# TiCoSS: Tightening the Coupling Between Semantic Segmentation and Stereo Matching Within a Joint Learning Framework

Guanfeng Tang, Zhiyuan Wu, *Graduate Student Member, IEEE*, Jiahang Li, *Member, IEEE*, Ping Zhong, *Member, IEEE*, Wei Ye, Xieyuanli Chen, *Member, IEEE*, Huimin Lu, *Member, IEEE*, and Rui Fan, *Senior Member, IEEE*

***Abstract*—Semantic segmentation and stereo matching, respectively analogous to the ventral and dorsal streams in our human brain, are two key components of autonomous driving perception systems. Addressing these two tasks with separate networks is no longer the mainstream direction in developing computer vision algorithms, particularly with the recent advances in large vision models and embodied artificial intelligence. The trend is shifting towards combining them within a joint learning framework, especially emphasizing feature sharing between the two tasks. The major contributions of this study lie in comprehensively tightening the coupling between semantic segmentation and stereo matching. Specifically, this study makes three key contributions: (1) a tightly coupled, gated feature fusion strategy, (2) a hierarchical deep supervision strategy, and (3) a coupling tightening loss function. The combined use of these technical contributions results in TiCoSS, a state-of-the-art joint learning framework that simultaneously tackles semantic segmentation and stereo matching. Through extensive experiments on the KITTI, vKITTI2, and Cityscapes datasets, along with both qualitative and quantitative analyses, we validate the effectiveness of our developed strategies and loss function. Our approach demonstrates superior performance compared to prior arts, with a notable increase in mean intersection over union by over 9%.**

***Note to Practitioners*—TiCoSS is a robust and effective joint learning framework that can simultaneously tackle semantic segmentation and stereo matching tasks. This work aims to improve semantic segmentation performance by exploring the potential complementarity and tightening the coupling between these two tasks. In the future, we plan to further improve the efficiency of the framework, so as to enable its real-time performance on resource-constrained hardware.**

***Index Terms*—Semantic segmentation, stereo matching, autonomous driving, computer vision, joint learning.**

Received 10 February 2025; revised 11 May 2025; accepted 19 June 2025. Date of publication 7 July 2025; date of current version 25 July 2025. This article was recommended for publication by Associate Editor W. Zhang and Editor V. Villani upon evaluation of the reviewers' comments. This work was supported in part by the National Natural Science Foundation of China under Grant 62473288, Grant 62233013, Grant 62403478, Grant 62176184, and Grant 62272489; in part by the Fundamental Research Funds for the Central Universities; in part by the NIO University Programme (NIO UP); in part by the Xiaomi Young Talents Program; and in part by the Young Elite Scientists Sponsorship Program by CAST under Grant 2023QNRC001. *(Corresponding author: Rui Fan.)*

Guanfeng Tang and Jiahang Li are with the College of Electronics and Information Engineering, Tongji University, Shanghai 201804, China (e-mail: gftang@tongji.edu.cn; lijiahang617@tongji.edu.cn).

Zhiyuan Wu is with the Department of Engineering, King's College London, WC2R 2LS London, U.K. (e-mail: zhiyuan.1.wu@kcl.ac.uk).

Ping Zhong is with the Department of Computer Science and Technology, Central South University, Changsha 410017, China (e-mail: ping.zhong@csu.edu.cn).

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

Xieyuanli Chen and Huimin Lu are with the College of Intelligence Science and Technology, National University of Defense Technology, Changsha 410082, China (e-mail: chenxieyuanli@hotmail.com; lhmnew@nudt.edu.cn).

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

This article has supplementary downloadable material available at https://doi.org/10.1109/TASE.2025.3586286, provided by the authors.

Digital Object Identifier 10.1109/TASE.2025.3586286

## I. Introduction

### A. Background

VISUAL environment perception serves as a fundamental and front-end module in robotic systems [1]. Semantic segmentation and stereo matching are two key visual environment perception tasks [2]. The former, akin to the ventral stream in our brain, provides a pixel-level understanding of the scene [3], while the latter, akin to the dorsal stream in our brain, mimics human binocular vision to acquire depth information [4], which is crucial for 3D geometry reconstruction. These two tasks collaborate to deliver both contextual and geometric information, resulting in a comprehensive scene understanding that significantly enhances the capabilities of robotic systems [5].

Nevertheless, previous studies [6], [7], [8] address these two tasks with separate networks, which limits their potential to share informative contextual and geometric information [9]. For instance, stereo matching networks can occasionally produce ambiguous disparity estimations, particularly in texture-less and occluded regions [10]. Semantic segmentation can provide pixel-level scene understanding results, which help resolve such ambiguities and ultimately lead to more reliable disparity estimations [11]. In addition, semantic segmentation networks often struggle to distinguish clear object boundaries, particularly in complex driving scenarios, due to the lack of spatial geometric information [12], [13]. A common solution for improved semantic segmentation performance is to employ

feature-fusion networks equipped with duplex encoders to extract heterogeneous features from RGB-X data [14], where "X" provides spatial geometric information, such as the depth images generated from LiDAR point clouds [15] and surface normal maps obtained through depth-to-normal translation [3]. However, the availability and quality of "X" significantly influence the overall performance of semantic segmentation and can potentially limit the practical deployment of feature-fusion networks [5].

Therefore, in recent years, the simultaneous learning, deployment, and inference of both tasks have become a mainstream [16], [17], [18], because a unified joint learning framework can process contextual and geometric information more comprehensively. It also enables end-to-end training of the entire system, capable of tackling the challenges posed by both tasks [19], [20]. Consequently, this joint learning approach can enhance the overall performance of both semantic segmentation and stereo matching, and outperform models trained separately for each task [5].

### B. Existing Challenges and Motivation

The performance of a feature-fusion semantic segmentation network is heavily influenced by the employed strategy for heterogeneous feature fusion [21], [22]. Currently, the bottleneck lies in the simplistic and indiscriminate fusion of heterogeneous features, which often causes conflicting learning representations and erroneous segmentation results [23]. Taking the state-of-the-art (SoTA) joint learning method $S^3$M-Net [5] as an example, its adopted feature fusion strategy essentially performs an element-wise summation between contextual and geometric feature maps at each stage. These feature maps are then directly fed into subsequent layers without filtering out irrelevant information, leading to a loose coupling between the semantic segmentation and stereo matching tasks within the encoder. Furthermore, as the network goes deeper, such an indiscriminate feature fusion strategy tends to diminish the proportion of informative geometric features in the decoder's input [24], potentially leading to unsatisfactory semantic segmentation performance.

Additionally, due to the vanishing gradient problem, existing joint learning frameworks often suffer from slow convergence during training. A common solution is to employ the deep supervision (DS) strategies that incorporate additional pathways to achieve gradient propagation [25]. Nonetheless, existing DS strategies employed in feature-fusion networks typically overlook the potential interactions between the main and side auxiliary classifiers, which can limit the overall semantic segmentation performance within our joint learning framework.

In the loss function aspect, previous joint learning frameworks, such as SegStereo [17] and SSNet [26], typically compute the losses for two tasks independently, and supervise the entire training process by simply minimizing the weighted sum of these losses. Such loss function fails to leverage the potential complementarity between the two tasks at the output level, which also results in a loose coupling.

Prior arts, such as $S^3$M-Net [5] and DSNet [9], primarily focus on introducing a joint learning framework that performs semantic segmentation and stereo matching simultaneously. However, exploring the potential complementarity and tightening the coupling between these two tasks have received relatively limited attention in this research area and warrant further investigation.

### C. Contributions

To address the aforementioned limitations, we introduce **Ti**ghtly-**Co**upled Semantic **S**egmentation and **S**tereo Matching Network (**TiCoSS**), an end-to-end joint learning approach that focuses primarily on improving the coupling between stereo matching and semantic segmentation, which has not been emphasized in previous relevant studies. Our proposed TiCoSS introduces three new techniques: (1) a tightly-coupled, gated feature fusion (TGF) strategy, which utilizes a series of selective inheritance gates (SIGs) to propagate useful contextual and geometric information from the preceding layer to the current layer, resulting in a tightly-coupled encoder; (2) a hierarchical deep supervision (HDS) strategy that uses the fused feature maps with the highest resolution to guide deep supervision throughout each branch, as these features contain the most abundant local spatial details; (3) a novel coupling tightening (CT) loss, consisting of a widely used stereo matching loss presented in the study [8], the semantic consistency-guided (SCG) loss introduced in the study [5], a disparity inconsistency-aware (DIA) loss that leverages disparity estimation results to help distinguish clearer object boundaries, and a deep supervision consistency constraint (DSCC) loss which employs the Kullback-Leibler (KL) divergence to improve prediction consistency across outputs from all deep supervision branches. These contributions collectively advance $S^3$M-Net, and results in TiCoSS, a new, powerful, and tightly-coupled joint learning framework that simultaneously performs robust and accurate semantic segmentation and stereo matching tasks. Extensive experiments conducted on the vKITTI2 [27], KITTI 2015 [28], and Cityscapes [29] datasets unequivocally demonstrate the effectiveness of the aforementioned contributions and the superior performance of TiCoSS over other SoTA approaches.

In summary, the main contributions of this article include:

- The TGF strategy, which propagates useful contextual and geometric information from the preceding layer to the current layer, enabling more effective feature fusion for semantic segmentation;
- The HDS strategy, which uses the fused features with the richest local spatial details to guide deep supervision across each branch;
- The DIA loss and the DSCC loss that tighten the coupling between the two tasks, thereby further improving the semantic segmentation performance.

### D. Article Structure

The remainder of this article is organized as follows: Sect. II reviews related prior arts. Sect. III introduces our proposed TiCoSS. Sect. IV compares our network with other SoTA approaches and presents the ablation studies. Finally, we conclude this article and provide recommendations for feature work in Sect. V.

## II. Literature Review

### A. Semantic Segmentation

Semantic segmentation has been a long-standing challenge in the fields of computer vision and robotics over the past decade [1]. SoTA networks generally fall into two groups: (1) single-modal networks (with a single encoder) and (2) feature-fusion networks (with multiple encoders) [3]. Early efforts primarily focused on encoder-decoder architectures for pixel-level classification. Representative examples include the DeepLab series [30], as well as Transformer-based networks [7], [31], [32]. The encoder extracts hierarchical, contextual feature maps from input images, while the decoder generates segmentation maps by upsampling and combining feature maps from different encoder layers. Nonetheless, these networks are limited in effectively combining heterogeneous features extracted from different sources (or modalities) of visual information, which makes it challenging to produce accurate segmentation results in scenarios with poor lighting and illumination conditions [3]. This has led researchers to focus on feature-fusion networks that can effectively fuse heterogeneous features extracted from multiple sources (or modalities) of visual information. This problem is often referred to as "RGB-X semantic segmentation", where "X" represents the additional modality (or source) of visual information, in addition to the RGB images. The most representative feature-fusion networks include convolutional neural network (CNN)-based ones, such as RTFNet [33] and the SNE-RoadSeg series [3], [24], [25], as well as Transformer-based ones, such as OFF-Net [34], RoadFormer [14], and DFormer [35]. In this article, we design TiCoSS based on $S^3$M-Net, with a special emphasis on exploring more effective solutions for tighter coupling between semantic segmentation and stereo matching.

### B. Stereo Matching

Owing to recent advancements in deep learning techniques, end-to-end deep stereo matching networks [8], [36], [37], [38] have dramatically outperformed traditional explicit programming-based stereo matching algorithms. PSMNet [36] introduces a spatial pyramid to capture multi-scale information and employs a series of 3D convolutional layers to aggregate both local and global contexts for cost computation. To address the high computational cost of 3D convolutions, researchers have sought ways to balance efficiency and accuracy in stereo matching. LEA-Stereo [37], for instance, introduces the first neural architecture search (NAS) framework to stereo matching, enabling automated architecture optimization. RAFT-Stereo [8], a rectified stereo matching approach, developed based on RAFT [39], uses a series of gated recurrent units to iteratively refine correlation features and improve disparity estimation accuracy. CRE-Stereo [38] further advances this approach by introducing an adaptive group-wise correlation layer to mitigate the impact of rectification errors in stereo images, resulting in more accurate disparity estimation results. In this article, we primarily focus on improving semantic segmentation performance, and therefore, adopt the stereo matching approach used in $S^3$M-Net.

### C. Simultaneous Semantic Segmentation and Stereo Matching

Existing joint learning frameworks that simultaneously address these two tasks mainly focus on improving disparity accuracy by leveraging semantic information [9], [11], [16], [17], [18]. However, discussions on utilizing disparity information to enhance semantic segmentation performance at the feature level for joint learning remain limited, except for the aforementioned "RGB-X semantic segmentation". These prior arts often require extensive annotated training data or involve complex training strategies for joint learning. For example, SegStereo [17] necessitates an initial unsupervised training phase on the large-scale Cityscapes [29] dataset, followed by a subsequent supervised fine-tuning on the smaller KITTI [28], [40] datasets. Similarly, the approaches introduced in [11], [18], [41] require pre-training their spatial branches for stereo matching on the large-scale SceneFlow [42] dataset before fine-tuning both semantic and spatial branches on the KITTI [28] dataset. DSNet [9] adopts a different joint learning strategy by alternating the training of the semantic segmentation and stereo matching networks, with parameters of each network being frozen during the training of the other. Nevertheless, leveraging both contextual and geometric information can be challenging, as the shared features between these two tasks are not learned in an end-to-end manner. DispSegNet [16] utilizes an embedding learned from the semantic segmentation branch to refine the initial disparity estimations. RTS$^2$Net [18] relies on coarse-to-fine estimations in a multi-stage fashion for accurate disparity estimation. However, both methods only achieve limited improvement in semantic segmentation since they fail to leverage informative spatial details to enhance semantic segmentation performance. SSNet [26] employs a single encoder to extract shareable features for both tasks. However, as demonstrated in the study [43], such shareable features may not be suitable for both dense prediction and geometric vision tasks. $S^3$M-Net [5], MENet [12], and SG-RoadSeg [19] use two separate encoding branches to accomplish these two tasks simultaneously, but the weak coupling between these branches limits the integration of contextual and geometric information. In contrast to the aforementioned approaches, our proposed TiCoSS uses a tightly-coupled joint learning framework that effectively leverages both contextual and geometric information. Moreover, TiCoSS is trained in an end-to-end manner and is capable of learning accurate and robust semantic segmentation and stereo matching tasks simultaneously, even with limited training data.

## III. Methodology

### A. Framework Overview

The architecture of our proposed TiCoSS is illustrated in Fig. 1, containing three major technical contributions:

(1) A novel duplex, tightly-coupled encoder designed to selectively extract and fuse heterogeneous features, namely contextual features from RGB images and geometric features from disparity maps.

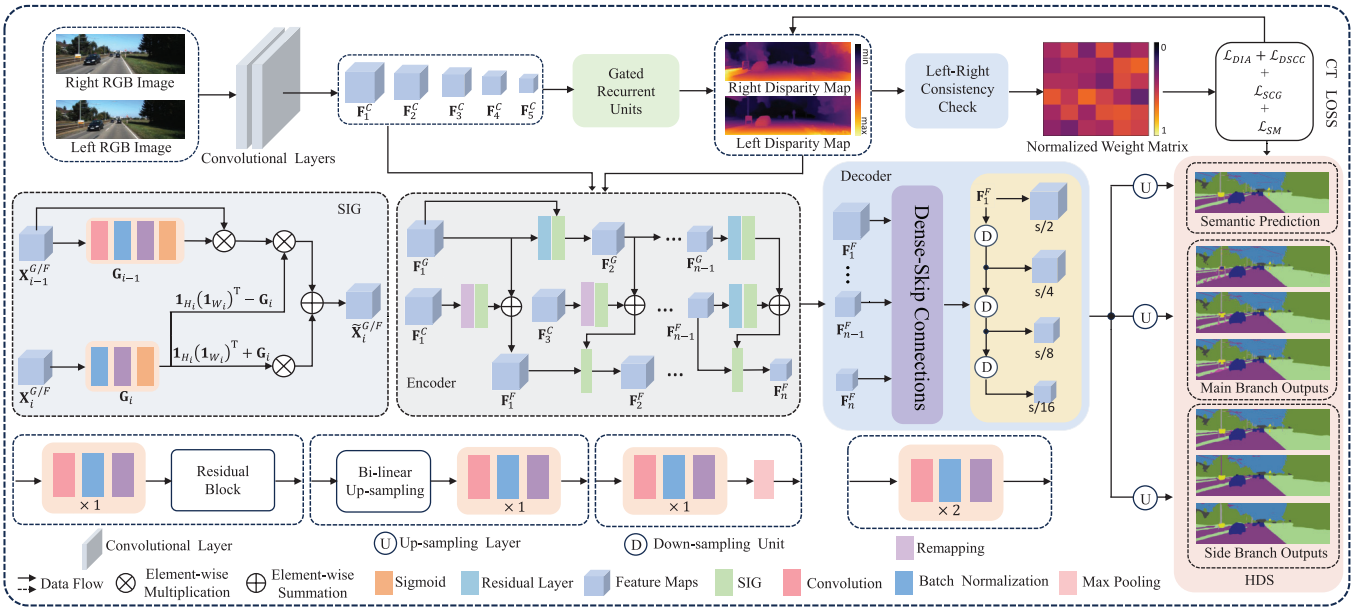


Fig. 1. The architecture of our proposed TiCoSS for end-to-end joint learning of semantic segmentation and stereo matching.

(2) A novel HDS strategy that leverages fused features with the richest local spatial details to guide deep supervision across each branch (auxiliary classifier).
(3) A CT loss that supervises the entire joint learning process and further tightens the coupling between semantic segmentation and stereo matching.

### B. Tightly-Coupled Gated Fusion Strategy

S$^3$M-Net [5] proposes an effective joint learning framework to simultaneously perform semantic segmentation and stereo matching. Despite achieving impressive results, these two tasks are loosely coupled. It merely employs the feature fusion strategy proposed in SNE-RoadSeg [3], where the geometric features extracted from disparity maps are indiscriminately fused into the contextual features extracted from RGB images via simplistic element-wise summation. The fused heterogeneous features are then treated as preceding contextual features and fed into subsequent layers without selective processing, which can potentially mislead the semantic segmentation task. This is primarily because the deeper geometric features contain irrelevant semantic information, and as the network goes deeper, the proportion of contextual features in the decoder's input tends to diminish [24].

Our TGF strategy is, therefore, designed to overcome this limitation by selectively complementing contextual features with informative geometric features, resulting in a tightly-coupled duplex encoder. The core of our TGF strategy is the SIGs, developed based on Gated Fully Fusion (GFF) [44], which fuse features from multiple scales using gates that control the propagation of useful information. This enables the features at each scale to be enhanced by both deeper, semantically stronger features and shallower, spatially richer features, significantly reducing noises during feature fusion. Nonetheless, GFF is primarily regarded as a late fusion strategy [45], as it performs feature fusion at the decision layer and requires multi-scale features to be generated prior to processing. Additionally, GFF focuses solely on fusing features extracted from RGB images across multiple scales and is not well-suited to fuse heterogeneous features which are extracted and fused progressively. In contrast, our proposed TGF strategy performs intermediate feature fusion during the encoding stage, enabling more interactions between heterogeneous features. Specifically, it utilizes a series of SIGs (see Fig. 1) to selectively inherit useful information in $\mathbf{X}_{i-1}^{G,F}$ from the previous layer into $\mathbf{X}_{i}^{G,F}$, the features at the current layer, where $i \in [1,n] \cap \mathbb{Z}$ denotes the layer number, and the superscripts '$G$' as well as '$F$' represent 'geometric' and 'fused' features,[1] respectively. Our SIG outputs $\tilde{\mathbf{X}}_{i}^{G,F}$ selectively inherit feature maps at the $i$-th layer using the following expression:

$$\begin{aligned}\tilde{\mathbf{X}}_{i}^{G,F} &= \Omega_i\left(\mathbf{X}_{i-1}^{G,F}, \mathbf{X}_{i}^{G,F}\right) = \left(\mathbf{1}_{H_i}(\mathbf{1}_{W_i})^{\top} + \mathbf{G}_i\right) \odot \mathbf{X}_{i}^{G,F} \\ &\quad + \left(\mathbf{1}_{H_i}(\mathbf{1}_{W_i})^{\top} - \mathbf{G}_i\right) \odot \left[\mathbf{G}_{i-1} \odot \mathcal{R}(\mathbf{X}_{i-1}^{G,F})\right], \end{aligned} \tag{1}$$

where $\Omega_i$ represents the SIG operation at the $i$-th layer, $\mathbf{1}_k$ denotes a column vector of ones, $\mathbf{G}_i \in [0,1]^{H_i \times W_i}$ represents a gate map that controls feature propagation, $\odot$ denotes the element-wise multiplication broadcasting in the channel dimension, and $\mathcal{R}$ represents the remapping operation, as detailed in the study [5].

Based on our proposed TGF strategy, the heterogeneous feature extraction and fusion process in our duplex encoder can be formulated as follows:

$$\mathbf{F}_{i}^{G} = \begin{cases} \mathcal{E}_{i}^{G}(\mathbf{D}^{L}), & i = 1 \\ \Omega_{i}^{G}(\mathbf{F}_{i-1}^{G}, \mathcal{E}_{i}^{G}(\mathbf{F}_{i-1}^{G})), & 1 < i \le n \end{cases} \tag{2}$$

[1]It is worth noting here that we use $\mathbf{X}_{i}^{F}$ instead of $\mathbf{X}_{i}^{C}$, where the superscript '$C$' denotes 'contextual' features, primarily because the branch processing RGB images progressively fuses the geometric features extracted from disparity maps to obtain fused features.

and

$$\mathbf{F}_i^F = \begin{cases} \mathcal{E}_i^F(\mathbf{F}_i^C) \oplus \mathbf{F}_i^G, & i = 1 \\ \Omega_i^F(\mathbf{F}_{i-1}^F, \mathcal{R}(\mathbf{F}_{2i-1}^C)) \oplus \mathbf{F}_i^G, & 1 < i \leq \frac{n+1}{2} \\ \Omega_i^F(\mathbf{F}_{i-1}^F, \mathcal{E}_i^F(\mathbf{F}_{i-1}^F)) \oplus \mathbf{F}_i^G, & \frac{n+1}{2} < i \leq n, \end{cases} \tag{3}$$

where $\mathbf{D}^L$ denotes the estimated disparity map, $\mathbf{F}_i^C$, $\mathbf{F}_i^G$, and $\mathbf{F}_i^F$ respectively represent the extracted contextual feature maps in the left view, geometric feature maps, and fused feature maps at the $i$-th layer, $\mathcal{E}_i^G$ and $\mathcal{E}_i^F$ denote the geometric and fused features encoding operations at the $i$-th layer, respectively, and $\oplus$ represents the element-wise summation. Considering that the shallower features between the two tasks have similar numbers of channels, we make the contextual feature maps of the first three layers share weights with the feature maps extracted from the stereo matching network in our practical implementation. Our proposed TGF strategy selectively propagates useful information to subsequent layers, reducing indiscriminate heterogeneous feature fusion which can severely mislead semantic segmentation as the network goes deeper, thus achieving tighter coupling between these two relevant perception tasks. Further theoretical analyses of the TGF strategy are provided in the supplement.

### C. Hierarchical Deep Supervision

After tightening the coupling between semantic segmentation and stereo matching during the feature encoding stage, we turn our focus towards the feature decoding process. We first revisit the deep supervision strategies employed in SNE-RoadSeg+ [25] and UNet 3+ [46]. The former applies deep supervision to the decoded features with the highest resolution, whereas the latter achieves this goal on the deepest decoded features at each resolution. Despite the effectiveness of these two prior approaches in addressing challenges such as vanishing gradients and slow model convergence, the deep supervision strategies employed by them are not comprehensive (the former emphasizes enhancing local details, while the latter focuses on improving consistency across multi-scale segmentation predictions). Ideally, they should be used in conjunction to complement each other for improved results. Therefore, our proposed HDS strategy combines the strengths of both methods and demonstrates superior performance compared to each individually.

A straightforward way to combine the strengths of these two methods is to apply deep supervision strategies simultaneously to both the main and side branches, enabling the network to leverage features from shallow layers (containing rich local details) and deep layers (being semantically strong). However, this method can not fully exploit the potential complementarity between the main and side auxiliary classifiers, since the decoded features are exclusively derived from adjacently-connected ones. Additionally, as discussed in Sect. III-B, decoded features in deep layers also lack informative spatial details that are present in shallow layers, limiting the performance of DS strategies in our multi-task learning framework. Thus, to improve the interactions among auxiliary classifiers and provide local spatial details for decoded features in deep layers, we utilize the decoded, fused feature maps $\mathbf{F}_1^F$ (containing rich, fine-grained local spatial details that are essential for semantic segmentation) at the highest resolution to guide the feature decoding process in the side branches. Specifically, for the $l$-th auxiliary classifier within the side branch, we utilize a feature dynamic alignment (FDA) block, which is composed of $l$ downsampling units to progressively align channel dimensions and spatial resolutions between a pair of features at different layers. Each downsampling unit comprises a $3 \times 3$ convolutional layer with a stride of 2, followed by a batch normalization (BN) layer and a rectified linear unit (ReLU) activation layer. Compared to the simple max-pooling operation which may disrupt the original feature representation, our FDA achieves a smoother feature alignment. The feature maps obtained by downsampling $\mathbf{F}_1^F$ are then concatenated with the deepest decoded features at the corresponding layers. This downsampling output not only guides the decoding process but also serves as the input for the subsequent downsampling unit, thereby preserving fine-grained local details to the greatest extent. Since the outputs of the side branches are obtained by directly upsampling the deepest features at each layer, this strategy does not significantly impact training efficiency and memory usage. Moreover, the auxiliary classifiers of the main and side branches collaboratively provide additional pathways for gradients to propagate more efficiently from the output layers to their corresponding layers, thereby accelerating the convergence of our model.

### D. Coupling Tightening Loss for Multi-Task Joint Learning

Compared to the tasks focused solely on either semantic segmentation or stereo matching, the loss function employed in our joint learning framework aims to supervise both tasks simultaneously [5]. Our proposed CT loss function is expressed as follows:

$$\mathcal{L}_{CT} = \alpha \mathcal{L}_{DIA} + \beta \mathcal{L}_{DSCC} + \mathcal{L}_{SCG} + \mathcal{L}_{SM}. \tag{4}$$

where the DIA loss $\mathcal{L}_{DIA}$ measures disparity inconsistency, the DSCC loss $\mathcal{L}_{DSCC}$ measures the segmentation consistency across outputs from all deep supervision branches, the SCG loss $\mathcal{L}_{SCG}$ denotes the semantic consistency-guided (SCG) loss proposed in the study [5], $\mathcal{L}_{SM}$, a prevalently used loss function to supervise the training of the stereo matching network, is detailed in the study [5], and the weight factors $\alpha$ and $\beta$ are determined through extensive ablation studies, as detailed in Sect. IV-E.

*1) Disparity Inconsistency-Aware Loss:* In the previous study [5], the developed SCG loss focuses mainly on enforcing semantic consistency to reduce segmentation errors caused by occlusions, neglecting the fact that occlusions can also lead to inconsistent disparity estimations, which are unfortunately under-explored when designing the loss. Thus, we further incorporate the disparity inconsistency in it to form a more advanced loss function to train our TiCoSS. Specifically, we define a weight matrix $\mathbf{W} \in \mathbb{R}^{H \times W}$, drawing from the concept of left-right consistency check in stereo matching, where

$$\mathbf{W}(\mathbf{p}) = \mathbf{D}^L(\mathbf{p}) - \mathbf{D}^R(\mathbf{p} - (\mathbf{D}^L(\mathbf{p}; 0))) \tag{5}$$

denotes the weight at the given pixel $\mathbf{p}$, and $\mathbf{D}^L$ as well as $\mathbf{D}^R$ represent the left and right disparity maps, respectively. $\mathbf{W}^N \in \mathbb{R}^{H\times W}$, a normalized weight matrix, is then yielded, where

$$\mathbf{W}^N(\mathbf{p}) = \frac{1}{1+e^{-|\mathbf{W}(\mathbf{p})|}} \in [0,1]. \tag{6}$$

A higher normalized weight corresponds to a greater inconsistency between a given pair of left and right disparities, indicating the need for more attention during the training of a semantic segmentation model. Our proposed DIA loss is, therefore, formulated as follows:

$$\mathcal{L}_{DIA} = \sum_{i=1}^{n}\left\{-\frac{1}{N}\sum_{j=1}^{N}\sum_{k=1}^{C}\left[\mathbf{W}^N(\mathbf{p})\, y_k(\boldsymbol{p})\log(\hat{y}_k(\boldsymbol{p}))\right]\right\}, \tag{7}$$

where $n$ denotes the number of deep supervision branches, $N$ represents the total number of pixels, $C$ denotes the total number of classes, $\hat{y}_k(\boldsymbol{p})$ represents the predicted probability of pixel $\boldsymbol{p}$ belonging to class $k$, and $y_k(\boldsymbol{p})$ denotes the ground-truth label for $\boldsymbol{p}$ in class $k$.

*2) Deep Supervision Consistency Constraint Loss:* In prior studies [25], [46], the relationship among prediction maps generated by different auxiliary classifiers was not considered, which may lead to semantic inconsistencies across scales. To address this issue, we draw inspiration from the study [47] to design the following DSCC loss, which utilizes the KL divergence to measure the prediction differences across scales:

$$\mathcal{L}_{DSCC} = \sum_{r=1}^{L}\sum_{\substack{s=1\\ s\neq r}}^{L}\left[-\frac{1}{N}\sum_{i=1}^{N}\hat{y}_k^r(\boldsymbol{p})\log\frac{\hat{y}_k^r(\boldsymbol{p})}{\hat{y}_k^s(\boldsymbol{p})}\right], \tag{8}$$

where $L$ denotes the total number of auxiliary classifiers. Further theoretical analyses of the DSCC loss are provided in the supplement.

## IV. Experiments

In this article, we conduct extensive experiments to evaluate the performance of our introduced TiCoSS both quantitatively and qualitatively. The following subsections detail the used datasets, experimental setup, evaluation metrics, and the comprehensive evaluation of our proposed method.

### A. Datasets

Since our network training requires both semantic and disparity annotations, we utilize the following three public datasets to conduct the experiments:

- The vKITTI2 [27] dataset contains virtual replicas (providing 15 semantic classes) of five sequences from the KITTI dataset. Dense ground-truth disparity maps are obtained through depth rendering using a virtual engine. Following the study [5], we employ 700 stereo image pairs, along with their semantic and disparity annotations, to evaluate the effectiveness and robustness of our proposed TiCoSS, where 500 pairs are utilized for model training and the remaining 200 pairs are used for model validation.
- The KITTI 2015 [28] dataset contains 400 stereo image pairs captured in real-world driving scenarios. Half of these pairs have ground truth annotations, while the remaining half do not. This dataset provides 19 semantic classes (consistent with those in the Cityscapes [29] dataset). Sparse ground-truth disparity maps are obtained using a Velodyne HDL-64E LiDAR. In our experiments, we split the data with a 7:3 ratio for training and testing, respectively.
- The Cityscapes [29] dataset is a widely used urban scene understanding dataset, containing 2,975 stereo images for model training and 500 stereo images for model validation, with well-annotated semantic annotations. It is noteworthy that the corresponding depth information is obtained using ViTAStereo [43], since depth ground truth is unavailable.

### B. Experimental Setup

Our experiments are conducted on an NVIDIA RTX 3090 GPU with a batch size of 1. We set the maximum disparity search range to 192 pixels. All images are cropped to 512 $\times$ 256 pixels before being processed by the network. We utilize the AdamW optimizer for model training, with epsilon and weight decay set to $10^{-8}$ and $10^{-5}$, respectively. The initial learning rate is set to $2\times10^{-4}$. Training lasts for 100,000 iterations on the vKITTI2 dataset, 20,000 iterations on the KITTI 2015 dataset, and 50,000 iterations on the Cityscapes dataset. Standard data augmentation techniques are applied to improve model robustness.

### C. Evaluation Metrics

Following the study [5], we quantify the performance of semantic segmentation using seven metrics: (1) accuracy (Acc), (2) mean accuracy (mAcc), (3) precision (Pre), (4) recall (Rec), (5) mean F1-score (mFSc), (6) mean intersection over union (mIoU), and (7) frequency-weighted intersection over union (fwIoU). We calculate each of these metrics on a per-image basis before averaging them across the entire dataset. Moreover, we quantify the performance of stereo matching using two metrics: (1) average end-point error (EPE) and (2) percentage of error pixels (PEP) at tolerance levels of 1.0 and 3.0 pixels, respectively.

### D. Comparison With State-of-the-Art Methods

*1) Semantic Segmentation Performance:* The qualitative experimental results on the KITTI, vKITTI2, and Cityscapes datasets are presented in Figs. 2, 3, and 4, respectively, while the quantitative experimental results on these three datasets are given in Tables I, II, and III, respectively. We also compare the semantic segmentation performance of TiCoSS and the baseline S$^3$M-Net using the mmsegmentation framework,[2] with the results provided in Table IV.

These results suggest that TiCoSS outperforms all other SoTA single-modal and feature-fusion networks (including both CNN-based and Transformer-based methods) across most

[2]The mmsegmentation framework is available at https://github.com/open-mmlab/mmsegmentation

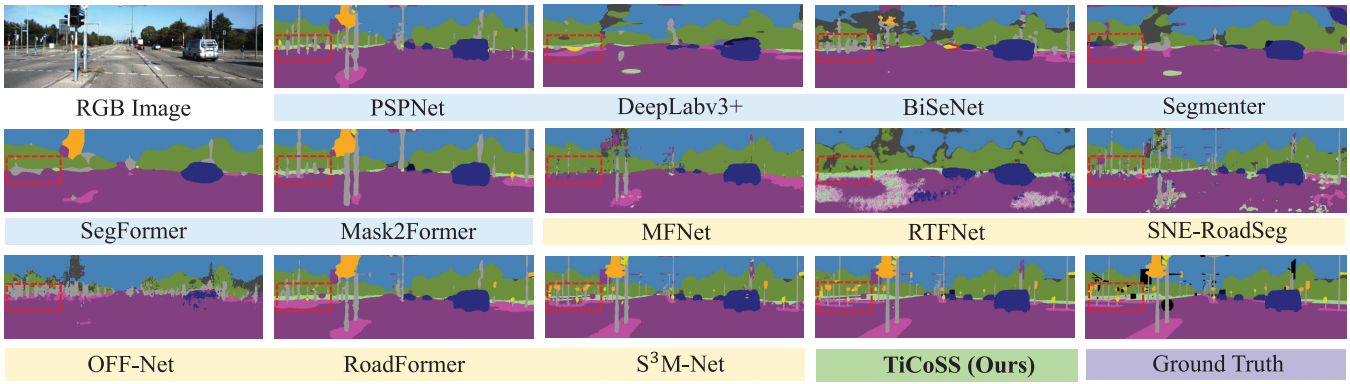


Fig. 2. Qualitative experimental results achieved by the SoTA semantic segmentation networks on the KITTI 2015 [28] dataset.

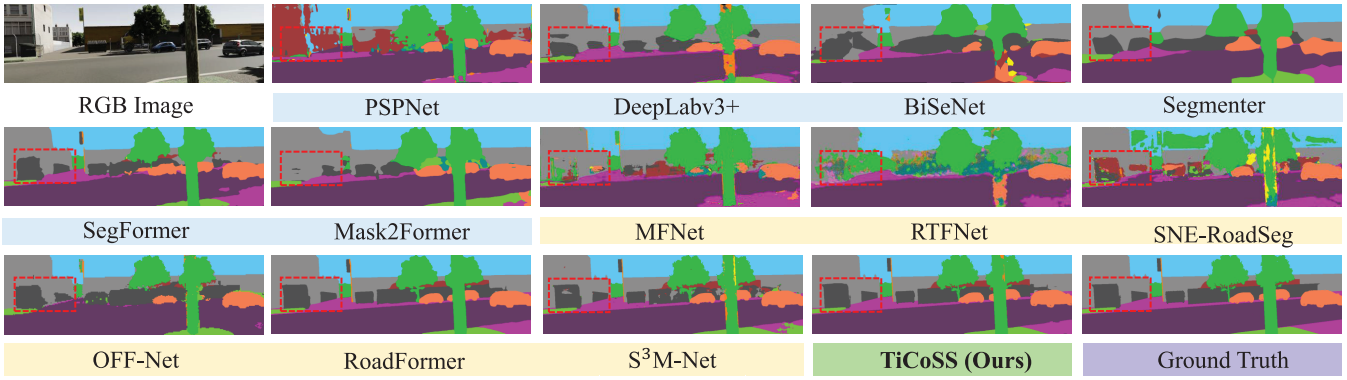


Fig. 3. Qualitative experimental results achieved by the SoTA semantic segmentation networks on the vKITTI2 [27] dataset.

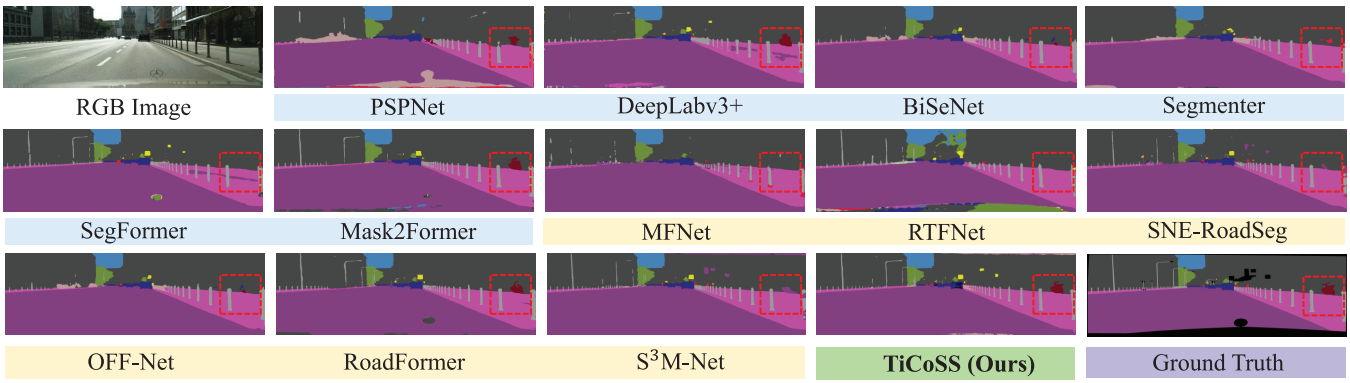


Fig. 4. Qualitative experimental results achieved by the SoTA semantic segmentation networks on the Cityscapes [29] dataset.

evaluation metrics on these three datasets. Specifically, compared with S$^3$M-Net, the SoTA joint learning method, TiCoSS demonstrates substantial improvements on the KITTI dataset, achieving increases of 9.68% in mAcc, 10.57% in mIoU, 2.71% in fwIoU, and 1.20% in mFSc, respectively.

Similarly, on the vKITTI2 dataset, it outperforms other networks across all evaluation metrics, with improvements of 3.88% in mAcc, 5.08% in mIoU, 0.62% in fwIoU, and 0.26% in mFSc, respectively. On the Cityscapes dataset, it outperforms other networks in most evaluation metrics, with improvements of 1.99% in mAcc, 4.22% in mIoU, 1.60% in Pre, 2.12% in Rec, and 3.00% in mFSc. Particularly, as observed in Figs. 2 and 3, TiCoSS achieves more accurate predictions on distant regions as well as object boundaries and is capable of providing more fine-grained semantic segmentation details compared to S$^3$M-Net.

We attribute these improvements to the tighter coupling between the two tasks within our joint learning framework, which effectively leverages the informative geometric information extracted from our predicted disparity maps. This enhances the integration of contextual and geometric features through our proposed TGF strategy, further improving the ability to handle heterogeneous features.

*2) Stereo Matching Performance:* The qualitative experimental results on the KITTI and vKITTI2 datasets are illustrated in Figs. 5 and 6, respectively, while the quantitative experimental results on these two datasets are provided in Table V. Since the primary focus of this study is to improve

TABLE I

QUANTITATIVE COMPARISONS OF SOTA SEMANTIC SEGMENTATION NETWORKS ON THE KITTI 2015 [28] DATASET. THE BEST RESULTS ARE SHOWN IN BOLD FONT. THE SYMBOL ↑ INDICATES THAT A HIGHER VALUE CORRESPONDS TO BETTER PERFORMANCE

| Networks | Publication | Type | Acc (%) ↑ | mAcc (%) ↑ | mIoU (%) ↑ | Pre (%) ↑ | Rec (%) ↑ | mFSc (%) ↑ |
|---|---|---|---|---|---|---|---|---|
| DeepLabv3+ [30] | ECCV'18 | Single-Modal | 82.33 | 50.15 | 42.79 | 83.85 | 87.18 | 84.59 |
| BiSeNet V2 [48] | IJCV'21 | | 73.68 | 41.66 | 32.71 | 68.35 | 81.79 | 72.37 |
| Segmenter [31] | ICCV'21 | | 84.53 | 50.77 | 43.63 | 82.99 | 87.15 | 84.41 |
| SegFormer [7] | NeurIPS'21 | | 88.28 | 59.23 | 51.39 | 87.15 | 90.85 | 88.46 |
| Mask2Former [32] | CVPR'22 | | 84.35 | 54.33 | 45.87 | 84.74 | 89.12 | 85.92 |
| RTFNet [33] | RAL'19 | Feature-Fusion | 71.61 | 39.26 | 30.35 | 69.52 | 85.16 | 74.28 |
| SNE-RoadSeg [3] | ECCV'20 | | 79.46 | 51.91 | 41.56 | 81.45 | 87.05 | 82.91 |
| OFF-Net [34] | ICRA'22 | | 75.84 | 40.13 | 33.13 | 77.48 | 72.19 | 70.62 |
| RoadFormer [14] | TIV'24 | | 90.05 | 62.34 | 55.13 | 91.65 | 91.39 | 91.11 |
| RoadFormer+ [23] | TIV'24 | | 91.35 | 66.29 | 57.69 | 91.24 | 92.07 | 91.47 |
| DFormer [35] | ICLR'24 | | 90.59 | 69.01 | 58.18 | 91.20 | 93.58 | 91.96 |
| DispSegNet [16] | RA-L'18 | Joint Learning | 88.03 | 62.54 | 53.44 | 88.22 | 92.53 | 89.68 |
| DSNet [9] | ICRA'19 | | 89.48 | 64.44 | 52.83 | 89.41 | 93.25 | 90.73 |
| SG-RoadSeg [19] | ICRA'24 | | 86.51 | 61.10 | 52.02 | 88.46 | 80.10 | 84.96 |
| $S^3$M-Net [5] | TIV'24 | | 90.66 | 65.90 | 57.80 | 90.85 | 93.55 | 91.80 |
| **TiCoSS (Ours)** | TASE'25 | | **91.90** | **71.97** | **63.63** | **92.43** | **94.10** | **92.90** |

TABLE II

QUANTITATIVE COMPARISONS OF SOTA SEMANTIC SEGMENTATION NETWORKS ON THE VKITTI2 [27] DATASET. THE BEST RESULTS ARE SHOWN IN BOLD FONT. THE SYMBOL ↑ INDICATES THAT A HIGHER VALUE CORRESPONDS TO BETTER PERFORMANCE

| Networks | Publication | Type | Acc (%) ↑ | mAcc (%) ↑ | mIoU (%) ↑ | Pre (%) ↑ | Rec (%) ↑ | mFSc (%) ↑ |
|---|---|---|---|---|---|---|---|---|
| DeepLabv3+ [30] | ECCV'18 | Single-Modal | 92.19 | 63.15 | 56.90 | 89.00 | 92.71 | 90.16 |
| BiSeNet V2 [48] | IJCV'21 | | 81.77 | 51.07 | 44.45 | 83.23 | 82.19 | 80.67 |
| Segmenter [31] | ICCV'21 | | 90.39 | 60.33 | 52.99 | 88.05 | 87.89 | 87.70 |
| SegFormer [7] | NeurIPS'21 | | 94.75 | 70.56 | 64.98 | 93.57 | 93.62 | 93.46 |
| Mask2Former [32] | CVPR'22 | | 89.29 | 64.58 | 57.14 | 90.75 | 87.23 | 87.19 |
| RTFNet [33] | RAL'19 | Feature-Fusion | 85.22 | 49.47 | 42.59 | 83.74 | 89.17 | 84.41 |
| SNE-RoadSeg [3] | ECCV'20 | | 83.64 | 60.85 | 52.56 | 83.44 | 81.66 | 77.77 |
| OFF-Net [34] | ICRA'22 | | 90.84 | 61.51 | 55.27 | 89.24 | 86.71 | 86.15 |
| RoadFormer [14] | TIV'24 | | 97.54 | 86.58 | 80.83 | 96.99 | 96.86 | 96.91 |
| RoadFormer+ [23] | TIV'24 | | 98.45 | 89.79 | 84.03 | 96.36 | 96.94 | 97.77 |
| DFormer [35] | ICLR'24 | | 97.79 | 89.06 | 85.54 | 96.41 | 94.67 | 95.58 |
| DispSegNet [16] | RA-L'18 | Joint Learning | 97.15 | 82.63 | 80.24 | 97.39 | 97.66 | 97.46 |
| DSNet [9] | ICRA'19 | | 95.71 | 78.71 | 77.47 | 97.20 | 96.66 | 96.81 |
| SG-RoadSeg [19] | ICRA'24 | | 95.17 | 85.57 | 80.91 | 86.10 | 88.50 | 86.31 |
| $S^3$M-Net [5] | TIV'24 | | 98.32 | 88.24 | 84.18 | 98.37 | 98.28 | 98.31 |
| **TiCoSS (Ours)** | TASE'25 | | **98.69** | **91.66** | **88.46** | **98.55** | **98.67** | **98.57** |

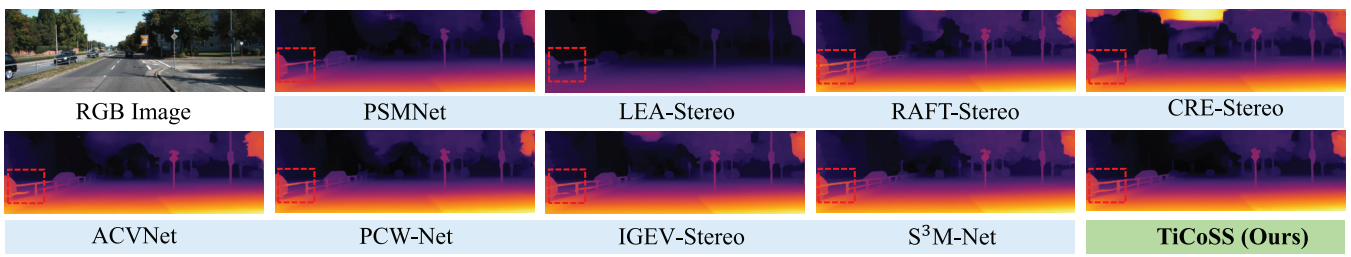


Fig. 5. Qualitative experimental results achieved by the SoTA stereo matching networks on the KITTI 2015 [28] dataset.

semantic segmentation performance, the stereo matching performance of TiCoSS is slightly better than that of $S^3$M-Net. Specifically, compared to $S^3$M-Net, TiCoSS demonstrates improvements of 3.64% in EPE and 2.47% in PEP 3.0 on the KITTI dataset. Additionally, on the vKITTI2 dataset, it achieves improvements of 5.26% in EPE and 1.26% in

TABLE III

QUANTITATIVE COMPARISONS OF SOTA SEMANTIC SEGMENTATION NETWORKS ON THE CITYSCAPES [29] DATASET. THE BEST RESULTS ARE SHOWN IN BOLD FONT. THE SYMBOL ↑ INDICATES THAT A HIGHER VALUE CORRESPONDS TO BETTER PERFORMANCE. VALUES MARKED WITH "-" DENOTE THAT THE CORRESPONDING METRICS WERE NOT REPORTED IN THE ORIGINAL PAPER

| Networks | Publication | Type | Acc (%) ↑ | mAcc (%) ↑ | mIoU (%) ↑ | Pre (%) ↑ | Rec (%) ↑ | mFSc (%) ↑ |
|---|---|---|---|---|---|---|---|---|
| DeepLabv3+ [30] | ECCV'18 | Single-Modal | 84.11 | 62.75 | 50.15 | 71.12 | 57.34 | 59.96 |
| BiSeNet V2 [48] | IJCV'21 | | 85.61 | 61.16 | 49.65 | 71.00 | 57.61 | 62.50 |
| Segmenter [31] | ICCV'21 | | 90.19 | 79.87 | 62.76 | 76.90 | 80.99 | 77.19 |
| SegFormer [7] | NeurIPS'21 | | 88.12 | 73.79 | 62.48 | 75.50 | 78.81 | 71.02 |
| Mask2Former [32] | CVPR'22 | | 87.97 | 78.80 | 64.29 | 75.01 | 78.26 | 74.71 |
| RTFNet [33] | RAL'19 | Feature-Fusion | 88.41 | 67.16 | 50.19 | 69.41 | 70.07 | 67.40 |
| SNE-RoadSeg [3] | ECCV'20 | | 86.60 | 75.47 | 63.01 | 70.56 | 74.19 | 75.58 |
| OFF-Net [34] | ICRA'22 | | 76.88 | 60.03 | 43.11 | 70.68 | 71.01 | 69.65 |
| RoadFormer [14] | TIV'24 | | 87.11 | 75.49 | 62.20 | 74.49 | 76.84 | 75.50 |
| RoadFormer+ [23] | TIV'24 | | 90.40 | 78.86 | 64.18 | 77.19 | 80.06 | 76.60 |
| DFormer [35] | ICLR'24 | | **91.37** | 80.16 | 65.59 | 76.60 | 79.65 | 77.41 |
| DispSegNet [16] | RA-L'18 | Joint Learning | 84.07 | 62.18 | 56.10 | 70.00 | 60.92 | 61.16 |
| DSNet [9] | ICRA'19 | | 84.01 | 62.25 | 52.79 | 69.56 | 63.74 | 61.26 |
| MENet [12] | TITS'21 | | 93.39 | 69.53 | 61.50 | - | - | - |
| SG-RoadSeg [19] | ICRA'24 | | 85.80 | 66.95 | 54.18 | 74.10 | 67.92 | 70.10 |
| $S^3$M-Net [5] | TIV'24 | | 88.47 | 77.30 | 62.59 | 75.20 | 77.48 | 76.30 |
| **TiCoSS (Ours)** | TASE'25 | | 90.70 | **81.76** | **68.36** | **78.43** | **81.76** | **79.74** |

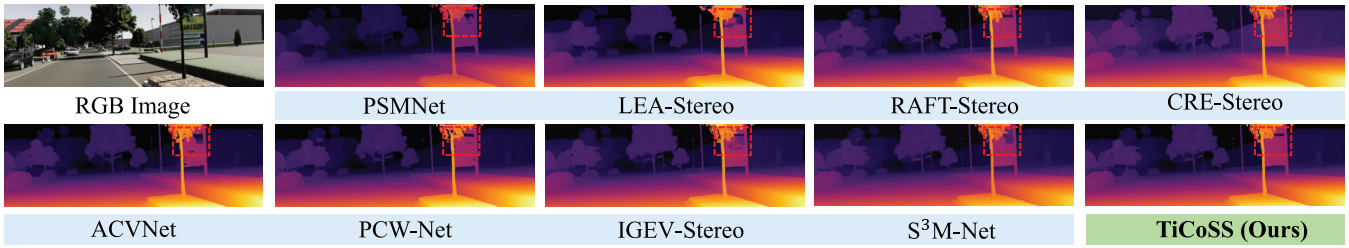


Fig. 6. Qualitative experimental results achieved by the SoTA stereo matching networks on the vKITTI2 [27] dataset.

TABLE IV

QUANTITATIVE COMPARISONS BETWEEN TICOSS AND THE BASELINE $S^3$M-NET, EVALUATED USING THE MMSEGMENTATION FRAMEWORK. THE BEST RESULTS ARE SHOWN IN BOLD FONT. THE SYMBOL ↑ INDICATES THAT A HIGHER VALUE CORRESPONDS TO BETTER PERFORMANCE

| Methods | mIoU (%) ↑ | | |
|---|---|---|---|
| | vKITTI2 | KITTI 2015 | Cityscapes |
| $S^3$M-Net [5] | 84.00 | 41.54 | 55.40 |
| **TiCoSS (Ours)** | **87.94** | **47.66** | **62.16** |

PEP 1.0. These improvements can be attributed to the tighter coupling between the two tasks, resulting in more comprehensive geometric features with informative contextual information compared to $S^3$M-Net. Additionally, by minimizing the CT loss, our model focuses more on areas with inconsistent disparities and achieves improved performance in occluded regions, as depicted in Fig. 6.

## E. Ablation Studies

*1) Heterogeneous Feature Fusion Strategy:* Extensive experiments are conducted on the KITTI 2015 [28] dataset to validate the effectiveness of our proposed TGF strategy, compared with two other SoTA feature fusion strategies: adaptive spatial feature fusion (ASFF) [52] and GFF [44]. As presented in Table VI, TGF outperforms both ASFF and GFF, with an increase in mIoU by up to 7.93%, and an increase in fwIoU by 8.26%. These compelling results demonstrate the capability of TGF for effective feature extraction and for selective feature propagation. Additionally, we demonstrate the rationality of employing feature fusion strategies within each encoding branch. It is evident that all models perform best when feature fusion is adopted in both branches, demonstrating the necessity to selectively extract and fuse heterogeneous features within both encoding branches.

*2) Deep Supervision Strategy:* We first compare the performance of our proposed HDS strategy with single-resolution deep supervision (SDS) [25], full-scale deep supervision (FDS) [46], and the combination of them (referred to SDS + FDS). As presented in Table VII, our proposed HDS achieves the SoTA performance across both evaluation metrics, with improvements of up to 5.59% in mIoU and 2.02% in fwIoU. These compelling results can be attributed to the improved interactions among auxiliary classifiers, leveraging fused features with the richest local spatial details in the main branch

TABLE V

COMPARISONS OF SOTA STEREO MATCHING NETWORK ON THE KITTI 2015 [28] AND VKITTI2 [27] DATASETS. THE SYMBOL ↓ INDICATES THAT A LOWER VALUE CORRESPONDS TO BETTER PERFORMANCE. THE BEST RESULTS ARE SHOWN IN BOLD FONT

| Networks | Publications | vKITTI2 [27] | | | KITTI 2015 [28] | | |
|---|---|---|---|---|---|---|---|
| | | EPE (pixels) ↓ | PEP (%) ↓ | | EPE (pixels) ↓ | PEP (%) ↓ | |
| | | | > 1 pixel | > 3 pixels | | > 1 pixel | > 3 pixels |
| PSMNet [36] | CVPR'18 | 0.68 | 10.31 | 3.77 | 0.74 | 16.12 | 2.61 |
| LEA-Stereo [37] | NeurIPS'20 | 0.83 | 13.33 | 4.84 | 0.83 | 18.67 | 3.22 |
| RAFT-Stereo [8] | 3DV'21 | 0.40 | 5.88 | 2.67 | 0.60 | 10.78 | 1.96 |
| CRE-Stereo [38] | CVPR'22 | 0.63 | 10.35 | 3.90 | 0.92 | 19.68 | 3.35 |
| ACVNet [49] | CVPR'22 | 0.61 | 9.41 | 3.45 | 0.68 | 13.93 | 2.10 |
| PCW-Net [50] | ECCV'22 | 0.63 | 9.45 | 3.49 | 0.70 | 14.81 | 2.43 |
| IGEV-Stereo [51] | CVPR'23 | 0.47 | 7.15 | 3.09 | 0.62 | 12.15 | 1.99 |
| DispSegNet [16] | RA-L'18 | 0.50 | 6.37 | 2.91 | 0.81 | 14.47 | 2.69 |
| DSNet [9] | ICRA'19 | 0.64 | 8.42 | 3.82 | 0.73 | 15.10 | 2.92 |
| $S^3$M-Net [5] | TIV'24 | 0.38 | 5.56 | **2.55** | 0.55 | **10.02** | 1.62 |
| **TiCoSS (Ours)** | TASE'25 | **0.34** | **5.43** | 2.58 | **0.54** | 10.39 | **1.60** |

TABLE VI

QUANTITATIVE COMPARISONS BETWEEN OUR PROPOSED TGF STRATEGY AND TWO SOTA FEATURE FUSION STRATEGIES, ASFF AND GFF, ONTHE KITTI 2015 [28] DATASET. "**Baseline**": $S^3$M-NET W/O SCG LOSS [5]. THE SYMBOL ↑ INDICATES THAT A HIGHER VALUE CORRESPONDS TO BETTER PERFORMANCE. THE BEST RESULTS ARE SHOWN IN BOLD FONT

| Feature Fusion Strategies | Disparity Branch | RGB Branch | mIoU (%) ↑ | fwIoU (%) ↑ |
|---|---|---|---|---|
| Baseline | | | 54.33 | 83.44 |
| Baseline + ASFF [52] | ✓ | | 54.72 | 83.10 |
| | | ✓ | 54.16 | 78.25 |
| | ✓ | ✓ | 55.92 | 83.88 |
| Baseline + GFF [44] | ✓ | | 57.13 | 84.33 |
| | | ✓ | 57.66 | 83.69 |
| | ✓ | ✓ | 58.03 | 84.39 |
| Baseline + TGF (**Ours**) | ✓ | | 55.80 | 84.30 |
| | | ✓ | 58.44 | 82.87 |
| | ✓ | ✓ | **59.06** | **84.72** |

TABLE VII

ABLATION STUDY ON OUR HDS STRATEGY ON THE KITTI 2015 [28] DATASET. "**Baseline**": $S^3$M-NET [5] ENHANCED BY OUR TGF STRATEGY. THE SYMBOL ↑ INDICATES THAT A HIGHER VALUE CORRESPONDS TO BETTER PERFORMANCE. THE BEST RESULTS ARE SHOWN IN BOLD FONT

| Methods | mIoU (%) ↑ | fwIoU (%) ↑ |
|---|---|---|
| Baseline | 59.06 | 84.72 |
| Baseline + SDS [25] | 60.01 | 84.79 |
| Baseline + FDS [46] | 60.62 | 85.20 |
| Baseline + SDS + FDS | 60.86 | 85.59 |
| Baseline + HDS (**Ours**) | **62.36** | **86.33** |

to guide deep supervision across side branches. Additionally, it is noteworthy that the straightforward combination of SDS and FDS results in a marginal improvement compared to using FDS alone. Additionally, SDS is adopted only at the highest resolution, where features have a low number of channels, resulting in almost no additional memory usage. Therefore, we incorporate SDS into our HDS. Furthermore, we conduct an additional ablation study to evaluate the performance of HDS when utilizing guidance features in different layers and of various types. As shown in Table VIII, the fused features at the shallowest layer result in the best performance compared to another two. We attribute this superior performance to the rich, fine-grained local spatial details that $\mathbf{F}_1^F$ contains, which are essential for semantic segmentation.

TABLE VIII

ABLATION STUDY ON THE SELECTION OF THE GUIDANCE FEATURES EMPLOYED IN OUR HDS STRATEGY ON THE KITTI 2015 [28] DATASET. "**FeatureLayer**": THE LAYER INDEX OF THE GUIDANCE FEATURE, WITH OPTIONS OF 1, 2, AND 3. "**GF**": GEOMETRIC FEATURES, "**CF**": CONTEXTUAL FEATURES, "**FF**": FUSED FEATURES, OBTAINED BY PERFORMING AN ELEMENT-WISE SUMMATION OF THE GEOMETRIC AND CONTEXTUAL FEATURES. THE SYMBOL ↑ INDICATES THAT A HIGHER VALUE CORRESPONDS TO BETTER PERFORMANCE. THE BEST RESULTS ARE SHOWN IN BOLD FONT

| Feature Layer | | | Method | | | mIoU (%) ↑ | fwIoU (%) ↑ |
|---|---|---|---|---|---|---|---|
| 1 | 2 | 3 | GF | CF | FF | | |
| ✓ | | | ✓ | | | 61.74 | **86.88** |
| ✓ | | | | ✓ | | 62.09 | 86.48 |
| ✓ | | | | | ✓ | **62.36** | 86.33 |
| | ✓ | | ✓ | | | 61.18 | 85.52 |
| | ✓ | | | ✓ | | 61.12 | 86.85 |
| | ✓ | | | | ✓ | 62.11 | 86.31 |
| | | ✓ | ✓ | | | 56.67 | 83.68 |
| | | ✓ | | ✓ | | 59.05 | 84.11 |
| | | ✓ | | | ✓ | 60.26 | 85.60 |

*3) Loss Function:* We systematically incorporate each element into the CT loss to assess its influence on the overall performance. Experimental results presented in Table IX validate the effectiveness of each component of our proposed CT loss. It is evident that the DIA and DSCC losses make significant contributions to semantic segmentation. Notably,

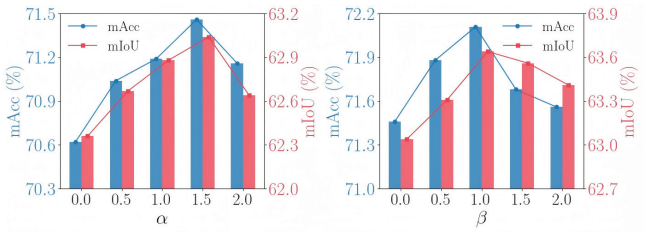


Fig. 7. The selection of hyperparameters $\alpha$ and $\beta$ within the CT loss on the KITTI 2015 [28] dataset.

TABLE IX

ABLATION STUDY TO VALIDATE THE EFFECTIVENESS OF THE THREE SEMANTIC SEGMENTATION LOSSES WITHIN OUR CT LOSS ON THE KITTI 2015 [28] DATASET. THE SYMBOL ↑ INDICATES THAT A HIGHER VALUE CORRESPONDS TO BETTER PERFORMANCE. THE BEST RESULTS ARE SHOWN IN BOLD FONT

| DIA | DCSS | SCG | mIoU (%) ↑ | fwIoU (%) ↑ |
|---|---|---|---|---|
| | | | 62.36 | 86.33 |
| ✓ | | | 63.04 | 86.48 |
| | ✓ | | 62.88 | 85.98 |
| | | ✓ | 62.75 | 86.66 |
| ✓ | ✓ | | 63.54 | 86.50 |
| ✓ | | ✓ | 63.15 | 86.7 |
| | ✓ | ✓ | 62.99 | 86.59 |
| ✓ | ✓ | ✓ | **63.63** | **86.68** |

when the entire joint learning framework is trained by minimizing the CT loss, TiCoSS achieves the best performance on the KITTI dataset, attaining a maximum mIoU of 63.63%. Additionally, to maximize the effectiveness of our proposed loss function, we first conduct an ablation study on the selection of loss weight $\alpha$ in (7). Fig. 7 shows the mAcc and mIoU values with respect to different $\alpha$ within the range of 0.0 to 2.0. It can be obviously observed that when $\alpha = 1.5$, TiCoSS achieves the best overall performance for both evaluation metrics. Following the selection of $\alpha$, we conduct another ablation study to determine $\beta$ in (8). Fig. 7 demonstrates that $\beta = 1.0$ is the best choice. Further weight tuning is possible, but it should be approached cautiously, especially when dealing with limited data to avoid over-fitting.

*4) All Contributions:* We explore the rationality of each contribution adopted in our TiCoSS. As presented in Table X, we sequentially incorporate each novel component to assess its impact on the overall performance. Please note that our CT loss is based on HDS strategy. Therefore, no ablation study needs to be conducted solely with the CT loss. The results demonstrate that each component employed in our model contributes to an improvement in the overall performance, and the network achieves its peak performance when all novel contributions (TGF, HDS, and CT loss) are leveraged, demonstrating the effectiveness of our design.

### F. Efficiency Analysis

Additionally, TiCoSS contains 385.05 million trainable parameters and requires 308.86 GFLOPs to process an image with a resolution of 512 × 256 pixels. When deployed on an NVIDIA GeForce RTX 3090 GPU paired with an Intel Core i7-13700KF processor, it achieves an inference speed of 0.30 seconds per image, with a memory usage of around 5.82 GB. We believe that the further reduction of TiCoSS's computational complexity is essential for deployment on resource-constrained hardware.

TABLE X

ABLATION STUDY ON THE THREE CONTRIBUTIONS ON THE KITTI 2015 [28] DATASET. THE SYMBOL ↑ INDICATES THAT A HIGHER VALUE CORRESPONDS TO BETTER PERFORMANCE. THE BEST RESULTS ARE SHOWN IN BOLD FONT

| TGF | HDS | CT loss | mIoU (%) ↑ | mFSc (%) ↑ |
|---|---|---|---|---|
| ✓ | | | 59.06 | 91.26 |
| | ✓ | | 58.49 | 92.10 |
| ✓ | ✓ | | 62.36 | 92.74 |
| | ✓ | ✓ | 61.18 | 92.46 |
| ✓ | ✓ | ✓ | **63.63** | **92.90** |

### G. Additional Experiments

We conduct several additional experiments, as detailed in the supplement to comprehensively validate the effectiveness of TiCoSS. Specifically, we evaluate TiCoSS's robustness by conducting extensive experiments under various weather conditions and challenging scenarios. Moreover, we submit our results to the KITTI Semantics benchmark to compare TiCoSS with methods whose codes are not publicly available. These additional qualitative and quantitative experimental results further demonstrate the superior performance of TiCoSS.

## V. CONCLUSION AND FUTURE WORK

This article introduced TiCoSS, a novel, high-performing, and state-of-the-art joint learning framework designed to tighten the coupling between the semantic segmentation and stereo matching tasks. We made three key contributions in this work: (1) an effective feature fusion strategy and a tightly-coupled duplex encoder, leveraging the informative spatial information to enhance the semantic segmentation task, (2) a novel hierarchical deep supervision strategy that improves the interactions among all auxiliary classifiers, and (3) a joint learning loss that focuses on further tightening the coupling of these two tasks at the output level. The effectiveness of each contribution was validated through extensive experiments.

Despite its superior performance over existing approaches, TiCoSS still requires both semantic and disparity annotations, and collecting data with such ground truth remains a labor-intensive process. Thus, exploring semi-supervised or few-shot semantic segmentation methods, and un/self-supervised stereo matching methods is a promising avenue for our future research.

## REFERENCES

[1] Ö. Ciftcioglu et al., "Studies on visual perception for perceptual robotics," in *Proc. Int. Conf. Robot. Autom. (ICRA)*, vol. 2, 2006, pp. 352–359.

[2] F. Wei and W. Wang, "A method for designing the perception module of autonomous vehicles using stereo depth and semantic segmentation," in *Proc. 24th Int. Conf. Control, Autom. Syst. (ICCAS)*, Oct. 2024, pp. 1423–1428.

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

[4] 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.

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

[6] J. Huang, J. Li, S. Vityazev, A. Dvorkovich, and R. Fan, "DepthMatch: Semi-supervised RGB-D scene parsing through depth-guided regularization," *IEEE Signal Process. Lett.*, early access, Jun. 2, 2025, doi: 10.1109/LSP.2025.3575640.

[7] E. Xie et al., "SegFormer: Simple and efficient design for semantic segmentation with transformers," in *Proc. Adv. Neural Inf. Process. Sys. (NIPS)*, vol. 34, Dec. 2021, pp. 12077–12090.

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

[9] W. Zhan, X. Ou, Y. Yang, and L. Chen, "DSNet: Joint learning for scene segmentation and disparity estimation," in *Proc. Int. Conf. Robot. Autom. (ICRA)*, May 2019, pp. 2946–2952.

[10] X. Guan, W. Tong, S. Jiang, P. Z. H. Sun, E. Q. Wu, and G. Chen, "Multistage pixel-visibility learning with cost regularization for multi-view stereo," *IEEE Trans. Autom. Sci. Eng.*, vol. 20, no. 2, pp. 751–762, Apr. 2023.

[11] Z. Wu, X. Wu, X. Zhang, S. Wang, and L. Ju, "Semantic stereo matching with pyramid cost volumes," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 7483–7492.

[12] X. Zhang, Y. Chen, H. Zhang, S. Wang, J. Lu, and J. Yang, "When visual disparity generation meets semantic segmentation: A mutual encouragement approach," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 3, pp. 1853–1867, Mar. 2021.

[13] Y. Sun, W. Zuo, P. Yun, H. Wang, and M. Liu, "FuseSeg: Semantic segmentation of urban scenes based on RGB and thermal data fusion," *IEEE Trans. Autom. Sci. Eng.*, vol. 18, no. 3, pp. 1000–1011, Jul. 2021.

[14] J. Li, Y. Zhang, P. Yun, G. Zhou, Q. Chen, and R. Fan, "RoadFormer: Duplex transformer for RGB-normal semantic road scene parsing," *IEEE Trans. Intell. Vehicles*, vol. 9, no. 7, pp. 5163–5172, Jul. 2024, doi: 10.1109/TIV.2024.3388726.

[15] J. Zhang, H. Liu, K. Yang, X. Hu, R. Liu, and R. Stiefelhagen, "CMX: Cross-modal fusion for RGB-X semantic segmentation with transformers," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 12, pp. 14679–14694, Dec. 2023.

[16] J. Zhang, K. A. Skinner, R. Vasudevan, and M. Johnson-Roberson, "DispSegNet: Leveraging semantics for end-to-end learning of disparity estimation from stereo imagery," *IEEE Robot. Autom. Lett.*, vol. 4, no. 2, pp. 1162–1169, Apr. 2019.

[17] G. Yang, H. Zhao, J. Shi, Z. Deng, and J. Jia, "SegStereo: Exploiting semantic information for disparity estimation," in *Proc. Eur. Conf. Comput. Vis.*, Aug. 2018, pp. 636–651.

[18] P. L. Dovesi et al., "Real-time semantic stereo matching," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2020, pp. 10780–10787.

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

[20] M.-J. Lee et al., "SG-RoadSeg+: End-to-end freespace detection upgraded at data, feature, and loss levels," *IEEE Trans. Instrum. Meas.*, early access, Jun. 16, 2025, doi: 10.1109/TIM.2025.3579733.

[21] W. Zhou, Y. Xiao, W. Yan, and L. Yu, "CMPFFNet: Cross-modal and progressive feature fusion network for RGB-D indoor scene semantic segmentation," *IEEE Trans. Autom. Sci. Eng.*, vol. 21, no. 4, pp. 5523–5533, Oct. 2024.

[22] 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.

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

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

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

[26] D. Jia, Y. Pang, J. Cao, and J. Pan, "SSNet: A joint learning network for semantic segmentation and disparity estimation," *Vis. Comput.*, vol. 41, no. 1, pp. 423–435, Apr. 2024.

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

[28] M. Menze and A. Geiger, "Object scene flow for autonomous vehicles," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3061–3070.

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

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

[31] R. Strudel, R. Garcia, I. Laptev, and C. Schmid, "Segmenter: Transformer for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 7262–7272.

[32] 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. (CVPR)*, Jun. 2022, pp. 1290–1299.

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

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

[35] B. Yin, X. Zhang, Z. Li, L. Li, M. Cheng, and Q. Hou, "DFormer: Rethinking RGBD representation learning for semantic segmentation," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, Jan. 2024.

[36] J. R. Chang and Y. S. Chen, "Pyramid stereo matching network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5410–5418.

[37] X. Cheng et al., "Hierarchical neural architecture search for deep stereo matching," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 33, Jan. 2020, pp. 22158–22169.

[38] J. Li et al., "Practical stereo matching via cascaded recurrent network with adaptive correlation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 16263–16272.

[39] Z. Teed and J. Deng, "RAFT: Recurrent all-pairs field transforms for optical flow," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, Aug. 2020, pp. 402–419.

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

[41] S. Chen, Z. Xiang, C. Qiao, Y. Chen, and T. Bai, "SGNet: Semantics guided deep stereo matching," in *Proc. Asian Conf. Comput. Vis. (ACCV)*, 2020, pp. 106–122.

[42] N. Mayer et al., "A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4040–4048.

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

[44] X. Li, H. Zhao, L. Han, Y. Tong, S. Tan, and K. Yang, "Gated fully fusion for semantic segmentation," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 7, pp. 11418–11425.

[45] S. Y. Boulahia, A. Amamra, M. R. Madi, and S. Daikh, "Early, intermediate and late fusion strategies for robust deep learning-based multimodal action recognition," *Mach. Vis. Appl.*, vol. 32, no. 6, p. 121, Nov. 2021.

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

[47] D. Li and Q. Chen, "Dynamic hierarchical mimicking towards consistent optimization objectives," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 7639–7648.

[48] C. Yu, C. Gao, J. Wang, G. Yu, C. Shen, and N. Sang, "BiSeNet v2: Bilateral network with guided aggregation for real-time semantic segmentation," *Int. J. Comput. Vis.*, vol. 129, no. 11, pp. 3051–3068, Nov. 2021.
[49] G. Xu, J. Cheng, P. Guo, and X. Yang, "Attention concatenation volume for accurate and efficient stereo matching," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 12981–12990.
[50] Z. Shen, Y. Dai, X. Song, Z. Rao, D. Zhou, and L. Zhang, "PCW-Net: Pyramid combination and warping cost volume for stereo matching," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2022, pp. 280–297.
[51] G. Xu, X. Wang, X. Ding, and X. Yang, "Iterative geometry encoding volume for stereo matching," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 21919–21928.
[52] S. Liu, D. Huang, and Y. Wang, "Learning spatial fusion for single-shot object detection," 2019, *arXiv:1911.09516*.

**Guanfeng Tang** is currently pursuing the B.E. degree with Tongji University, where he will pursue the Ph.D. degree with the MIAS Group, supervised by Prof. Rui Fan. His research interests include computer vision, robotics, and deep learning, with a particular emphasis on joint learning.

**Zhiyuan Wu** (Graduate Student Member, IEEE) received the B.E. degree in automation from Tongji University, Shanghai, China, in 2024. He is currently pursuing the Ph.D. degree with the Robot Perception Lab, King's College London, supervised by Dr. Shan Luo. His research interests include multi-modal perception, 3-D vision, and robot manipulation.

**Jiahang Li** (Member, IEEE) received the B.E. degree in automation from Taiyuan University of Technology in 2021 and the M.Sc. degree in control science and engineering from Tongji University in 2025. His research interests include computer vision and deep learning.

**Ping Zhong** (Member, IEEE) received the B.S. degree from the National University of Defense Technology, Changsha, China, in 2004, and the Ph.D. degree in communication engineering from Xiamen University, China, in 2011. From 2012 to 2013, she was a Post-Doctoral Researcher with the Computer Laboratory, University of Cambridge. She is currently an Associate Professor with the Department of Computer Science and Technology, Central South University. Her research interests include autonomous uncrewed systems and machine learning.

**Wei Ye** received the Ph.D. degree in computer science from the Institut für Informatik, Ludwig-Maximilians-Universität München, Munich, Germany, in 2018. He is currently a Tenure-Track Professor with the College of Electronics and Information Engineering, Tongji University; the Frontier Science Center for Intelligent Autonomous Systems, Ministry of Education, Shanghai; and Shanghai Innovation Institute. From 2018 to 2020, he was a Post-Doctoral Researcher with the DYNAMO Laboratory, University of California at Santa Barbara. Before that, he was a Researcher with the Department of AI Platform, Tencent, China. His research interests include data mining, graph machine learning, deep learning, and network science.

**Xieyuanli Chen** (Member, IEEE) received the B.E. degree in electrical engineering and automation from Hunan University in 2015, the master's degree in robotics from the National University of Defense Technology in 2017, and the Ph.D. degree from the Photogrammetry and Robotics Laboratory, University of Bonn. He is an Associate Professor at the National University of Defense Technology, China. His research interests include SLAM and robot perception. He serves as an Associate Editor for IEEE ROBOTICS AND AUTOMATION LETTERS, ICRA, and IROS.

**Huimin Lu** (Member, IEEE) received the B.E. degree in automation and the M.E. and Ph.D. degrees in control science and engineering from the National University of Defense Technology, Changsha, China, in 2003, 2005, and 2010, respectively. He joined the College of Intelligence Science and Technology, National University of Defense Technology, in 2010, where he is currently a Professor. His research interests include mobile robotics, mainly robot vision, multi-robot coordination, and human–robot interaction.

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