S^{3}M-Net: Joint Learning of Semantic Segmentation and Stereo Matching for Autonomous Driving

Zhiyuan Wu, Yi Feng, Chuang-Wei Liu, Fisher Yu, *Member, IEEE*, Qijun Chen, *Senior Member, IEEE*, and Rui Fan, *Senior Member, IEEE*

***Abstract*—Semantic segmentation and stereo matching are two essential components of 3D environmental perception systems for autonomous driving. Nevertheless, conventional approaches often address these two problems independently, employing separate models for each task. This approach poses practical limitations in real-world scenarios, particularly when computational resources are scarce or real-time performance is imperative. Hence, in this article, we introduce S^{3}M-Net, a novel joint learning framework developed to perform semantic segmentation and stereo matching simultaneously. Specifically, S^{3}M-Net shares the features extracted from RGB images between both tasks, resulting in an improved overall scene understanding capability. This feature sharing process is realized using a feature fusion adaption (FFA) module, which effectively transforms the shared features into semantic space and subsequently fuses them with the encoded disparity features. The entire joint learning framework is trained by minimizing a novel semantic consistency-guided (SCG) loss, which places emphasis on the structural consistency in both tasks. Extensive experimental results conducted on the vKITTI2 and KITTI datasets demonstrate the effectiveness of our proposed joint learning framework and its superior performance compared to other state-of-the-art single-task networks. Our project webpage is accessible at mias.group/S3M-Net.**

***Index Terms*—Autonomous driving, environmental perception, joint learning, semantic segmentation, stereo matching.**

I. Introduction

3D environmental perception stands as a critical and foundational aspect of autonomous driving [1], [2]. Semantic segmentation and stereo matching are two key functionalities in 3D environmental perception systems [3], [4], [5]. The former provides a comprehensive pixel-level understanding of the environment, while the latter simulates human binocular vision to acquire accurate and dense depth information [6]. The combined utilization of both functionalities has become the mainstream approach in recent years [7], [8], [9], [10], [11], [12].

In recent years, the research focus in semantic segmentation has shifted from single-modal networks [13], [14], [15], [16], [17], [18] with a single encoder to feature-fusion networks with dual encoders [19], [20], [21], [22]. The latter type of networks extract heterogeneous features from RGB-X data, where "X" can represent various forms of spatial geometric information, e.g., depth images generated from LiDAR point clouds and surface normal maps obtained through depth-to-normal translation [23]. These heterogeneous features are subsequently fused to achieve a more comprehensive understanding of the environment [21]. However, a critical drawback of feature-fusion networks is their dependency on the availability of the "X" data, which can pose limitations in scenarios where LiDARs are not present. Additionally, when the accuracy of the "X" data is not satisfactory, such as due to variations in camera-LiDAR calibration, the fusion of these heterogeneous features can potentially lead to a degradation in the overall performance of semantic segmentation [24]. While a stereo camera can serve as a practical and cost-effective alternative to LiDARs for depth information acquisition, the incorporation of a separate stereo matching network introduces additional computations, and therefore, poses difficulties in achieving real-time processing speeds for the entire system [9]. Moreover, stereo matching and semantic segmentation share the same input, and the representations from RGB images can be more informative when they are jointly learned by both tasks.

The joint learning of multiple interconnected 3D environmental perception tasks introduces a form of regularization that has demonstrated superiority over uniform complexity penalization in reducing over-fitting [25]. Furthermore, rather than employing separate models for semantic segmentation and stereo matching, joint learning can potentially reduce computational complexity [7], [8], [9], [10], [11], [12], as shared learning representations can be used for both tasks. This can be advantageous in real-time or resource-constrained applications. Moreover, joint learning enables end-to-end optimization of the entire system, allowing the model to adapt to the specific challenges of both tasks simultaneously. Consequently, this can lead to improved performance when compared to models trained separately for each task [12]. In addition, stereo matching

Manuscript received 28 December 2023; accepted 15 January 2024. Date of publication 23 January 2024; date of current version 29 April 2024. This work was supported in part by the Science and Technology Commission of Shanghai Municipal under Grant 22511104500, in part by the National Natural Science Foundation of China under Grant 62233013, in part by the Fundamental Research Funds for the Central Universities, and in part by Xiaomi Young Talents Program. *(Corresponding author: Rui Fan.)*

Zhiyuan Wu, Yi Feng, Chuang-Wei Liu, Qijun Chen, and Rui Fan are with the College of Electronics & Information Engineering, Shanghai Research Institute for Intelligent Autonomous Systems, The State Key Laboratory of Intelligent Autonomous Systems, and Frontiers Science Center for Intelligent Autonomous Systems, Tongji University, Shanghai 201804, China (e-mail: gwu@tongji.edu.cn; fengyi@ieee.org; cwliu@tongji.edu.cn; qjchen@tongji.edu.cn; rui.fan@ieee.org).

Fisher Yu is with the Department of Information Technology and Electrical Engineering, ETH Zürich, 8092 Zürich, Switzerland (e-mail: i@yf.io).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TIV.2024.3357056.

Digital Object Identifier 10.1109/TIV.2024.3357056

can occasionally produce ambiguously estimated disparities, particularly in texture-less or occluded regions [26]. Semantic segmentation can provide informative contextual information that helps disambiguate such cases, ultimately leading to more reliable disparity estimations [9]. Regrettably, the joint learning of semantic segmentation and stereo matching, especially within feature-fusion networks or when faced with a scarcity of training samples, has received relatively limited attention in this research area and calls for further investigation.

Therefore, in this article, we present **S**emantic **S**egmentation and **S**tereo **M**atching **Net**work (**S^{3}M-Net**), a joint framework to simultaneously predict both semantic and disparity information. S^{3}M-Net begins with the extraction of features from stereo images. These features are then processed by a multi-level gate recurrent unit (GRU) operator to generate a disparity map. Simultaneously, these features are shared with the semantic segmentation task via a feature fusion adaptation (FFA) module. Building upon our prior work SNE-RoadSeg [3], we extract additional features from the estimated disparity map. Finally, a densely-connected skip connection decoder is employed to decode the fused features and generate the semantic predictions. S^{3}M-Net is trained in a fully supervised manner by minimizing a semantic consistency-guided (SCG) joint learning loss. Extensive experiments conducted on the vKITTI2 [27] and KITTI 2015 [28] datasets unequivocally demonstrate the effectiveness and superior performance of our proposed S^{3}M-Net.

In summary, the main contributions of this article include:

- S^{3}M-Net, a joint learning framework designed to address semantic segmentation and stereo matching simultaneously, where both tasks collaboratively leverage the features extracted from RGB images, enhancing the overall understanding of the driving scenario;
- A feature fusion adaption module to transform the shared feature maps into semantic space and subsequently fuse them with encoded disparity features;
- A semantic consistency-guided loss function to supervise the training process of the joint learning framework, emphasizing on the structural consistency in both tasks.

The remainder of this article is organized as follows: Section II provides a review of related work. Section III introduces our proposed S^{3}M-Net. Section IV presents the experimental results and compares our framework with other state-of-the-art (SoTA) approaches. In Section V, we discuss the advantages and limitations of our method. Finally, we conclude this article in Section VI.

II. Literature Review

A. Semantic Segmentation

Semantic segmentation has been a long-standing problem in the field of computer vision over the past decade [6], [29]. The SoTA networks in this research area can generally be classified into two categories: (1) single-modal networks with a single encoder and (2) feature-fusion networks with multiple encoders [3], [30], [31]. In the early attempts to tackle semantic segmentation, researchers primarily focused on encoder-decoder architectures for pixel-level classification. Notable examples include SegNet [13], U-Net [14], PSP-Net [15], the DeepLab series [16], [17], and Transformer-based networks [32], [33], [34]. The encoder extracts hierarchical deep features from the input image, while the decoder produces the segmentation map by upsampling and combining the features from different encoder layers. However, these networks are limited in their ability to effectively combine deep features extracted from different modalities (or sources) of visual information. As a result, they often struggle to produce accurate segmentation results in challenging scenarios characterized by poor lighting and illumination conditions [3]. Therefore, researchers have turned their focus towards feature-fusion networks that can effectively integrate deep features learned from multiple modalities (or sources) of visual information. This problem is commonly 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 based on convolutional neural networks (CNNs) include FuseNet [19], MFNet [35], RTFNet [20], and our previous works SNE-RoadSeg series [3], [21]. Furthermore, Transformer-based RGB-X semantic segmentation networks, such as OFF-Net [22] and RoadFormer [24], have been recently introduced. In this article, we design our S^{3}M-Net based on the SNE-RoadSeg architecture and explore more effective solutions for the feature fusion operation.

B. Stereo Matching

Conventional explicit programming-based stereo matching algorithms (local, global, and semi-global) generally consist of four main procedures: (1) cost computation, (2) cost aggregation, (3) disparity optimization, and (4) disparity refinement [26]. The performance of these algorithms has been significantly outperformed by end-to-end deep stereo networks, thanks to the recent advancements in deep learning techniques. PSMNet [36], GwcNet [37], AANet [38], LEA-Stereo [39], RAFT-Stereo [40], and CRE-Stereo [41] are six representative end-to-end deep stereo networks proposed in recent years. PSMNet [36] employs a spatial pyramid to capture multi-scale information and employs multiple 3D convolutional layers to exploit both local and global contexts for cost computation. GwcNet [37], on the other hand, builds upon the foundation of PSMNet by constructing the cost volume via group-wise correlation, thereby enhancing the 3D stacked hourglass network. In light of the computational demands of 3D convolutions, researchers have actively sought ways to minimize the trade-off between efficiency and accuracy in stereo matching. For example, LEA-Stereo [39] introduces the neural architecture search (NAS) [42] technique to stereo matching. This pioneering approach results in the first end-to-end hierarchical NAS framework for deep stereo matching. RAFT-Stereo [40], a rectified stereo matching method that draws inspiration from the optical flow estimation network RAFT [43], leverages the RAFT architecture to perform accurate and real-time stereo matching inference. The network utilizes recurrent structures to refine correlation features and enhance the disparity estimation accuracy. CRE-Stereo [41],

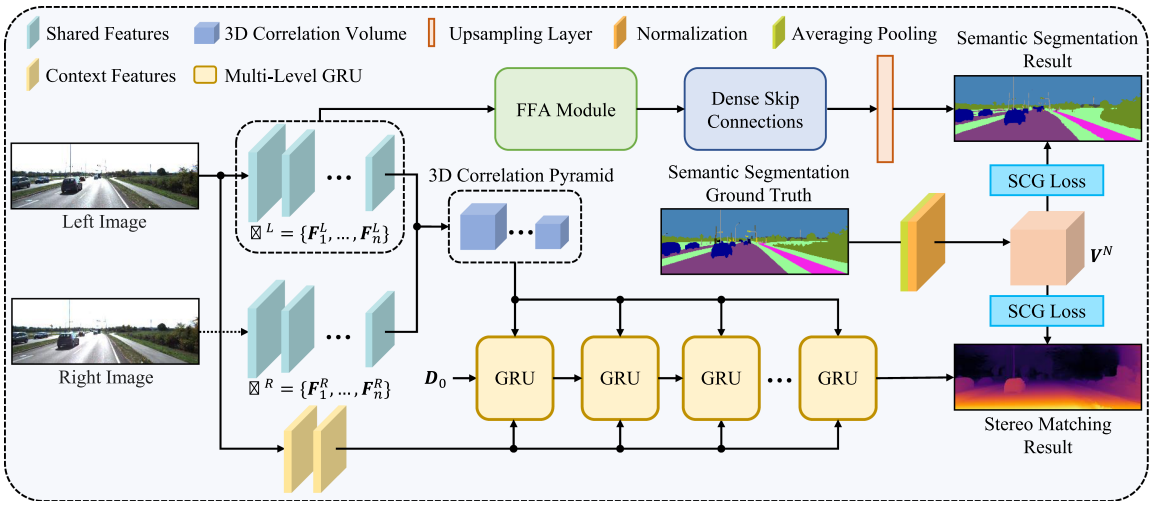

Fig. 1. Architecture of our proposed S^{3}M-Net for end-to-end joint learning of semantic segmentation and stereo matching.

another recent prior art based on recurrent refinement (to update disparities in a coarse-to-fine manner) and adaptive group correlation (to mitigate the impact of erroneous rectification), yields more compelling disparity estimation results. In this article, we develop our S^{3}M-Net based on the RAFT-Stereo architecture.

C. Multi-Task Joint Learning for Semantic Segmentation and Stereo Matching

Existing frameworks that jointly address semantic segmentation and stereo matching generally focus on improving disparity accuracy by leveraging semantic information [7], [8], [9], [10], [11], [12], while the discussion regarding the utilization of disparity information to enhance semantic segmentation at the feature level for joint learning remains limited, except for the exploration of "RGB-X semantic segmentation" discussed in Section II-A. Nevertheless, these prior arts either require a large amount of well-annotated training data or involve intricate training strategies for the joint learning of both tasks. For instance, SegStereo [7] and DispSegNet [8] require an initial unsupervised training phase on the large-scale Cityscapes [44] dataset, followed by a subsequent supervised fine-tuning on the smaller KITTI 2012 and 2015 [28], [45] datasets. Similarly, the studies presented in [9], [11], [12] involve the pre-training of their spatial branches (performing stereo matching) on the large-scale SceneFlow [46] dataset, followed by the fine-tuning of both semantic and spatial branches on the KITTI 2012 and 2015 datasets [28], [45]. DSNet [10] adopts a different joint learning strategy in which the training alternates between the semantic segmentation and stereo matching networks, with each network being frozen during the training of the other. However, achieving simultaneous convergence of the two networks can be challenging, as the shared features are not learned in an end-to-end manner. Additionally, we were unable to locate publicly available source code (in PyTorch or TensorFlow) for these prior arts, and re-implementations carry the risk of introducing errors. In contrast to the aforementioned approaches, our proposed S^{3}M-Net is trained in an end-to-end fashion and capable of jointly learning semantic segmentation and stereo matching even when the training data are limited.

III. Methodology

As illustrated in Fig. 1, our proposed S^{3}M-Net consists of five main components:

1) Joint encoder to extract shared features from RGB images;
2) Multi-level GRU update operator to refine disparity maps;
3) Feature fusion adaptation module to transform shared features into the semantic space and fuse them with features extracted from the disparity maps;
4) Densely-connected skip connection decoder to decode fused features and produce final semantic predictions;
5) Semantic consistency-guided loss to supervise the entire joint learning process.

A. Joint Encoder

Given a pair of well-rectified stereo images $\boldsymbol{I}^L$, $\boldsymbol{I}^R \in \mathbb{R}^{H\times W\times 3}$, where H and W denote their height and width, respectively, we employ a joint encoder consisting of a series of residual blocks and downsampling layers to extract features $\mathcal{F}^L = \{\boldsymbol{F}_1^L, \ldots, \boldsymbol{F}_n^L\}$ and $\mathcal{F}^R = \{\boldsymbol{F}_1^R, \ldots, \boldsymbol{F}_n^R\}$ from $\boldsymbol{I}^L$ and $\boldsymbol{I}^R$, respectively. $\mathcal{F}^L$ is subsequently shared with the semantic segmentation task.

B. Multi-Level GRU Update Operator

Using the features $\mathcal{F}^L$ and $\mathcal{F}^R$ extracted by the joint encoder, we first construct an initial 3D correlation volume $\boldsymbol{C}_1 \in \mathbb{R}^{H\times W\times W}$ as follows:

$$\boldsymbol{C}_1(i,j,k) = \boldsymbol{F}_n^L(i,j,:)\cdot \boldsymbol{F}_n^R(i,k,:), \tag{1}$$

where i represents the i-th row, and j and k represent to the j-th and k-th columns in the left and right shared feature maps, respectively. We then construct a pyramid of 3D correlation

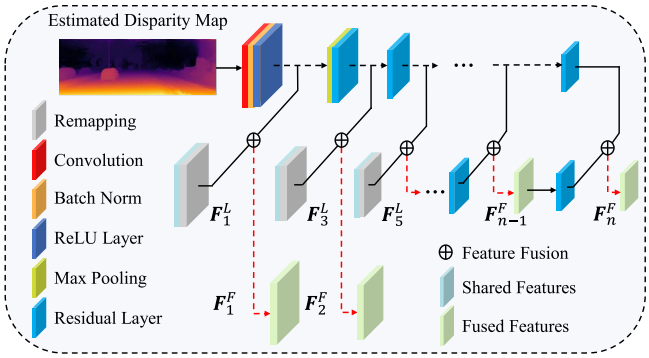

Fig. 2. Illustration of our proposed FFA module.

volumes $\mathcal{C} = \{\boldsymbol{C}_1, \ldots, \boldsymbol{C}_m\}$ by downsampling $\boldsymbol{C}_1$ with average pooling operations. The m-th 3D correlation volume $\boldsymbol{C}_m \in \mathbb{R}^{H \times W \times \frac{W}{2^{m-1}}}$ is constructed from the $(m-1)$-th 3D correlation volume $\boldsymbol{C}_{m-1}$ using 1D average pooling with a kernel size of 2 and a stride of 2. Inspired by RAFT-Stereo [40], we adopt a multi-level GRU update operator to refine a sequence of disparity maps $\mathcal{D} = \{\boldsymbol{D}_1, \ldots, \boldsymbol{D}_n\}$, where $\boldsymbol{D}_i \in \mathbb{R}^{H \times W}$ $(i = 1, \ldots, n)$. This refinement process is performed in a coarse-to-fine manner, starting from an initial disparity map $\boldsymbol{D}_0$ in which all disparities are initialized to 0.

C. Feature Fusion Adaptation Module

In stereo matching, a lower number of channels, e.g., the 256 channels utilized in RAFT-Stereo [40], is often sufficient for capturing relevant features for 1D correspondence search, especially when considering computational efficiency. On the other hand, semantic segmentation requires pixel-level classification and a more in-depth scene understanding. It benefits from complex feature representations that can capture fine-grained details and object boundaries, making a larger number of channels, e.g., the 2048 channels employed in SNE-RoadSeg [3], advantageous for this task. Therefore, we introduce the FFA module to align the channels and resolutions between the disparity and semantic feature maps during joint learning.

As illustrated in Fig. 2, given the left shared feature maps $\mathcal{F}^L = \{\boldsymbol{F}_1^L, \ldots, \boldsymbol{F}_n^L\}$ and the disparity map pyramid $\mathcal{D} = \{\boldsymbol{D}_1, \ldots, \boldsymbol{D}_n\}$, we obtain the adapted fused feature sequence $\mathcal{F}^F = \{\boldsymbol{F}_1^F, \ldots, \boldsymbol{F}_n^F\}$ using our proposed FFA module, which can be formulated as follows:

$$\boldsymbol{F}_i^F = \mathcal{A}_i(\mathcal{F}^L) \oplus \mathcal{E}_i^D(\boldsymbol{D}_n), \tag{2}$$

where $\mathcal{E}^D$ denotes the disparity map encoding operation, $\oplus$ denotes the feature fusion operation, and $\mathcal{A}_i$ is defined as our feature adaptation operation, as formulated as follows:

$$\mathcal{A}_i(\mathcal{F}^L) = \begin{cases} \mathcal{R}(\boldsymbol{F}_{2i-1}^L), & i \leq \frac{(n+1)}{2} \\ \mathcal{E}(\boldsymbol{F}_{i-1}^F \oplus \mathcal{E}_{i-1}^D(\boldsymbol{D}_n)), & i > \frac{(n+1)}{2} \end{cases}, \tag{3}$$

where $\mathcal{R}$ represents the remapping operation from the shared feature space to the semantic feature space, and $\mathcal{E}$ represents the encoding operation for the semantic feature maps.

Specifically, for the remapping operation $\mathcal{R}$, we employ 3×3 convolutional layers with a stride of 2 and padding of 1, each followed by a batch normalization layer and a rectified linear unit (ReLU) activation layer, adapting the feature map channels to 64, 256, and 512, respectively. Regarding the disparity encoding operation $\mathcal{E}^D$, we employ ResNet-152 [47] as the backbone network to extract features from the last disparity map $\boldsymbol{D}_n$. In ResNet-152, the first block consists of a convolutional layer, a batch normalization layer, and a ReLU activation layer. Then, a max pooling layer and four residual layers are sequentially applied to progressively increase the number of feature map channels.

Similarly, we utilize the residual block for the encoding operation $\mathcal{E}$ on the semantic feature maps, resulting in feature maps with 1024 and 2048 channels. The fused features $\mathcal{F}^F$ contain both texture and spatial geometric information, thereby enhancing semantic scene understanding. We conduct an ablation study for different feature fusion modules in Section IV-F.

D. Densely-Connected Skip Connection Decoder

We employ the decoder introduced in our previous work SNE-RoadSeg [3] to decode the fused features and generate the semantic prediction. In this encoder, three convolutional layers in the feature extractor and the upsampling layer share the same parameters: a 3×3 kernel size, a stride of 1, and a padding of 1. In the final layer, features are upsampled to create the prediction map with N channels, where N denotes the number of semantic classes.

E. Semantic Consistency-Guided Joint Learning Loss

The loss function employed in our joint learning framework should guide the supervision of both the semantic segmentation and stereo matching tasks. Gradient smoothness between the disparity and semantic segmentation maps typically aligns closely, particularly at inter-class boundaries, where traditional training strategies tend to result in more errors due to factors such as occlusion and reflection. In light of this, we propose an SCG loss function to supervise the entire joint learning process, which leverages semantic consistency to optimize the training of S^{3}M-Net.

Given the ground-truth semantic segmentation map $\boldsymbol{M}^G \in \mathbb{R}^{H \times W}$, Each pixel $\boldsymbol{p}$ of $\boldsymbol{M}^G$ can be written as follows:

$$\boldsymbol{M}^G(\boldsymbol{p}) \in \{1, \ldots, C\}, \tag{4}$$

where C refers to the number of the semantic classes. We construct an extended 3D volume $\boldsymbol{V}^{3D} \in \mathbb{R}^{H \times W \times C}$ using the following expression:

$$\boldsymbol{V}_c^{3D}(\boldsymbol{p}) = \delta(\boldsymbol{M}^G(\boldsymbol{p}), c), \tag{5}$$

where c represents the c-th channel in the volume, and δ denotes the Kronecker Delta function [48]. As a result, each channel of the volume can be regarded as a binary segmentation map of the c-th class. To emphasize semantic consistency, We use an

average pooling operation for each channel to obtain the inter-class volume $\boldsymbol{V}^I \in \mathbb{R}^{H\times W\times C}$:

$$\boldsymbol{V}^I = \mathcal{P}(\boldsymbol{V}^{3D}), \tag{6}$$

where $\mathcal{P}$ denotes the average pooling operation. Furthermore, we apply a normalization operation:

$$\boldsymbol{V}^N(\boldsymbol{p}) = e^{-(2\boldsymbol{V}^I(\boldsymbol{p})-1)^2}. \tag{7}$$

to obtain a normalized volume $\boldsymbol{V}^N \in \mathbb{R}^{H\times W\times C}$. We then map $\boldsymbol{V}^N$ to a semantic consistency-guided weight map $\boldsymbol{W} \in \mathbb{R}^{H\times W}$ through:

$$\boldsymbol{W}(\boldsymbol{p}) = \max_c \left\{\boldsymbol{V}_c^N(\boldsymbol{p})\right\}. \tag{8}$$

The total loss function

$$\mathcal{L}_{scg} = \mathcal{L}_{ss} + \mathcal{L}_{sm} \tag{9}$$

consists of an SCG semantic segmentation loss $\mathcal{L}_{ss}$ and an SCG stereo matching loss $\mathcal{L}_{sm}$. $\mathcal{L}_{ss}$ is formulated as follows:

$$\mathcal{L}_{ss} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C}[(1-\alpha)+\alpha\boldsymbol{W}(\boldsymbol{p})]y_c(\boldsymbol{p})\log(\hat{y}_c(\boldsymbol{p})), \tag{10}$$

where N denotes the pixel number, C represents the class number, $\hat{y}_c(\boldsymbol{p})$ denotes the predicted probability of $\boldsymbol{p}$ belonging to class c, $y_c(\boldsymbol{p})$ represents the ground-truth label for $\boldsymbol{p}$ in class c, and α denotes the loss weight. Based on the ablation study detailed in Section IV-F, we set the value of α to 0.1. Moreover, $\mathcal{L}_{sm}$ is formulated as follows:

$$\mathcal{L}_{sm} = \sum_{i=1}^{N}[(1-\alpha)+\alpha\boldsymbol{W}(\boldsymbol{p})]\gamma^{N-i}\left\|\boldsymbol{D}^G - \boldsymbol{D}_i\right\|_1, \tag{11}$$

where $\boldsymbol{D}^G$ represents the ground-truth disparity map and $\boldsymbol{D}_i$ denotes the i-th disparity map in $\mathcal{D}$. α is set to 0.1 and γ is set to 0.9.

IV. Experiments

In this article, we conduct extensive experiments to evaluate the performance of our developed S^{3}M-Net both quantitatively and qualitatively. The following subsections provide details on the used datasets, experimental set-up, evaluation metrics, and the comprehensive evaluation of our proposed method.

A. Datasets

Since the training of our network requires both semantic and disparity annotations, we employ two public datasets to evaluate its performance: the vKITTI2 [27] dataset (synthetic yet large-scale) and the KITTI 2015 [28] dataset (real-world yet modest-scale). Their details are as follows:

- The vKITTI2 dataset contains virtual replicas of five sequences from the KITTI dataset. It provides 15 classes for the semantic segmentation tasks. Dense ground-truth disparity maps are acquired through depth rendering using a virtual engine. In our experiments, we randomly select 700 pairs of stereo images, along with their semantic and disparity annotations to evaluate the effectiveness and robustness of our proposed S^{3}M-Net, where 500 pairs are used for model training and the remaining 200 pairs are used for model validation.
- The KITTI 2015 dataset contains 400 pairs of stereo images captured in real-world driving scenarios, with 200 pairs containing ground truth and the other 200 pairs lacking ground truth. It provides 19 classes for the semantic segmentation tasks (in alignment with the Cityscapes [44] dataset). Sparse disparity ground truth is obtained using a Velodyne HDL-64E LiDAR. In our experiments, we allocate 70% of the dataset for training, while the remaining portion is used as the test set.

B. Experimental Setup

Our experiments are conducted on an NVIDIA RTX 3090 GPU. The batch size is set to 1. The maximum disparity search range is set to 192 pixels. All images are cropped to 1000 $\times$ 320 pixels before feeding into the network. We utilize the AdamW [54] optimizer for model training, setting the epsilon and weight decay parameters to 10^{-8} and 10^{-5}, respectively. The initial learning rate is set to 2×10^{-4}. Training lasts for 100 K iterations on the vKITTI2 dataset and 20 K iterations on the KITTI 2015 dataset. We employ traditional data augmentation techniques to enhance the robustness of the models.

C. Evaluation Metrics

Since our proposed S^{3}M-Net simultaneously performs semantic segmentation and stereo matching, we evaluate the performance of both tasks in our experiments.

We utilize seven evaluation metrics to quantify the performance of semantic segmentation: (1) accuracy (Acc), (2) mean accuracy (mAcc), (3) mean intersection over union (mIoU), (4) frequency-weighted intersection over union (fwIoU) [55], (5) precision (Pre), (6) recall (Rec), and (7) F1-score (FSc).

Additionally, we use two evaluation metrics: (1) the average end-point error (EPE) and (2) the percentage of error pixels (PEP), setting the tolerance for the latter to 1.0 and 3.0 pixels, respectively, to quantify the performance of stereo matching.

D. Semantic Segmentation Performance

The qualitative experimental results on the vKITTI2 and KITTI datasets are presented in Figs. 3 and 4, respectively, while the quantitative experimental results on the vKITTI2 and KITTI datasets are given in Tables I and II, respectively. These results suggest that S^{3}M-Net outperforms other SoTA single-modal and feature-fusion networks (including CNN-based and Transformer-based methods) across all evaluation metrics on both datasets. Specifically, it is noteworthy that when the entire joint learning framework is trained by minimizing our proposed SCG loss, S^{3}M-Net achieves the best performance on the KITTI dataset across all evaluation metrics except for Pre. Compared with SoTA methods, it shows improvements of 5.71% in mAcc, 4.84% in mIoU, 1.35% in fwIoU, and 0.76% in FSc, respectively. Similarly, it outperforms other networks on the vKITTI2 dataset

Fig. 3. Qualitative experimental results of semantic segmentation on the vKITTI2 [27] dataset: (a) RGB images; (b)–(k) semantic segmentation results achieved by SegNet [13], U-Net [14], PSPNet [15], DeepLabv3+ [17], HRNet [49], BiSeNet V2 [50], Segmenter [32], SegFormer [33], Mask2Former [34], and DDRNet [51], respectively; (l)–(q) semantic segmentation results achieved by FuseNet [19], MFNet [35], RTFNet [20], SNE-RoadSeg [3], OFF-Net [22], and RoadFormer [24], respectively; (r)–(s) semantic segmentation results achieved by our proposed S^{3}M-Net w/o and w/ the use of the SCG loss, respectively; (t) ground truth annotations.

in most evaluation metrics, with improvements of 1.72% in fwIoU and 1.44% in FSc. However, for mAcc, mIoU, and Rec, its performance is comparable to that of S^{3}M-Net trained without using the SCG loss. Additionally, it is obvious that feature-fusion networks consistently outperform single-modal networks, particularly under challenging weather and lighting conditions. This observation aligns with our expectations, as feature-fusion networks leverage both RGB images and disparity maps, allowing them to effectively learn informative spatial geometric representations. However, SoTA feature-fusion networks may exhibit higher error rates in distant regions. For instance, FuseNet and SNE-RoadSeg demonstrate poor performance in the sky. We attribute this phenomenon to the deep structure of the encoders, where distinguishing distant objects using disparity features becomes challenging, and the feature fusion process amplifies the influence of the disparity feature. In contrast, within our proposed joint learning framework, we can extract more informative features benefiting from both tasks, irrespective of dataset size. This improvement is likely due to the fact that joint learning of multiple interconnected tasks introduces a form of regularization, which has shown its superiority over uniform complexity penalization in reducing over-fitting.

E. Stereo Matching Performance

The qualitative experimental results on the vKITTI2 and KITTI datasets are given in Figs. 5 and 6, respectively, while

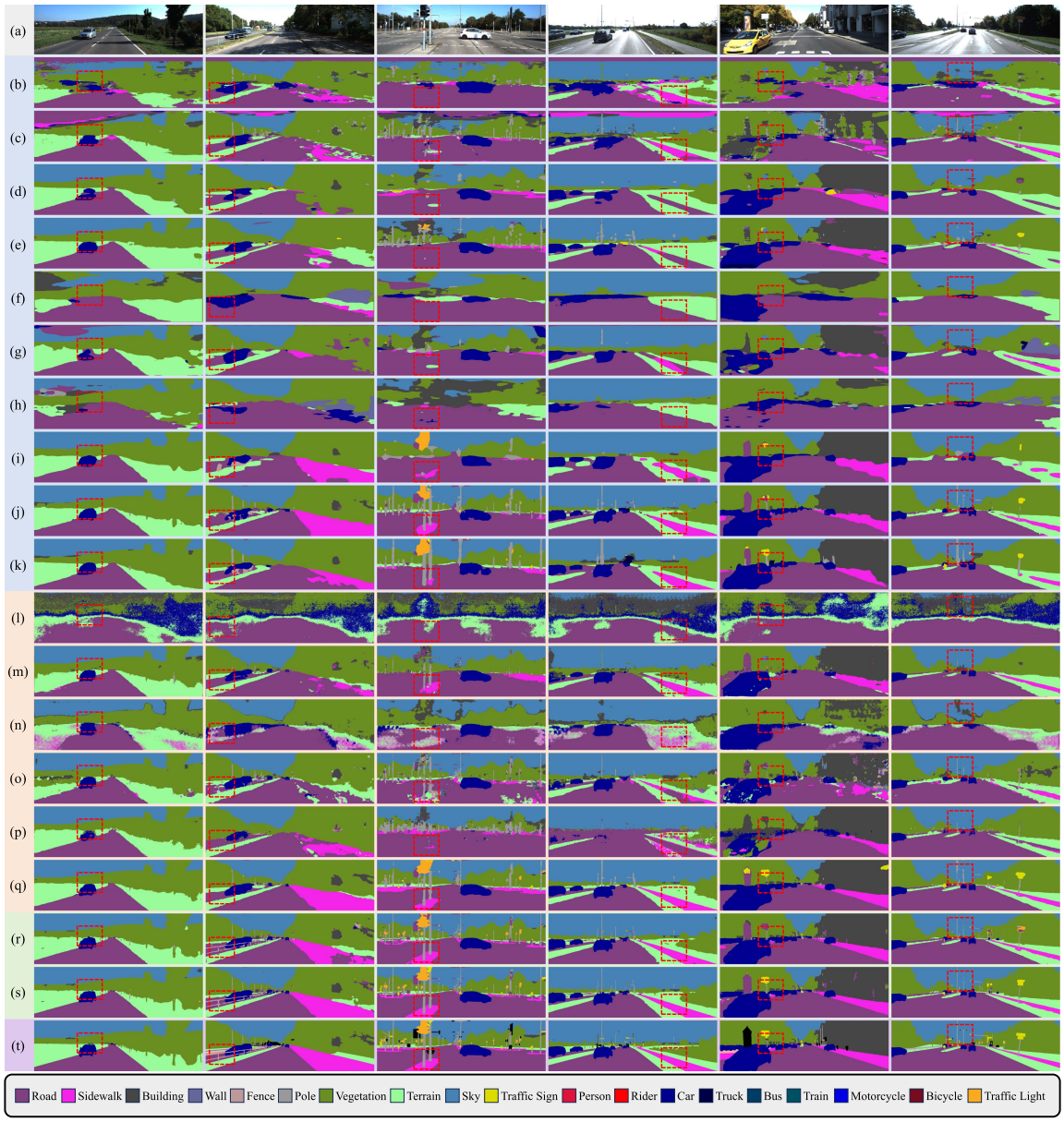

Fig. 4. Qualitative experimental results of semantic segmentation on the KITTI 2015 [28] dataset: (a) RGB images; (b)–(k) semantic segmentation results achieved by SegNet [13], U-Net [14], PSPNet [15], DeepLabv3+ [17], HRNet [49], BiSeNet V2 [50], Segmenter [32], SegFormer [33], Mask2Former [34], and DDRNet [51], respectively; (l)–(q) semantic segmentation results achieved by FuseNet [19], MFNet [35], RTFNet [20], SNE-RoadSeg [3], OFF-Net [22], and RoadFormer [24], respectively; (r)–(s) semantic segmentation results achieved by our proposed S^{3}M-Net w/o and w/ the use of the SCG loss, respectively; (t) ground truth annotations.

the quantitative experimental results on the vKITTI2 and KITTI datasets are presented in Tables III and IV, respectively. These results suggest that S^{3}M-Net outperforms other SoTA stereo matching networks across all evaluation metrics on both datasets. Specifically, S^{3}M-Net trained with and without using the SCG loss achieves the top and second-best overall performances, respectively. S^{3}M-Net, when trained without using the SCG loss, demonstrates improvements of 2.50%–71.32% in EPE, 4.17%–64.19% in PEP 1.0, and 4.49%–67.97% in PEP 3.0. On the other hand, S^{3}M-Net, when trained with the SCG loss, shows improvements of 5.00%–72.06% in EPE, 5.44%–64.38% in PEP 1.0, and 4.49%–69.83% in PEP 3.0. We attribute these improvements to the feature sharing and fusion strategies applied in S^{3}M-Net. First, sharing features with the semantic segmentation task allows S^{3}M-Net to learn stereo matching effectively even with limited training data. Second, as discussed above, stereo matching can sometimes produce ambiguous disparity estimations, especially in occluded or texture-less areas. The pursuit of semantic consistency helps resolve such ambiguities, leading to more reliable disparity estimation results. In Fig. 6, it is evident that regions lacking disparity ground truth frequently have substantial errors. Previous stereo matching algorithms have endeavored to tackle this issue through knowledge distillation with pre-trained models [38]. Nevertheless, our S^{3}M-Net successfully overcomes this challenge by leveraging semantic information.

TABLE I
COMPARISONS OF SOTA SEMANTIC SEGMENTATION NETWORKS ON THE VKITTI2 [27] DATASET

Category	Networks	Acc (%) ↑	mAcc (%) ↑	mIoU (%) ↑	fwIoU (%) ↑	Pre (%) ↑	Rec (%) ↑	FSc (%) ↑
Single-Modal	SegNet [13]	59.29	32.54	23.93	48.17	66.10	66.73	61.11
	U-Net [14]	62.71	37.65	29.83	55.10	75.80	67.67	65.26
	PSPNet [15]	76.26	53.53	44.81	69.30	81.55	79.68	75.38
	DeepLabv3+ [17]	92.19	63.15	56.90	87.15	89.00	92.71	90.16
	HRNet [49]	74.79	40.82	32.47	63.23	73.69	76.50	73.39
	BiSeNet V2 [50]	81.77	51.07	44.45	74.71	83.23	82.19	80.67
	Segmenter [32]	90.39	60.33	52.99	83.47	88.05	87.89	87.70
	SegFormer [33]	94.75	70.56	64.98	90.49	93.57	93.62	93.46
	Mask2Former [34]	89.29	64.58	57.14	83.84	90.75	87.23	87.19
	DDRNet [51]	70.80	40.32	32.10	61.44	76.35	71.67	70.57
Feature-Fusion	FuseNet [19]	49.42	31.21	22.56	41.07	79.39	50.67	47.50
	MFNet [35]	76.22	51.50	43.41	68.82	82.46	78.65	73.80
	RTFNet [20]	85.22	49.47	42.59	77.69	83.74	89.17	84.41
	SNE-RoadSeg [3]	83.64	60.85	52.56	75.14	83.44	81.66	77.77
	OFF-Net [22]	90.84	61.51	55.27	84.69	89.24	86.71	86.15
	RoadFormer [24]	97.54	86.58	80.83	95.34	96.99	96.86	96.91
	S^{3}M-Net	98.27	**88.28**	**84.25**	96.92	98.29	**98.32**	98.28
	S^{3}M-Net w/ SCG loss	**98.32**	88.24	84.18	**96.98**	**98.37**	98.28	**98.31**

The symbol ↑ indicates that a higher value corresponds to better performance. The best results are shown in bold font.

TABLE II
COMPARISONS OF SOTA SEMANTIC SEGMENTATION NETWORKS ON THE KITTI 2015 [28] DATASET

Category	Networks	Acc (%) ↑	mAcc (%) ↑	mIoU (%) ↑	fwIoU (%) ↑	Pre (%) ↑	Rec (%) ↑	FSc (%) ↑
Single-Modal	SegNet [13]	59.63	31.98	22.61	43.98	55.25	67.49	57.29
	U-Net [14]	69.02	41.15	30.64	55.65	69.11	77.65	71.04
	PSPNet [15]	80.03	44.97	38.15	68.62	79.29	82.66	79.59
	DeepLabv3+ [17]	82.33	50.15	42.79	72.43	83.85	87.18	84.59
	HRNet [49]	63.42	31.68	22.78	49.40	55.10	67.71	57.21
	BiSeNet V2 [50]	73.68	41.66	32.71	60.55	68.35	81.79	72.37
	Segmenter [32]	84.53	50.77	43.63	74.72	82.99	87.15	84.41
	SegFormer [33]	88.28	59.23	51.39	80.53	87.15	90.85	88.46
	Mask2Former [34]	84.35	54.33	45.87	75.56	84.74	89.12	85.92
	DDRNet [51]	62.12	31.61	22.63	48.15	57.09	68.98	59.07
Feature-Fusion	FuseNet [19]	41.79	19.05	11.38	27.53	44.14	44.35	37.68
	MFNet [35]	81.02	48.13	40.70	70.42	82.85	85.73	82.36
	RTFNet [20]	71.61	39.26	30.35	57.98	69.52	85.16	74.28
	SNE-RoadSeg [3]	79.46	51.91	41.56	69.22	81.45	87.05	82.91
	OFF-Net [22]	75.84	40.13	33.13	64.02	77.48	72.19	70.62
	RoadFormer [24]	90.05	62.34	55.13	83.40	**91.65**	91.39	91.11
	S^{3}M-Net	90.01	62.48	54.33	83.44	88.96	93.52	90.65
	S^{3}M-Net w/ SCG loss	**90.66**	**65.90**	**57.80**	**84.53**	90.85	**93.55**	**91.80**

The symbol ↑ indicates that a higher value corresponds to better performance. The best results are shown in bold font.

F. Ablation Studies

In this subsection, we first conduct an ablation study on the selection of loss weight α in (11). Fig. 7 shows the quantitative experimental results with respect to different α in the range of 0.0 to 0.4 for both semantic segmentation and stereo matching. It can be obvious that when $\alpha = 0.1$, S^{3}M-Net achieves the best overall performance for both tasks. Further weight tuning is possible, but it should be approached cautiously, especially when dealing with limited data to avoid over-fitting.

Furthermore, we conduct an additional ablation study on the feature fusion strategy in our proposed FFA module. As shown in Table V, when using the addition operation to fuse heterogeneous features, the FFA module consistently achieves the best performance across all evaluation metrics, compared to other feature fusion strategies, including concatenation, cross feature module (CFM) [56], dynamic dilated pyramid module (DDPM) [57], separation-and-aggregation gate (SA Gate) [58], and softmax weighted sum (SWS) used in AysmFusion [59], CEN [60], and TokenFusion [61].

V. DISCUSSION

The experimental results shown in Section IV provide strong support for the claims made in Section I. First, the joint learning of semantic segmentation and stereo matching, two interconnected environmental perception tasks, using our proposed S^{3}M-Net introduces a form of regularization that has shown its effectiveness in reducing overfitting, particularly in scenarios where training data are limited. Secondly, this end-to-end joint learning framework yields improved performance when compared to the models trained separately for each task. Finally, the pursuit of

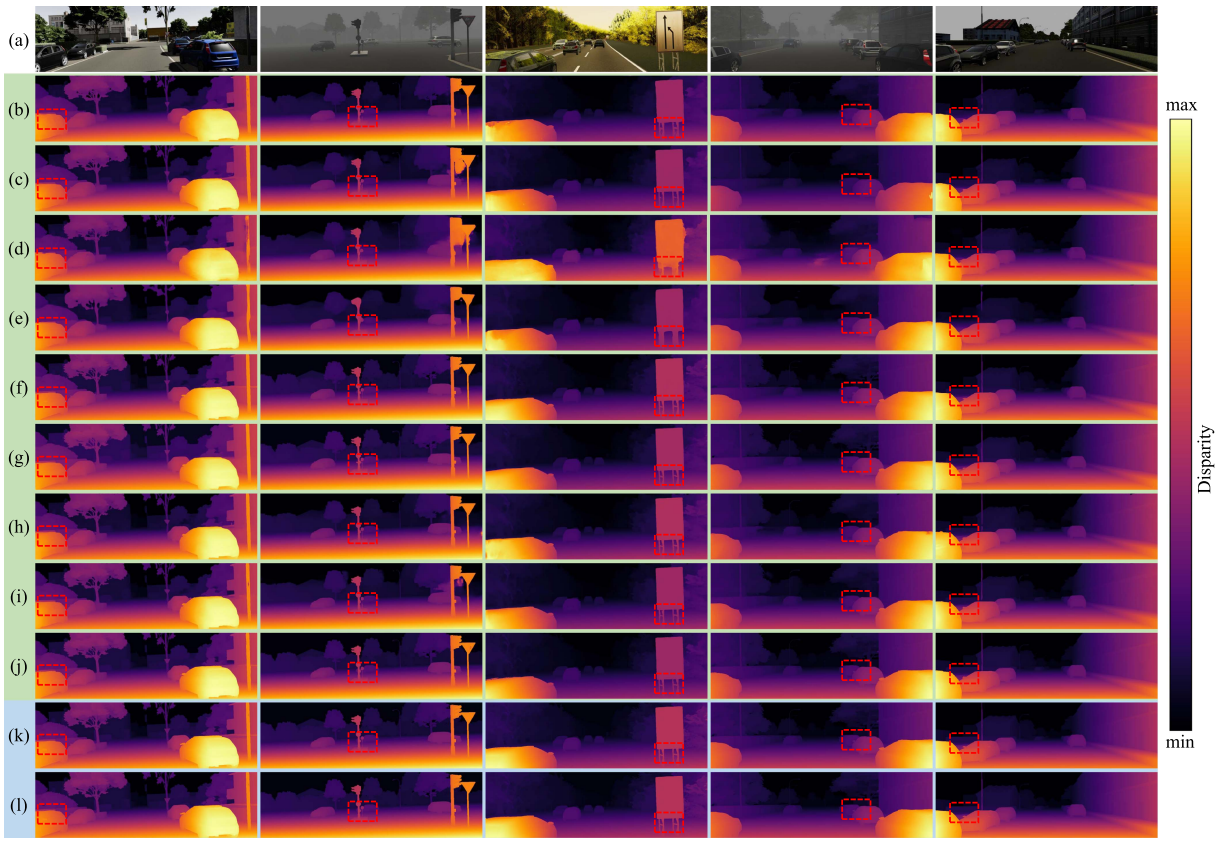

Fig. 5. Qualitative experimental results of stereo matching on the vKITTI2 [27] dataset: (a) left RGB images; (b)–(j) disparity maps estimated using PSMNet [36], GwcNet [37], AANet [38], LEA-Stereo [39], RAFT-Stereo [40], CRE-Stereo [41], ACVNet [52], PCWNet [53], and IGEV-Stereo [4], respectively; (k)–(l) disparity maps estimated using our proposed S^{3}M-Net w/o and w/ the use of the SCG loss, respectively.

TABLE III
COMPARISONS OF SOTA STEREO MATCHING NETWORK ON THE VKITTI2 [27] DATASET

Networks	EPE (pixels) ↓	PEP (%) ↓	
		> 1 pixel	> 3 pixels
PSMNet [36]	0.68	10.31	3.77
GwcNet [37]	0.65	9.72	3.69
AANet [38]	1.36	15.61	6.98
LEA-Stereo [39]	0.83	13.33	4.84
RAFT-Stereo [40]	0.40	5.88	2.67
CRE-Stereo [41]	0.63	10.35	3.90
ACVNet [52]	0.61	9.41	3.45
PCW-Net [53]	0.63	9.45	3.49
IGEV-Stereo [4]	0.47	7.15	3.09
S^{3}M-Net	0.39	5.59	2.55
S^{3}M-Net w/ SCG loss	**0.38**	**5.56**	**2.55**

The symbol ↓ indicates that a lower value corresponds to better performance. The best results are shown in bold font.

TABLE IV
COMPARISONS OF SOTA STEREO MATCHING NETWORK ON THE KITTI 2015 [28] DATASET

Networks	EPE (pixels) ↓	PEP (%) ↓	
		> 1 pixel	> 3 pixels
PSMNet [36]	0.74	16.12	2.61
GwcNet [37]	0.68	14.21	2.01
AANet [38]	1.10	22.67	5.37
LEA-Stereo [39]	0.83	18.67	3.22
RAFT-Stereo [40]	0.60	10.78	1.96
CRE-Stereo [41]	0.92	19.68	3.35
ACVNet [52]	0.68	13.93	2.10
PCW-Net [53]	0.70	14.81	2.43
IGEV-Stereo [4]	0.62	12.15	1.99
S^{3}M-Net	0.56	10.33	1.72
S^{3}M-Net w/ SCG loss	**0.55**	**10.02**	**1.62**

The symbol ↓ indicates that a lower value corresponds to better performance. The best results are shown in bold font.

semantic consistency in joint learning helps reduce ambiguous disparity estimations in texture-less or occluded regions. We believe that our proposed S^{3}M-Net can be readily deployed in autonomous vehicles after addressing the following limitations:

- S^{3}M-Net requires both semantic and disparity annotations, and collecting data with such ground truth remains a labor-intensive process. Therefore, the exploration of potential solutions such as semi-supervised or low/few-shot semantic segmentation and un/self-supervised stereo matching is a promising avenue for future research.
- S^{3}M-Net achieves a processing speed of 0.66 fps when processing input RGB images with a resolution of 1248 × 384 pixels. We believe that further computational efficiency optimizations are necessary before deploying S^{3}M-Net in autonomous vehicles.

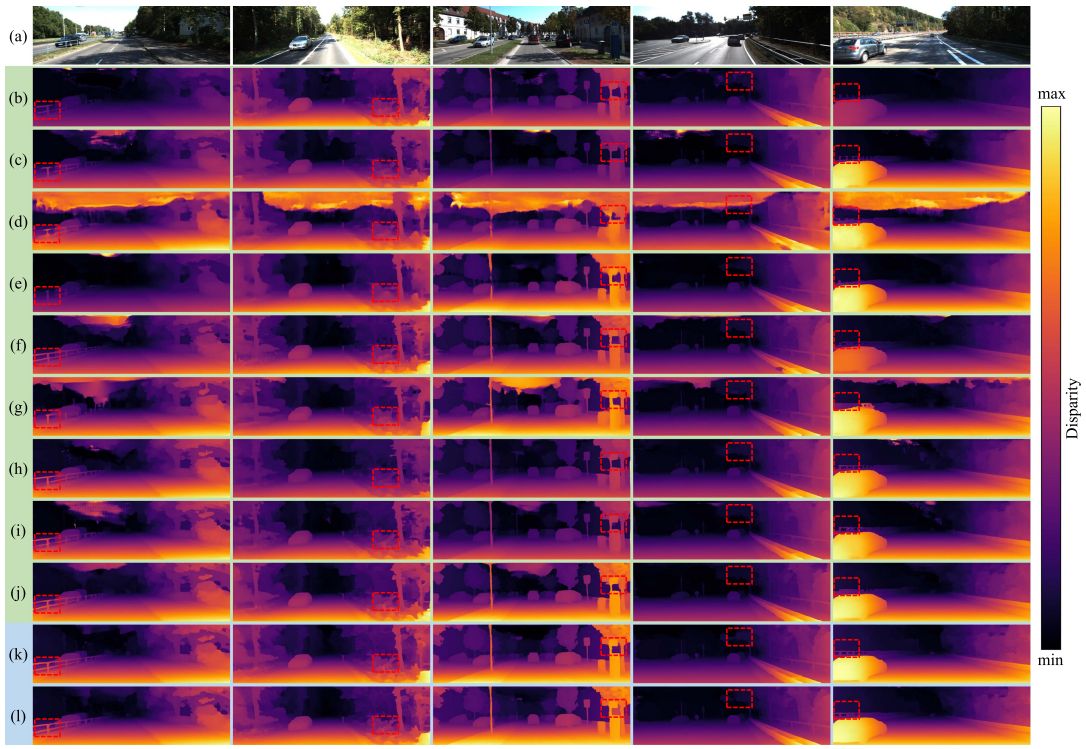

Fig. 6. Qualitative experimental results of stereo matching on the KITTI 2015 [28] dataset: (a) left RGB images; (b)–(j) disparity maps estimated using PSMNet [36], GwcNet [37], AANet [38], LEA-Stereo [39], RAFT-Stereo [40], CRE-Stereo [41], ACVNet [52], PCWNet [53], and IGEV-Stereo [4], respectively; (k)–(l) disparity maps estimated using our proposed S^{3}M-Net w/o and w/ the use of the SCG loss, respectively.

TABLE V
ABLATION STUDY ON FEATURE FUSION STRATEGY IN OUR FFA MODULE ON THE KITTI [28] 2015 DATASET

Fusion Strategy	Acc (%) ↑	mAcc (%) ↑	mIoU (%) ↑	fwIoU (%) ↑	Pre (%) ↑	Rec (%) ↑	FSc (%) ↑
Addition	**90.01**	**62.48**	**54.33**	**83.44**	**88.96**	93.52	**90.65**
Concatenation	86.88	57.43	48.40	78.50	85.96	**93.71**	88.92
CFM [56]	86.87	57.41	48.77	79.13	85.41	92.05	87.63
DDPM [57]	86.52	58.65	49.51	78.14	85.56	93.34	88.44
SA Gate [58]	87.62	61.77	52.10	80.55	86.74	92.31	88.52
SWS [59], [60], [61]	87.94	58.11	49.64	80.25	86.63	93.34	89.20

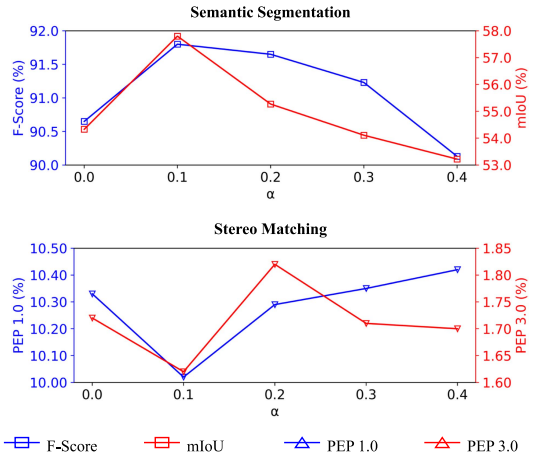

Fig. 7. Ablation study on the selection of α in the SCG joint learning loss function on the KITTI 2015 [28] dataset.

VI. CONCLUSION

This article introduced S^{3}M-Net, an effective solution for joint learning of semantic segmentation and stereo matching. We have made three significant contributions in this work: (1) the development of an entire joint learning framework that shares features between both tasks and fuses heterogeneous features to improve semantic segmentation, (2) a feature fusion adaption module designed to enable effective feature sharing between the two tasks, and (3) a semantic consistency-guided joint learning loss that emphasizes structural consistency in both tasks. We conducted extensive experiments on the vKITTI2 (synthetic and large) and KITTI (real-world and small) datasets to validate the effectiveness of our framework, the FFA module, and the training loss. Our results demonstrate the superior performance of our approach compared to all other existing methods.

REFERENCES

[1] B. Ranft and C. Stiller, "The role of machine vision for intelligent vehicles," *IEEE Trans. Intell. Veh.*, vol. 1, no. 1, pp. 8–19, Mar. 2016.

[2] D. L. Fisher, M. Lohrenz, D. Moore, E. D. Nadler, and J. K. Pollard, "Humans and intelligent vehicles: The hope, the help, and the harm," *IEEE Trans. Intell. Veh.*, vol. 1, no. 1, pp. 56–67, Mar. 2016.

[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.*, 2020, pp. 340–356.

[4] 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.*, 2023, pp. 21919–21928. .

[5] Z. Rao et al., "Masked representation learning for domain generalized stereo matching," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 5435–5444.

[6] R. Fan, S. Guo, and M. J. Bocus, *Autonomous Driving Perception*. Berlin, Germany: Springer, 2023.

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

[8] 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. Automat. Lett.*, vol. 4, no. 2, pp. 1162–1169, Apr. 2019.

[9] 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.*, 2019, pp. 7484–7493.

[10] W. Zhan, X. Ou, Y. Yang, and L. Chen, "DSNet: Joint learning for scene segmentation and disparity estimation," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2019, pp. 2946–2952.

[11] P. L. Dovesi et al., "Real-time semantic stereo matching," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2020, pp. 10780–10787.

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

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

[14] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.- Assist. Intervention*, 2015, pp. 234–241.

[15] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2881–2890.

[16] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*.

[17] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 801–818.

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

[19] C. Hazirbas, L. Ma, C. Domokos, and D. Cremers, "FuseNet: Incorporating depth into semantic segmentation via fusion-based CNN architecture," in *Proc. 13th Asian Conf. Comput. Vis.*, 2017, pp. 213–228.

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

[21] 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.*. 2021, pp. 1140–1145.

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

[23] J. Yang, B. Xue, Y. Feng, D. Wang, R. Fan, and Q. Chen, "Three-filters-to-normal: Revisiting discontinuity discrimination in depth-to-normal translation," *IEEE Trans. Automat. Sci. Eng.*, 2024, doi: 10.1109/TASE.2024.3355941.

[24] J. Li, Y. Zhang, P. Yun, G. Zhou, Q. Chen, and R. Fan, "RoadFormer: Duplex transformer for RGB-normal semantic road scene parsing," 2023, *arXiv:2309.10356*.

[25] R. Fan et al., "One-vote veto: Semi-supervised learning for low-shot glaucoma diagnosis," *IEEE Trans. Med. Imag.*, vol. 42, no. 12, pp. 3764–3778, Dec. 2023, doi: 10.1109/TMI.2023.3307689.

[26] R. Fan, X. Ai, and N. Dahnoun, "Road surface 3D reconstruction based on dense subpixel disparity map estimation," *IEEE Trans. Image Process.*, vol. 27, no. 6, pp. 3025–3035, Jun. 2018.

[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/CVF Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3061–3070.

[29] U. Michieli, M. Biasetton, G. Agresti, and P. Zanuttigh, "Adversarial learning and self-teaching techniques for domain adaptation in semantic segmentation," *IEEE Trans. Intell. Veh.*, vol. 5, no. 3, pp. 508–518, Sep. 2020.

[30] Y. Yang, C. Shan, F. Zhao, W. Liang, and J. Han, "On exploring shape and semantic enhancements for RGB-X semantic segmentation," *IEEE Trans. Intell. Veh.*, early access, Jul. 17, 2023, doi: 10.1109/TIV.2023.3296219.

[31] J. Fan, F. Wang, H. Chu, X. Hu, Y. Cheng, and B. Gao, "MLFNet: Multi-level fusion network for real-time semantic segmentation of autonomous driving," *IEEE Trans. Intell. Veh.*, vol. 8, no. 1, pp. 756–767, Jan. 2023.

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

[33] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: Simple and efficient design for semantic segmentation with transformers," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 12077–12090.

[34] 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.

[35] 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.

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

[37] X. Guo, K. Yang, W. Yang, X. Wang, and H. Li, "Group-wise correlation stereo network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3273–3282.

[38] H. Xu and J. Zhang, "AANet: Adaptive aggregation network for efficient stereo matching," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 1959–1968.

[39] X. Cheng et al., "Hierarchical neural architecture search for deep stereo matching," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 22158–22169.

[40] 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.

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

[42] T. Elsken, J. H. Metzen, and F. Hutter, "Neural architecture search: A survey," *J. Mach. Learn. Res.*, vol. 20, no. 1, pp. 1997–2017, 2019.

[43] Z. Teed and J. Deng, "RAFT: Recurrent all-pairs field transforms for optical flow," in *Proc. 16th Eur. Conf. Comput. Vis.*, 2020, pp. 402–419.

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

[45] 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.*, 2012, pp. 3354–3361.

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

[47] 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.

[48] S. Okada, M. Ohzeki, and S. Taguchi, "Efficient partition of integer optimization problems with one-hot encoding," *Sci. Rep.*, vol. 9, no. 1, 2019, Art. no. 13036.

[49] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 5693–5703.

[50] 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, pp. 3051–3068, 2021.
[51] Y. Hong, H. Pan, W. Sun, and Y. Jia, "Deep dual-resolution networks for real-time and accurate semantic segmentation of road scenes," 2021, *arXiv:2101.06085*.
[52] 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.*, 2022, pp. 12981–12990.
[53] 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.*, 2022, pp. 280–297.
[54] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. Int. Conf. Learn. Representations*, 2018.
[55] H. Zhao, X. Puig, B. Zhou, S. Fidler, and A. Torralba, "Open vocabulary scene parsing," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2002–2010.
[56] J. Wei, S. Wang, and Q. Huang, "F^3 Net: Fusion, feedback and focus for salient object detection," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 12321–12328.
[57] Y. Pang, L. Zhang, X. Zhao, and H. Lu, "Hierarchical dynamic filtering network for RGB-D salient object detection," in *Proc. 16th Eur. Conf. Comput. Vis.*, 2020, pp. 235–252.
[58] 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.
[59] Y. Wang, F. Sun, M. Lu, and A. Yao, "Learning deep multimodal feature representation with asymmetric multi-layer fusion," in *Proc. 28th ACM Int. Conf. Multimedia*, 2020, pp. 3902–3910.
[60] Y. Wang, W. Huang, F. Sun, T. Xu, Y. Rong, and J. Huang, "Deep multimodal fusion by channel exchanging," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 4835–4845.
[61] Y. Wang, X. Chen, L. Cao, W. Huang, F. Sun, and Y. Wang, "Multimodal token fusion for vision transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 12186–12195.

Zhiyuan Wu is currently working toward the B.E. degree with the MIAS Group, Tongji University, Shanghai, China, supervised by Prof. Rui Fan. His research interests include computer vision and deep learning, with a particular emphasis on stereo matching and feature fusion.

Yi Feng received the B.E. degree in automation in 2022 from Tongji University, Shanghai, China, where he is currently working toward the M.Sc. degree, supervised by Prof. Rui Fan, with the MIAS Group,College of Electronics and Information Engineering. His research interests include computer vision and deep learning.

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, supervised by Prof. Rui Fan, with the MIAS Group. His research interests include computer stereo vision, especially for unsupervised approaches and long-term learning.

Fisher Yu (Member, IEEE) received the Ph.D. degree from Princeton University, Princeton, NJ, USA. He is currently an Assistant Professor with ETH Zürich, Zürich, Switzerland. He became a postdoctoral Researcher with UC Berkeley. He now leads the Visual Intelligence and Systems (VIS) group, ETH Zürich. His research interests include the junction of machine learning, computer vision, and robotics.

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 Electronics and Information Engineering, Tongji University. His research interests include robotics control, environmental perception, and understanding of mobile robots, and bioinspired control.

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 (supervisors: Prof. John G. Rarity and Prof. Naim Dahnoun) in electrical and electronic engineering from the University of Bristol, Bristol, U.K., in 2018. He was a Research Associate (supervisor: Prof. Ming Liu) with the Hong Kong University of Science and Technology from 2018 to 2020 and a Postdoctoral Scholar-Employee (supervisors: Prof. Linda M. Zangwill and Prof. David J. Kriegman) with the University of California San Diego, San Diego, CA, USA, from 2020 and 2021. He began his faculty career as a Full Research Professor with the College of Electronics & Information Engineering, Tongji University, Shanghai, China, in 2021, and was then promoted to a Full Professor with the same college, as well as with the Shanghai Research Institute for Intelligent Autonomous Systems in 2022. His research interests include computer vision, deep learning, and robotics. Dr. Fan was an Associate Editor for ICRA'23, IROS'23 and IROS'24, and as a senior program committee Member of AAAI'23/24. He is the general Chair of the AVVision community and organized several impactful workshops and special sessions in conjunction with WACV'21, ICIP'21/22/23, ICCV'21, and ECCV'22. He was honored by being included in the Stanford University List of Top 2% Scientists Worldwide in both 2022 and 2023, recognized on the Forbes China List of 100 Outstanding Overseas Returnees in 2023, and acknowledged as one of Xiaomi Young Talents in 2023.