remote sensing

MDPI

Article

Voxel- and Bird's-Eye-View-Based Semantic Scene Completion for LiDAR Point Clouds

Li Liang [1], Naveed Akhtar [2], Jordan Vice [1] and Ajmal Mian [1,*]

[1] Department of Computer Science and Software Engineering, The University of Western Australia, 35 Stirling Hwy, Crawley, WA 6009, Australia; li.liang@research.uwa.edu.au (L.L.); jordan.vice@uwa.edu.au (J.V.)
[2] School of Computing and Information Systems, The University of Melbourne, Parkville, VIC 3052, Australia; naveed.akhtar1@unimelb.edu.au
* Correspondence: ajmal.mian@uwa.edu.au

Abstract: Semantic scene completion is a crucial outdoor scene understanding task that has direct implications for technologies like autonomous driving and robotics. It compensates for unavoidable occlusions and partial measurements in LiDAR scans, which may otherwise cause catastrophic failures. Due to the inherent complexity of this task, existing methods generally rely on complex and computationally demanding scene completion models, which limits their practicality in downstream applications. Addressing this, we propose a novel integrated network that combines the strengths of 3D and 2D semantic scene completion techniques for efficient LiDAR point cloud scene completion. Our network leverages a newly devised lightweight multi-scale convolutional block (MSB) to efficiently aggregate multi-scale features, thereby improving the identification of small and distant objects. It further utilizes a layout-aware semantic block (LSB), developed to grasp the overall layout of the scene to precisely guide the reconstruction and recognition of features. Moreover, we also develop a feature fusion module (FFM) for effective interaction between the data derived from two disparate streams in our network, ensuring a robust and cohesive scene completion process. Extensive experiments with the popular SemanticKITTI dataset demonstrate that our method achieves highly competitive performance, with an mIoU of 35.7 and an IoU of 51.4. Notably, the proposed method achieves an mIoU improvement of 2.6 % compared to previous methods.

Keywords: LiDAR; 3D point cloud; 3D semantic scene completion; convolution

Citation: Liang, L.; Akhtar, N.; Vice, J.; Mian, A. Voxel- and Bird's-Eye-View-Based Semantic Scene Completion for LiDAR Point Clouds. *Remote Sens.* **2024**, *16*, 2266. https://doi.org/10.3390/rs16132266

Academic Editor: Wen Liu

Received: 15 April 2024
Revised: 21 May 2024
Accepted: 19 June 2024
Published: 21 June 2024

1. Introduction

Three-dimensional scene understanding is a crucial task in autonomous driving, which requires a detailed understanding of both the three-dimensional geometry and semantic information of an environment. In this domain, light detection and ranging (LiDAR) sensors play a pivotal role due to their superior capability to capture accurate distance measurements over longer ranges when compared to traditional visual cameras. However, despite their advantages, the data collected by LiDAR are characteristically sparse, which poses significant challenges for machines to achieve a complete understanding of the environment. To address this, the task of 3D semantic scene completion aims at reconstructing and semantically categorizing the entire environment using the limited data available, thereby bridging the gap in scene understanding capabilities between machines and humans.

Three-dimensional point cloud semantic scene completion is a considerably challenging task, particularly for large-scale outdoor scenarios. Nevertheless, there has been a remarkable interest in research in this direction recently [1–6]. The underlying efforts aim to enhance the precision and efficiency of reconstructing and semantically interpreting large-scale outdoor environments from sparse 3D point cloud data, which marks a significant stride towards improving the environmental comprehension of autonomous

systems. Still, existing scene completion techniques face obvious limitations [1–6]. For instance, Yan et al. [3] introduced JS3C-Net, which combines a segmentation network and a scene completion module to achieve semantic scene completion. Initially, the segmentation network is utilized to conduct preliminary semantic segmentation, and then the scene completion module utilizes the outcomes of the segmentation network to construct a comprehensive voxel representation of the entire scene. Additionally, a point-voxel interaction (PVI) module is introduced to facilitate the transfer of shape-specific knowledge between the incomplete point clouds and the completed voxels generated in the initial steps. Nevertheless, this approach introduces the risk of perpetuating initial segmentation errors throughout the completion phase, potentially resulting in the significant corruption of valuable information.

More recently, Cheng et al. [4] presented S3CNet, a network leveraging sparse convolutions to effectively learn features from sparse point cloud data. They also developed a 2D version of S3CNet, incorporating a multi-view fusion strategy aimed at improving the 3D semantic scene completion performance. Later, Xia et al. [6] designed SCPNet to improve the representation learning of the single-frame model by transferring dense, relation-based semantic knowledge from a multi-frame teacher to a single-frame student via a dense-to-sparse knowledge distillation. A completion label rectification is further proposed to remove traces of dynamic objects in the completion labels. However, the approach initially utilizes a completion network to derive completion results, followed by employing a segmentation network to obtain the segmentation outcomes, which considerably increases the model size. This is further exacerbated by the absence of downsampling operations in the model. These aspects collectively contribute to significantly increasing the memory and computational footprint of the model.

In this work, we propose an integrated approach that merges a 3D semantic scene completion network (3D SSCNet) with a 2D semantic scene completion network (2D SSCNet), enhancing semantic scene completion from both 3D and 2D perspectives. For the proposed network, as shown in Figure 1, our design includes an efficient multi-scale convolution block (MSB) within the 3D SSCNet that aggregates features across multiple scales, leveraging rich contextual information for enhancing accuracy in small objects, distant objects, and dense scenes within outdoor scenarios. A layout-aware semantic block (LSB) is also developed in the 3D SSCNet to grasp the overall layout of the scene to precisely guide the reconstruction and recognition of features. Additionally, the introduction of a 2D SSCNet complements this framework by providing additional semantic scene information, further bolstering the capabilities of the model. A pivotal component of our approach is the feature fusion module (FFM), designed for the effective integration of the 3D and 2D features from their respective networks. This combination facilitates superior scene completion. We conduct extensive experiments with the major outdoor scene completion benchmark, SemanticKITTI [7]. Experimental results on the SemanticKITTI validation dataset indicate that our method achieves highly competitive results, outperforming current state-of-the-art methods by a significant margin of 2.6 mIoU.

Our contributions are summarized as follows:

- We propose an integrated network that merges a 3D SSCNet with a 2D SSCNet. For the former, a highly efficient MSB is devised to segment small, distant, and dense objects. Moreover, an LSB is developed to grasp the overall layout information of the outdoor scenes.
- We propose the 2D SSCNet to process bird's-eye-view (BEV) features of the scene, which deliver precise spatial layout information in the two-dimensional space, thereby enhancing the overall performance of 3D semantic scene completion.
- We propose FFM for an improved interaction of the information from the 3D SSCNet and the 2D SSCNet, where the strengths of the other can enhance each set of features.

The organization of the paper is as follows: Section 2 presents a comprehensive review of the related works, outlining the key contributions and methodologies previously employed in this research area. Section 3 details the methodology applied in our study,

describing the experimental design. Section 4 discusses the experimental results, providing a thorough analysis of the data and highlighting significant findings. Finally, Section 5 presents the conclusion, summarizing the main outcomes of the research.

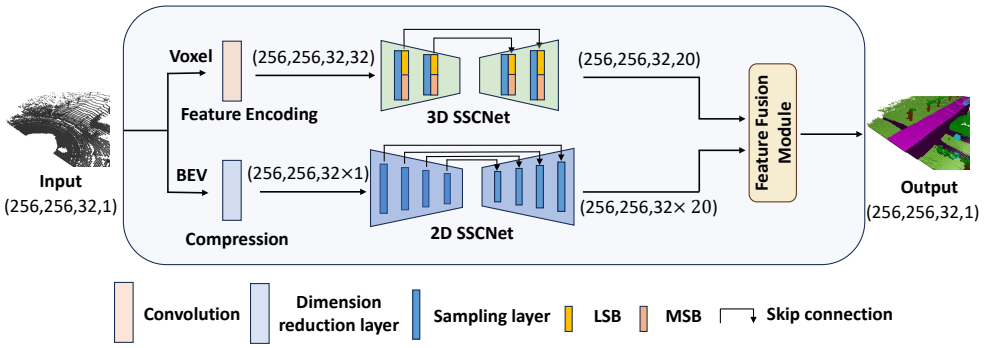

Figure 1. The network structure of the proposed method. Our method employs 3D and 2D semantic scene completion networks (SSCNets) for voxel and bird's-eye-view (BEV) feature recognition, respectively. In the upper stream, the feature encoding block is used to extract robust voxel features, which serve as the input to the 3D SSCNet, containing the layout-aware semantic block (LSB) and multi-scale convolutional block (MSB). In the lower part, the voxel feature is compressed into BEV features and then passes through the 2D SSCNet. Finally, the feature fusion module is adopted to enhance the interaction of data from the two streams.

2. Related Work

In the field of 3D semantic scene completion, existing techniques can be categorized as follows: (*i*) image-based methods, (*ii*) point-based methods, (*iii*) voxel-based methods, and (*iv*) multi-modality-based methods. Each category represents a unique methodology for processing and interpreting 3D data, with distinct advantages and challenges.

2.1. Image-Based Methods

In image-based techniques for semantic scene completion, many proposed methods in literature employ convolutional neural networks or transformers to process multi-modal data, like RGB and depth images [1,8–20]. For instance, Song et al. [1] proposed an end-to-end 3D convolutional network, introducing a dilation-based context module for efficient large-receptive-field context learning. Guo et al. [8] introduced VVNet to bridge the gap between 2D geometric features and 3D context by projecting 2D CNN-extracted features onto a volumetric space. Wang et al. [9] introduced the first generative adversarial network (GAN) for 3D semantic scene completion. This innovative approach leverages adversarial learning in both the output and latent spaces to improve results. Building on this, Wang et al. [10] presented ForkNet, which utilizes a unique architecture that features a single encoder and three generators that share a latent space for further incorporating geometric information by employing strategic connections. This design addresses inaccuracies in ground truth annotations, while multiple discriminators refine the details and realism of the completed scenes. Zhang et al. [11] developed the Cascaded Context Pyramid Network, aiming to limit the large-memory requirement and capture multi-scale spatial contexts by integrating local geometry details and multi-scale 3D contexts in a self-cascaded manner. Similarly, Cao et al. [14] developed MonoScene, which performs semantic completion from single RGB images by projecting 2D features into 3D and enhancing context awareness. Lastly, Li et al. [15] introduced VoxFormer, which employs occupancy prediction and self-attention mechanisms for depth correction and semantic segmentation, respectively. Though effective, these techniques are not tailored to deal with raw point cloud data, which limits their applicability in practical outdoor settings, especially those related to autonomous driving applications.

2.2. Point-Based Methods

These methods [2,21–28] directly apply convolution or transformer operations to original point clouds, leveraging their distance measurements and fine semantic descriptions for semantic scene completion. Roldao et al. [2] introduced LMSCNet, which uses lighter 2D convolutions to process point clouds, thus reducing the heavy computation caused by voxelization. Wang et al. [22] constructed octree-based CNNs (O-CNNs) with U-Net-like structures, where a novel output-guided skip connection is introduced to better preserve the geometric information in the input while also being robust to the input noise. This method reports high computational and memory efficiency and supports the construction of a deep network structure for 3D CNNs. Subsequently, Zhang et al. [24] developed a point cloud semantic scene completion network (PCSSC-Net), which utilizes a patch-based contextual encoder to learn point-level, patch-level, and scene-level geometric features, facilitating comprehensive information extraction from partial inputs through a divide-and-conquer strategy. Rist et al. [25] designed Deep Implicit Functions (DIFs) with spatial support derived from a 2D multi-resolution grid to produce a representation for both the geometry and semantics of 3D scenes. Note that this continuous representation avoids a trade-off between the achievable spatial output resolution and the extent of the 3D scene. Furthermore, Xiong et al. [26] proposed UltraLiDAR, which uses a sparse-to-dense data reconstruction pipeline to enhance data density and a zero-shot scheme to improve the generalization ability of the trained detection models. Xu et al. [27] presented CasFusionNet, deploying the following: (*i*) a global completion module (GCM) to produce an unsampled and completed but coarse point set, (*ii*) a semantic segmentation module (SSM) to predict the per-point semantic labels for the completed points, and (*iii*) a local refinement module (LRM), which further refines the coarse completed points and the associated labels. While point-based methods can effectively process point clouds, their high computational load and unstructured data processing remain open challenges.

2.3. Voxel-Based Methods

These methods [3–6,29–32] partition point clouds into 3D voxels and then perform convolution or transformer (or both) operations on the voxels to perform 3D semantic scene completion. Zhang et al. [29] introduced the concept of Spatial Group Convolution (SGC), utilizing sparse convolutions to efficiently process voxels by grouping them spatially, significantly reducing computational demands. Dai et al. [30] developed ScanComplete, which leverages a fully convolutional completion network for processing 3D scenes with arbitrary spatial extents, and employs a coarse-to-fine strategy to capture both local and global structures accurately. Zou et al. [31] proposed an Up-to-Down network (UDNet) to achieve large-scale semantic scene completion with an encoder–decoder architecture for voxel grids. The UDNet aggregates multi-scale context information to improve labeling coherence by UD blocks and expands the receptive field while preserving detailed geometric information through the atrous spatial pyramid pooling module. Furthermore, the designed multi-scale fusion mechanism efficiently aggregates global background information and improves semantic completion accuracy. Yan et al. [3] proposed JS3C-Net, which combines semantic segmentation and semantic scene completion: initially segmenting the scene semantically and then completing the scene's voxel representation. This method includes a PVI module for transferring shape-aware knowledge between incomplete point clouds and their voxel counterparts. Li et al. [32] proposed a novel hybrid architecture combining a discriminative model and a generative model, along with a tailored training paradigm for implicit representation. The learned shape embeddings are treated as dense boundary values that constrain the semi-supervised signed distance learning function. Recently, Xia et al. [6] designed SCPNet, aiming to enhance single-frame model representations by distilling dense relational semantic knowledge from a multi-frame teacher model to the single-frame student by introducing a label rectification technique to remove the traces of dynamic objects in completion labels. Whereas voxel-based techniques have computational

advantages over point-based methods, these gains are offset by a loss of information due to the discretization of the 3D space.

2.4. Multi-Modality-Based Methods

Semantic scene completion is achieved in [33–50] by adopting convolutions, transformers, or a combination of both, on different modalities like RGB images, depth images, and truncated signed distance functions (TSDFs), where the distance to the closest TSDF surface is computed at given 3D locations (usually voxel centers). Liu et al. [33] proposed a disentangled framework, sequentially carrying out 2D semantic segmentation, 2D–3D projection, and 3D semantic scene completion. In their technique, explicit semantic segmentation boosts performance and flexible fusion of sensor data brings extensibility. Guedes et al. [34] fused color and depth data to correct depth image inaccuracies. Garbade et al. [36] proposed a two-stream approach that leverages depth information and semantic information for 3D semantic scene completion. This method designs three-channel encoding for the semantic volume, which is not only memory efficient but also results in higher accuracies compared to single-channel encoding. It remains competitive with memory-expensive one-hot encoding. Chen et al. [38] devised a geometry-based strategy to embed depth information with low-resolution voxel representation, where a 3D sketch-aware feature embedding is proposed to explicitly encode geometric information. In addition, a lightweight 3D Sketch Hallucination module within an effective semantic scene completion framework is also devised to guide the inference of occupancy and semantic labels via a semi-supervised structure prior to the learning strategy. Cai et al. [42] developed the Scene-Instance-Scene Network (SISNet) to iteratively perform scene-to-instance and instance-to-scene conversions to encode object surrounding context and capture fine-grained object details within 3D scenes. Wang et al. [45] devised FFNet to address inconsistencies in RGB-D data and the uncertainty measurements of depth data during the fusion process of these two. Their method uses a frequency fusion module and a correlation descriptor to obtain the explicit correlation of the RGB-D feature and enhance its structural information, respectively. Recently, Wang et al. [48] proposed a Cleaner Self (CleanerS) framework, which addresses the noise problem in depth cameras by distilling intermediate supervision from a teacher network to a student network. Overall, multi-modal methods can benefit from the strengths of different data modalities. However, it is crucial to control the computational complexity of such methods for pragmatic reasons.

3. Methodology

The primary objective of 3D semantic scene completion is to infer complete geometric and semantic scene information from data that are both incomplete and sparse. The goal here is to reconstruct a complete scene by utilizing merely a single frame of point cloud data. Given an input point cloud $P \in \mathbb{R}^{N\times 3}$, we partition it into a grid of 3D voxels with the proportional dimensions $L \times W \times H$, indicating length, width, and height, respectively. The task is to accurately assign a class to each voxel within this specified grid. This means that, for each voxel, it needs to be determined whether the voxel is vacant or contains an object belonging to one of C semantic categories, represented as $c \in 0, 1, 2, \ldots, C-1$.

3.1. Methodology Overview

Our proposed network architecture for the task, presented in Figure 1, synergistically fuses 2D and 3D data representations through an integrated approach. It combines a 2D semantic scene completion network (SSCNet) and a 3D SSCNet to elevate the precision of semantic scene completion. The 3D SSCNet leverages multi-scale convolutional blocks (MSBs) that are not only parameter-efficient but are also capable of capturing rich local and global geometric features while maintaining a low memory footprint. Furthermore, our proposed layout-aware semantic blocks (LSBs) within the network are specifically tailored to comprehend the overall layout of the scene, which serves to accurately direct the feature reconstruction and recognition processes. Concurrently, the 2D SSCNet employs

standard techniques found in existing image-based networks. Moreover, we present an innovative feature fusion module that is designed to integrate voxel-based characteristics from the 3D SSCNet with the 2D BEV attributes sourced from the 2D SSCNet. This strategic fusion facilitates the incorporation of essential structural and layout information from the 2D domain into the 3D semantic scene completion process, thereby enhancing the overall accuracy and depth of scene understanding. In the forthcoming sections, we will provide an in-depth discussion of the 3D SSCNet. This will be complemented by a comprehensive exploration of the feature fusion module, elucidating its role and significance in bridging the gap between two-dimensional and three-dimensional scene analysis.

3.2. 3D Semantic Scene Completion Network

While advancements in 3D semantic scene completion have been notable, existing methodologies exhibit certain limitations. For instance, SCPNet [6] adopts a sequential approach where scene completion precedes semantic segmentation. Although this strategy improves the overall semantic scene completion fidelity, it also substantially increases the number of network parameters. Additionally, the absence of downsampling layers in SCPNet necessitates considerable memory allocation, which is a bottleneck for efficient processing. In contrast, JS3C-Net [3] inverts this order by initiating the process with semantic segmentation, subsequently using the segmented features to infer the scene completion. This approach, while computationally economical, introduces the risk of perpetuating initial segmentation errors throughout the completion phase, potentially amplifying the loss of valuable information. Furthermore, S3CNet [4] innovatively employs sparse tensor operations to effectively construct the representation of the scene through integrated 2D and 3D modalities. This technique aims to achieve semantic completion across both dimensions. However, despite its computational efficiency, the 3D module of S3CNet may not capture the full geometric information effectively, from local details to global context, which is critical for a nuanced and comprehensive semantic scene completion. To address these challenges, we propose the incorporation of the 3D layout-aware semantic block (LSB) and the multi-scale convolutional block (MSB), which can augment the capability of the network for the reconstruction of both the semantic content and structural integrity of various scenes. The structures of the LSB and MSB are shown in Figure 2.

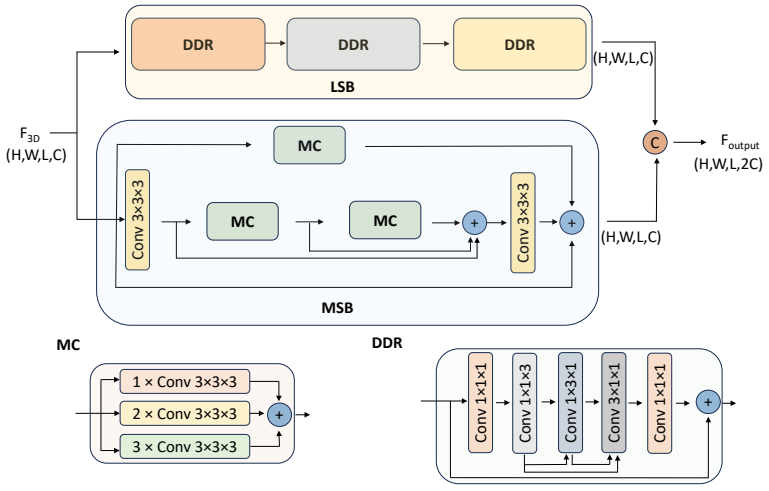

Figure 2. The structure of a semantic scene completion block within the 3D SSCNet. The upper part is the LSB, which utilizes three-dimensional decomposition residual (DDR) blocks with progressively increasing dilations to capture spatial layout context. The lower part is the MSB, composed of a series of $3 \times 3 \times 3$ convolutions that operate at different scales to extract a diverse range of geometric features from the input features, F_{3D}. The outputs from the LSB and the MSB are then concatenated as the final output feature, F_{output}.

3.2.1. Layout-Aware Semantic Block

Recent studies [51,52] have highlighted the importance of capturing comprehensive contextual and layout information for effective 3D semantic segmentation, pointing to the challenges posed by the inherent sparsity of 3D voxels and their computational demands. To deal with these challenges and enhance context acquisition within a scene, our approach introduces the LSB. This block incorporates a set of dimensional decomposition residual (DDR) blocks [37], each employing a distinct dilation rate to capture extensive scene layout information efficiently, which was inspired by prior research [38]. Specifically, the DDR block is designed to offer a computational advantage by decomposing the conventional computation in 3D CNNs. In detail, the computational cost for a traditional block in a 3D CNN is quantified as $C^{in} \times C^{out} \times k \times k \times k$ for C^{in} and C^{out} input and output channels, respectively, with $k \times k \times k$ specifying the kernel size. By breaking down kernel operations into three successive layers, i.e., $1 \times 1 \times k$, $1 \times k \times 1$, and $k \times 1 \times 1$, the DDR block significantly reduces the computational memory demands from $C^{in} \times C^{out} \times \{k \times k \times k\} \rightarrow C^{in} \times C^{out} \times \{3k\}$. This reduces the parameter count to one-third of that required by a standard $3 \times 3 \times 3$ kernel in a 3D CNN while maintaining the capacity of the network for detailed spatial layout recognition. This strategic architectural decision enables our framework to utilize parameters more efficiently, ensuring that the network can expand its receptive field without significantly increasing the computational memory load. The detailed formula for the LSB is as follows:

$$F_{out1} = \sigma(W_{d1} F_{3D} + b_1), \tag{1}$$

$$F_{out2} = \sigma(W_{d2} F_{out1} + b_2), \tag{2}$$

$$F_{out3} = \sigma(W_{d3} F_{out2} + b_3). \tag{3}$$

where $d1$, $d2$, and $d3$ are dilation rates to modify the receptive fields of the layers, represented by W_{d1}, W_{d2}, and W_{d3}, respectively. The input feature map F_{3D} is processed through these layers, with biases b_1, b_2, and b_3 added at each stage. The nonlinear activation function σ is applied to produce outputs F_{out1}, F_{out2}, and F_{out3}, with F_{out3} being the final output of the block. The increasing dilation factor enhances the ability of each layer to capture global contextual information from the input, expanding the effective receptive field and improving the capacity of the model to interpret the scene comprehensively.

3.2.2. Multi-Scale Convolutional Block

In SCPNet [6], the approach to 3D semantic scene completion involves a multi-path block featuring convolution kernels of varying sizes, $3 \times 3 \times 3$, $5 \times 5 \times 5$, and $7 \times 7 \times 7$, to capture multi-scale information. While this design allows for a wider receptive field, the deployment of larger convolutions, i.e., greater than $5 \times 5 \times 5$, significantly increases computational demands. To address this challenge and improve efficiency, we introduce the multi-scale convolutional block (MSB) based on the multi-path block within the SCPNet [6], which can extract multi-scale context in a computationally efficient manner, drawing inspiration from the principles underlying the VGG architecture [53]. Our solution involves substituting the larger convolution kernels with sequential applications of the $3 \times 3 \times 3$ convolution layers; specifically, two iterations mimic the effect of a single $5 \times 5 \times 5$ convolution, and three iterations approximate the effect of a $7 \times 7 \times 7$ convolution. This adjustment dramatically reduces computational costs from $5^3 \times C^2 = 125C^2$ for a single $5 \times 5 \times 5$ convolution to $2 \times (3^3) \times C^2 = 54C^2$ with our method, and from $7^3 \times C^2 = 343C^2$ for a $7 \times 7 \times 7$ to $3 \times (3^3) \times C^2 = 81C^2$. By adopting this strategy, not only does our framework efficiently gather critical local and global contextual information but it also ensures no significant increase in computational overhead and memory usage. Therefore, this adaptation permits a more nuanced and comprehensive analysis of semantic scenes, leveraging the benefits of multi-scale processing while circumventing the limitations associated with the use of large convolutional kernels. The MSB in our model incorporates multi-scale convolution (MC)

layers that utilize convolutions with varying kernel sizes to extract features from different receptive fields. The formula of the MC is shown below:

$$F_{out} = \sum_{i=1}^{N} W_i F_{in}. \quad (4)$$

where W_i is the kernel weights at scale i, F_{in} and F_{out} are the input and output feature maps.

3.3. 2D Semantic Scene Completion Network

To enhance the semantic scene completion performance of our proposed 3D semantic scene completion framework, we integrate a specialized lightweight 2D semantic scene completion network, inspired by S3CNet [5]. This network leverages a 2D encoder–decoder architecture optimized for semantic scene completion. Utilizing a bird's-eye-view (BEV) approach enables the transformation of the 3D scene into a 2D feature map with dimensions $L \times W$. This transformation effectively converts the complex 3D spatial data into an image-like format, making them amenable to conventional 2D convolutional neural network (CNN) methodologies. Thus, the process initiates with the deployment of one convolutional layer, specifically designed to extract preliminary voxel features from the 3D data. This is followed by a crucial feature dimension reduction phase, where the initial voxel features are compressed into BEV features. These BEV feature representations not only retain spatial and semantic information but also significantly reduce computational load and memory requirements. Therefore, the 2D SSCNet can facilitate data processing and enhance the overall capability of our framework to achieve precise and computationally efficient semantic scene completion.

In the 2D BEV network framework, we strategically deploy a series of 2D convolution layers, each developed to expand the receptive field of the network. The architecture undergoes a systematic progression through four downsampling stages, wherein each stage halves the resolution, thereby incrementally simplifying the level of detail for efficient feature processing. To further augment the capability of the network and ensure a more comprehensive feature integration, we integrate skip connections across the architecture. Subsequently, the refined outputs generated by the 2D SSCNet are put into an advanced feature fusion module.

3.4. Feature Fusion Module

We implement a novel, two-stage feature fusion module (FFM) to combine data from the two previously derived feature maps. The FFM fuses the detailed 2D SSCNet outputs with the geometric 3D SSCNet features. This fusion process leverages the spatial and semantic information richness of 3D data and the spatial layout information of 2D data to achieve improved performance in semantic scene completion tasks. The structure of the FFM network is illustrated in Figure 3. The first stage, termed the feature exchange stage, facilitates separate pathways for the features derived from each prior network. This architecture is crucial for enabling a thorough and efficient exchange of information between the two- and three-dimensional domains, ensuring that the strengths of each pathway are used to enhance the other. The second stage facilitates the feature fusion, where previously segregated features in the feature exchange stage are intricately recombined into their original dimensional contexts. This recombination is achieved through a channel mixing and embedding strategy, which is essential to enhance model comprehension and to represent complex spatial relationships more accurately.

This two-stage approach ensures that our framework not only maintains the integrity of the distinct dimensional features but also leverages the unique advantages of each. By facilitating an extensive exchange of information in the feature exchange stage and by employing a sophisticated fusion technique in the feature fusion stage, our model achieves a more holistic and nuanced understanding of the scene. Notably, these two stages are designed to complement each other and cannot function independently. Consequently, this enhances the overall capacity of the model to perform semantic scene completion tasks

with greater accuracy and efficiency, capturing the full complexity and richness of spatial relationships within the scene.

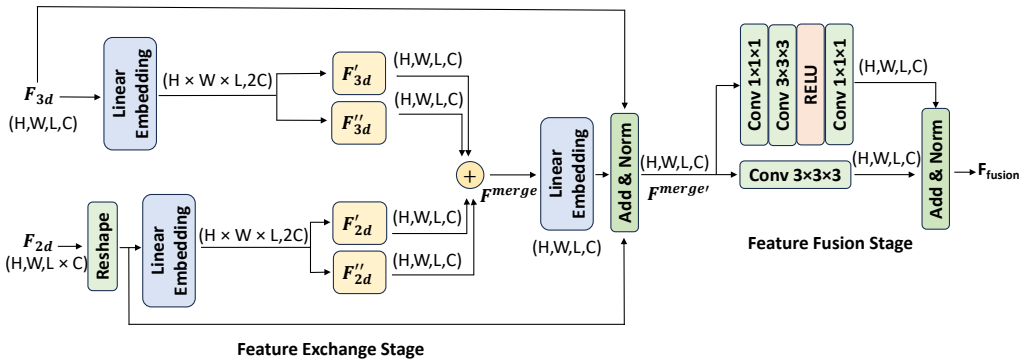

Figure 3. The network architecture of feature fusion module (FFM). The FFM is structured into two distinct stages: the feature *exchange* stage and the feature *fusion* stage. In the feature exchange stage, 3D (F_{3d}) and 2D (F_{2d}) features are first linearly embedded to obtain F'_{3d}, F''_{3d}, F'_{2d}, and F''_{2d} features. The derived features are then fed through an addition operator, with the combined 3D (F_{3d}) and 2D (F_{2d}) features then serving as the final output of the feature exchange stage. The feature fusion stage then processes these combined features through an intuitive set of convolutional layers and operations to achieve a comprehensive fusion of multi-dimensional features.

3.4.1. Feature Exchange Stage

Our proposed FFM is critical for integrating multi-dimensional feature sets. It is designed to enhance the semantic scene completion by enabling an efficient exchange and fusion of feature maps from different dimensions. In the feature exchange stage, features from the 3D SSCNet (F_{3d}) and the 2D SSCNet (F_{2d}) undergo a structured exchange. Here, each set of features is bifurcated into two pathways facilitated by a linear embedding, which generates feature vectors F'_{nd} and F''_{nd} given an n-dimensional feature set. Mathematically, the transformation for the n-dimensional features can be articulated as follows:

$$F'_{nd}, F''_{nd} = Linear(F_{nd}), \tag{5}$$

where the function *Linear* embodies the linear embedding process. The feature vectors derived from the 3D and 2D SSCNets are combined into F_{merge}, creating a robust platform for dynamic information exchange. This structured exchange is instrumental in fostering a synergistic relationship between the modalities, enhancing the effectiveness of the model. Consequently, the vectors from each modality (F'_{2d} and F''_{2d}, F'_{3d} and F''_{3d}) are combined, serving as the input into a linear projection layer, followed by normalization to yield an integrated feature representation F_{merge}. Then, the feature maps F_{merge}, F_{3d}, and F_{2d} are finally combined to form the input of the feature fusion stage F'_{merge}, i.e.,

$$F_{merge} = Add(F'_{nd}, F''_{nd}), \tag{6}$$

$$F'_{merge} = N\big(P\big(F_{3d} + F_{2d} + F_{merge}\big)\big). \tag{7}$$

where P signifies the linear projection operation that maps the combined feature vectors to a higher-level representation, and N represents the normalization operation that stabilizes the learning process by normalizing the feature distributions. By fusing the detailed spatial geometric information features captured by the 3D SSCNet with the textural and spatial layout contextual information identified by the 2D SSCNet, our framework ensures that the output is not only detail-rich but also guarantees the accuracy of the semantic structure of the scene.

3.4.2. Feature Fusion Stage

During the feature fusion stage, our methodology employs a sophisticated approach to ensure the accurate fusion of features from both 3D and 2D SSCNets. This stage first uses a $3 \times 3 \times 3$ convolution layer, which is adept at efficiently extracting and combining features across different channels. This step improves segmentation precision by extracting contextual information from neighboring regions. Drawing upon the principles established in advanced neural network architectures [54,55], this stage incorporates an element: a depth-wise $3 \times 3 \times 3$ convolution layer. We then establish a skip-connection structure that significantly augments the feature integration process and enables the accurate merging of feature vectors. By adopting this, our network capitalizes on the wealth of contextual information inherent in the scene, thus fostering a more robust and comprehensive representation of semantic information. This fusion strategy amplifies the descriptive capabilities of the model while maintaining computational efficiency, which is essential for the scalability of real-world applications.

3.5. Overall objective

The overall loss function of our network intuitively combines two pivotal components: the cross-entropy loss (CE) and the Lovasz-softmax loss. The Lovasz-softmax loss is designed to optimize the mean intersection-over-union (mIoU), a key performance metric in semantic segmentation, and can be mathematically represented as

$$\mathcal{L}_{Lovasz} = \frac{1}{|C|} \sum_{c=1}^{C} J(e(c)), \tag{8}$$

where J signifies the Lovasz extension of the IoU, manifesting as a piecewise linear function that directly targets the minimization of the mIoU error, and $e(c)$ denotes the vector of errors for each class c within the set of classes C. The cross-entropy loss function, a staple in classification tasks, is utilized in parallel, minimizing the prediction error and thus improving accuracy, and is given as

$$\mathcal{L}_{CE} = -\sum_{i} y_i \log(\hat{y}_i), \tag{9}$$

where $\hat{y}_i$ and y_i represent the predicted probability distribution and the ground truth label distribution for the ith element. Integrating these loss functions, the overall objective function for our model is formulated as

$$\mathcal{L} = \mathcal{L}_{CE} + \beta \mathcal{L}_{Lovasz}. \tag{10}$$

where β is the balancing coefficient that modulates the influence of each loss component. This calibration allows us to maintain a balance between optimizing for overall accuracy and directly improving the mIoU metric, thereby leading to an accurate model that is also aligned with the core evaluation criteria for semantic scene completion tasks. Through this dual-component loss function, the network is trained to achieve an understanding of the scene, enabling it to classify various elements accurately within the semantic landscape.

4. Experiments

The proposed network is rigorously tested on the SemanticKITTI dataset [7], a comprehensive LiDAR-based benchmark for semantic scene completion. This section outlines the characteristics of the dataset and the metrics utilized for evaluation. Next, we delve into the specific implementation details of the proposed network, laying the groundwork for its operational understanding. Then, we report our detailed experimental results and compare them to the existing methods that are also validated on SemanticKITTI. Finally, we conduct a series of ablation studies designed to evaluate the impact of individual components of the proposed approach on overall performance. Through this systematic exploration, we

aim to substantiate the efficacy of our approach and its contribution to the advancement of semantic scene completion tasks.

4.1. Datasets and Evaluation Metrics

Datasets. In our study, we rigorously assess the efficacy of our proposed methodology using the widely recognized semantic scene completion benchmark, i.e., SemanticKITTI [7]. SemanticKITTI presents a comprehensive dataset composed of LiDAR (Velodyne HDL-64E laser scanner, Velodyne Lidar, San Jose, CA, USA) point clouds, annotated at the point level. Specifically tailored for semantic scene completion tasks, SemanticKITTI provides ground truth semantic labels across a sequence of successive point cloud frames. Adhering to the established guidelines, we select individual frames from the raw point cloud data, confined within $[0 \sim 51.2, \text{m}, -25.6 \sim 25.6, \text{m}, -2 \sim 4.4, \text{m}]$ spatial dimensions. This selected frame is subsequently partitioned into voxels with a resolution of 0.2 m, resulting in a detailed scene representation encapsulated within a grid of $256 \times 256 \times 32$ voxels. Our objective is to reconstruct an exhaustive scene model within these incomplete and partial scenes. The SemanticKITTI dataset encompasses a total of 22 sequences and has 19 distinct semantic categories designated for both training and evaluation purposes. For training, we utilize Sequences 0–7 and 9–10, which collectively offer a compilation of 3834 scans. Sequence 8, comprising 815 scans, is allocated as our validation dataset. Sequences 11–21, including 3901 scans, served as test datasets. This follows the same protocol as LMSCNet [2].

Evaluation metrics. To evaluate our proposed L3D-SCN framework and compare our results to existing works, we deploy a set of evaluation metrics for semantic scene completion, as established by Song et al. [34]. Our evaluation focuses on two critical aspects: (*i*) the accurate geometry reconstruction of the scene, and (*ii*) the precise semantic segmentation of each voxel. Thus, the core metrics we deploy for our assessment of semantic scene completion performance are the intersection-over-union (IoU), reflecting the geometric completion performance of our reconstructions, and the mean intersection-over-union (mIoU), evaluating the semantic accuracy across multiple categories. These metrics are computed as follows:

$$IoU_i = \frac{TP_i}{TP_i + FN_i + FP_i}, \tag{11}$$

$$mIoU_i = \frac{1}{C} \sum_{c=1}^{C} IoU_i, \tag{12}$$

where TP_i, FP_i, and FN_i represent the true positives, false positives, and false negatives for the ith class, providing an exact measure of our model per class. C is the total number of classes. These metrics provide a holistic evaluation of the ability of our framework to not only complete but also semantically understand scenes, thus offering a robust evaluation of our contribution to the semantic scene completion domain.

4.2. Implementation Details

For the experimental evaluation of our proposed method, we employed one Nvidia GeForce 4090 GPU (Nvidia, Santa Clara, CA, USA), equipped with 24 GB of graphical RAM. Our training was conducted on a single GPU, enabled by the resource efficiency of our method. Throughout the experiments, each model was trained for a total of 40 epochs. The training protocol for achieving the best performance with our proposed network incorporated the Adam optimization algorithm, starting with a learning rate of 0.01 and a weight decay of 0.0001 to mitigate overfitting. Additionally, we integrate an exponential learning rate scheduler to iteratively reduce the learning rate.

4.3. Results

Our comparative analysis highlights the performance of our proposed method against six state-of-the-art methods that use the SemanticKITTI validation dataset: LMSCNet [2],

UDNet [31], Local-DIFs [25], SSA-SC [5], JS3C-Net [3], and SSC-RS [56]. The results are summarized in Table 1. Note that the performance of SCPNet on the validation set depends on the implementation of downsampling within its network architecture. Its mIoU is 33.1 when incorporating downsampling layers. In contrast, our proposed network architecture, which integrates two layers of downsampling, attains an mIoU of 35.7 on the same validation set. Although SCPNet achieves a higher mIoU of 37.2 without downsampling, surpassing our results, it requires significantly greater memory resources, rendering deployment on a single RTX 4090 GPU impractical.

Table 1. Quantitative results of semantic scene completion on the SemanticKITTI validation set. Best-performing single-frame results in each column are boldfaced. w/o denotes "without", and M denotes "millions".

Methods	mIoU	Completion	Precision	Recall	Parameters (M)	Car	Bicycle	Motorcycle	Truck	Other Vehicle	Person	Bicyclist	Motorcyclist	Road	Parking	Sidewalks	Other Ground	Building	Fence	Vegetation	Trunk	Terrain	Pole	Traffic Sign
LMSCNet-SS [2]	16.8	54.2	-	-	0.4	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
UDNet [31]	20.7	**58.9**	78.5	70.9	-	42.1	1.8	2.3	25.7	11.2	2.5	1.2	0.0	67.0	20.3	37.2	2.2	36.0	11.9	40.1	18.3	45.8	23.0	3.8
Local-DIFs [25]	26.1	57.8	-	-	-	**51.3**	4.3	3.3	32.3	10.6	15.7	**24.7**	0.0	71.2	31.8	43.8	3.3	38.6	13.6	40.1	19.6	50.6	25.7	14.0
SSA-SC [5]	24.5	58.2	78.5	69.3	41.0	47.0	9.2	7.4	39.7	19.1	6.3	3.2	0.0	72.8	21.0	44.3	4.1	**41.5**	15.2	41.9	22.0	49.5	17.9	4.4
JS3C-Net [3]	24.0	57.0	71.5	**73.5**	3.1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
SSC-RS [56]	24.8	58.6	78.5	69.8	23.0	46.8	1.5	6.9	41.5	19.8	6.2	1.5	0.0	**73.8**	26.6	45.3	2.1	41.0	15.8	**42.6**	22.2	50.6	17.9	4.6
S3CNet [4]	33.1	57.1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
SCPNet [6] w/o downsampling	**37.2**	49.9	-	-	-	50.5	**28.5**	**31.7**	**58.4**	**41.4**	**19.4**	19.9	**0.2**	70.5	**60.9**	52.0	20.2	34.1	33.0	35.3	33.7	51.9	**38.3**	**27.5**
SCPNet [6] w downsampling	33.1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Ours	35.7	51.4	**83.5**	57.2	40.0	46.8	20.7	26.7	49.6	41.1	8.3	4.4	0.0	73.7	58.8	**54.0**	**27.6**	36.4	**38.2**	40.3	**35.7**	**56.6**	37.7	21.4

Our proposed semantic scene completion method outperforms existing works in the key metric of mIoU. However, it exhibits lower performance in IoU when compared to these methods. This discrepancy can be attributed primarily to the enhanced ability of the method to accurately segment small, distant, and dense objects. These types of objects, while crucial for detailed scene analysis, constitute a relatively minor proportion of the total scene composition. Consequently, their improved segmentation does not substantially influence the overall IoU value. In detail, we demonstrate superior performance in accurately completing and segmenting smaller objects like bicycles, motorcycles, poles, and traffic signs. Thus, these results support the enhanced capabilities of our model in capturing and interpreting the nuances of small-scale features within the scene. Moreover, the proposed method exhibits remarkable improvements in segmenting larger object categories, including trucks, parking spaces, sidewalks, other ground types, fences, trunks, and terrain. This balanced proficiency indicates that our approach does not singularly prioritize small objects but extends its efficacy across objects of varying sizes and complexities.

We also employ the pre-trained models and official implementations of LMSCNet [2], SSA-SC [5], and SSC-RS [56] to generate visual, qualitative results on the SemanticKITTI validation set, which we use to compare our proposed approach against. The qualitative results are presented in Figure 4. The visual analysis reinforces the advantages of our model in segmenting planar categories like parking areas, other ground types, and terrain, which is supported by our quantitative findings. Furthermore, while other comparative methods occasionally produce distorted representations, our proposed approach consistently delivers accurate and recognizable scene reconstructions. Both the qualitative and quantitative results emphasize the comprehensive advancement of the proposed method in semantic scene completion, highlighting its potential in the field. Additionally, we present the failure cases in Figure 5. The inability of the proposed methods to accurately complete objects such as bicycles and bicyclists can be attributed to the insufficient number of examples in the training dataset. This limitation restricts the capacity of the model to learn the diverse features and variations associated with these objects, which is essential for accurate object completion.

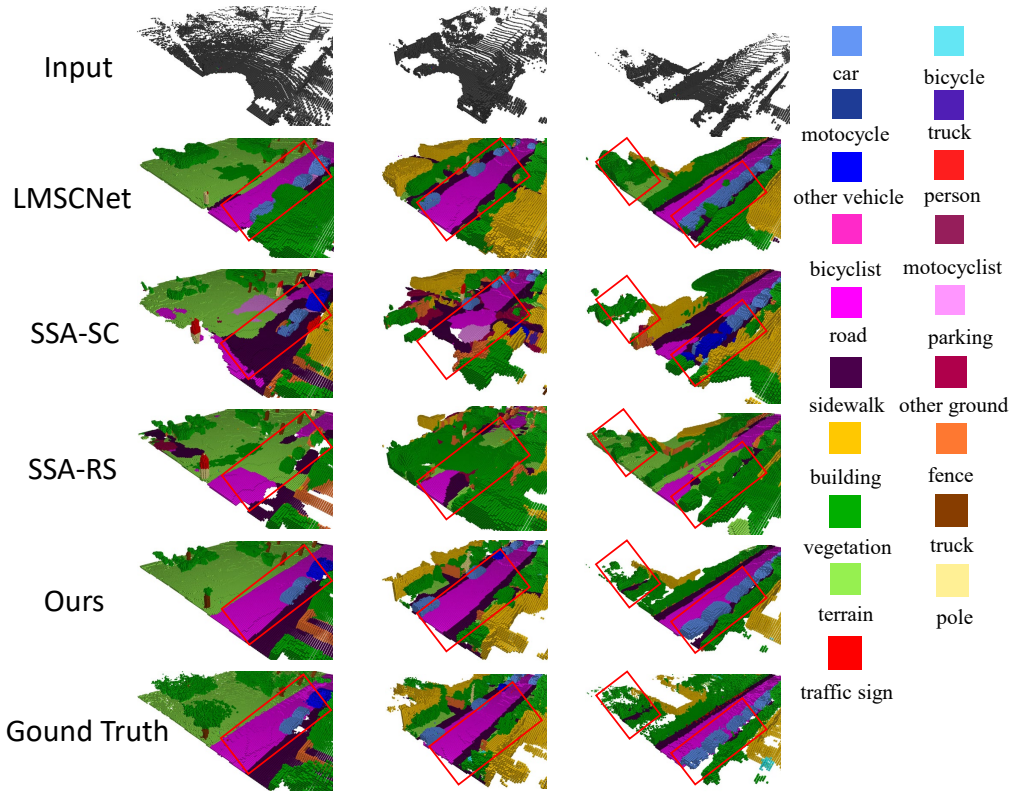

Figure 4. The visualization results on the SemanticKITTI validation set. The first row is the input data. The second row to the fifth row displays the results of LMSCNet [2], SSA-SC [5], SSA-RS [56], and the proposed method, respectively. The final row is the ground truth. We emphasize regions handled particularly well by our method using red boxes.

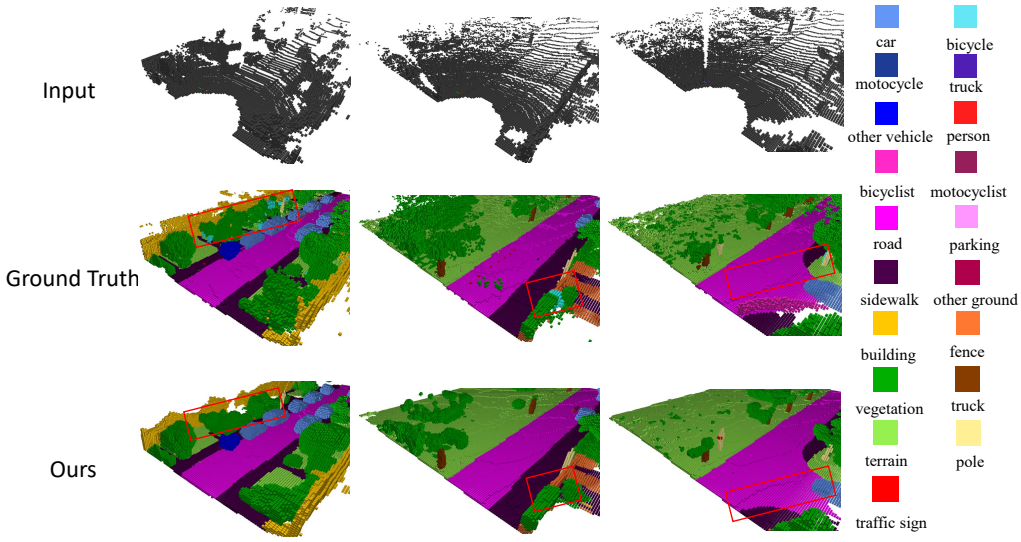

Figure 5. The visualization results of failure cases on the SemanticKITTI validation set. The first row is the input data, the second row is the ground truth, and the third row is the failure results of the proposed method.

In terms of parameter efficiency, LMSCNet [2], JS3C-Net [3], and SSC-RS [56] demonstrate higher efficiency compared to our proposed method. However, it is crucial to recognize that the performance of these models (LMSCNet [2], JS3C-Net [3], and SSC-RS [56]) falls below acceptable thresholds. This indicates that despite their lower parameter counts, their practical applicability might be constrained. Conversely, our method outperforms

these models and maintains fewer parameters than SSA-SC [5], which has 41 million parameters. Additionally, while JS3C-Net [3] records the longest runtime of 0.28 s per sample on the same GPU (NVIDIA RTX 4090), our method processes each sample in just 0.15 s. This suggests that our method not only delivers higher accuracy, but does so with greater parameter efficiency and speed. Such advantages are particularly valuable in scenarios computational resources are limited, highlighting the practical superiority of our approach.

However, there are two shortcomings in the current methodology. The first shortcoming is the lack of specific datasets. Currently, there is only one dataset, SemanticKITTI [7], in the 3D point cloud semantic scene completion. This limitation restricts the ability to train and test our models under varied conditions and may affect the generalizability of our approach. To address this, we plan to develop comprehensive datasets for 3D point cloud semantic scene completion. These datasets will support the advancement of our research and will be valuable resources for the community. By publishing these datasets, we aim to enhance model performance and achieve scene completions that are closer to real-world scenarios. Another shortcoming is network efficiency. Our current model architecture has more parameters compared to LMSCNet [2], which could hinder its applicability in scenarios where computational resources are limited. To overcome this, we are in the process of designing a more lightweight network. This revision will aim to maintain or improve the accuracy of our model while significantly reducing its computational demands, thus facilitating easier deployment and application in practical settings.

4.4. Ablation Studies

Through a series of methodical experiments, we analyze the impact of distinct blocks proposed in our framework, aiming to quantify their respective roles in enhancing the overall performance of the model. We detail the results of our ablation studies in Table 2. Additionally, we complement our analysis with qualitative results, as shown in Figure 6. These visual representations allow us to illustrate the practical effect of each block on the semantic scene completion task, offering intuitive insights into how each component contributes to refining the model output. Through these evaluations, we report the significance of each proposed block within the overall proposed architecture, showcasing their collective synergy in achieving superior performance for the scene completion task.

Table 2. Experimental results for removing different modules from our overall scene completion framework, evaluated on the SemanticKITTI validation dataset. Note that w/o denotes "without".

Method	mIoU	Completion	Car	Bicycle	Motorcycle	Truck	Other Vehicle	Person	Bicyclist	Motorcyclist	Road	Parking	Sidewalks	Other Ground	Building	Fence	Vegetation	Trunk	Terrain	Pole	Traffic Sign
w/o LSB	28.9	44.9	41.8	8.9	3.2	38.8	19.8	7.8	3.1	7.3	70.6	56.8	49.3	30.6	32.4	37.9	33.8	28.5	53.3	17.4	5.8
w/o MSB	26.7	39.1	39.0	6.9	13.7	42.3	22.0	5.2	0.0	0.0	66.4	48.7	46.2	18.5	26.5	37.1	27.6	34.7	44.2	23.2	4.6
w/o 2D SSCNet	31.9	46.4	44.8	14.3	29.8	45.9	18.8	9.3	2.5	2.7	72.5	57.2	52.2	32.6	32.2	37.3	34.8	35.0	54.0	22.9	7.5
w/o FFM	33.9	49.0	45.2	30.1	33.3	50.6	38.5	9.4	1.6	0.0	71.4	56.3	50.6	27.6	34.2	35.6	36.7	36.3	55.6	25.2	5.2
Full Model	35.7	51.4	46.8	20.7	26.7	49.6	41.1	8.3	4.4	0.0	73.7	58.8	54.0	27.6	36.4	38.2	40.3	35.7	56.6	37.7	21.4

Effect of layout-aware semantic blocks. In our comprehensive analysis, we explore the impact of incorporating LSB within our framework, specifically examining its performance on the SemanticKITTI validation set. A detailed comparison is provided, contrasting scenarios with and without integrating the LSB. As illustrated in Table 2, the inclusion of LSBs markedly enhances the ability of our model to classify accurately and segment "planar" categories such as roads, sidewalks, other ground types, and vegetation. This improvement highlights the pivotal role of the LSB in enabling the model to effectively capture and interpret spatial and semantic scene information. Through this focused in-

vestigation, we demonstrate the significant value added by these blocks to our overall methodology, reinforcing the performance of our model for semantic scene completion.

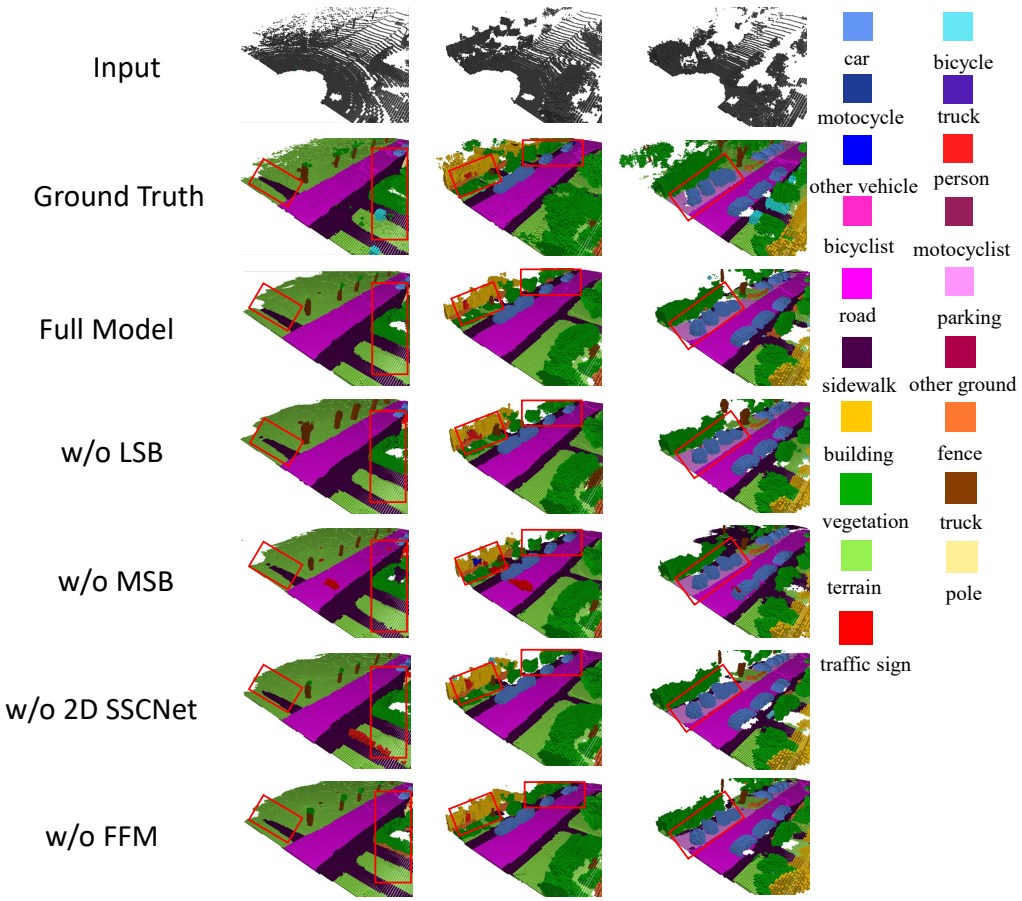

Figure 6. Qualitative results on the SemanticKITTI validation dataset for ablation studies. The first and second rows represent the input and the ground truth, respectively. From the third to final rows, we show qualitative results for the *full* proposed method with all components integrated, without (w/o) incorporating the LSB, without the MSB, bypassing the 2D SSCNet, and without deploying the multi-dimensional FFM, respectively.

Effect of multi-scale convolutional blocks. As part of our ablation studies, we scrutinized the impact of integrating the MSB into our framework. The comparative analysis, as detailed in Table 3, distinctly highlights the significant enhancements brought by these blocks. With the incorporation of MSBs, the performance of our model in semantic scene completion saw an increase from 26.7 mIoU to 35.7 mIoU, alongside an improvement in geometric scene completion from 39.1 IoU to 51.4 IoU. This substantial uplift in performance is particularly noteworthy across smaller object categories, such as bicycles, motorcycles, persons, and bicyclists, and "plane" categories like roads, parking areas, sidewalks, and other ground types. These improvements highlight the effectiveness of MSBs in capturing both nuanced local geometric details and global contextual scene information. Thus, the introduction of MSBs represents a significant advancement in our methodology and demonstrates their pivotal role in achieving superior performance, particularly in processing complex scenes with a diverse range of object sizes and spatial relationships.

Effect of 2D SSCNet. In our evaluation, we assess the impact of integrating a 2D SSCNet within our proposed framework. This comparative analysis, detailed in Table 4, highlights the performance differential when incorporating the 2D SSCNet versus its absence. The inclusion of 2D SSCNet notably enhances the performance of our method, yielding a 3.8 mIoU increase and a 5.0 IoU improvement for geometric completion. Table 2

demonstrates that the integration of 2D SSCNet significantly improves performance across various categories, including cars, bicycles, trucks, other vehicle types, poles, and traffic signs. These improvements suggest that the 2D SSCNet effectively complements the three-dimensional analysis by providing spatial layout information from the two-dimensional space in the semantic scene completion process. As a result, the addition of this component demonstrates the importance of leveraging two-dimensional insights to enrich the perception and interpretation of the model in three-dimensional spaces.

Effect of feature fusion model. The comparative analysis, as summarized in Table 5, showcases the improvements made when integrating the FFM in our method. This integration elevates the performance for semantic scene completion from 33.9 mIoU to 35.7 mIoU and for geometric completion from 49.0 IoU to 51.4 IoU. This evidences the effectiveness of our proposed FFM, which is designed with feature exchange and feature fusion stages. The feature exchange stage facilitates a comprehensive interchange of attributes between different modal inputs, while the feature fusion stage adeptly integrates these attributes, optimizing the overall semantic and geometric understanding of the scene. Thus, it is clear that these two stages are not only pivotal in enhancing the capability of the model to process multidimensional information fusion but also illustrate the substantial benefits of two such stages in improving the accuracy and detail of semantic scene completion tasks.

Table 3. Impact of multi-scale information blocks on the performance with SemanticKITTI val dataset.

Methods	mIoU	Completion	Car	Bicycle	Motorcycle	Truck	Other Vehicle	Person	Bicyclist	Motorcyclist	Road	Parking	Sidewalks	Other Ground	Building	Fence	Vegetation	Trunk	Terrain	Pole	Traffic Sign
Our Method w/o MSB	26.7	39.1	39.0	6.9	13.7	42.3	22.0	5.2	0.0	0.0	66.4	48.7	46.2	18.5	26.5	37.1	27.6	34.7	44.2	23.2	4.6
Our Method w MSB	35.7	51.4	46.8	20.7	26.7	49.6	41.1	8.3	4.4	0	73.7	58.8	54.0	27.6	36.4	38.2	40.3	35.7	56.6	37.7	21.4

Table 4. Impact of 2D SSCNet on the performance with SemanticKITTI val dataset.

Methods	mIoU	Completion	Car	Bicycle	Motorcycle	Truck	Other Vehicle	Person	Bicyclist	Motorcyclist	Road	Parking	Sidewalks	Other Ground	Building	Fence	Vegetation	Trunk	Terrain	Pole	Traffic Sign
Our Method w/o 2D SSCNet	31.9	46.4	44.8	14.3	29.8	45.9	18.8	9.3	2.5	2.7	72.5	57.2	52.2	32.6	32.2	37.3	34.8	35.0	54.0	22.9	7.5
Our Method w 2D SSCNet	35.7	51.4	46.8	20.7	26.7	49.6	41.1	8.3	4.4	0	73.7	58.8	54.0	27.6	36.4	38.2	40.3	35.7	56.6	37.7	21.4

Table 5. Impact of Feature Fusion Model on the performance with SemanticKITTI val dataset.

Methods	mIoU	Completion	Car	Bicycle	Motorcycle	Truck	Other Vehicle	Person	Bicyclist	Motorcyclist	Road	Parking	Sidewalks	Other Ground	Building	Fence	Vegetation	Trunk	Terrain	Pole	Traffic Sign
Our Method w/o FFM	33.9	49.0	45.2	30.1	33.3	50.6	38.5	9.4	1.6	0.0	71.4	56.3	50.6	27.6	34.2	35.6	36.7	36.3	55.6	25.2	5.2
Our Method w FFM	35.7	51.4	46.8	20.7	26.7	49.6	41.1	8.3	4.4	0	73.7	58.8	54.0	27.6	36.4	38.2	40.3	35.7	56.6	37.7	21.4

5. Conclusions

This work introduces an integrated framework for 3D semantic scene completion tasks, which intuitively combines the strengths of both 3D and 2D semantic scene completion networks (SSCNets). The parameter-efficient and effective multi-scale convolutional block (MSB) within 3D SSCNet is developed to aggregate multi-scale features. This enhances the capability of the network to accurately segment small, distant, and crowded objects, addressing a common shortcoming of the existing methods. Additionally, the 3D SSCNet

incorporates a layout-aware semantic block (LSB), which plays a crucial role in understanding the layout of the overall scene, facilitating the precise guidance of feature reconstruction and recognition. Concurrently, the 2D SSCNet is developed to process bird's-eye-view (BEV) features, augmenting the layout information for three-dimensional semantic scene completion. The FFM is designed to foster effective fusion between the 2D SSCNet and 3D SSCNet data. This module ensures a robust scene completion process by leveraging the combined strengths of 3D and 2D data, thereby overcoming the challenges associated with fusing disparate multimodalities. Our extensive evaluation, conducted on the widely recognized SemanticKITTI dataset, demonstrates the superior performance of our method when compared to the state-of-the-art.

Author Contributions: Conceptualization, L.L., N.A., J.V. and A.M.; methodology, L.L., N.A., J.V. and A.M.; software, L.L.; validation, L.L.; formal analysis, L.L.; investigation, L.L., N.A., J.V. and A.M.; resources, L.L., N.A. and A.M.; data curation, L.L.; writing—original draft preparation, L.L.; writing—review and editing, N.A., J.V. and A.M.; visualization and project administration, N.A. and A.M.; funding acquisition, N.A. and A.M. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by an Australian Research Council Discovery Early Career Award (project number DE230101058) funded by the Australian Government, the recipient of the project is Dr. Naveed Akhtar; and an Australian Research Council Future Fellowship Award (project number FT210100268) funded by the Australian Government, the recipient of the project is Professor Ajmal Mian.

Data Availability Statement: Data are contained within the article.

Acknowledgments: The authors are grateful to the editor and reviewers for their constructive comments, which significantly improved the work.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Song, S.; Yu, F.; Zeng, A.; Chang, A.X.; Savva, M.; Funkhouser, T. Semantic scene completion from a single depth image. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1746–1754.
2. Roldao, L.; de Charette, R.; Verroust-Blondet, A. Lmscnet: Lightweight multiscale 3d semantic completion. In Proceedings of the 2020 International Conference on 3D Vision (3DV), Fukuoka, Japan, 25–28 November 2020; pp. 111–119.
3. Yan, X.; Gao, J.; Li, J.; Zhang, R.; Li, Z.; Huang, R.; Cui, S. Sparse single sweep lidar point cloud segmentation via learning contextual shape priors from scene completion. In Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 2–9 February 2021; Volume 35, pp. 3101–3109.
4. Cheng, R.; Agia, C.; Ren, Y.; Li, X.; Bingbing, L. S3cnet: A sparse semantic scene completion network for lidar point clouds. In Proceedings of the Conference on Robot Learning, PMLR, London, UK, 8–11 November 2021; pp. 2148–2161.
5. Yang, X.; Zou, H.; Kong, X.; Huang, T.; Liu, Y.; Li, W.; Wen, F.; Zhang, H. Semantic segmentation-assisted scene completion for lidar point clouds. In Proceedings of the 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Prague, Czech Republic, 27 September–1 October 2021; pp. 3555–3562.
6. Xia, Z.; Liu, Y.; Li, X.; Zhu, X.; Ma, Y.; Li, Y.; Hou, Y.; Qiao, Y. SCPNet: Semantic Scene Completion on Point Cloud. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–19 June 2023; pp. 17642–17651.
7. Behley, J.; Garbade, M.; Milioto, A.; Quenzel, J.; Behnke, S.; Stachniss, C.; Gall, J. Semantickitti: A dataset for semantic scene understanding of lidar sequences. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 29 October–1 November 2019; pp. 9297–9307.
8. Guo, Y.X.; Tong, X. View-volume network for semantic scene completion from a single depth image. *arXiv* **2018**, arXiv:1806.05361.
9. Wang, Y.; Tan, D.J.; Navab, N.; Tombari, F. Adversarial semantic scene completion from a single depth image. In Proceedings of the 2018 International Conference on 3D Vision (3DV), Verona, Italy, 5–8 September 2018; pp. 426–434.
10. Wang, Y.; Tan, D.J.; Navab, N.; Tombari, F. Forknet: Multi-branch volumetric semantic completion from a single depth image. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 29 October–1 November 2019; pp. 8608–8617.
11. Zhang, P.; Liu, W.; Lei, Y.; Lu, H.; Yang, X. Cascaded context pyramid for full-resolution 3d semantic scene completion. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 29 October–1 November 2019; pp. 7801–7810.

12. Dai, A.; Diller, C.; Nießner, M. Sg-nn: Sparse generative neural networks for self-supervised scene completion of rgb-d scans. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–18 June 2020; pp. 849–858.
13. Wu, S.C.; Tateno, K.; Navab, N.; Tombari, F. Scfusion: Real-time incremental scene reconstruction with semantic completion. In Proceedings of the 2020 International Conference on 3D Vision (3DV), Fukuoka, Japan, 25–28 November 2020; pp. 801–810.
14. Cao, A.Q.; de Charette, R. Monoscene: Monocular 3d semantic scene completion. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 3991–4001.
15. Li, Y.; Yu, Z.; Choy, C.; Xiao, C.; Alvarez, J.M.; Fidler, S.; Feng, C.; Anandkumar, A. Voxformer: Sparse voxel transformer for camera-based 3d semantic scene completion. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 9087–9098.
16. Li, B.; Sun, Y.; Jin, X.; Zeng, W.; Zhu, Z.; Wang, X.; Zhang, Y.; Okae, J.; Xiao, H.; Du, D. StereoScene: BEV-Assisted Stereo Matching Empowers 3D Semantic Scene Completion. *arXiv* **2023**, arXiv:2303.13959.
17. Jiang, H.; Cheng, T.; Gao, N.; Zhang, H.; Liu, W.; Wang, X. Symphonize 3D Semantic Scene Completion with Contextual Instance Queries. *arXiv* **2023**, arXiv:2306.15670.
18. Miao, R.; Liu, W.; Chen, M.; Gong, Z.; Xu, W.; Hu, C.; Zhou, S. Occdepth: A depth-aware method for 3d semantic scene completion. *arXiv* **2023**, arXiv:2302.13540.
19. Hayler, A.; Wimbauer, F.; Muhle, D.; Rupprecht, C.; Cremers, D. S4C: Self-Supervised Semantic Scene Completion with Neural Fields. *arXiv* **2023**, arXiv:2310.07522.
20. Mei, J.; Yang, Y.; Wang, M.; Zhu, J.; Zhao, X.; Ra, J.; Li, L.; Liu, Y. Camera-based 3D Semantic Scene Completion with Sparse Guidance Network. *arXiv* **2023**, arXiv:2312.05752.
21. Rist, C.B.; Schmidt, D.; Enzweiler, M.; Gavrila, D.M. Scssnet: Learning spatially-conditioned scene segmentation on lidar point clouds. In Proceedings of the 2020 IEEE Intelligent Vehicles Symposium (IV), Las Vegas, NV, USA, 19 October–13 November 2020; pp. 1086–1093.
22. Wang, P.S.; Liu, Y.; Tong, X. Deep octree-based CNNs with output-guided skip connections for 3D shape and scene completion. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 14–19 June 2020; pp. 266–267.
23. Nie, Y.; Hou, J.; Han, X.; Nießner, M. Rfd-net: Point scene understanding by semantic instance reconstruction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 4608–4618.
24. Zhang, S.; Li, S.; Hao, A.; Qin, H. Point cloud semantic scene completion from rgb-d images. In Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 2–9 February 2021; Volume 35, pp. 3385–3393.
25. Rist, C.B.; Emmerichs, D.; Enzweiler, M.; Gavrila, D.M. Semantic scene completion using local deep implicit functions on lidar data. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *44*, 7205–7218. [CrossRef] [PubMed]
26. Xiong, Y.; Ma, W.C.; Wang, J.; Urtasun, R. Learning Compact Representations for LiDAR Completion and Generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 1074–1083.
27. Xu, J.; Li, X.; Tang, Y.; Yu, Q.; Hao, Y.; Hu, L.; Chen, M. Casfusionnet: A cascaded network for point cloud semantic scene completion by dense feature fusion. In Proceedings of the AAAI Conference on Artificial Intelligence, Washington, DC, USA, 7–14 February 2023; Volume 37, pp. 3018–3026.
28. Li, H.; Dong, J.; Wen, B.; Gao, M.; Huang, T.; Liu, Y.H.; Cremers, D. DDIT: Semantic Scene Completion via Deformable Deep Implicit Templates. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 1–6 October 2023; pp. 21894–21904.
29. Zhang, J.; Zhao, H.; Yao, A.; Chen, Y.; Zhang, L.; Liao, H. Efficient semantic scene completion network with spatial group convolution. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 733–749.
30. Dai, A.; Ritchie, D.; Bokeloh, M.; Reed, S.; Sturm, J.; Nießner, M. Scancomplete: Large-scale scene completion and semantic segmentation for 3d scans. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4578–4587.
31. Zou, H.; Yang, X.; Huang, T.; Zhang, C.; Liu, Y.; Li, W.; Wen, F.; Zhang, H. Up-to-Down Network: Fusing Multi-Scale Context for 3D Semantic Scene Completion. In Proceedings of the 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Prague, Czech Republic, 27 September–1 October 2021; pp. 16–23.
32. Li, P.; Shi, Y.; Liu, T.; Zhao, H.; Zhou, G.; Zhang, Y.Q. Semi-supervised implicit scene completion from sparse LiDAR. *arXiv* **2021**, arXiv:2111.14798.
33. Liu, S.; Hu, Y.; Zeng, Y.; Tang, Q.; Jin, B.; Han, Y.; Li, X. See and think: Disentangling semantic scene completion. *Adv. Neural Inf. Process. Syst.* **2018**, *31*, 261–272.
34. Guedes, A.B.S.; de Campos, T.E.; Hilton, A. Semantic scene completion combining colour and depth: Preliminary experiments. *arXiv* **2018**, arXiv:1802.04735.
35. Li, J.; Liu, Y.; Yuan, X.; Zhao, C.; Siegwart, R.; Reid, I.; Cadena, C. Depth based semantic scene completion with position importance aware loss. *IEEE Robot. Autom. Lett.* **2019**, *5*, 219–226. [CrossRef]

36. Garbade, M.; Chen, Y.T.; Sawatzky, J.; Gall, J. Two stream 3d semantic scene completion. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Long Beach, CA, USA, 16–17 June 2019.
37. Li, J.; Liu, Y.; Gong, D.; Shi, Q.; Yuan, X.; Zhao, C.; Reid, I. Rgbd based dimensional decomposition residual network for 3d semantic scene completion. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 7693–7702.
38. Chen, X.; Lin, K.Y.; Qian, C.; Zeng, G.; Li, H. 3d sketch-aware semantic scene completion via semi-supervised structure prior. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 4193–4202.
39. Li, S.; Zou, C.; Li, Y.; Zhao, X.; Gao, Y. Attention-based multi-modal fusion network for semantic scene completion. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 11402–11409.
40. Liu, Y.; Li, J.; Yan, Q.; Yuan, X.; Zhao, C.; Reid, I.; Cadena, C. 3D gated recurrent fusion for semantic scene completion. *arXiv* **2020**, arXiv:2002.07269.
41. Li, J.; Wang, P.; Han, K.; Liu, Y. Anisotropic convolutional neural networks for RGB-D based semantic scene completion. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *44*, 8125–8138. [CrossRef] [PubMed]
42. Cai, Y.; Chen, X.; Zhang, C.; Lin, K.Y.; Wang, X.; Li, H. Semantic scene completion via integrating instances and scene in-the-loop. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 324–333.
43. Li, J.; Ding, L.; Huang, R. Imenet: Joint 3d semantic scene completion and 2d semantic segmentation through iterative mutual enhancement. *arXiv* **2021**, arXiv:2106.15413.
44. Dourado, A.; Guth, F.; de Campos, T. Data augmented 3d semantic scene completion with 2d segmentation priors. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–8 January 2022; pp. 3781–3790.
45. Wang, X.; Lin, D.; Wan, L. Ffnet: Frequency fusion network for semantic scene completion. In Proceedings of the AAAI Conference on Artificial Intelligence, Philadelphia, PA, USA, 22 February–1 March 2022; Volume 36, pp. 2550–2557.
46. Tang, J.; Chen, X.; Wang, J.; Zeng, G. Not all voxels are equal: Semantic scene completion from the point-voxel perspective. In Proceedings of the AAAI Conference on Artificial Intelligence, Philadelphia, PA, USA, 22 February–1 March 2022; Volume 36, pp. 2352–2360.
47. Fu, R.; Wu, H.; Hao, M.; Miao, Y. Semantic scene completion through multi-level feature fusion. In Proceedings of the 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Kyoto, Japan, 23–27 October 2022; pp. 8399–8406.
48. Wang, F.; Zhang, D.; Zhang, H.; Tang, J.; Sun, Q. Semantic Scene Completion with Cleaner Self. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 867–877.
49. Dong, H.; Ma, E.; Wang, L.; Wang, M.; Xie, W.; Guo, Q.; Li, P.; Liang, L.; Yang, K.; Lin, D. Cvsformer: Cross-view synthesis transformer for semantic scene completion. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 1–6 October 2023; pp. 8874–8883.
50. Cao, H.; Behnke, S. SLCF-Net: Sequential LiDAR-Camera Fusion for Semantic Scene Completion using a 3D Recurrent U-Net. *arXiv* **2024**, arXiv:2403.08885.
51. Hou, Y.; Zhu, X.; Ma, Y.; Loy, C.C.; Li, Y. Point-to-voxel knowledge distillation for lidar semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 21–24 June, 2022; pp. 8479–8488.
52. Tang, L.; Zhan, Y.; Chen, Z.; Yu, B.; Tao, D. Contrastive boundary learning for point cloud segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 8489–8499.
53. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
54. Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J.M.; Luo, P. SegFormer: Simple and efficient design for semantic segmentation with transformers. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 12077–12090.
55. Li, J.; Hassani, A.; Walton, S.; Shi, H. Convmlp: Hierarchical convolutional mlps for vision. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 6307–6316.
56. Mei, J.; Yang, Y.; Wang, M.; Huang, T.; Yang, X.; Liu, Y. SSC-RS: Elevate LiDAR Semantic Scene Completion with Representation Separation and BEV Fusion. In Proceedings of the 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Detroit, MI, USA, 1–5 October 2023; pp. 1–8.