Fully Exploiting Vision Foundation Model's Profound Prior Knowledge for Generalizable RGB-Depth Driving Scene Parsing

Sicen Guo, *Graduate Student Member, IEEE*, Tianyou Wen, Chuang-Wei Liu, Qijun Chen, *Senior Member, IEEE*, and Rui Fan, *Senior Member, IEEE*

***Abstract*—Recent vision foundation models (VFMs), typically based on Vision Transformer (ViT), have significantly advanced numerous computer vision tasks. Despite their success in tasks focused solely on RGB images, the potential of VFMs in RGB-depth driving scene parsing remains largely under-explored. In this article, we take one step toward this emerging research area by investigating a feasible technique to fully exploit VFMs for generalizable RGB-depth driving scene parsing. Specifically, we explore the inherent characteristics of RGB and depth data, thereby presenting a Heterogeneous Feature Integration Transformer (HFIT). This network enables the efficient extraction and integration of comprehensive heterogeneous features without re-training ViTs. Relative depth prediction results from VFMs, used as inputs to the HFIT side adapter, overcome the limitations of the dependence on depth maps. Our proposed HFIT demonstrates superior performance compared to all other traditional single-modal and data-fusion scene parsing networks, pre-trained VFMs, and ViT adapters on the Cityscapes and KITTI Semantics datasets. We believe this novel strategy paves the way for future innovations in VFM-based data-fusion techniques for driving scene parsing.**

***Index Terms*—Vision foundation models (VFMs), transformer, computer vision, scene parsing, data-fusion.**

I. Introduction

THE rise of vision foundation models (VFMs) [1], [2], [3], [4] has heralded a new era in computer vision. Traditional fundamental computer vision tasks, such as semantic segmentation [1], object detection [5], and depth estimation [3], [6], have all started turning to VFMs for more compelling performance. This feasibility generally arises from the VFM's stronger capability in extracting profound, informative features through a range of related computer vision tasks [1], [2], [3]. Despite several attempts made in RGB-depth (often abbreviated as RGB-D) driving scene parsing [7], [8], [9], existing prior arts based on VFMs are sketchy, and the achieved results remain unsatisfactory. Therefore, this article primarily focuses on fully exploiting VFMs' profound prior knowledge for RGB-D driving scene parsing.

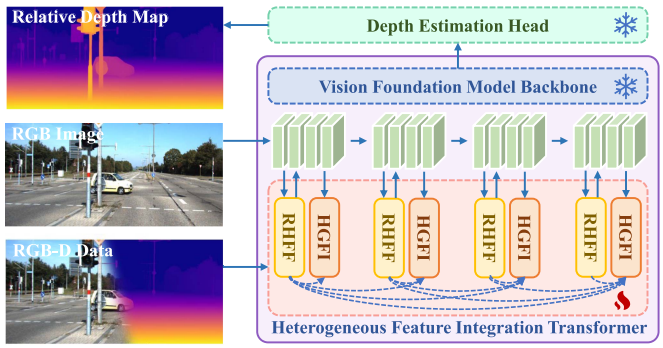

Fig. 1. An overview of our proposed **HFIT**. Through the interaction between recalibrated heterogeneous feature fusion (RHFF) modules and holistic gated feature integration (HGFI) modules, heterogeneous features are fully integrated with the profound prior features.

The integration of RGB images and depth maps has been proven to significantly improve the performance of driving scene parsing [10], [11], [12], [13]. However, a significant limitation of these fusion models is their reliance on depth maps, which restricts their applicability in scenarios where LiDARs are unavailable. Stereo cameras provide a practical alternative for obtaining depth information [14], [15]. However, even supervised methods fall short of achieving LiDAR-level accuracy, and the performance gap is even wider for unsupervised and self-supervised approaches [16]. Furthermore, integrating these heterogeneous characteristics can degrade scene parsing performance when the depth maps lack sufficient precision, e.g., due to inaccuracies in camera-LiDAR calibration [17], [18], [19]. Surprisingly, VFMs like Depth Anything V2 [20] have demonstrated exceptional effectiveness not only in extracting informative deep features but also in learning geometric information and performing relative depth estimation. Since these methods perceive only relative spatial relationships and cannot directly derive exact depth values, many researchers assume that

Received 8 November 2024; revised 14 December 2024; accepted 21 December 2025. Date of publication 5 January 2026; date of current version 26 February 2026. This work was supported in part by the National Natural Science Foundation of China under Grant 62473288 and Grant 62233013, in part by the Fundamental Research Funds for the Central Universities, and in part by the Xiaomi Young Talents Program. *(Corresponding author: Rui Fan.)*

The authors are with the College of Electronic and Information Engineering, Shanghai Institute of Intelligent Science and Technology, Shanghai Research Institute for Intelligent Autonomous Systems, Shanghai Key Laboratory of Intelligent Autonomous Systems, State Key Laboratory of Autonomous Intelligent Unmanned Systems, and Frontiers Science Center for Intelligent Autonomous Systems (Ministry of Education), Tongji University, Shanghai 201804, China (e-mail: guosicen@tongji.edu.cn; tmi@tongji.edu.cn; cwliu@tongji.edu.cn; qjchen@tongji.edu.cn; rui.fan@ieee.org).

Our source code is publicly available at https://mias.group/HFIT.

This article has supplementary downloadable material available at https://doi.org/10.1109/TIV.2025.3650682, provided by the authors.

Digital Object Identifier 10.1109/TIV.2025.3650682

their impact on scene parsing is minimal [21], [22]. However, we contend that relative depth maps are sufficient to enhance the performance of scene parsing. Utilizing VFMs' relative depth estimation output to assist data-fusion models deserves greater attention. Thus, this article marks the initial exploration in this uncharted territory, specifically focusing on the utilization of not only the VFMs' deep features but also their relative depth outputs.

The main challenges for data-fusion models to have spatial understanding ability are in the following aspects. First, VFMs have limited capacity to understand depth maps as they are trained only on RGB images [2], [3], [20]. Consequently, directly inputting depth maps into VFMs results in poor performance. Most existing data-fusion models typically use dual encoders with separate trainable parameters to extract heterogeneous features from the RGB-D pairs [7], [10], [11]. Nevertheless, this strategy not only increases computational load but also overlooks the inherent characteristics between heterogeneous features. Therefore, developing an architecture that can leverage VFMs' informative prior knowledge while enabling the extraction and integration of spatial features with minimum resource utilization, stands as an under-explored area requiring further investigation. Additionally, RGB images are abundant in global semantic cues [23], making them suitable for feature extraction via pre-trained Vision Transformers (ViTs), while depth images provide gradient-related semantics (local details) [24], which can be captured more efficiently through convolutional neural networks (CNNs). Thus, side adapters may be the optimal architecture choice. Freezing pre-trained ViTs and solely updating adapters not only leverages the profound prior knowledge of ViTs but also enables the model to extract spatial features with minimal resource usage [25], [26]. Although it may not achieve the performance of full fine-tuning strategies, its significantly lower number of trainable parameters and reduced resource consumption provide compelling advantages over full fine-tuning.

Additionally, strategies for fusing prior features from ViTs with heterogeneous features from the adapter are crucial for generalizable RGB-D driving scene parsing [27], [28]. Indiscriminate fusion strategies, such as direct overlaying interactions between modalities [12], not only introduce noise into the feature space but also exhibit the inadaptability of RGB-based models to mixed-modality inputs [29], [30]. Since traditional adapter inputs are typically RGB images, indiscriminate feature fusion [25], [31] is relatively harmless. However, if the adapter extracts heterogeneous information, indiscriminate fusion strategies will potentially drown out valid complementary information due to different semantic spaces [32], [33]. Therefore, it becomes essential to explore the inherent differences in heterogeneous feature characteristics, emphasizing their respective strengths and complementing each other based on reliability.

In addition to integrating heterogeneous features, the fusion of multi-level features with varying degrees of semantics has often been overlooked [11], [12], [23], [27]. As the network depth increases, semantic clues become richer, while fine-grained information, such as boundary details, tends to be lost [34]. SNE-RoadSeg [10] and UNet++ [35] employ densely connected skip connections to enhance comprehensive holistic feature integration. While this approach incorporates both fine-grained and coarse-grained details, it fails to address the semantic gap between features extracted at different network depths. These limitations underscore the importance of assessing the significance of each feature vector within the feature map and aggregating the information accordingly.

To tackle the limitations discussed above, we introduce **H**eterogeneous **F**eature **I**ntegration **T**ransformer (**HFIT**) (see Fig. 1). HFIT investigates the utilization of VFMs for generalizable RGB-D driving scene parsing, thereby unleashing the power of profound prior knowledge embedded within these pre-trained VFMs. We begin by performing relative depth estimation [20]. Then, we develop a side adapter to capture robust and multi-scale spatial pyramid features from RGB-D pairs and adaptively integrate them with the prior features from VFMs. HFIT adapter starts with a duplex spatial prior extractor (DSPE), which aims to capture local spatial priors from the RGB-D data. Then, we develop a recalibrated heterogeneous feature fusion module, which supplements VFMs with multi-scale spatial geometric features. Finally, we design a holistic gated feature integration module, which first selectively integrates features from multiple levels and then fuses multi-scale features from VFMs and the adapter. Extensive experiments have been conducted on publicly available datasets [36], [37], and unequivocally demonstrate the superior performance of our proposed HFIT. Our contributions are as follows:

- To the best of our knowledge, this study is the first to investigate the adaptation of VFMs for RGB-D driving scene parsing. HFIT can directly extract and integrate heterogeneous features without re-training ViTs. By leveraging relative depth prediction results from VFMs as inputs to the HFIT adapter, our approach overcomes the limitations of the dependence on depth maps.
- We introduce a DSPE strategy to jointly capture local semantics from RGB-D data, enabling HFIT to effectively learn heterogeneous features.
- We design an RHFF module to collaboratively recalibrate spatial priors, enhancing the comprehensiveness of feature fusion.
- We propose an HGFI strategy to combine complementary strengths of multi-level features, equipping HFIT with enhanced fine-grained feature representation capabilities.

II. Related Work

A. Conventional RGB-D Semantic Segmentation Networks

Considering the feature fusion stage of state-of-the-art (SoTA) RGB-D semantic segmentation networks, we can categorize them into three main types: early fusion, intermediate fusion, and late fusion [38]. Early fusion methods typically combine RGB and depth images at the input stage. While this approach is straightforward, it falls short in achieving a comprehensive understanding of the environment [38]. Intermediate fusion techniques [10] extract and fuse these heterogeneous features from RGB-D pairs using duplex encoders. Late fusion approaches [12], [39], [40] utilize two parallel encoders, with a primary focus on feature fusion within the decoder. Although

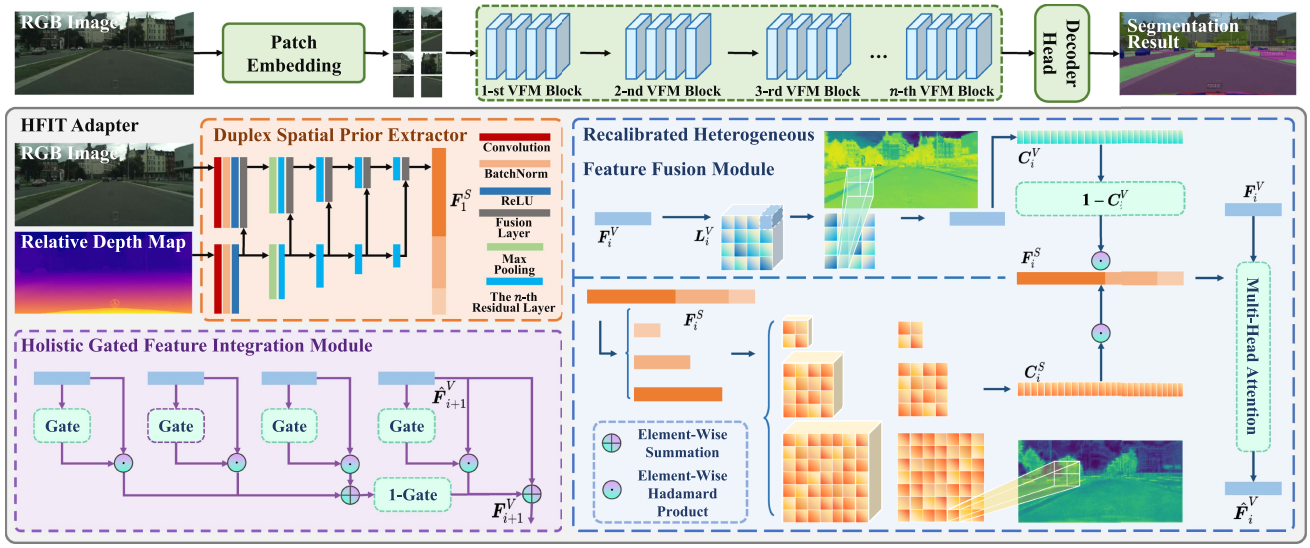

Fig. 2. An illustration of our proposed **HFIT**, consisting of (1) a plain ViT, (2) a duplex spatial prior extractor, (3) recalibrated heterogeneous feature fusion modules, and (4) holistic gated feature integration modules.

these strategies have proven to be more effective than single-modal methods, they still face challenges such as insufficient utilization of depth features and high computational costs associated with multiple feature extractors [30].

B. Feature Extraction and Fusion Strategies

Feature extraction and fusion play essential roles in advancing various computer vision tasks [41], [42], [43], [44]. Traditional fusion strategies perform effectively when features exhibit relevance and consistency. However, when dealing with multi-scale, multi-level, and multi-modality features, indiscriminate extraction and fusion can introduce redundancies and noise due to the semantic gap [17]. To address these limitations, the separation-and-aggregation gate (SAGate) [45] recalibrates and aggregates heterogeneous features to generate enhanced selective representations specifically for segmentation. Additionally, several studies [10], [35] have addressed this issue by introducing densely connected skip connections to integrate fine-grained and coarse-grained information. The cross feature module (CFM) [46] incorporates adaptive aggregation mechanisms to selectively retain useful feature maps and refine multi-level features while filtering out background noise. Nevertheless, these models often overlook the semantic gap between hierarchical features extracted at different network depths, which may result in incomplete fusion of high-level and low-level details. Hence, our primary focus in this article is on addressing this critical limitation.

C. Vision Foundation Models

Vision foundation models have emerged as a groundbreaking approach in artificial intelligence, demonstrating superior learning capabilities compared to traditional models. The segment anything model (SAM) [1] is a recently proposed VFM for image segmentation. Trained on a massive dataset consisting of over 1 billion masks from 11 million images, SAM is built to generalize across diverse segmentation tasks. By performing discriminative self-supervised learning at image and patch levels, DINOv2 [2] learns all-purpose visual features for various downstream tasks. Depth Anything V1 [3] inherits profound semantic priors from the pre-trained DINOv2 backbone via a simple feature alignment constraint, achieving superior performance in downstream tasks. Depth Anything V2 [20] produces finer and more robust depth predictions than the first version by mitigating the distributional shift and limited diversity of synthetic data. Larger models naturally excel at integrating vision features, making them well-suited for our RGB-D driving scene parsing tasks. However, they also come with increased memory requirements and higher training costs. To strike a balance between performance, resource efficiency, and experimental versatility, we select Large DINOv2, Large Depth Anything V1, and Large Depth Anything V2 as the backbones for our proposed HFIT.

III. Methodology

A. Overall Workflow

As shown in Fig. 2, our HFIT comprises two parts: the first is a plain ViT, which consists of a patch embedding layer followed by L encoder layers, and the second is our proposed HFIT adapter, which contains (1) a DSPE to capture local spatial features from the input RGB-D pairs, (2) RHFF modules to inject recalibrated spatial priors into the ViT, and (3) HGFI modules to integrate holistic hierarchical features and to reconstruct multi-level and multi-scale features.

First, in the ViT branch, the RGB image $\boldsymbol{I}^R \in \mathbb{R}^{B\times 3\times H\times W}$ is fed into the patch embedding layer to extract features with a resolution reduced to 1/16 of the original image. These patches are then flattened and projected into D-dimensional tokens, which, after being added with the position embedding, pass through L encoder layers. Simultaneously, in the adapter branch, the RGB image and its corresponding depth map $\boldsymbol{I}^D \in$

$\mathbb{R}^{B\times 3\times H\times W}$ pass through the DSPE to generate a feature pyramid $\{\boldsymbol{F}^S_{1,\frac{1}{8}}, \boldsymbol{F}^S_{1,\frac{1}{16}}, \boldsymbol{F}^S_{1,\frac{1}{32}}\}$. These pyramid features are then flattened and concatenated into $\boldsymbol{F}^S_1 \in \mathbb{R}^{(\sum_{t=2}^{4}\frac{HW}{S_t^2})\times D}$ as input for subsequent feature interaction, where $S_t = 2^{t+1}(t \in [2,4]\cap\mathbb{Z})$ denotes the corresponding stride number. Specifically, given the number of interactions N (usually $N = 4$), we evenly split the Transformer encoders into N blocks, each containing L/N encoder layers. At the i-th stage, the spatial priors $\boldsymbol{F}^S_i$ are first injected into the Transformer features $\boldsymbol{F}^V_i \in \mathbb{R}^{\frac{HW}{16^2}\times D}$ via the RHFF module, which adaptively emphasizes and diminishes complementary information to achieve comprehensive fusion. Then, the HGFI module combines multi-level features and reorganizes multi-scale features from the ViT's single-scale features at the end of each stage. After N bidirectional feature interactions, the heterogeneous features from two branches are aggregated and passed to the back-end decoders for dense prediction tasks.

B. Duplex Spatial Prior Extractor

Recent studies have shown that convolutions can help Transformers better capture local spatial semantics, underscoring their significance in computer vision tasks that necessitate clear object boundaries [47]. Building on this idea, we propose the DSPE module, which is designed to capture local spatial contexts from RGB-D pairs in conjunction with the patch embedding layer. As depicted in Fig. 2, the DSPE employs dual branches to extract spatial priors from RGB images and relative depth maps. By incorporating RGB images as complementary inputs, the DSPE extracts more comprehensive local spatial patterns than using relative depth maps alone. EfficientNet [48], known for its superior performance in extracting detailed visual features, outperforms other hierarchical CNN models such as the ResNet [49] series. Hence, we construct the DSPE using two identical EfficientNet blocks, enabling the efficient extraction of spatial priors from RGB-D pairs.

$\boldsymbol{I}^R \in \mathbb{R}^{B\times 3\times H\times W}$ and $\boldsymbol{I}^D \in \mathbb{R}^{B\times 3\times H\times W}$ are independently fed into duplex EfficientNet [48] blocks, generating multi-scale feature pyramids $\mathcal{F}^R_1 = \{\boldsymbol{F}^R_{1,\frac{1}{8}}, \boldsymbol{F}^R_{1,\frac{1}{16}}, \boldsymbol{F}^R_{1,\frac{1}{32}}\}$ and $\mathcal{F}^D_1 = \{\boldsymbol{F}^D_{1,\frac{1}{8}}, \boldsymbol{F}^D_{1,\frac{1}{16}}, \boldsymbol{F}^D_{1,\frac{1}{32}}\}$, where $\boldsymbol{F}^{R,D}_{1,\frac{1}{2^l}} \in \mathbb{R}^{\frac{H}{S_l}\times\frac{W}{S_l}\times C_l}$ represents the features at the l-th stage, C_l and $S_l = 2^{l+1}(l \in [2,4]\cap\mathbb{Z})$ denote the corresponding channel and stride numbers, respectively. Subsequently, $\boldsymbol{F}^R_{1,\frac{1}{2^l}}$ and $\boldsymbol{F}^D_{1,\frac{1}{2^l}}$ are merged via element-wise summation, followed by 1×1 convolution layers to reduce the output to D channels, yielding heterogeneous features $\mathcal{F}^S_1 = \{\boldsymbol{F}^S_{1,\frac{1}{8}}, \boldsymbol{F}^S_{1,\frac{1}{16}}, \boldsymbol{F}^S_{1,\frac{1}{32}}\}$, where $\boldsymbol{F}^S_{1,\frac{1}{2^t}} \in \mathbb{R}^{\frac{H}{S_t}\times\frac{W}{S_t}\times D}$. These three-scale features are then flattened and concatenated to form the heterogeneous spatial prior $\boldsymbol{F}^S_1 \in \mathbb{R}^{(\sum_{t=2}^{4}\frac{HW}{S_t^2})\times D}$, which is utilized in subsequent heterogeneous feature fusion stages.

C. Recalibrated Heterogeneous Feature Fusion Module

The primary challenge in feature extraction for RGB-D scene parsing lies in the efficient fusion and integration of diverse heterogeneous features obtained from multiple data sources [24], [50], [51]. As discussed in Section II, previous studies [12], [27], [28] combined these heterogeneous features without adequately addressing their inherent differences. To tackle these limitations, RHFF selectively highlights key regions while diminishing the focus on less significant areas. Original heterogeneous spatial priors are adaptively recalibrated through dynamic weighting and seamlessly integrated into the ViT via a multi-scale deformable attention mechanism.

Before injecting heterogeneous features, it is crucial to distinguish and quantify the reliability of each feature. Convolutional layers are initially employed to generate a logit map $\boldsymbol{L}^V_i \in \mathbb{R}^{\frac{H}{16}\times\frac{W}{16}\times C}$ for Transformer features through the following process:

$$\boldsymbol{L}^V_i = \mathrm{ReLU}\left(\mathrm{BN}\left(\underset{1\times 1}{\mathrm{Conv}}\left(\mathrm{Reshape}(\boldsymbol{F}^V_i)\right)\right)\right), \tag{1}$$

where "$\mathrm{BN}(\cdot)$" denotes BatchNorm operation, "$\mathrm{Reshape}$" operation reshapes the two-dimensional matrix $\boldsymbol{F}^V_i \in \mathbb{R}^{\frac{HW}{16^2}\times D}$ into a three-dimensional tensor $\boldsymbol{F}^V_i \in \mathbb{R}^{\frac{H}{16}\times\frac{W}{16}\times D}$, and C denotes the segmentation classes. After modeling interdependencies across channel dimensions, the reliabilities are contrasted to compute the confidence map $\boldsymbol{C}^V_{i,\frac{1}{16}} \in \mathbb{R}^{\frac{H}{16}\times\frac{W}{16}\times 1}$ using the following expression:

$$\boldsymbol{C}^V_{i,\frac{1}{16}} = \sigma\left(\underset{k\times k,r}{\mathrm{Conv}}\left(-\boldsymbol{L}^V_i \odot \log \boldsymbol{L}^V_i\right)\right), \tag{2}$$

where $\odot$ represents the element-wise Hadamard product between two metrics, $\sigma(\cdot)$ denotes the Sigmoid operation, and "$\underset{k\times k,r}{\mathrm{Conv}}$" denotes atrous convolutional layers with kernel size k and dilation rate r. Each element in $\boldsymbol{C}^V_i$ indicates the confidence level of a pixel relative to its segmentation class. A higher confidence level suggests the pixel significantly contributes to accurate segmentation, while a lower confidence level implies redundancy, which should be either ignored or supplemented with relevant information. To align the feature pyramid $\boldsymbol{F}^S_i \in \mathbb{R}^{(\sum_{t=2}^{4}\frac{HW}{S_t^2})\times D}$ across three resolutions, we downsample and upsample $\boldsymbol{C}^V_{i,\frac{1}{16}}$ to $\boldsymbol{C}^V_{i,\frac{1}{8}} \in \mathbb{R}^{\frac{H}{8}\times\frac{W}{8}\times 1}$ as well as $\boldsymbol{C}^V_{i,\frac{1}{32}} \in \mathbb{R}^{\frac{H}{32}\times\frac{W}{32}\times 1}$. These features are then concatenated to yield the confidence map $\boldsymbol{C}^V_i$ as follows:

$$\boldsymbol{C}^V_i = \left[\boldsymbol{C}^V_{i,\frac{1}{8}}; \boldsymbol{C}^V_{i,\frac{1}{16}}; \boldsymbol{C}^V_{i,\frac{1}{32}}\right] \in \mathbb{R}^{\left(\sum_{t=2}^{4}\frac{HW}{S_t^2}\right)\times 1}. \tag{3}$$

In the adapter branch, the spatial prior $\boldsymbol{F}^S_i \in \mathbb{R}^{(\sum_{t=2}^{4}\frac{HW}{S_t^2})\times D}$ is first divided into three-scale feature maps $\boldsymbol{F}^S_{i,\frac{1}{2^t}} \in \mathbb{R}^{\frac{H}{S_t}\times\frac{W}{S_t}\times D}$. Then, we calculate the confidence map of the heterogeneous spatial prior $\boldsymbol{C}^S_i$ as follows:

$$\boldsymbol{L}^S_{i,\frac{1}{2^t}} = \mathrm{ReLU}\left(\mathrm{BN}\left(\underset{1\times 1}{\mathrm{Conv}}\left(\mathrm{Reshape}(\boldsymbol{F}^S_{i,\frac{1}{2^t}})\right)\right)\right), \tag{4}$$

$$\boldsymbol{C}^S_{i,\frac{1}{2^t}} = \mathrm{Reshape}'\left(\sigma\left(\underset{k\times k,r}{\mathrm{Conv}}\left(-\boldsymbol{L}^S_{i,\frac{1}{2^t}} \odot \log \boldsymbol{L}^S_{i,\frac{1}{2^t}}\right)\right)\right), \tag{5}$$

$$\boldsymbol{C}^S_i = \left[\boldsymbol{C}^S_{i,\frac{1}{8}}; \boldsymbol{C}^S_{i,\frac{1}{16}}; \boldsymbol{C}^S_{i,\frac{1}{32}}\right] \in \mathbb{R}^{\left(\sum_{t=2}^{4}\frac{HW}{S_t^2}\right)\times 1}, \tag{6}$$

where "Reshape′" expands the three-dimensional tensor into a two-dimensional matrix. In contrast to previous works like SAGate [45], CFM [46], and DDPM [52], which mainly focus on feature commonality, our RHFF module delves deeper into complementary and redundant features. Since a plain ViT may struggle to confidently segment certain regions, these areas should be supplemented with heterogeneous spatial priors. Additionally, these priors need to be validated for reliability to avoid introducing redundant or incorrect information. Upon such inspirations, we use these two confidence maps to adaptively recalibrate the heterogeneous spatial priors as follows:

$$\boldsymbol{F}_i^S = \left(1 - \boldsymbol{C}_i^V\right) \odot \boldsymbol{C}_i^S \odot \boldsymbol{F}_i^S, \tag{7}$$

where $(1 - \boldsymbol{C}_i^V)$ selectively activates spatial priors that are deficient in Transformer features, and $\boldsymbol{C}_i^S$ emphasizes high-reliability regions within spatial priors.

After adaptively recalibrating and purifying the heterogeneous spatial priors, we utilize a cross-attention mechanism to inject $\boldsymbol{F}_i^S$ into the Transformer feature $\boldsymbol{F}_i^V$. The process can be described as:

$$\hat{\boldsymbol{F}}_i^V = \boldsymbol{F}_i^V + \boldsymbol{\gamma}_i \,\mathrm{MHA}\left(\mathrm{LN}\left(\boldsymbol{F}_i^V\right), \mathrm{LN}\left(\boldsymbol{F}_i^S\right)\right), \tag{8}$$

where "LN$(\cdot)$" denotes LayerNorm operation [53], and "MHA$(\cdot)$" is a multi-head attention operation [54]. In addition, we introduce a learnable vector $\boldsymbol{\gamma}^i \in \mathbb{R}^D$ [31], [55] to ensure that the feature distribution of $\boldsymbol{F}_i^V$ remains largely unchanged. This allows for effective utilization of the pre-trained weights of the ViT while maintaining stable feature integration.

D. Holistic Gated Feature Integration Module

Integrating low-level features to recover detailed information lost in high-level features is a natural approach [44], [56]. Unfortunately, directly combining them suffers from the semantic gap and will drown useful information [34]. To this end, we propose a holistic integration module to selectively extract features from multiple levels. Specifically, lower-level features, rich in fine-grained details, are used to enhance features at higher levels, while gate mechanisms regulate the flow of relevant information.

After injecting heterogeneous spatial priors into the plain ViT, we obtain the output feature $\hat{\boldsymbol{F}}_{i+1}^V$ by passing $\hat{\boldsymbol{F}}_i^V$ through the backbone layers of the i-th block. Then, we utilize the gate mechanism to control the feature flow as follows:

$$\boldsymbol{F}_{i+1}^V = \left(1+\boldsymbol{G}_{i+1}^V\right) \odot \hat{\boldsymbol{F}}_{i+1}^V + \left(1-\boldsymbol{G}_{i+1}^V\right) \odot \sum_{l=1}^{i} \boldsymbol{G}_l^V \odot \hat{\boldsymbol{F}}_l^V, \tag{9}$$

$$\boldsymbol{F}_i^S = \left(1+\boldsymbol{G}_i^S\right) \odot \boldsymbol{F}_i^S + \left(1-\boldsymbol{G}_i^S\right) \odot \sum_{l=1, l\neq i}^{i} \boldsymbol{G}_l^S \odot \boldsymbol{F}_l^S. \tag{10}$$

Features from lower levels are propagated to the i-th stage only when $\boldsymbol{G}_l^{V,S}$ has high values and $\boldsymbol{G}_i^{V,S}$ has low values. In addition to regulating the flow of useful information to appropriate stages, the gates can effectively prevent information redundancy by updating features only when the current information is insufficient or irrelevant. $\boldsymbol{G}_l^V \in \mathbb{R}^{\frac{HW}{16^2} \times D}$ and $\boldsymbol{G}_l^S \in \mathbb{R}^{(\sum_{t=2}^{4} \frac{HW}{S_t^2}) \times D}$ are calculated as follows:

$$\boldsymbol{G}_l^V = \mathrm{Reshape}'\left(\sigma\left(\underset{1\times1}{\mathrm{Conv}}\left(\mathrm{Reshape}\left(\hat{\boldsymbol{F}}_l^V\right)\right)\right)\right), \tag{11}$$

$$\boldsymbol{G}_{l,\frac{1}{2^t}}^S = \mathrm{Reshape}'\left(\sigma\left(\mathrm{DW}\left(\mathrm{Reshape}\left(\boldsymbol{F}_{i,\frac{1}{2^t}}^S\right)\right)\right)\right), \tag{12}$$

$$\boldsymbol{G}_i^S = \left[\boldsymbol{G}_{i,\frac{1}{8}}^S; \boldsymbol{G}_{i,\frac{1}{16}}^S; \boldsymbol{G}_{i,\frac{1}{32}}^S\right], \tag{13}$$

where "DW$(\cdot)$" represents a set of depth-wise convolutions with varying kernel sizes, designed to accommodate the diverse receptive field requirements of different feature groups.

Moreover, semantic representation often exhibits biases due to modality differences. To alleviate this problem, we employ a module comprising a cross-attention layer and a feed-forward network (FFN) to unify spatial priors with Transformer features. This process can be formulated as:

$$\boldsymbol{F}_{i+1}^S = \hat{\boldsymbol{F}}_i^S + \mathrm{FFN}\left(\mathrm{LN}\left(\hat{\boldsymbol{F}}_i^S\right)\right), \tag{14}$$

$$\hat{\boldsymbol{F}}_i^S = \boldsymbol{F}_i^S + \mathrm{MHA}\left(\mathrm{LN}\left(\boldsymbol{F}_i^S\right), \mathrm{LN}\left(\boldsymbol{F}_{i+1}^V\right)\right). \tag{15}$$

Moreover, unlike traditional Transformer architecture [57], [58], which employs the attention mechanism solely on single-scale features, our HGFI strategy incorporates multi-resolution features at scales of 1/8, 1/16, and 1/32. The generated feature $\boldsymbol{F}_{i+1}^S$ is then passed to the next RHFF module.

IV. Experiments

A. Datasets, Implementation Details, and Evaluation Metrics

We utilize two public datasets to evaluate the performance of our proposed methods: Cityscapes [36] and KITTI Semantics [37] datasets. Our model was trained for 20,000 iterations on an NVIDIA RTX 3090 GPU. Images were randomly cropped to a size of 448×448 pixels during the training process. Standard data augmentation techniques, such as random color adjustments, photometric distortion, rescaling, and flipping, were also used. We employed five widely used metrics to evaluate the scene parsing performance: mean F_1-score (mFsc), mean intersection over union (mIoU), average accuracy (aAcc), mean precision (mPre), and mean recall (mRec). More details are given in the supplementary material.

B. Comparisons With State-of-the-Art Methods

As illustrated in Figs. 3 and 4, HFIT consistently outperforms other scene parsing methods on the KITTI Semantics and Cityscapes datasets. As depicted in Figs. 3 and 4, our HFIT achieves segmentation results with clear boundaries, effectively separating adjacent objects such as vehicles, pedestrians, and buildings. Additionally, HFIT demonstrates significant improvements in segmenting small and distant categories, including streetlights, traffic signs, and fences. In dense, visually complex regions like intersections, HFIT ensures robust semantic consistency, accurately segmenting objects such as sidewalks,

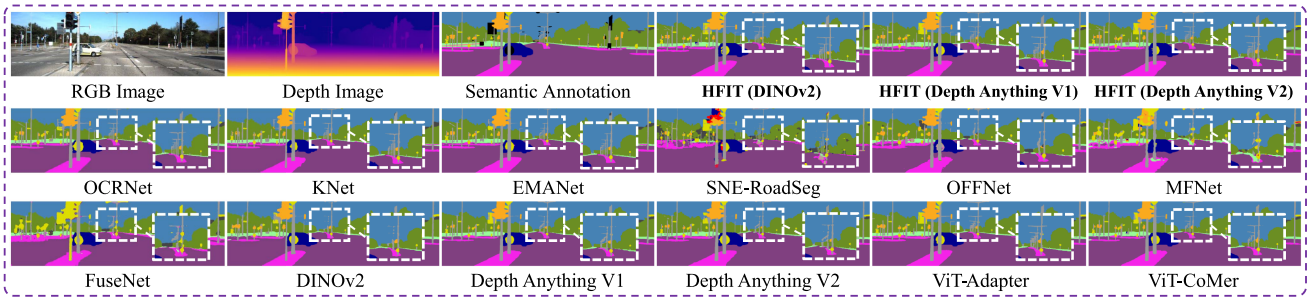

Fig. 3. Qualitative comparisons with SoTA scene parsing approaches on the KITTI Semantics [37] dataset.

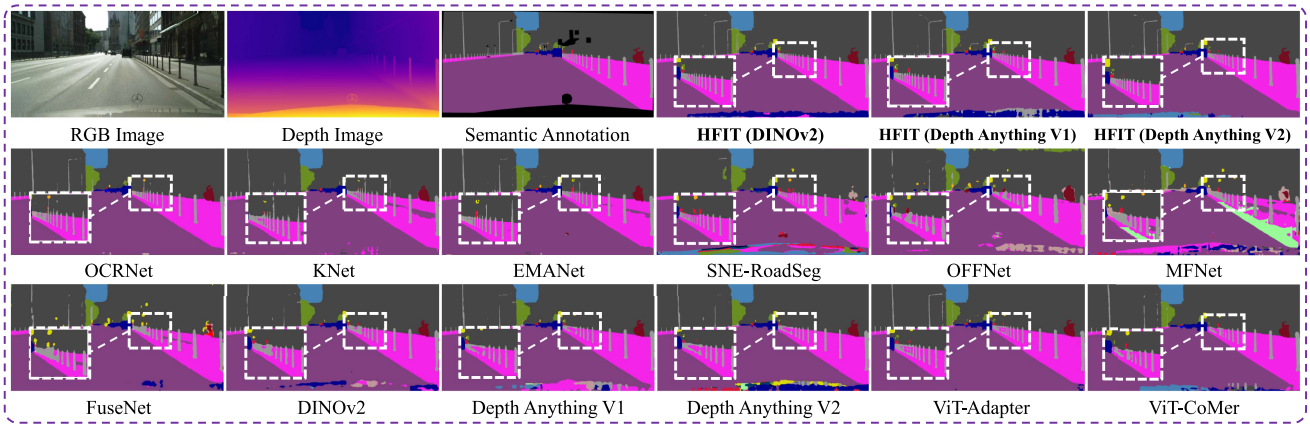

Fig. 4. Qualitative comparisons with SoTA scene parsing approaches on the Cityscapes [36] dataset.

TABLE I
QUANTITATIVE COMPARISONS WITH SOTA SCENE PARSING METHODS ON THE CITYSCAPES DATASET

Backbone	Type	Methods	mFsc (%) ↑	mIoU (%) ↑	aAcc (%) ↑	mPre (%) ↑	mRec (%) ↑	error range (%) ↓
Traditional	Single-Modal	OCRNet [59]	79.78	67.29	94.99	79.73	77.91	1.23
		KNet [60]	74.43	62.25	93.50	81.58	71.74	0.59
		EMANet [61]	72.10	61.52	93.91	78.85	69.69	0.67
	Data-Fusion	SNE-RoadSeg [10]	54.04	42.52	89.34	67.01	49.78	2.32
		OFFNet [11]	59.22	49.16	64.33	71.21	53.41	1.54
		MFNet [12]	51.50	39.20	89.39	62.69	46.56	0.81
		FuseNet [13]	53.28	40.40	89.79	63.63	49.57	2.03
VFM	Single-Modal w/o Adapter	DINOv2 [2]	88.11	80.23	96.12	89.29	88.91	0.52
		Depth Anything V1 [3]	89.48	81.11	96.30	90.26	88.66	**0.29**
		Depth Anything V2 [20]	89.00	80.31	96.03	89.51	88.27	0.41
	Single-Modal w/ Adapter	ViT-Adapter [25]	89.69	82.11	96.53	90.91	89.31	0.35
		ViT-CoMer [31]	88.84	80.02	96.15	89.21	88.01	0.61
	Data-Fusion w/ Adapter (**ours**)	DINOv2 [2]	91.32	84.53	97.15	91.42	**91.19**	0.42
		Depth Anything V1 [3]	**91.64**	**84.74**	**97.09**	**92.27**	90.26	0.64
		Depth Anything V2 [20]	91.06	84.59	96.94	91.87	90.49	0.52

vegetation, and poles, while minimizing overlap and misclassification.

Tables I and II highlight the best results in bold, with ↑ indicating that higher values correspond to improved performance. To ensure a thorough evaluation, we conduct additional experiments by averaging the results from three independent runs for each metric. These results indicate that our proposed HFIT outperforms all other traditional RGB-D methods and single-modal methods, with an increase in mIoU by 17.45-45.54% on the Cityscapes dataset and an increase in mIoU by up

TABLE II
QUANTITATIVE COMPARISONS WITH SOTA SCENE PARSING METHODS ON THE KITTI SEMANTICS DATASET

Backbone	Type	Methods	mFsc (%) ↑	mIoU (%) ↑	aAcc (%) ↑	mPre (%) ↑	mRec (%) ↑	error range (%) ↓
Traditional	Single-Modal	OCRNet [59]	79.77	62.24	93.26	81.40	68.70	0.54
		KNet [60]	74.56	56.47	91.22	84.74	62.79	0.72
		EMANet [61]	67.39	57.03	90.42	80.28	63.71	0.67
	Data-Fusion	SNE-RoadSeg [10]	67.74	41.28	89.92	70.39	50.81	1.03
		OFFNet [11]	55.12	39.24	91.65	67.47	46.26	1.23
		MFNet [12]	58.55	37.97	88.65	69.70	42.98	0.64
		FuseNet [13]	50.91	34.12	84.38	50.94	42.68	0.75
VFM	Single-Modal w/o Adapters	DINOv2 [2]	86.82	78.63	95.49	87.84	86.39	0.34
		Depth Anything V1 [3]	86.62	77.85	95.63	89.41	84.66	**0.22**
		Depth Anything V2 [20]	87.05	79.30	95.77	89.61	86.04	0.45
	Single-Modal w/ Adapters	ViT-Adapter [25]	87.28	79.28	95.91	88.27	87.23	0.23
		ViT-CoMer [31]	85.13	76.54	95.47	88.93	83.77	0.44
	Data-Fusion w/ Adapters (**ours**)	DINOv2 [2]	87.58	79.91	**96.22**	**89.41**	86.88	0.52
		Depth Anything V1 [3]	**88.07**	**80.24**	95.82	89.37	**87.27**	0.51
		Depth Anything V2 [20]	87.23	79.38	95.88	89.13	86.25	0.63

to 46.12% on the KITTI Semantics dataset. Compared to VFM-based algorithms, HFIT demonstrates greater performance, delivering a 2.63-4.72% increase in mIoU on the Cityscapes dataset and a 0.94-2.39% improvement on the KITTI dataset. These results underscore HFIT's superior visual representation capabilities and reliability in real-world driving scenarios. They also validate the hypothesis proposed in Section I, demonstrating that incorporating heterogeneous spatial information significantly enhances the scene parsing performance of VFMs.

Compared to traditional RGB-D methods, single-modal VFMs demonstrate significant advantages, largely due to their profound prior knowledge, which enables them to achieve competitive results in constrained settings. Additionally, we select ViT-Adapter [25] and ViT-CoMer [31] as the baselines of single-modal VFMs with adapters. ViT-Adapter enhances ViT's ability to capture multi-scale information by adding spatial prior, injector, and extractor modules. ViT-CoMer efficiently handles multi-scale interactions, balancing complex scene comprehension with memory efficiency. Notably, ViT-Adapter [25] outperforms standard VFMs, attributed to its dynamic fine-tuning mechanism, which allows for better refinement of feature representations. Nevertheless, ViT-CoMer's [31] performance is underwhelming. This may be due to its bidirectional summation feature fusion mechanism, which fails to bridge the semantic gap between two distinct feature types.

Additionally, we also validate the compatibility of HFIT with various cutting-edge VFMs. Among them, Depth Anything V1 [3] achieves the best performance, with the highest mIoU of 84.74%. However, an interesting observation emerged during our experiments. Despite the evaluation matrices having comparable values on the Cityscapes [36] and KITTI Semantics [37] datasets, the performance improvements achieved by HFIT compared to VFM-based methods are relatively less significant on the KITTI Semantics dataset. This performance gap can be attributed to KITTI's smaller dataset size and limited diversity, which restricts the ability of large models like HFIT to fully utilize their representational capacity. As a result, the model is more prone to overfitting and exhibits reduced generalization on the KITTI dataset.

TABLE III
ABLATION STUDY ON DSPE WITH DIFFERENT INPUTS

Input Type	mFsc (%) ↑	mIoU (%) ↑	aAcc (%) ↑
RGB	90.39	83.07	96.76
Depth	90.53	83.85	96.74
Relative Depth	90.42	83.82	96.78
RGB+Depth	91.02	84.10	**96.99**
RGB+Relative Depth	**91.11**	**84.22**	96.97

C. Ablation Studies

As shown in Table III, we first evaluate HFIT's performance using different input strategies. Single-modal configurations, using either RGB or depth images alone, achieve mIoU scores of 83.07% and 83.85%, respectively. Inputting both RGB and depth images achieves superior performance, with an mIoU of 84.10%, validating the importance of modality complementation. Notably, the performance difference between using scaled metric depth obtained from RAFT-Stereo [62] and relative depth derived from the VFM is minimal, within a margin of 0.12%. This observation further supports the hypothesis presented in Section I.

To investigate the most effective structure for the DSPE, we substitute the basic blocks with alternatives, as detailed in Table IV. Compared to earlier models like the ResNet [49] series and the more recent Swin Transformer [57] series, EfficientNet [48] demonstrates superior performance, achieving the highest mIoU of 83.82%. This suggests that the lightweight

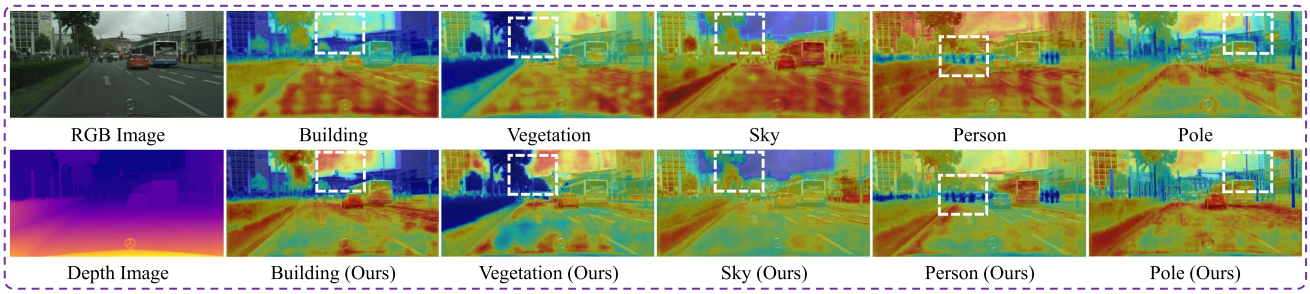

Fig. 5. Probability maps for different classes generated by VFMs and HFIT, where blue indicates high prediction confidence and red represents low prediction confidence.

TABLE IV
ABLATION STUDY ON DSPE BLOCKS

Input Type	Params (M)	mFsc (%) ↑	mIoU (%) ↑	aAcc (%) ↑
ResNet-18	23.34	90.55	83.31	96.87
ResNet-34	43.56	90.78	83.68	96.90
ResNet-50	50.95	90.81	83.76	96.90
Swin-T	56.51	90.76	83.65	96.86
Swin-S	99.15	90.84	83.78	96.88
Swin-B	175.45	90.69	83.56	96.90
EfficientNet-B0	6.35	90.53	83.32	96.81
EfficientNet-B1	9.21	90.69	83.59	**96.91**
EfficientNet-B7	79.09	**90.85**	**83.82**	96.89

TABLE V
ABLATION STUDY ON RHFF MODULES

RGB Weight	Depth Weight	mFsc (%) ↑	mIoU (%) ↑	aAcc (%) ↑
		90.24	82.83	96.79
✓		90.31	82.98	96.77
	✓	90.43	83.14	96.83
✓	✓	**90.96**	**84.01**	**96.97**

design of EfficientNet is well-suited for effectively extracting spatial priors.

Theoretically, higher confidence and greater complementarity in features should result in improved fusion. To validate this hypothesis, we experimentally evaluate the contributions of the RGB weights $(1 - \boldsymbol{C}_i^V)$ and depth weights $\boldsymbol{C}_i^S$ in the RHFF module. As expected, removing either component leads to a significant decline in HFIT's performance (see Table V). Moreover, we compare the probability maps produced by VFMs and HFIT. As shown in Fig. 5, the probability maps generated by HFIT demonstrate more reliable and confident predictions for each category, resulting in clearer and sharper probability distributions. This improvement is particularly evident for categories such as buildings, riders, and cars, especially when compared to predictions at neighboring pixels.

We also investigate the impact of holistic gated feature integration on both Transformer and CNN features. As presented in Table VI, the baseline setup, which relies solely on the cross-attention mechanism for feature fusion, achieves an mIoU

TABLE VI
ABLATION STUDY ON HGFI MODULES

HGFI-ViT	HGFI-Adapter	mFsc (%) ↑	mIoU (%) ↑	aAcc (%) ↑
		89.91	82.36	96.73
✓		90.44	83.17	96.80
	✓	90.30	82.96	96.77
✓	✓	**91.04**	**84.11**	**96.94**

that is 1.75% lower than the full HGFI strategy. As expected, the holistic gated feature integration of ViT features $\hat{\boldsymbol{F}}_{i+1}^V$ and spatial priors $\boldsymbol{F}_i^S$ from the side adapter significantly enhances HFIT's performance.

V. CONCLUSION AND FUTURE WORK

This article discussed a new challenge in computer vision: fully exploiting profound prior knowledge of VFMs for RGB-D driving scene parsing. Our contributions can be summarized in four key aspects: the successful development of an RGB-D scene parsing structure and the introduction of three novel components in the side adapter—DSPE, RHFF, and HGFI. Experimental results demonstrated the significant advantages of HFIT over existing single-modal and data-fusion algorithms, emphasizing its effectiveness and potential for autonomous driving applications.

Looking ahead, advancements in multi-modal VFMs for autonomous driving present new opportunities for robust scene understanding and decision-making. Incorporating emerging techniques like LiDAR-LLM [63] could enable HFIT to simultaneously process LiDAR and depth-based cues, enhancing 3D spatial comprehension in driving environments. Additionally, integrating large language models could improve HFIT's contextual understanding by interpreting ambiguous or complex visual inputs through language-informed perspectives. Furthermore, drawing inspiration from DriveLM's [64] graph-based reasoning for sequential decision-making, a promising future direction for HFIT lies in its evolution from a static model to a dynamic system capable of localization, prediction, and action planning. Such advancements would not only increase HFIT's adaptability and responsiveness to real-time challenges but also pave the way for smarter and more interactive autonomous driving systems. Further discussions are provided in the supplementary materials.

REFERENCES

[1] A. Kirillov et al., "Segment anything," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2023, pp. 4015–4026.

[2] M. Oquab et al., "DINOv2: Learning robust visual features without supervision," *Trans. Mach. Learn. Res. J.*, pp. 1–31, 2024.

[3] L. Yang, B. Kang, Z. Huang, X. Xu, J. Feng, and H. Zhao, "Depth anything: Unleashing the power of large-scale unlabeled data," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 10371–10381.

[4] Y. Feng et al., "ViPOcc: Leveraging visual priors from vision foundation models for single-view 3D occupancy prediction," in *Proc. AAAI Conf. Artif. Intell.*, 2025, pp. 3004–3012.

[5] G. Han and S.-N. Lim, "Few-shot object detection with foundation models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 28608–28618.

[6] C.-W. Liu, Q. Chen, and R. Fan, "Playing to vision foundation model's strengths in stereo matching," *IEEE Trans. Intell. Veh.*, early access, Sep. 25, 2024, doi: 10.1109/TIV.2024.3467287.

[7] M. Shvets, D. Zhao, M. Niethammer, R. Sengupta, and A. C. Berg, "Joint depth prediction and semantic segmentation with multi-view SAM," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2024, pp. 1328–1338.

[8] Z. Chen et al., "Bridging the domain gap: Self-supervised 3D scene understanding with foundation models," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 36, 2024, pp. 79226–79239.

[9] S. Guo, Z. Long, Z. Wu, Q. Chen, I. Pitas, and R. Fan, "LIX: Implicitly infusing spatial geometric prior knowledge into visual semantic segmentation for autonomous driving," *IEEE Trans. Image Process.*, vol. 34, pp. 7250–7263, 2025.

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

[11] C. Min et al., "ORFD: A dataset and benchmark for OFF-road freespace detection," in *Proc. Int. Conf. Robot. Automat.*, 2022, pp. 2532–2538.

[12] 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.*, 2017, pp. 5108–5115.

[13] Hazirbas et al., "FuseNet: Incorporating depth into semantic segmentation via fusion-based CNN architecture," in *Proc. Asian Conf. Comput. Vis.*, 2017, pp. 213–228.

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

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

[16] C.-W. Liu, Y. Zhang, Q. Chen, I. Pitas, and R. Fan, "These maps are made by propagation: Adapting deep stereo networks to road scenarios with decisive disparity diffusion," *IEEE Trans. Image Process.*, vol. 34, pp. 1516–1528, 2025.

[17] Z. Wu, Y. Feng, C.-W. Liu, F. Yu, Q. Chen, and R. Fan, "S^3M-Net: Joint learning of semantic segmentation and stereo matching for autonomous driving," *IEEE Trans. Intell. Veh.*, vol. 9, no. 2, pp. 3940–3951, Feb. 2024.

[18] Z. Huang, Y. Zhang, Q. Chen, and R. Fan, "Online, target-free LiDAR-camera extrinsic calibration via cross-modal mask matching," *IEEE Trans. Intell. Veh.*, vol. 10, no. 5, pp. 3531–3542, May 2025.

[19] Z. Huang et al., "Environment-driven online LiDAR-camera extrinsic calibration," *IEEE Trans. Automat. Sci. Eng.*, vol. 22, pp. 24164–24176, 2025.

[20] L. Yang et al., "Depth anything V2," in *Proc. Adv. Neural Inf. Process. Syst.*, 2024, pp. 21875–21911.

[21] L. Hoyer, D. Dai, Y. Chen, A. Köring, S. Saha, and L. Van Gool, "Three ways to improve semantic segmentation with self-supervised depth estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 11130–11140.

[22] S. Saha et al., "Learning to relate depth and semantics for unsupervised domain adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 8197–8207.

[23] D. Seichter, M. Köhler, B. Lewandowski, T. Wengefeld, and H.-M. Gross, "Efficient RGB-D semantic segmentation for indoor scene analysis," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2021, pp. 13525–13531.

[24] J. Wang et al., "Learning common and specific features for RGB-D semantic segmentation with deconvolutional networks," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 664–679.

[25] Z. Chen et al., "Vision transformer adapter for dense predictions," in *Proc. 11th Int. Conf. Learn. Representations*, 2023.

[26] H. Chen et al., "Conv-adapter: Exploring parameter efficient transfer learning for ConvNets," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 1551–1561.

[27] H. Zhou et al., "CANet: Co-attention network for RGB-D semantic segmentation," *Pattern Recognit.*, vol. 124, 2022, Art. no. 108468.

[28] X. Zhou et al., "FANet: Feature aggregation network for RGBD saliency detection," *Signal Process., Image Commun.*, vol. 102, 2022, Art. no. 116591.

[29] 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. Veh.*, vol. 9, no. 7, pp. 5163–5172, Jul. 2024.

[30] J. Huang et al., "RoadFormer+: Delivering RGB-X scene parsing through scale-aware information decoupling and advanced heterogeneous feature fusion," *IEEE Trans. Intell. Veh.*, vol. 10, no. 5, pp. 3156–3165, May 2025, doi: 10.1109/TIV.2024.3448251.

[31] C. Xia, X. Wang, F. Lv, X. Hao, and Y. Shi, "ViT-CoMer: Vision transformer with convolutional multi-scale feature interaction for dense predictions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 5493–5502.

[32] Z. Wu, S. Gobichettipalayam, B. Tamadazte, G. Allibert, D. P. Paudel, and C. Demonceaux, "Robust RGB-D fusion for saliency detection," in *Proc. Int. Conf. 3D Vis.*, 2022, pp. 403–413.

[33] P. Sun, W. Zhang, H. Wang, S. Li, and X. Li, "Deep RGB-D saliency detection with depth-sensitive attention and automatic multi-modal fusion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 1407–1417.

[34] X. Li et al., "Gated fully fusion for semantic segmentation," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 7, 2020, pp. 11418–11425.

[35] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: Redesigning skip connections to exploit multiscale features in image segmentation," *IEEE Trans. Med. Imag.*, vol. 39, no. 6, pp. 1856–1867, Jun. 2020.

[36] M. Cordts et al., "The CityScapes dataset for semantic urban scene understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 3213–3223.

[37] M. Menze and A. Geiger, "Object scene flow for autonomous vehicles," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3061–3070.

[38] Y. Zhang et al., "Deep multimodal fusion for semantic image segmentation: A survey," *Image Vis. Comput.*, vol. 105, 2021, Art. no. 104042.

[39] A. Valada, J. Vertens, A. Dhall, and W. Burgard, "AdapNet: Adaptive semantic segmentation in adverse environmental conditions," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2017, pp. 4644–4651.

[40] Y. Cheng, R. Cai, Z. Li, X. Zhao, and K. Huang, "Locality-sensitive deconvolution networks with gated fusion for RGB-D indoor semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3029–3037.

[41] E. Hassan et al., "Learning feature fusion in deep learning-based object detector," *J. Eng.*, vol. 2020, no. 1, 2020, Art. no. 7286187.

[42] W. Zhou, Q. Guo, J. Lei, L. Yu, and J.-N. Hwang, "ECFFNet: Effective and consistent feature fusion network for RGB-T salient object detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 3, pp. 1224–1235, Mar. 2022.

[43] H. Jiang, J. Wang, Z. Yuan, Y. Wu, N. Zheng, and S. Li, "Salient object detection: A discriminative regional feature integration approach," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 2083–2090.

[44] Z. Zhang et al., "ExFuse: Enhancing feature fusion for semantic segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 269–284.

[45] X. Chen et al., "Bi-directional cross-modality feature propagation with separation-and-aggregation gate for RGB-D semantic segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 561–577.

[46] J. Wei, S. Wang, and Q. Huang, "F^3Net: Fusion, feedback and focus for salient object detection," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 7, 2020, pp. 12321–12328.

[47] K. Yuan, S. Guo, Z. Liu, A. Zhou, F. Yu, and W. Wu, "Incorporating convolution designs into visual transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 579–588.

[48] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 6105–6114.

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

[50] 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. Automat. Lett.*, vol. 5, no. 4, pp. 5558–5565, Oct. 2020.
[51] G. Xu et al., "THCANet: Two-layer hop cascaded asymptotic network for robot-driving road-scene semantic segmentation in RGB-D images," *Digit. Signal Process.*, vol. 136, 2023, Art. no. 104011.
[52] Y. Pang et al., "Hierarchical dynamic filtering network for RGB-D salient object detection," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 235–252.
[53] J. L. Ba et al., "Layer normalization," *Stat*, vol. 1050, p. 21, 2016.
[54] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Masked-attention mask transformer for universal image segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 1290–1299.
[55] S. Huang, Z. Lu, R. Cheng, and C. He, "FaPN: Feature-aligned pyramid network for dense image prediction," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 864–873.
[56] H. Fu and G. Qiu, "Integrating low-level and semantic features for object consistent segmentation," *Neurocomputing*, vol. 119, pp. 74–81, 2013.
[57] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 10012–10022.
[58] E. Xie et al., "SegFormer: Simple and efficient design for semantic segmentation with transformers," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, vol. 34, pp. 12077–12090.
[59] Y. Yuan et al., "Object-contextual representations for semantic segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 173–190.
[60] W. Zhang et al., "K-Net: Towards unified image segmentation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, vol. 34, pp. 10326–10338.
[61] 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.*, 2019, pp. 9167–9176.
[62] L. Lipson, Z. Teed, and J. Deng, "RAFT-Stereo: Multilevel recurrent field transforms for stereo matching," in *Proc. Int. Conf. 3D Vis.*, 2021, pp. 218–227.
[63] S. Yang et al., "Lidar-LLM: Exploring the potential of large language models for 3D LiDAR understanding," in *Proc. AAAI Conf. Artif. Intell.*, vol. 39, no. 9, 2025, pp. 9247–9255.
[64] C. Sima et al., "DriveLM: Driving with graph visual question answering," in *Proc. Eur. Conf. Comput. Vis.*, 2024, pp. 256–274.

Sicen Guo (Graduate Student Member, IEEE) received the B.E. degree in automation, in 2022, from Tongji University, Shanghai, China, where she is currently working toward the Ph.D. degree with MIAS Group. Her research interests include stereo matching and semantic segmentation.

Tianyou Wen is currently working toward the B.E. degree in electrical engineering with Tongji University, Shanghai, China. His research interests include semantic segmentation and dedicated to using advanced spatial perception technologies to address real-world challenges.

Chuang-Wei Liu received the B.E. degree in automation, in 2020, from Tongji University, Shanghai, China, where he is currently working toward the Ph.D. degree with MIAS Group, supervised by Prof. Rui Fan. His research focuses on computer stereo vision, especially for unsupervised and long-term learning.

Qijun Chen (Senior Member, IEEE) received the B.S. degree in automation from the Huazhong University of Science and Technology, Wuhan, China, in 1987, the M.S. degree in information and control engineering from Xi'an Jiaotong University, Xi'an, China, in 1990, and the Ph.D. degree in control theory and control engineering from Tongji University, Shanghai, China, in 1999. He is currently a Full Professor with the College of Electronic and Information Engineering, Tongji University. His research focuses on robotics.

Rui Fan (Senior Member, IEEE) received the B.Eng. degree in automation from the Harbin Institute of Technology, Harbin, China, in 2015, and the Ph.D. degree in electrical and electronic engineering from the University of Bristol, Bristol, U.K., in 2018. He was a Research Associate with the Hong Kong University of Science and Technology, Hong Kong, from 2018 to 2020 and Postdoctoral Scholar-Employee with the University of California San Diego, La Jolla, CA, USA, from 2020 to 2021. He began his faculty career as a Full Research Professor with the College of Electronic and Information Engineering, Tongji University, Shanghai, China, in 2021. He was promoted to a Full Professor and attained Tenure with the College of Electronic and Information Engineering and Shanghai Research Institute for Intelligent Autonomous Systems, in 2022 and 2024, respectively. His research interests include computer vision, deep learning, and robotics, with a specific focus on humanoid visual perception under the two-streams hypothesis. Prof. Fan was an Associate Editor for IEEE TRANSACTIONS ON AUTOMATION SCIENCE AND ENGINEERING, ICRA'23/25 and IROS'23/24, Area Chair of ICIP'24, and Senior Program Committee Member of AAAI'23/24/25/26. He organized several impactful workshops and special sessions in conjunction with WACV'21, ICIP'21/22/23, ICCV'21/25, and ECCV'22. He was honored by being included in the Stanford University List of Top 2% Scientists Worldwide between 2022 and 2025, recognized on the Forbes China List of 100 Outstanding Overseas Returnees in 2023, acknowledged as one of Xiaomi Young Talents in 2023, and awarded the Shanghai Science & Technology 35 Under 35 honor in 2024 as its youngest recipient.