surface normals, showing its effectiveness even in strong architectures. On KITTI-Road with FuseNet, TerrFlat increases IoU by 11.2% compared with depth-based encoding, confirming its general applicability across networks and datasets.

To evaluate semantic–geometric fusion, SFFM is applied to networks supporting explicit geometric input recalibration, including MFNet, FuseNet, and SNE-RoadSeg. As shown in the Δ_{Fsc} and Δ_{IoU} columns of Table I, SFFM consistently improves performance across datasets. Minor drops, such as in MFNet with depth input, arise from feature redundancy or dataset bias but do not affect the overall trend.

3) Qualitative Results: Representative qualitative results are shown in Fig. 3. All visualizations are obtained using the TerrFlat-Seg framework to ensure consistent fusion across geometric inputs. Compared with depth, transformed disparity, HHA, and surface normals, TerrFlat generates smoother and more coherent freespace boundaries. In both structured urban and unstructured off-road scenes, it maintains long-range stability and captures subtle elevation variations often

Fig. 3. Examples of TerrFlat-Seg results on the KITTI-Road and ORFD dataset with respect to different geometric inputs: (i) Ground-truth; (ii) Depth; (iii) Transformed disparity; (iv) HHA; (v) Surface normal; (vi) TerrFlat (Ours). (a)–(f) illustrate various challenging scenarios: (a) a road scene with sloped surroundings; (b–c) distant vehicles under complex illumination; (d) an unstructured muddy off-road terrain; (e) a snowy environment with strong reflectance; (f) a suburban road with irregular boundaries.

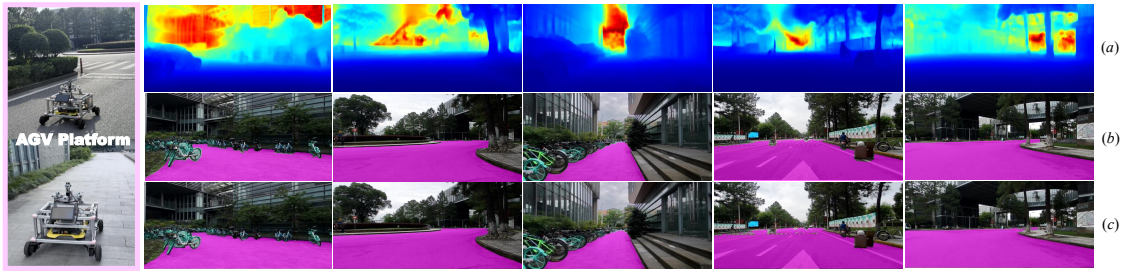

Fig. 4. Representative visual results of TerrFlat-Seg on the AGV platform: (a) depth maps; (b) manual annotations; (c) predicted freespace areas.

Fig. 5. Quantitative comparison of different geometric inputs on 50 manually annotated real-world frames: (a) Depth; (b) Transformed disparity; (c) HHA; (d) Surface normal; (e) TerrFlat (Ours).

missed by other representations.

Although the quantitative gains of TerrFlat over surface normal inputs within the same TerrFlat-Seg framework (SNE-RoadSeg+TerrFlat+SFFM) are modest (0.7% Fsc and 0.2% IoU on ORFD), the qualitative comparison in Fig. 3(*v*)–(*vi*) clearly shows its advantage. TerrFlat delivers more stable and visually consistent segmentation, effectively suppressing spurious detections from surface irregularities or illumination changes. These results confirm that TerrFlat provides a more robust and generalizable geometric representation even within identical fusion settings.

D. Real-World Deployment

We evaluate TerrFlat-Seg in real-world conditions using a self-developed automated guided vehicle (AGV) equipped with a monocular camera (IM244, $1,920 \times 1,080$, 120° FOV) and a Livox MID-360 LiDAR. The LiDAR provides dense point clouds projected onto the image plane to generate depth maps, ensuring accurate geometric alignment between visual and spatial domains. The network, trained on KITTI-Road (Sects. V-A and V-B), is deployed without fine-tuning, enabling zero-shot evaluation of cross-domain generalization and real-world robustness.

For quantitative assessment, 50 frames are manually annotated using the open-source LabelMe [43] and evaluated with Fsc and IoU. As shown in Fig. 4, TerrFlat-Seg achieves 88.7% Fsc and 81.3% IoU, outperforming other geometric inputs, demonstrating strong geometric consistency and robustness to real-world noise. Representative qualitative comparisons on several scenes are presented in Fig. 5, including depth maps, manual annotations, and TerrFlat-Seg predictions, further confirming the framework's stability and deployment potential in outdoor environments.

VI. CONCLUSION

We propose TerrFlat, a physics-driven geometric representation encoding road geometry along three interpretable dimensions, enabling structured freespace reasoning. TerrFlat-Seg fuses these geometric cues with RGB semantics via SFFM, improving interpretability, robustness, and cross-domain generalization. Extensive experiments on structured and unstructured terrains, including real-world AGV tests, validate its effectiveness. This work establishes a physics-aware perception paradigm beyond conventional data-driven heuristics. Nevertheless, reliance on accurate depth maps makes the approach sensitive to noise or incompleteness. Future work will focus on developing robust geometric estimation techniques and exploring adaptive multi-modal fusion strategies to further enhance reliability under dynamic or low-texture conditions.

REFERENCES

[1] A. Gupta *et al.*, "Embodied intelligence via learning and evolution," *Nat. Commun.*, vol. 12, no. 1, p. 5721, 2021.

[2] E. Milli *et al.*, "Multi-modal multi-task (3MT) road segmentation," *IEEE Robot. Autom. Lett.*, vol. 8, no. 9, pp. 5408–5415, 2023.

[3] H. Wang *et al.*, "Self-supervised drivable area and road anomaly segmentation using rgb-d data for robotic wheelchairs," *IEEE Robot. Autom. Lett.*, vol. 4, no. 4, pp. 4386–4393, 2019.

[4] F. Kong *et al.*, "Trajectory optimization for drone logistics delivery via attention-based pointer network," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 4, pp. 4519–4531, 2022.

[5] L. Annamalai *et al.*, "EventASEG: An event-based asynchronous segmentation of road with likelihood attention," *IEEE Robot. Autom. Lett.*, 2024.

[6] H. Zhao *et al.*, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 2881–2890.

[7] C. Hazirbas *et al.*, "Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture," in *Proc. Asian Conf. Comput. Vis. (ACCV)*. Springer, 2017, pp. 213–228.

[8] J. Yang *et al.*, "Semantic segmentation for autonomous driving," in *Autonomous Driving Perception: Fundamentals and Applications*. Springer, 2023, pp. 101–137.

[9] J.-H. Hwang *et al.*, "How to relieve distribution shifts in semantic segmentation for off-road environments," *IEEE Robot. Autom. Lett.*, 2025, DOI: 10.1109/LRA.2025.3551536.

[10] Y. Li *et al.*, "Lidar for autonomous driving: The principles, challenges, and trends for automotive lidar and perception systems," *IEEE Signal Process. Mag.*, vol. 37, no. 4, pp. 50–61, 2020.

[11] S. Sun *et al.*, "MIMO radar for advanced driver-assistance systems and autonomous driving: Advantages and challenges," *IEEE Signal Process. Mag.*, vol. 37, no. 4, pp. 98–117, 2020.

[12] X. Dai *et al.*, "TIRNet: Object detection in thermal infrared images for autonomous driving," *Appl. Intell.*, vol. 51, no. 3, pp. 1244–1261, 2021.

[13] G. Hua *et al.*, ""Where Does the Devil Lie?": Multimodal multitask collaborative revision network for trusted road segmentation," *IEEE Trans. Multimedia*, vol. 26, pp. 10 306–10 317, 2024.

[14] R. Strudel *et al.*, "Segmenter: Transformer for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2021, pp. 7262–7272.

[15] R. Fan *et al.*, "SNE-Roadseg: Incorporating surface normal information into semantic segmentation for accurate freespace detection," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Springer, 2020, pp. 340–356.

[16] J. Yang *et al.*, "Three-Filters-to-Normal+: Revisiting discontinuity discrimination in depth-to-normal translation," *IEEE Trans. Autom. Sci. Eng.*, vol. 22, pp. 895–904, 2024.

[17] Z. Xue *et al.*, "Learning to simulate complex scenes for street scene segmentation," *IEEE Trans. Multimedia*, vol. 24, pp. 1253–1265, 2021.

[18] W. Zhang *et al.*, "A multi-resolution fusion model incorporating color and elevation for semantic segmentation," *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, vol. 42, pp. 513–517, 2017.

[19] S. Gupta *et al.*, "Learning rich features from rgb-d images for object detection and segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Springer, 2014, pp. 345–360.

[20] R. Fan *et al.*, "Pothole detection based on disparity transformation and road surface modeling," *IEEE Trans. Image Process.*, vol. 29, pp. 897–908, 2019.

[21] S. Gupta *et al.*, "Indoor scene understanding with rgb-d images: Bottom-up segmentation, object detection and semantic segmentation," *Int. J. Comput. Vis.*, vol. 112, pp. 133–149, 2015.

[22] Y. Sun *et al.*, "Rtfnet: Rgb-thermal fusion network for semantic segmentation of urban scenes," *IEEE Robot. Autom. Lett.*, vol. 4, no. 3, pp. 2576–2583, 2019.

[23] Q. Ha *et al.*, "MFNet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*. IEEE, 2017, pp. 5108–5115.

[24] Y. Feng *et al.*, "SNE-RoadSegV2: Advancing heterogeneous feature fusion and fallibility awareness for freespace detection," *IEEE Trans. Instrum. Meas.*, 2025, DOI: 10.1109/TIM.2025.3545498.

[25] Y. Chang *et al.*, "Fast road segmentation via uncertainty-aware symmetric network," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*. IEEE, 2022, pp. 11 124–11 130.

[26] J. Zhang *et al.*, "CMX: Cross-modal fusion for RGB-X semantic segmentation with transformers," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 12, pp. 14 679–14 694, 2023.

[27] F. Xue *et al.*, "Evidence-based real-time road segmentation with RGB-D data augmentation," *IEEE Trans. Intell. Transp. Syst.*, 2025, DOI: 10.1109/TITS.2024.3509140.

[28] J. Li *et al.*, "RoadFormer: Duplex transformer for RGB-normal semantic road scene parsing," *IEEE Trans. Intell. Veh.*, vol. 9, no. 7, pp. 5163–5172, 2024.

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

[30] S. Donaldson, *Riemann surfaces*. Oxford University Press, 2011.

[31] J. Long *et al.*, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2015, pp. 3431–3440.

[32] V. Badrinarayanan *et al.*, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, 2017.

[33] O. Ronneberger *et al.*, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Med. Image Comput. Comput.-Assisted Interv. (MICCAI)*. Springer, 2015, pp. 234–241.

[34] L.-C. Chen *et al.*, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 801–818.

[35] J. Wang *et al.*, "Deep high-resolution representation learning for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3349–3364, 2020.

[36] B. Cheng *et al.*, "Masked-attention mask transformer for universal image segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 1290–1299.

[37] Z. Chen *et al.*, "Vision transformer adapter for dense predictions," *arXiv preprint arXiv:2205.08534*, 2022, Available: https://arxiv.org/abs/2205.08534.

[38] C. Xia *et al.*, "Vit-CoMer: Vision transformer with convolutional multi-scale feature interaction for dense predictions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2024, pp. 5493–5502.

[39] J. Fritsch *et al.*, "A new performance measure and evaluation benchmark for road detection algorithms," in *Proc. IEEE Intell. Transp. Syst. Conf. (ITSC)*, 2013.

[40] H. Alhaija *et al.*, "Augmented reality meets computer vision: Efficient data generation for urban driving scenes," *Int. J. Comput. Vis.*, vol. 126, no. 9, pp. 961–972, 2018.

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

[42] Y. Liu *et al.*, "An improved analysis of stochastic gradient descent with momentum," *Adv. Neural Inf. Process. Syst.*, vol. 33, pp. 18 261–18 271, 2020.

[43] B. C. Russell *et al.*, "Labelme: a database and web-based tool for image annotation," *Int. J. Comput. Vis.*, vol. 77, no. 1, pp. 157–173, 2008.