2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 1-5, 2023. Detroit, USA

SSC-RS: Elevate LiDAR Semantic Scene Completion with Representation Separation and BEV Fusion

Jianbiao Mei[1], Yu Yang[1], Mengmeng Wang[1], Tianxin Huang[1], Xuemeng Yang[2] and Yong Liu[1,†]

***Abstract*— Semantic scene completion (SSC) jointly predicts the semantics and geometry of the entire 3D scene, which plays an essential role in 3D scene understanding for autonomous driving systems. SSC has achieved rapid progress with the help of semantic context in segmentation. However, how to effectively exploit the relationships between the semantic context in semantic segmentation and geometric structure in scene completion remains under exploration. In this paper, we propose to solve outdoor SSC from the perspective of representation separation and BEV fusion. Specifically, we present the network, named SSC-RS, which uses separate branches with deep supervision to explicitly disentangle the learning procedure of the semantic and geometric representations. And a BEV fusion network equipped with the proposed Adaptive Representation Fusion (ARF) module is presented to aggregate the multi-scale features effectively and efficiently. Due to the low computational burden and powerful representation ability, our model has good generality while running in real-time. Extensive experiments on SemanticKITTI demonstrate our SSC-RS achieves state-of-the-art performance. Code is available at https://github.com/Jieqianyu/SSC-RS.git.**

I. Introduction

In recent years, 3D scene understanding, one of the most important functions of perception systems in autonomous driving, has attracted extensive studies and achieved rapid progress. When working with large-scale outdoor scene understanding, Semantic Scene Completion (SSC) aims to predict the semantic occupancy of each voxel of the entire 3D scene from the sparse LiDAR scans, including the completion of certain regions. Due to the ability to recover geometric structure, SSC can facilitate further applications like 3D object detection, which usually suffer from the sparsity and incompleteness (caused by occlusions or far distance from sensors) of the LiDAR point cloud. However, it's challenging to precisely estimate the semantics and geometry of the whole 3D real-world scene from partial observations due to the complex outdoor scenarios such as various shapes/sizes and occlusions.

Following the pioneering work SSCNet [1], some existing outdoor SSC methods [2], [3] exploit a single U-Net network, e.g., a heavy dense 3D convolution network to predict semantics and geometry jointly. However, they usually involve unnecessary calculations and extra memory and computation overhead, especially when the input voxel resolution is large since there are lots of empty voxels in the 3D scene.

[1]The authors are with the Institute of Cyber-Systems and Control, Zhejiang University, Hangzhou, 310027, China. (Yong Liu† is the corresponding author, email: yongliu@iipc.zju.edu.cn)

[2]The authors are with the Shanghai Artificial Intelligence Laboratory, Shanghai, China.

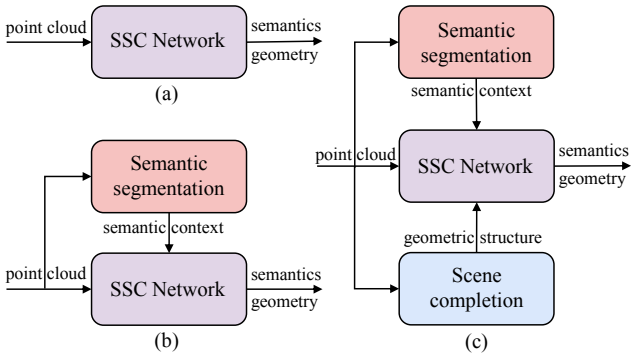

Fig. 1. (a) Consider the semantic context and geometric structure in a hybrid manner. (b) Consider the semantic context and geometric structure in a semi-hybrid manner. (c) Disentangling the learning procedure of semantic context and geometric structure explicitly.

On the other way, some methods [4]–[7] utilize the semantic information in the segmentation to assist outdoor SSC by combining the semantic completion network with the segmentation network. We found that most outdoor SSC methods consider the semantic context (semantic representation) and geometry structure (geometric representation) in a hybrid (Fig. 1 (a)) or semi-hybrid manner (Fig. 1 (b)). And how to effectively learn semantic/geometric representations and exploit their relationship remains unexplored.

In this paper, we explore the solutions to the outdoor SSC from the perspective of representation separation and BEV (Bird's-Eye View) fusion (Fig. 1 (c)). We propose to explicitly disentangle the learning procedure of semantic context and geometric structure and fuse them in the BEV, which is demonstrated to be a kind of success in 3D object detection and segmentation [8]–[13]. Our main insights are: (1) Semantic context and geometric structure complement each other and are vital for SSC tasks. Recovering the geometry details according to the semantics is easy, and the completed shapes help identify semantic categories. (2) Explicitly disentangling the representations can facilitate and accelerate the learning procedure of semantic context and geometric structure. (3) Compared to dense feature fusion in 3D space, BEV fusion is more convenient and efficient.

Specifically, we design separate branches, i.e., semantic and completion branches, for semantic/geometric representations according to their intrinsic properties. We also develop a BEV fusion network to aggregate the two types of representations from the two branches. We use a sparse 3D CNN [14] to encode the multi-level semantic context and a tiny dense 3D CNN to obtain the multi-scale geometric

2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) | 978-1-6654-9190-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/IROS55552.2023.10341742

Authorized licensed use limited to: Southeast University. Downloaded on August 18,2026 at 08:41:27 UTC from IEEE Xplore. Restrictions apply.

structures. In addition, we apply deep supervision on both branches to facilitate representation learning. Furthermore, to obtain selective cues from semantic/geometric representations and fuse the semantic context and geometric details sufficiently, we propose an Adaptive Representation Fusion (ARF) module in the BEV fusion network. Due to the lower computational burden and more powerful representation ability, our model runs in real-time and has good generality. Experiments on SemanticKITTI dataset (hidden test) show that our approach achieves state-of-the-art performance (rank the 2rd by mIoU and 1rd in terms of completion metric (IoU) on the public SSC benchmark[1]). Our contributions are summarized as follows:

• We develop the SSC-RS network to solve the outdoor SSC problem from the perspective of representation (semantic/geometric) separation and BEV fusion.

• We design two separate branches for multi-level semantic context and multi-scale geometric structures according to their properties. And deep supervision is applied to facilitate the learning procedure.

• We use 2D CNN as the BEV fusion network to aggregate semantic/geometric representations. And an Adaptive Representation Fusion (ARF) module in the BEV fusion network is proposed to fuse the semantic context and geometric details sufficiently.

• Due to the lower computational burden and more powerful representation ability, SSC-RS has low decay and achieves state-of-the-art performance on the SemanticKITTI benchmark.

II. Related Work

A. Semantic scene completion

The indoor SSC method has developed rapidly with the emergence of indoor benchmarks such as SUNCG [1] and NYU [15]. Existing methods use different types of geometrical inputs cooperated with corresponding network architectures to complete indoor SSC. For example, [1], [16], [17] process the depth maps with 3D CNNs end-to-end. [18]–[22] take the RGB-D images with 2D-3D CNNs to explore the modality complementarity. [6], [23]–[25] encode the truncated signed distance function (TSDF) representations with the volume network architectures. [26]–[30] process the point clouds with the point-based network to achieve semantic scene completion continuously.

Since SemanticKITTI [31] introduces a large-scale outdoor benchmark for SSC tasks, several outdoor SSC methods have emerged. Following the pioneering work SSCNet [1], [2] exploits a single U-Net framework to process segmentation and completion simultaneously, resulting in extra computation overhead of empty voxels. Some methods [3], [32] use sparse convolutions or introduce 2D CNN to solve the above problem. For example, LMSCNet [3] appends a 3D decoder after the lightweight 2D backbone, and Zhang et al. [32] designs a sparse CNN with dense deconvolution layers. Some solutions focusing on multi-view fusion [6], and local implicit functions [30] are also explored.

Besides, some methods [4], [5] exploit semantic segmentation to assist SSC. JS3C-Net [4] inserts a semantic segmentation network before SSC, and SSA-SC [5] injects the features from the segmentation branch into the completion branch hierarchically. In this work, we propose SSC-RS for large-scale outdoor SSC from the perspective of representation separation and BEV fusion. And different from [33], which designs a cascaded network to implement the complementary between scene completion and semantic segmentation for indoor SSC, our SSC-RS exploits the multi-scale context. And our parallel feature fusion in BEV is more convenient and efficient.

B. BEV perception in segmentation

BEV perception indicates vision algorithms in the sense of the BEV view representation for autonomous driving [34], which has been explored for a variety of tasks such as LiDAR detection [11], [35], LiDAR segmentation [9], [10], [13], and sensor fusion [12], [36]. SalsaNet [8] projects point clouds into BEV feature maps, and PolarNet [9] proposes a polar BEV representation for semantic segmentation. Panoptic-PHNet [10] exploits BEV features to enhance the segmentation and perform instance grouping in BEV. Panoptic-Polarnet [13] uses a polar BEV representation to implement semantic segmentation and class-agnostic instance clustering. In SSC tasks, S3CNet [6] designs a 2D S3CNet to predict the 2D SSC Image in the BEV of the input point cloud. And SSA-SC [5] take the 2D CNN as the semantic completion network to simultaneously predict the semantics and geometry. Similar to SSA-SC, we also use the 2D CNN as the BEV fusion network to efficiently provide the semantic occupancy of the entire 3D scene. And different from SSA-SC, we explicitly disentangle the learning process of semantic context and geometric structure and take the BEV network as a fusion network. Also, we design an adaptive representation module for aggregating the judicious cues in BEV sufficiently.

III. Method

A. Overview

In this paper, we explore the solutions to LiDAR semantic scene completion from the perspective of representation separation and BEV fusion. Specifically, we design two separate branches to encode semantic and geometric representations, respectively (Sec. III-B). Both branches are compact and lightweight. The semantic branch is a stack of 3D sparse convolutions for learning multi-scale semantic context. The completion branch uses several dense 3D convolutions to acquire multi-scale geometry structures from different stages. Based on the representation separation, the BEV fusion network equipped with the proposed ARF module is presented to aggregate informative multi-level features from semantic/completion branches for final semantic scene completion results (Sec. III-C). Fig. 2 illustrates the overall architecture of the proposed SSC-RS.

[1] https://codalab.lisn.upsaclay.fr/competitions/7170#results

Authorized licensed use limited to: Southeast University. Downloaded on August 18,2026 at 08:41:27 UTC from IEEE Xplore. Restrictions apply.

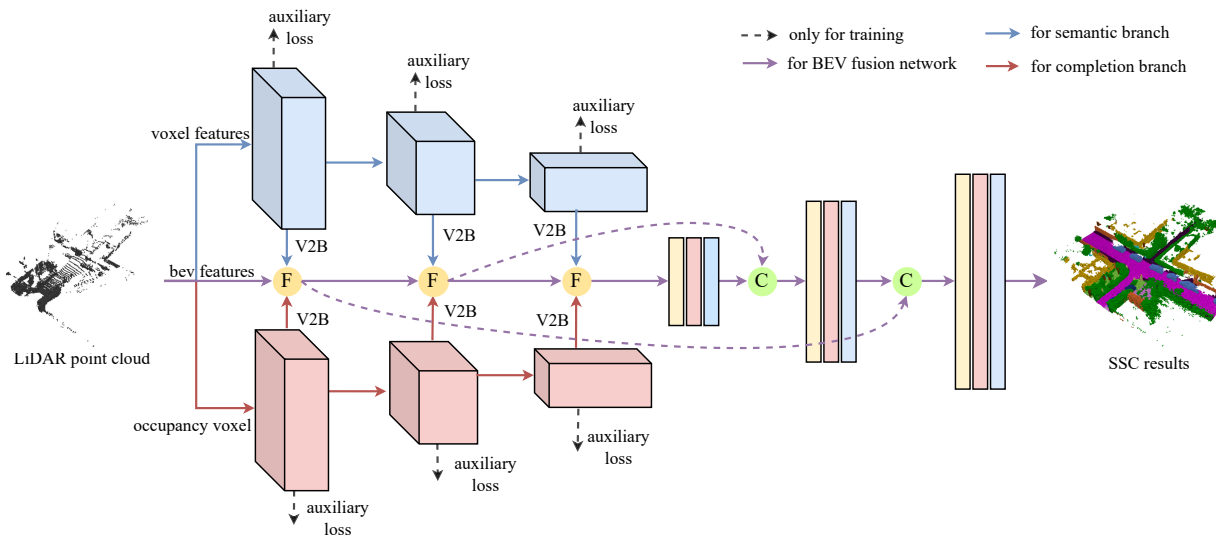

Fig. 2. The overview of the proposed SSC-RS. Two branches (semantic/completion branches) are used to learn semantic and geometric representations separately. Both branches are supervised by multi-level auxiliary losses, which will be removed during inference. The multi-scale semantic representations from the semantic branch (blue, sparse 3D CNN) and geometric representations from the completion branch (red, dense 3D CNN) will be fused by the adaptive representation (ARF) module in the BEV fusion network (purple). 'F' denotes the ARF module, and 'C' indicates concatenation along the channel dimension. V2B represents the projection from voxel to BEV.

B. Semantic-completion Representation Separation

As discussed above, semantic context and geometry structures are vital cues for semantic scene completion tasks. Thus, according to the inherent properties of these two types of clues, we design separate architectures for learning semantic and geometric representation, respectively.

Semantic Representation: To encode multi-scale semantic context and improve semantic accuracy, we introduce a compact semantic branch consisting of a voxelization layer and three sparse encoder blocks sharing a similar architecture as shown in Fig. 3 (a). The voxelization layer takes the point cloud $P \in \mathbb{R}^{N\times 3}$ in the range of $[R_x, R_y, R_z]$ as input and outputs sparse voxel features $F_V \in \mathbb{R}^{M\times C}$ with a dense spatial resolution of $L \times W \times H$. It discretizes a point $p_i = (x_i, y_i, z_i)$ to its voxel index V_i through:

$$V_i = (\lfloor x_i/s \rfloor, \lfloor y_i/s \rfloor, \lfloor z_i/s \rfloor) \tag{1}$$

where s is the voxelization resolution and $\lfloor \cdot \rfloor$ is a floor function. Since a occupied voxel could contain multiple points, the voxel features f_{V_m} indexed by $V_m \in \mathbb{Z}^{L\times W\times H}$ are aggregated by:

$$f_{V_m} = \mathrm{R}_f \left\{ \underset{V_p = V_m}{\mathrm{A}_f} \left(\mathrm{MLP}(f_p) \right) \right\} \tag{2}$$

where A_f is aggregation function (e.g. max function) and R_f denotes MLPs for dimension reduction. We concatenate the point coordinates, distance offset from the center of the voxel where the point locates, and reflection intensity as the point features f_p.

After the voxelization layer, the voxel features are fed into three cascade sparse encoder blocks to obtain sparse semantic features $(F_{s,1}, F_{s,2}, F_{s,3})$. Each sparse encoder block consists of a residual block [37] with sparse convolutions and an SGFE module developed in [38]. The SGFE module exploits multi-scale sparse projections and attentive scale selection to enhance the voxel-wise features with more geometric guidance and downscales the features' dense resolution by factor 2.

Also, similar to [38], we adopt multi-scale sparse supervision to facilitate the learning of semantic context, as shown in Fig. 2. Specifically, during the training stage, we attach lightweight MLPs as the auxiliary heads after each encoder block to get the semantic predictions of valid voxels. The voxelized semantic labels at different scales are generated by SSC labels according to occupancy grids. Note that point-wise semantic labels are unnecessary since we only apply voxel-wise supervision and compute loss on the valid voxels to avoid unnecessary computation and memory usage. We use lovasz loss [39] and cross-entropy loss to optimize the semantic branch. The semantic loss L_s is the summation of the loss of each stage, which can be expressed as:

$$L_s = \sum_{i=1}^{3} (L_{lovasz,i} + L_{ce,i}) \tag{3}$$

Note that the auxiliary heads are removed on the inference stage for efficiency, and our semantic branch only contains 1.45 M parameters.

Geometric Representation: The completion branch takes the occupancy voxels $O_V \in \mathbb{R}^{1\times L\times W\times H}$ generated by the LiDAR point cloud, indicating if voxels are occupied by laser measurements. It outputs multi-scale dense completion features $(F_{c,1}, F_{c,2}, F_{c,3})$ for more geometry details. Since the completion branch aims only to complete the semantic-agnostic scene, i.e., binary completion, we design a shallow architecture with dense 3D convolutions to obtain the geometry details of the scene. As shown in Fig. 3 (b), the com-

Authorized licensed use limited to: Southeast University. Downloaded on August 18,2026 at 08:41:27 UTC from IEEE Xplore. Restrictions apply.

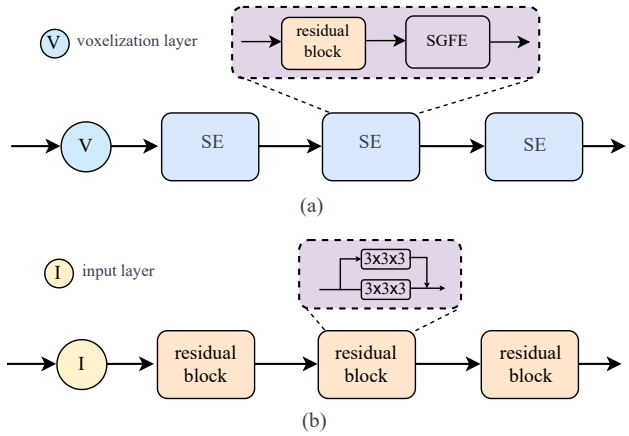

Fig. 3. (a) The architecture of the semantic branch. 'SE' denotes the sparse encoder block. (b) The overview of the completion branch. The max-pooling layers are inserted before each residual block.

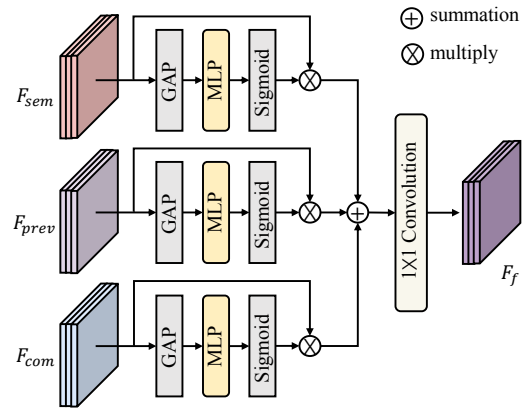

Fig. 4. Our designed adaptive representation fusion module. GAP denotes global average pooling.

pletion branch consists of an input layer and three residual blocks. The input layer is a dense 3D convolution with kernel size $7\times7\times7$ for a large receptive field, and the residual block is a stack of dense 3D convolutions with kernel size $3\times3\times3$. Also, the max-pooling layers are applied before each residual block to downscale the size of the feature map by factor 2. Similar to the semantic branch, deep supervision is used to enhance the multi-scale geometric representation. To this end, we attach MLPs as auxiliary heads after each block to obtain the binary prediction indicating the occupancy of the completed scene. And the training loss L_c for this branch is computed by:

$$L_c = \sum_{i=1}^{3} (L_{lovasz,i} + L_{bce,i}) \tag{4}$$

where i denotes the i-th stage of the completion branch and L_{bce} indicates the binary cross-entropy loss. During the inference, the auxiliary heads are removed. And due to the lightweight design (0.31 M parameters), the computational overhead of the completion branch is small (7.93G MACs with input shape $256\times256\times32$).

C. BEV Fusion Network

Since using dense 3D convolutions to fuse dense 3D feature maps brings a significant overhead on memory usage and slows down the running speed greatly, inspired by the success of BEV perception in 3D object detection and semantic segmentation, we develop a BEV fusion network to aggregate the multi-scale sparse semantic representations $(F_V, F_{s,1}, F_{s,2}, F_{s,3})$ and dense geometric representations $(O_V, F_{c,1}, F_{c,2}, F_{c,3})$ from the BEV.

We first elaborate on the BEV projection of features from semantic/completion branches. For sparse semantic features $F_{s,*}$, we first generate the BEV indices from the voxel indices. Then similar to the voxelization layer in the semantic branch, we use the aggregation function (max function) to aggregate the features with the same BEV index to get the sparse BEV features. Finally, according to the BEV indices and sparse BEV features, we generate the dense BEV features $(F^b_{s,0} \in \mathbb{R}^{C_0\times H\times W}, F^b_{s,1} \in \mathbb{R}^{C_1\times(H//2)\times(W//2)}, F^b_{s,2} \in \mathbb{R}^{C_2\times(H//4)\times(W//4)}, F^b_{s,3} \in \mathbb{R}^{C_3\times(H//8)\times(W//8)})$. Compared to [5], which stacks the dense 3D semantic features along the z-axis for BEV features, our used projection method is more efficient and requires less memory overhead. For dense features $F_{c,*}$ from the completion branch, we simply stack the dense 3D features along the z-axis and reduce the feature dimensions with 2D convolutions for dense BEV features $(F^b_{c,0}, F^b_{c,1}, F^b_{c,2}, F^b_{c,3})$, which keep the same dimensions as $(F^b_{s,0}, F^b_{s,1}, F^b_{s,2}, F^b_{s,3})$.

Similar to [5], our BEV fusion network is U-Net architecture with 2D convolutions. The encoder consists of an input layer and four residual blocks. Each residual block reduces the resolution size of input features by 2 to keep the same resolution as the semantic/completion features. The concatenation of features $F^b_{s,0}$ and $F^b_{c,0}$ is first fed into the input layer and then into the first residual block. Before the next residual block, an Adaptive Representation Fusion (ARF, detailed below) module takes the previous stage's output and semantic/geometric representations at the same scale as inputs and outputs the fused features containing informative semantic context and geometric structure. The decoder upscales the compressed features from the encoder three times by a factor of two at a time through skip connections. And the last convolution of the decoder outputs the SSC prediction $Y \in \mathbb{R}^{((C_n+1)\cdot L)\times H\times W}$, where C_n is the number of semantic classes. The prediction Y is further reshaped as the size of $((C_n+1)\times L\times H\times W)$, representing the semantic occupancy prediction of each voxel of the completed scene. Unlike the semantic/completion branches, we only apply supervision to the final prediction. Both lovasz loss and cross-entropy loss are used to compute the BEV loss L_{bev}:

$$L_{bev} = (L_{lovasz} + L_{ce}) \tag{5}$$

Adaptive Representation Fusion Module: Directly concatenating representations from different sources (semantic/completion/BEV branches) similar to SSA-SC implies an equal preference for these representations. However, we

Authorized licensed use limited to: Southeast University. Downloaded on August 18,2026 at 08:41:27 UTC from IEEE Xplore. Restrictions apply.

usually need selective cues from different sources. To better fuse the semantic and geometric representations, we design an adaptive representation fusion module for the BEV fusion network. Fig. 4 illustrates the detailed procedure of our ARF module. Let $F_{prev}, F_{sem}, F_{com}$ represent features from the previous stage, features from the semantic branch, and features from the completion branch, respectively. We first compute channel attention for features $F_{prev}/F_{sem}/F_{com}$ to weight the feature channels adaptively. Then the weighted features are summed and passed into a 1×1 convolution to obtain the fused features F_f. The procedure is formulated as:

$$\begin{aligned} F_f = \phi\{&\sigma[\mathrm{MLP}(\mathrm{AvgPool}(F_{prev}))] * F_{prev} \\ &+\sigma[\mathrm{MLP}(\mathrm{AvgPool}(F_{sem}))] * F_{sem} \\ &+\sigma[\mathrm{MLP}(\mathrm{AvgPool}(F_{com}))] * F_{com}\} \end{aligned} \tag{6}$$

where σ denotes the *sigmoid* function. ϕ is the 1×1 convolution.

D. Multi-task learning

We train the whole network end-to-end. The multi-task loss L_{total} is expressed as:

$$L_{total} = 3 \cdot L_{bev} + L_s + L_c \tag{7}$$

where L_{bev} is the BEV loss defined in Sec. III-C. L_s, L_c are the semantic loss and completion loss defined in Sec. III-B.

IV. Experiments

In this section, we introduce the implementation details of the proposed SSC-RS and conduct extensive experiments on the large-scale outdoor scenarios dataset SemanticKITTI [31] to show that SSC-RS achieves state-of-the-art performance. Also, we provide visualizations and qualitative analysis to demonstrate the effectiveness of our model. Moreover, ablation studies on semantic/geometric representation, ARF module, and multi-scale supervision are given to validate proposed components.

A. Datasets and Metrics

Datasets SemanticKITTI [31] is based on the KITTI odometry dataset [40], which collects 22 LiDAR sequences with 20 classes in the scenes of autonomous driving using a Velodyne HDL-64 laser scanner. According to the official setting for semantic scene completion, sequences from 00 to 10, except 08 (3834 scans), are for training, sequence 08 (815 scans) is for validation, and the rest (3901 scans) is for testing. The voxelized groud-truth labels with resolution $256 \times 256 \times 32$ of train and validation set are provided for the users.

Metrics Following [1], we compute the Intersection-over-Union (IoU) for scene completion (ignoring semantics) and mIoU of $C_n = 19$ classes (no "unlabeled" class) for semantic scene completion as the evaluation protocol. The mIoU is calculated by:

$$mIoU = \frac{1}{C_n} \sum_{c=1}^{C_n} \frac{TP_c}{TN_c + FP_c + FN_c} \tag{8}$$

where TP_c, TN_c, FP_c, and FN_c denote true positive, true negative, false positive, and false negative for class c.

B. Implementation Details

According to the official protocols, the range $[R_x, R_y, R_z]$ of input point cloud is set $[0 \sim 51.2m, -25.6 \sim 25.6m, -2 \sim 4.4m]$, the voxelization resolution s is $0.2m$, and the spatial resolution is $(L = 256, W = 256, H = 32)$. The input point cloud is augmented by randomly x-y flipping during the training procedure. And we use Adam optimizer [41] with an initial learning rate of 0.001 ($\beta_1 = 0.9, \beta_2 = 0.999$) to train SSC-RS end-to-end. The model is trained for 40 epochs on a single NVIDIA 3090 with batch size 2.

C. Comparison with the state-of-the-art.

Quantitative Results. We compare with the state-of-the-art on SemanticKITTI test set. We submit the results to the official test server to evaluate the performance of our proposed SSC-RS. Table I shows that our SSC-RS achieves the best performance on the completion metric IoU (59.7%) and ranks 2rd in terms of the scene completion metric mIoU (24.2%). Our SSC-RS also has low latency and runs in real-time (16.7 fps). UDNet [2], which adopts dense 3D CNN, has comparable performance on IoU to SSC-RS, while SSC-RS surpasses UDNet by 4.7% on mIoU and has lower latency. And compared to the semantic segmentation-assistant method SSA-SC [5], SSC-RS obtains 0.9% improvement on IoU and 0.7% on mIoU, which demonstrates the effectiveness of our proposed semantic/geometric representation separation. J3SC-Net [4] attaches semantic scene completion after segmentation. And our SSC-RS also outperforms J3SC-Net by 0.4% on mIoU, especially 3.1% on IoU. We notice that our SSC-RS has lower performance on mIoU than S3CNet [6]. We explain that the local geometric loss in S3CNet helps a lot, especially on small objects such as persons, bicycles, and motorcycles, while we don't make a special design on that. Notably, SSC-RS performs better on IoU (14.1%) and runs $\sim 14\times$ faster than S3CNet.

Qualitative Results. We provide the visualizations on SemanticKITTI validation as illustrated in Fig. 5. We also visualize the results of SSA-SC [5] and JS3C-Net [4] for comparison. From Fig 5, we can see that our SSC-RS predicts more accurate SSC results, especially on "plane" classes and large objects such as cars, consistent with the results in Table. I. While we also notice our SSA-SC fails to complete some hard samples, such as small objects, which is also difficult for most methods. We believe that some special designs for local geometry, such as geometry loss in S3CNet [6], help solve the problem, which also will be our future work.

D. Ablation Study

We conduct detailed ablation studies to validate the effectiveness of our proposed components. All experiments are trained with the same training configurations and tested on the SemanticKITTI validation set with a single GPU.

Impact of the semantic representation. We remove the semantic branch to explore the effectiveness of semantic representation. As Table II shows, without the semantic branch, the performance drops from 24.8% to 21.9% on mIoU (line

Authorized licensed use limited to: Southeast University. Downloaded on August 18,2026 at 08:41:27 UTC from IEEE Xplore. Restrictions apply.

TABLE I

COMPARISON OF PUBLISHED METHODS ON THE OFFICIAL SEMANTICKITTI [31] BENCHMARK (HIDDEN TEST). OUR NETWORK SURPASSES ALL THE PUBLISHED METHODS ON COMPLETION METRICS (IOU) AND RANKS 2rd IN TERMS OF THE SEMANTIC SEGMENTATION METRICS (MIOU). THE SPEED OF LMSCNET, JS3CNET, SSA-SC, AND OUR SSC-RS IS TESTED ON THE SAME DEVICE(NVIDIA 1080 TI).

Approach	IoU	road	sidewalk	parking	other-ground	building	car	truck	bicycle	motorcycle	other-vehicles	vegetation	trunk	terrain	person	bicyclist	motorcyclist	fence	pole	traffic-sign	mIoU	FPS
LMSCNet [3]	55.3	64.0	33.1	24.9	3.2	38.7	29.5	2.5	0.0	0.0	0.1	40.5	19.0	30.8	0.0	0.0	0.0	20.5	15.7	0.5	17.0	8.5
LMSCNet-singlescale [3]	56.7	64.8	34.7	29.0	4.6	38.1	30.9	1.5	0.0	0.0	0.8	41.3	19.9	32.1	0.0	0.0	0.0	21.3	15.0	0.8	17.6	-
Local-DIFs [30]	57.7	67.9	42.9	**40.1**	11.4	40.4	34.8	4.4	3.6	2.4	4.8	42.2	26.5	39.1	2.5	1.1	0.0	29.0	21.3	17.5	22.7	-
JS3C-Net [4]	56.6	64.7	39.9	34.9	14.1	39.4	33.3	**7.2**	14.4	8.8	12.7	43.1	19.6	40.5	8.0	5.1	0.4	30.4	18.9	15.9	23.8	1.7
S3CNet [6]	45.6	42.0	22.5	17.0	7.9	52.2	31.2	6.7	**41.5**	**45.0**	**16.1**	39.5	**34.0**	21.2	**45.9**	**35.8**	**16.0**	**31.3**	**31.0**	**24.3**	**29.5**	1.2
UDNet [2]	59.4	62.0	35.1	28.2	9.1	39.5	33.9	3.8	0.8	0.4	4.4	40.9	23.2	32.3	0.5	0.3	0.3	24.4	18.8	13.1	19.5	13.7
SSA-SC [5]	58.8	72.2	43.7	37.4	10.9	43.6	**36.5**	5.7	13.9	4.6	7.4	43.5	25.6	41.8	4.4	2.6	0.7	30.7	14.5	6.9	23.5	**20.0**
SSC-RS(ours)	**59.7**	**73.1**	**44.4**	38.6	**17.4**	**44.6**	36.4	5.3	10.1	5.1	11.2	**44.1**	26.0	**41.9**	4.7	2.4	0.9	30.8	15.0	7.2	24.2	16.7

TABLE II

ABLATION STUDIES ON SEMANTICKITT VALIDATION. BEV, SEM, AND COM DENOTE THE BEV FUSION NETWORK, SEMANTIC BRANCH, AND COMPLETION BRANCH. MSS INDICATES MULTI-SCALE SUPERVISION (DEEP SUPERVISION) ON THE SEMANTIC/COMPLETION BRANCH.

Variants	BEV	SEM	COM	ARF	MSS	IoU	mIoU	Precision	Recall	road	sidewalk	parking	other-ground	building	car	truck	bicycle	motorcycle	other-vehicles	vegetation	trunk	terrain	person	bicyclist	motorcyclist	fence	pole	traffic-sign
full model	✓	✓	✓	✓	✓	**58.6**	**24.8**	78.5	69.8	**73.8**	**45.3**	26.6	2.1	41.0	**46.8**	**41.5**	1.5	6.9	**19.8**	**42.6**	**22.2**	50.6	6.2	1.5	0.0	**15.8**	**17.9**	**4.6**
BEV only	✓	×	×	×	×	57.1	22.0	80.6	66.3	73.5	43.5	26.1	**3.2**	39.7	44.6	20.5	2.4	5.3	12.6	40.5	17.9	49.6	4.7	1.0	0.0	12.8	16.4	4.5
w/o COM	✓	✓	×	✓	✓	58.3	23.9	80.3	68.0	72.7	45.1	18.9	0.6	40.3	47.1	37.5	**5.4**	**10.2**	17.2	41.7	20.7	**51.1**	6.5	**2.7**	0.0	14.4	17.4	4.6
w/o SEM	✓	×	✓	✓	✓	58.1	21.9	77.9	69.6	72.5	42.3	25.1	1.3	39.3	44.9	22.2	1.0	4.6	15.4	40.0	18.5	49.9	5.5	0.9	0.0	11.1	16.4	4.4
w/o ARF	✓	✓	✓	×	✓	58.4	24.4	79.8	68.6	73.4	45.0	**26.8**	3.2	**41.1**	46.5	36.5	5.9	8.8	18.5	41.5	20.3	50.8	**6.8**	2.0	0.0	14.5	17.2	4.2
w/o MSS	✓	✓	✓	✓	×	58.1	22.7	77.4	69.9	73.5	43.5	25.8	1.7	40.4	45.7	21.7	4.1	8.2	17.3	41.3	19.7	50.4	5.3	1.2	0.0	11.2	16.2	4.1

1 vs. line 4). Due to the sparse design, the inference speed is still fast (16.7 fps as shown in Table I) when equipped with the semantic branch. And compared with the baseline model with the BEV network only, the semantic representation and ARF module improve the performance by 1.9% on mIoU and 1.2% on IoU (line 2 vs. line 3). It shows that semantic representation plays a vital role in SSC tasks.

Effect of the geometric representation. We further show the effect of the geometric representation. Line 3 of Table II provides the results without the completion branch. The completion branch brings 0.3% improvement on IoU and 0.9% gains on mIoU (line 1 vs. line 3) with 0.31 M parameters only. And comparing line 4 with line 2, when fusing with geometric representation using the ARF module, the performance is boosted by 1% on IoU, demonstrating its effectiveness in improving the accuracy of scene completion.

Ablation study on ARF module. We remove the ARF module and directly concatenate the features from different sources as the fused features to show the effectiveness of our ARF module. As shown in Table. II, the ARF module (line 1 vs. line 5) boosts the performance by 0.4% on mIou and 0.2% on IoU, which demonstrates the ARF model can select judicious cues from different sources and fuse the representations effectively.

Multi-scale supervision (deep supervision). Finally, we demonstrate the multi-scale supervision (MSS) on both branches is vital to our SSC-RS. The last line of Table. II shows the detailed results. Without multi-scale supervision, the performance drops a lot (2.1% on mIou, 0.5% on IoU). It shows that MSS can effectively facilitate the learning procedure of representation separation, which is important for SSC tasks.

V. CONCLUSION

In this paper, we develop SSC-RS to solve outdoor large-scale semantic scene completion from the perspective of representation separation and BEV fusion. Two separate branches with deep supervision are devised to disentangle the learning procedure of semantic/geometric representations explicitly, And the BEV fusion network is designed to fuse the multi-level features effectively and efficiently. Furthermore, an adaptive representation fusion module in the BEV fusion network is proposed to facilitate the fusion procedure. Extensive experiments demonstrate our SSC-RS achieves state-of-the-art performance and runs in real time. We hope our work can provide a new perspective for the SSC community. And in the future, we will focus on local geometry learning to improve the performance on small objects and extend the work to more scenarios, such as indoor and monocular scenes.

Authorized licensed use limited to: Southeast University. Downloaded on August 18,2026 at 08:41:27 UTC from IEEE Xplore. Restrictions apply.

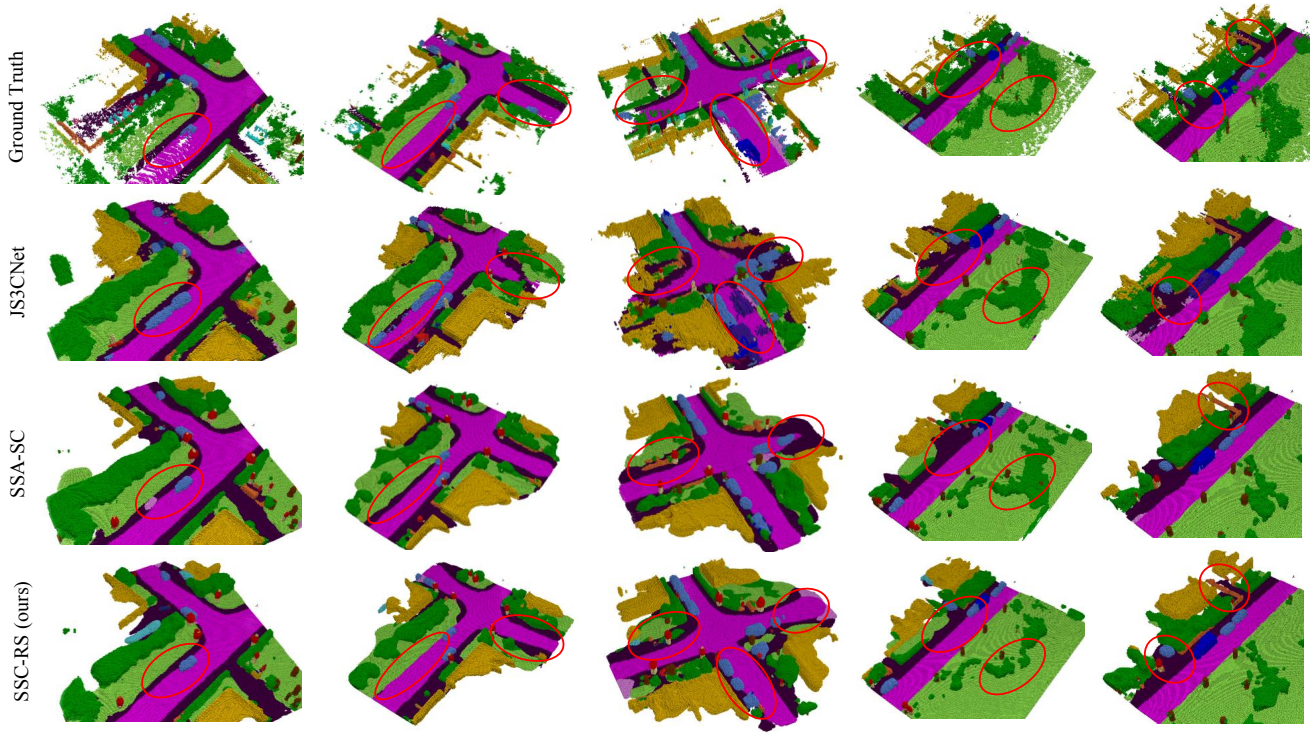

Fig. 5. Qualitative results on SemanticKITTI validation set. The visualizations demonstrate that our SSC-RS predicts more accurate SSC results, especially on "plane" classes and large objects such as cars. (labeled in red circles).

ACKNOWLEDGMENT

This work was supported by a Grant from the National Natural Science Foundation of China (No. U21A20484).

REFERENCES

[1] S. Song, F. Yu, A. Zeng, A. X. Chang, M. Savva, and T. Funkhouser, "Semantic scene completion from a single depth image," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1746–1754. 1, 2, 5

[2] H. Zou, X. Yang, T. Huang, C. Zhang, Y. Liu, W. Li, F. Wen, and H. Zhang, "Up-to-down network: Fusing multi-scale context for 3d semantic scene completion," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021, pp. 16–23. 1, 2, 5, 6

[3] L. Roldão, R. de Charette, and A. Verroust-Blondet, "Lmscnet: Lightweight multiscale 3d semantic completion," in *3DV 2020-International Virtual Conference on 3D Vision*, 2020. 1, 2, 6

[4] X. Yan, J. Gao, J. Li, R. Zhang, Z. Li, R. Huang, and S. Cui, "Sparse single sweep lidar point cloud segmentation via learning contextual shape priors from scene completion," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 4, 2021, pp. 3101–3109. 1, 2, 5, 6

[5] X. Yang, H. Zou, X. Kong, T. Huang, Y. Liu, W. Li, F. Wen, and H. Zhang, "Semantic segmentation-assisted scene completion for lidar point clouds," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021, pp. 3555–3562. 1, 2, 4, 5, 6

[6] R. Cheng, C. Agia, Y. Ren, X. Li, and L. Bingbing, "S3cnet: A sparse semantic scene completion network for lidar point clouds," in *Conference on Robot Learning*. PMLR, 2021, pp. 2148–2161. 1, 2, 5, 6

[7] M. Garbade, Y.-T. Chen, J. Sawatzky, and J. Gall, "Two stream 3d semantic scene completion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 0–0. 1

[8] E. E. Aksoy, S. Baci, and S. Cavdar, "Salsanet: Fast road and vehicle segmentation in lidar point clouds for autonomous driving," in *2020 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2019, pp. 926–932. 1, 2

[9] Y. Zhang, Z. Zhou, P. David, X. Yue, Z. Xi, B. Gong, and H. Foroosh, "Polarnet: An improved grid representation for online lidar point clouds semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9601–9610. 1, 2

[10] J. Li, X. He, Y. Wen, Y. Gao, X. Cheng, and D. Zhang, "Panoptic-phnet: Towards real-time and high-precision lidar panoptic segmentation via clustering pseudo heatmap," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 11 809–11 818. 1, 2

[11] A. H. Lang, S. Vora, H. Caesar, L. Zhou, J. Yang, and O. Beijbom, "Pointpillars: Fast encoders for object detection from point clouds," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12 697–12 705. 1, 2

[12] Z. Liu, H. Tang, A. Amini, X. Yang, H. Mao, D. Rus, and S. Han, "Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation," *arXiv preprint arXiv:2205.13542*, 2022. 1, 2

[13] Z. Zhou, Y. Zhang, and H. Foroosh, "Panoptic-polarnet: Proposal-free lidar point cloud panoptic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 13 194–13 203. 1, 2

[14] B. Graham, M. Engelcke, and L. Van Der Maaten, "3d semantic segmentation with submanifold sparse convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 9224–9232. 1

[15] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from rgbd images," in *European conference on computer vision*. Springer, 2012, pp. 746–760. 2

[16] A. Dai, D. Ritchie, M. Bokeloh, S. Reed, J. Sturm, and M. Nießner, "Scancomplete: Large-scale scene completion and semantic segmentation for 3d scans," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4578–4587. 2

[17] Y. Wang, D. J. Tan, N. Navab, and F. Tombari, "Forknet: Multi-branch volumetric semantic completion from a single depth image," in

Authorized licensed use limited to: Southeast University. Downloaded on August 18,2026 at 08:41:27 UTC from IEEE Xplore. Restrictions apply.

Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 8608–8617. 2

[18] S. Liu, Y. Hu, Y. Zeng, Q. Tang, B. Jin, Y. Han, and X. Li, "See and think: Disentangling semantic scene completion," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 2018, pp. 261–272. 2

[19] J. Li, Y. Liu, D. Gong, Q. Shi, X. Yuan, C. Zhao, and I. Reid, "Rgbd based dimensional decomposition residual network for 3d semantic scene completion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7693–7702. 2

[20] Y. Liu, J. Li, Q. Yan, X. Yuan, C. Zhao, I. Reid, and C. Cadena, "3d gated recurrent fusion for semantic scene completion," *arXiv preprint arXiv:2002.07269*, 2020. 2

[21] J. Li, K. Han, P. Wang, Y. Liu, and X. Yuan, "Anisotropic convolutional networks for 3d semantic scene completion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3351–3359. 2

[22] S. Li, C. Zou, Y. Li, X. Zhao, and Y. Gao, "Attention-based multi-modal fusion network for semantic scene completion," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 11 402–11 409. 2

[23] Y. Cai, X. Chen, C. Zhang, K.-Y. Lin, X. Wang, and H. Li, "Semantic scene completion via integrating instances and scene in-the-loop," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 324–333. 2

[24] X. Wang, M. R. Oswald, I. Cherabier, and M. Pollefeys, "Learning 3d semantic reconstruction on octrees," in *Pattern Recognition: 41st DAGM German Conference, DAGM GCPR 2019, Dortmund, Germany, September 10–13, 2019, Proceedings 41*. Springer, 2019, pp. 581–594. 2

[25] J. Li, Y. Liu, X. Yuan, C. Zhao, R. Siegwart, I. Reid, and C. Cadena, "Depth based semantic scene completion with position importance aware loss," *IEEE Robotics and Automation Letters*, vol. 5, no. 1, pp. 219–226, 2019. 2

[26] M. Zhong and G. Zeng, "Semantic point completion network for 3d semantic scene completion," in *ECAI 2020*. IOS Press, 2020, pp. 2824–2831. 2

[27] J. Tang, X. Chen, J. Wang, and G. Zeng, "Not all voxels are equal: Semantic scene completion from the point-voxel perspective," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 2, 2022, pp. 2352–2360. 2

[28] T. Huang, H. Zou, J. Cui, X. Yang, M. Wang, X. Zhao, J. Zhang, Y. Yuan, Y. Xu, and Y. Liu, "Rfnet: recurrent forward network for dense point cloud completion," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 12 508–12 517. 2

[29] X. Zhang, Y. Feng, S. Li, C. Zou, H. Wan, X. Zhao, Y. Guo, and Y. Gao, "View-guided point cloud completion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 15 890–15 899. 2

[30] C. B. Rist, D. Emmerichs, M. Enzweiler, and D. M. Gavrila, "Semantic scene completion using local deep implicit functions on lidar data," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 10, pp. 7205–7218, 2021. 2, 6

[31] J. Behley, M. Garbade, A. Milioto, J. Quenzel, S. Behnke, C. Stachniss, and J. Gall, "Semantickitti: A dataset for semantic scene understanding of lidar sequences," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 9297–9307. 2, 5, 6

[32] J. Zhang, H. Zhao, A. Yao, Y. Chen, L. Zhang, and H. Liao, "Efficient semantic scene completion network with spatial group convolution," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 733–749. 2

[33] J. Xu, X. Li, Y. Tang, Q. Yu, Y. Hao, L. Hu, and M. Chen, "Casfusionnet: A cascaded network for point cloud semantic scene completion by dense feature fusion," *arXiv preprint arXiv:2211.13702*, 2022. 2

[34] H. Li, C. Sima, J. Dai, W. Wang, L. Lu, H. Wang, E. Xie, Z. Li, H. Deng, H. Tian *et al.*, "Delving into the devils of bird's-eye-view perception: A review, evaluation and recipe," *arXiv preprint arXiv:2209.05324*, 2022. 2

[35] Z. Zhou, X. Zhao, Y. Wang, P. Wang, and H. Foroosh, "Centerformer: Center-based transformer for 3d object detection," in *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVIII*. Springer, 2022, pp. 496–513. 2

[36] Z. Li, W. Wang, H. Li, E. Xie, C. Sima, T. Lu, Y. Qiao, and J. Dai, "Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers," in *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IX*. Springer, 2022, pp. 1–18. 2

[37] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778. 3

[38] S. Xu, R. Wan, M. Ye, X. Zou, and T. Cao, "Sparse cross-scale attention network for efficient lidar panoptic segmentation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 3, 2022, pp. 2920–2928. 3

[39] M. Berman, A. Rannen Triki, and M. B. Blaschko, "The lovász-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4413–4421. 3

[40] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 2012, pp. 3354–3361. 5

[41] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014. 5

Authorized licensed use limited to: Southeast University. Downloaded on August 18,2026 at 08:41:27 UTC from IEEE Xplore. Restrictions apply.