CAAI Transactions on Intelligence Technology

WILEY

The Institution of Engineering and Technology

ORIGINAL RESEARCH OPEN ACCESS

CDFNet: Cross-Modal Deep Fusion for Monocular 3D Semantic Scene Completion

Xianjing Cheng[1] | Lintai Wu[2] | Jie Wen[3] | Tianhao Peng[4] | Yong Xu[5,6] | Zairong Wei[7]

[1]Department of Information Technology, Affiliated Hospital of Zunyi Medical University, Zunyi, China | [2]College of Engineering, Huaqiao University, Quanzhou, China | [3]School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen, China | [4]School of Automation Engineering, Moutai Institute, Renhuai, China | [5]Bio-Computing Research Center, Shenzhen Graduate School, Harbin Institute of Technology, Shenzhen, China | [6]Key Laboratory of Network Oriented Intelligent Computation, Shenzhen, China | [7]Department of Burns and Plastic Surgery, Affiliated Hospital of Zunyi Medical University, Zunyi, China

Correspondence: Zairong Wei (weizr@zmu.edu.cn)

Received: 14 October 2025 | **Revised:** 30 January 2026 | **Accepted:** 19 February 2026

Keywords: class imbalance | cross-modal learning | feature fusion | occupancy prediction | semantic scene completion

ABSTRACT

Semantic scene completion (SSC) aims to predict the semantic occupancy and geometry of 3D scenes. Recently, most studies focus on camera-based approaches due to the rich visual cues of images and the cost-effectiveness of cameras. However, these methods usually lack efficient fusion and fine-grained processing of cross-modal semantic information, resulting in sub-optimal performance. To address these issues, we propose a novel cross-modal semantic deep fusion framework. Unlike previous approaches, our method effectively integrates 2D textural, 2D spatial and 3D geometric knowledge from three different modalities to reconstruct complete 3D scenes. Specifically, we employ two encoders to extract 2D textural and 2D spatial features from RGB images and depth maps, which are then fused and lifted into 3D space via our tailored cross-modal semantic fusion module. In contrast to previous methods that encompass extensive redundant 3D voxel features, we design a lightweight voxel feature filter to efficiently eliminate these redundancies. Furthermore, 3D geometric features are extracted from the point cloud derived from the depth map. The 3D features from multiple modalities are deeply fused and further refined by a sparse-to-dense voxel completion module, which effectively enriches the semantic information. Besides, we propose a new evaluation metric more suitable for assessing the SSC task with class imbalance issues in the dataset. Extensive experiments show that our method achieves the state of the art in camera-based semantic scene completion. We will release the source code publicly.

1 | Introduction

Semantic scene completion is a critical 3D perception task which is widely applied in many fields [1–8], such as robotic navigation, autonomous driving and planning [9, 10]. Early studies, such as SSCNet [11] and LMSCNet [12], mainly focus on LiDAR-based approaches [13]. Nevertheless, capturing LiDAR point clouds is costly, and they are typically sparse, which hurts the performance of SSC. Recently, camera-based methods have attracted considerable attention since RGB images provide rich vision cues and are more readily available.

However, estimating the 3D scene from a single RGB image is a challenging problem due to limited observations of cameras [14]. MonoScene [15] is the first monocular method inferring 3D semantic scene from a single RGB image. It leverages the features line of sight projection (FLoSP) [16] to lift 2D features to 3D voxel features and integrate a 3D context relation prior to

Xianjing Cheng and Lintai Wu are co-first authors.

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2026 The Author(s). *CAAI Transactions on Intelligence Technology* published by John Wiley & Sons Ltd on behalf of The Institution of Engineering and Technology and Chongqing University of Technology.

enhance spatio-semantic awareness. The lifted 3D features often encompass much redundant information, introducing ambiguity and thus leading to poor performance. To address these issues, some recent methods [17–20] introduce depth information to reduce this ambiguity. On the one hand, however, these methods lack a deep and fine-grained fusion of RGB and depth features and thus still fail to effectively eliminate redundant 3D features. On the other hand, RGB and depth information provide only 2D information and lack 3D geometric knowledge of the scene, which is insufficient to accurately reconstruct a complete 3D scene.

In this paper, we propose CDFNet, a novel SSC framework that fully leverages cross-modal semantic information, that is, 2D textural, 2D spatial and 3D geometric knowledge from monocular RGB image, depth map and point cloud, respectively. Depth maps are characterised by abundant depth information, and the optimised features thereof exhibit rich 2D spatial information, which is defined as 2D spatial features. Point clouds are equipped with substantial 3D scene information, and their optimised features are characterised by structural features. In this work, we explore the fusion efficiency of three interrelated modalities in SSC, which is different from Voxformer [17] that focuses on RGB and depth map. Specifically, we first employ a 2D encoder to extract the textural features from the scene image. Since depth values offer rich 2D spatial information, we utilise a pre-trained model to estimate the depth map of the scene image, followed by a depth encoder for capturing the 2D spatial features. To effectively integrate these features, we design a cross-modal semantic fusion (CSF) module that is dedicated to the efficient fusion of RGB and depth map, where its optimised features embed extensive depth and semantic information, that is, depth-semantic features. First, we employ a SELayer [21] to encode the textural features and spatial features and then design a multi-path atrous convolution (MPAC) unit to extract both the coarse-grained and fine-grained semantic features. Next, we propose a depth semantic enhancement (DSE) unit to enrich 2D features with richer and more comprehensive 3D information, enabling them to better represent the complex spatial structures and semantic details of 3D scenes. The enhanced semantic features are further fed to a lightweight U-Net unit for extracting multi-scale 2D features, which are subsequently lifted into 3D voxel features following FLoSP [16].

However, conventional methods employ the FLoSP technique to lift multi-scale 2D features into 3D voxel features, resulting in numerous identical features within the 3D voxel grids. This redundancy introduces ambiguities and adversely affects subsequent optimization processes. To solve this problem, we devise a voxel feature filter (VFF) that removes redundant 3D features while preserving key ones, which effectively eliminates ambiguity.

In addition, since point cloud provides rich 3D geometric structure information of scene, we thus transform the depth map into a point cloud and utilise MLPs to extract 3D geometric knowledge of the scene. The 3D geometric features are then fused with lifted hybrid features. To get a complete and comprehensive semantic representation of 3D voxel features, we design a sparse-to-dense voxel feature completion (VFC) unit that aggregates 3D structure knowledge from the point cloud and 2D semantic feature filtered by the voxel feature filter and then propagates them to the voxels of the entire scene. Finally, we leverage a 3D segmentation sub-network followed by Cylinder3D [22] to predict the semantic scene. We conduct extensive experiments on a SemanticKITTI dataset [23]. The results show that our approach outperforms other state-of-the-art camera-based 3D semantic scene completion methods. Moreover, we also conduct ablation studies to validate the effectiveness of each component of the proposed method.

In summary, the core contributions of this work can be highlighted as follows:

1. We propose a monocular-based 3D semantic scene completion framework that effectively leverages 2D textural, 2D spatial and 3D geometric structure knowledge of the scene from three different modalities.
2. We introduce a cross-modal semantic fusion module that delicately integrates 2D textural and 2D spatial information of the scene while significantly enhancing the scene features through a series of meticulously crafted submodules.
3. We devise a voxel feature completion module that aggregates the 3D point-based geometric knowledge and the semantic features derived from the CSF module. The voxel feature completion module further enriches and refines the semantic features, ensuring a more comprehensive and precise representation of 3D scenes.
4. The proposed method achieves state-of-the-art performance in the camera-based semantic scene completion challenge on the SemanticKITTI dataset.

2 | Related Work

2.1 | Depth Estimation From RGB Images

In recent years, an increasing number of camera-based 3D semantic scene completion methods have utilised estimated depth maps to augment the semantic information of scenes [15, 24, 25]. Depth estimation technology has become increasingly mature with exploration over recent years. For example, Zoedepth [26] proposes a generalisable and high-precision depth estimation network that utilises the depth-wise separable convolution. DepthAnything [27] constructs a generalised framework for zero-shot depth estimation; it uses many large datasets with a wide variety of styles as training data and ensures robustness and generalisation across different scenes. This diversity of images in the training data helps the network to better understand and estimate more accurate depth in real-world applications. ECoDepth [28] proposes a diffusion-based network to conduct the diffusion process [29, 30] in the latent space of the image feature and achieves the satisfactory results in indoor and outdoor scenes.

Besides, stereo-based methods utilise triangulation on left and right image pairs for estimating more accurate depth. For example, GANet [31] proposes the semi-global guided

aggregation (SGA) and local-guided aggregation (LGA) modules in cost aggregation to estimate the disparities. AANet [32] constructs a multi-scale cost volume aggregated by six stacked adaptive aggregation modules to predict the disparities in which each module consists of three intra-scale aggregation modules and a cross-scale module. MobileStereoNet [33] proposes a lightweight network that employs 2D MobileNet blocks and extends them to 3D for stereo vision applications. RAFTStereo [34] extracts the correlation features from stereo image pairs and constructs a multi-level cost volume, and then the volume is fed to GRUs to update the disparities iteratively. Los [35] utilises a fine-grained local structure from monocular and binocular branches to guide the network learning and designs multi-level ConvGRU blocks to update the disparities. Selective-Stereo [36] is a expansion version of RAFT-Stereo [34] that extracts the attention maps from the contextual spatial attention module to update SRUs and then it iteratively updates the disparities by the local cost volume and the attention map.

Compared to monocular, stereo-based methods can produce more accurate and precise depth maps. Thus, we adopt a stereo-based methods named MobileStereoNet [33] to estimate the depth maps in this paper.

2.2 | 3D Semantic Scene Completion

Semantic occupancy prediction is an important task in autonomous driving, which can significantly improves the 3D perception ability of network in the outdoor scene. SSCNet [11] is the first SSC network that defines the 'SSC' task, and the survey done thoroughly can be reviewed in ref. [7].

Existing SSC methods are classified into LiDAR-based and camera-based methods. LiDAR-based methods use point cloud scanned by the LiDAR sensor as input to predict 3D semantic occupancy, for example, SGCNet [37] adopts the spatial group convolutions to improve the efficiency of feature extraction. LMSCNet [12] develops a lightweight network that projects the height of voxels into the channel dimension and subsequently employs a multi-scale 2D U-Net to predict 3D occupancies. UDNet [38] uses a 3D U-Net to predict an accurate semantic occupancy. Furthermore, some recent LiDAR-based solutions propose multi-view fusion [39], implicit functions [40] and knowledge distillation [41]. Besides, there are some methods that explore the relationships between semantic segmentation and scene completion. JS3CNet [42] improves the performance of scene completion with the aid of the semantic segmentation task. SCPNet [41] introduces knowledge distillation into the SSC task. However, LiDAR-based SSC methods usually adopt 3D convolutions to process voxel features from the point cloud, which demands a substantial amount of computational resources.

Recently, camera-captured images are easily deployed, cost-effective, and provide rich visual cues that strengthen the scene understanding of networks. Consequently, camera-based SSC methods have attracted increasing attention in the research community. MonoScene [15] is the first to use a monocular image to predict semantic classes. It employs FLoSP [16] to lift 2D features to 3D and utilise a 3D U-Net to process the voxel features. However, the 2D-to-3D projection in FLoSP [16] introduces redundant features to voxel grids. VoxFormer [17] follows the architecture of BEVFormer [43] and uses the results from depth prediction as queries and then performs deformable cross-attention to integrate the features from the image and depth map and leverages self-attention embedded mask tokens to predict 3D occupancy. TPVFormer [44] presents a tri-perspective view [45] to represent 3D voxelised features, and OccFormer [46] introduces a two-path transformer network for SSC. Symphonize [18] employs a sparse set of instance queries to predict the semantic scene. SGN [19] integrates the sparse semantic guidance from geometry and occupancy prediction from point cloud to construct seed features and then it aggregates the geometry prior and occupancy-aware features as voxel features. H2GFormer [47] uses the stereo image pairs to predict depth information to resolve the geometry ambiguity.

We focus on the monocular-based SSC method in this paper for predicting the accurate 3D semantic occupancies.

3 | Proposed Method

3.1 | Overview

Given a monocular RGB image of a scene, our goal is to reconstruct its dense and complete 3D semantic representation. Formally, let $\mathbf{I} \in \mathbb{R}^{H' \times W' \times 3}$ be the input monocular image, where H' and W' represent the height and width of the image, respectively. The desired output is a 3D voxel grid $Y = \{c_0, c_1, .., c_M\}^{H \times W \times Z}$, where H, W and Z denote the length, width and height of the voxel grid, and each voxel is assigned a semantic class label. There are totally $M + 1$ semantic classes and c_0 represents the free class. We aim to learn a network to predict the semantic classes for each voxel in the grid, thereby achieving a complete and semantically meaningful reconstruction of the scene.

The overall framework of our method is shown in Figure 1. First, to capture 2D spatial and 3D geometric information, we induce the depth map and the point cloud from the RGB scene image. The RGB image and depth map are fed into two separate 2D encoders to extract 2D textural features and spatial features, respectively. Furthermore, the point cloud is first voxelised and then input into MLPs to acquire 3D geometric structural features. Subsequently, these textural and spatial features are deeply integrated using a cross-modal semantic fusion module (CSF) and lifted into 3D space, with a voxel feature filter removing redundant features. Then, a voxel feature completion module further aggregates these 3D hybrid features from the CSF with the 3D geometric features and propagates them to unoccupied voxels. Finally, the complete and comprehensive aggregated features are input into a 3D segmentation network to infer the final result.

3.2 | Cross-Modal Semantic Fusion Module

In the CSF module, we first design two shallow-layer convolution networks that incorporate residual connections to extract

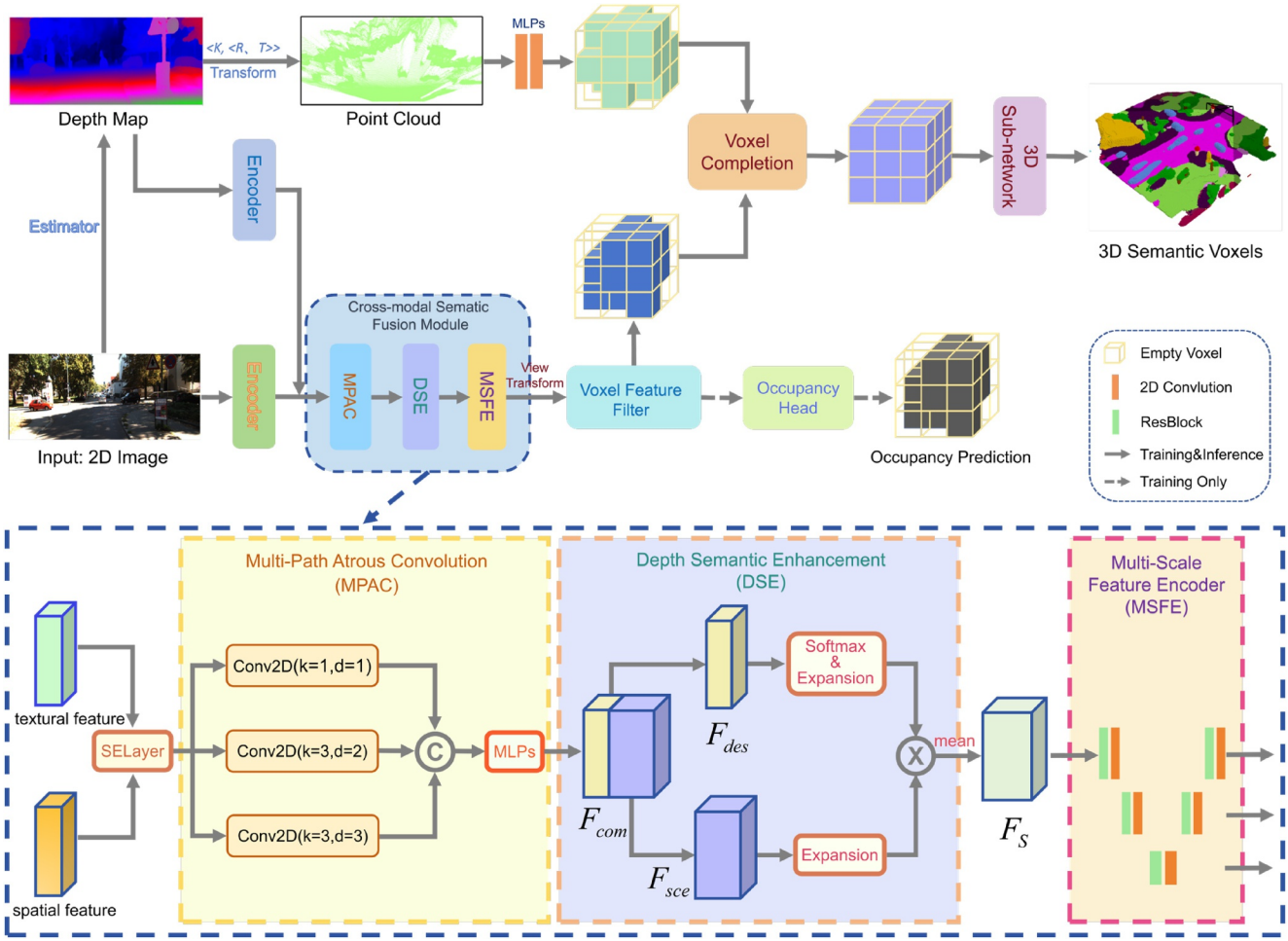


FIGURE 1 | Overall framework of CDFNet. Given a monocular RGB image, we first use a depth estimator to estimate a depth map and transform it into 3D space to generate a point cloud. The RGB image, depth map and point cloud are fed to their respective encoders to extract 2D and 3D features. We design a cross-modal semantic fusion module, a voxel feature filter and a voxel feature completion module to integrate and refine these cross-modal features. Finally, a 3D segmentation sub-network is employed to decode the complete scene and 3D semantic labels of voxels.

2D textural features from the monocular RGB image and 2D spatial features from the depth map, respectively. In order to enhance the semantic representation of the feature map, we employ a SELayer [21] to fuse the textural and spatial features. Then, the hybrid features are fed into the multi-path atrous convolution (MPAC) module for further optimisation. In our MPAC module, we employ three different dilated convolutions with two widely used kernel sizes $\{1, 3\}$ and three different dilated rates $\{1, 2, 3\}$ to refine semantic features as shown in Figure 1. Here, we use a small-kernel convolution to extract fine-grained local features and two convolutions with larger kernels to extract coarse-grained features. Followed by MLPs, we can get a refined semantic feature $\mathbf{F_{com}} \in \mathbb{R}^{(D+C)\times H\times W}$.

Unlike 2D images, 3D scenes require higher-dimensional and richer spatial information, necessitating that each dimension of the feature map contains detailed cues for accurate 3D reconstruction. Traditional 2D convolution and pooling operations focus on spatial relationships within the $H \times W$ dimensions but fail to adequately model and capture the intricate depth information along the D dimension. This limitation makes traditional 2D methods struggle to fully represent the detailed features in the depth direction when processing 3D scenes with complex depth information. To overcome this, the DSE unit leverages an attention mechanism to enhance the semantic features of the D dimension in the combined feature map $\mathbf{F}_{com}$, as shown in Figure 2. In this paper, the $H \times W$ dimensions and the D dimension are referred to as the 'plane dimension' and 'depth dimension,' respectively. The DSE unit first decomposes $\mathbf{F}_{com}$ into scene semantic features $\mathbf{F}_{sce} \in \mathbb{R}^{C\times H\times W}$ and scene semantic descriptors $\mathbf{F}_{des} \in \mathbb{R}^{D\times H\times W}$, where $\mathbf{F}_{sce}$ encodes the global structural representation of the scene. The descriptor $\mathbf{F}_{des}$ is then processed by a softmax operation along the depth dimension to generate depth semantic weights $\mathbf{F}_w \in \mathbb{R}^{D\times H\times W}$, which quantify the relative significance of various depth levels within the scene voxels. Next, $\mathbf{F}_w$ and $\mathbf{F}_{sce}$ are expanded to $\mathbf{F}_w^e \in \mathbb{R}^{C\times D\times H\times W}$ and $\mathbf{F}_{sce}^e \in \mathbb{R}^{C\times D\times H\times W}$, respectively. Finally, the elements of $\mathbf{F}_w^e$ and $\mathbf{F}_{sce}^e$ are fused by element-wise multiplication and averaging along the C dimension to obtain $\mathbf{F}_S \in \mathbb{R}^{D\times H\times W}$. This process is defined as follows:

$$\mathbf{F}_w = \mathtt{Softmax}(\mathbf{F}_{des}), \tag{1}$$

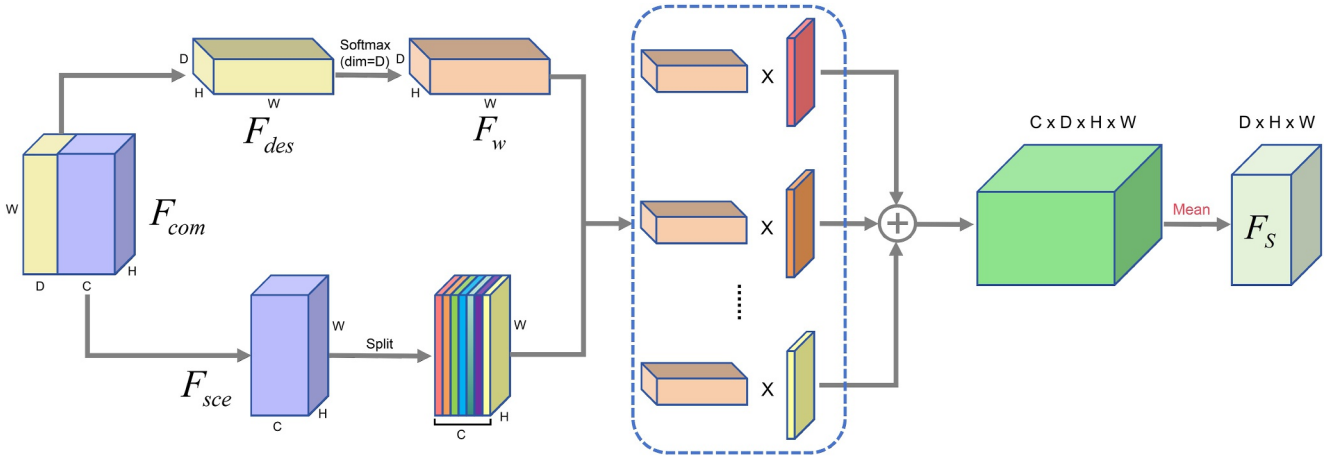


FIGURE 2 | The structure of the depth semantic enhancement module. We first split the feature map into the scene semantic features and their corresponding descriptors along the depth dimension. Subsequently, the descriptors enhance scene semantic features through element-wise multiplication and mean operations.

$$f_S(i) = \frac{1}{C}\sum_{j=1}^{C} w(i,j) * f(i,j), \tag{2}$$

where $f_S(i)$, $w(i,j)$ and $f(i,j)$ represent the elements of $\mathbf{F}_S$, $\mathbf{F}_w^e$ and $\mathbf{F}_{sce}^e$, respectively. $i = 1, 2, ..., D, j = 1, 2, ..., C$ (for simplicity, the subscript for the H and W dimensions is omitted).

In addition, to capture multi-scale features, we design a light-weight multi-scale feature encoder to extract three levels of 2D feature maps as shown in Figure 1. The encoder is a U-Net architecture consisting of three downsample and upsample blocks. In each downsample block, we embed one residual connection and use max-pooling operation to obtain global features. Compared to some recent methods [15, 17, 19, 46] that utilise ResNet50 or EfficientNet50 [48] as the feature encoder to extract four levels of features, our feature encoder extracts only three levels of features and achieves a better performance. This trick is validated by our ablation experiment depicted in Sec IV-C.

Upon obtaining the multi-scale 2D feature maps, we employ classified FLoSP [16] to lift 2D features into 3D voxel features, which is more efficient than LSS [49]. The resulting 3D feature map, denoted as $\mathbf{F}_{3D}$, is subsequently upsampled to align with the resolution of the original voxel grid.

3.3 | Voxel Feature Filter

Lifting 2D multi-scale feature maps to 3D voxel features is an ill-posed problem due to the scale ambiguity inherent in the 2D-to-3D projection. This often results in redundant voxels, which may introduce ambiguity in subsequent voxel feature aggregation.

To address this issue, we propose a lightweight voxel feature filter to eliminate these redundant 3D features. The voxel feature filter consists of two consecutive 3D residual convolution modules followed by a 3D occupancy head with one layer of 3D convolution, as shown in Figure 3.

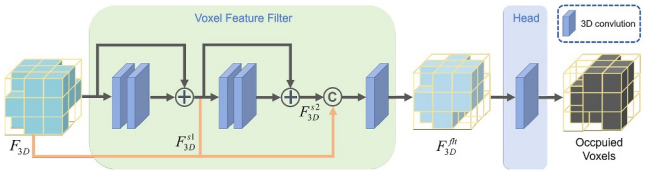


FIGURE 3 | Overview of the voxel feature filter. We concatenate two consecutive 3D voxel features optimised by residual connections with the original voxel feature and then we use an occupancy head with only a 3D convolution to predict the occupied voxels.

We first concatenate two intermediate voxel features $\mathbf{F}_{3D}^{s1}$, and $\mathbf{F}_{3D}^{s2}$ with $\mathbf{F}_{3D}$ to obtain the fused 3D feature map, denoted as $\mathbf{F}_{3D}^{flt}$. Next, we feed $\mathbf{F}_{3D}^{flt}$ into the 3D occupancy head and optimise the filter by predicting the occupancy probability of each voxel, which is supervised by the ground-truth 3D voxel occupancy with a binary cross-entropy loss:

$$\mathcal{L}_{occ} = -\sum_{i \in N}\left[(1-\hat{Y}_i^o)\log(1-Y_i^o) + \hat{Y}_i^o \log(Y_i^o)\right] \tag{3}$$

where $\hat{Y}^o$ represents the ground truth of 3D voxel occupancy, and Y^o denotes the predicted occupancy probability of voxel. Note that the occupancy head is only working during training and is disable during inference. Moreover, our voxel filter differs fundamentally from the semantic guidance mechanism employed in SGN [19]. Specifically, SGN [19] introduces a semantic guidance module (SGM) integrated into two sparse encoder blocks (SEB) to enhance perception performance. Each SEB comprises a sparse feature encoder and a sparse geometry feature encoder from ref. [50]. In contrast, our method only leverages the widely adopted 3D residual convolution, which does not introduce any additional computational overhead. Besides, SGN [19] mainly emphasises

the performance of semantic guidance, whereas our filter is dedicated to filtering redundant voxels, which is validated through ablation studies.

3.4 | Voxel Feature Completion

Unlike 2D RGB images and depth maps, which are inherently limited by their projection-based nature, point clouds preserve the full 3D geometry of the environment and provide an explicit and geometrically rich structural representation of 3D scenes, making them highly valuable for accurate and complete 3D scene reconstruction. To harness this advantage, we integrate point cloud data into our framework, leveraging its geometric fidelity to enhance and complement the 2D features extracted from RGB and depth modalities.

Specifically, as shown in Figure 1, we first transform the depth map into a point cloud with intrinsic and extrinsic parameters of the camera, which is described as follows:

$$\mathbf{P}^C = \mathbf{K}^{-1} \cdot \left(\mathbf{d} \times \mathbf{P}^I\right) \tag{4}$$

$$\mathbf{P}^W = [\mathbf{R}, \mathbf{T}]^{-1} \cdot \mathbf{P}^C \tag{5}$$

where $\mathbf{d}$ denotes the estimated depth map, $\mathbf{P}^I$, $\mathbf{P}^C$ and $\mathbf{P}^W$ represent the points under the homogeneous coordinates of pixels, camera and world, respectively. The intrinsic matrix $\mathbf{K}$ encompasses camera parameters, whereas the extrinsic matrix is composed of the rotation matrix $\mathbf{R}$ and the translation vector $\mathbf{T}$.

After the pixel-to-world transformation, we voxelise the point cloud and then the 3D points in each voxel are processed by MLPs to extract high-dimension features. Next, we develop a sparse-to-dense voxel feature completion module to aggregate the voxel features from point clouds and 2D images and then propagate the fused features to the unoccupied voxel as shown in Figure 4.

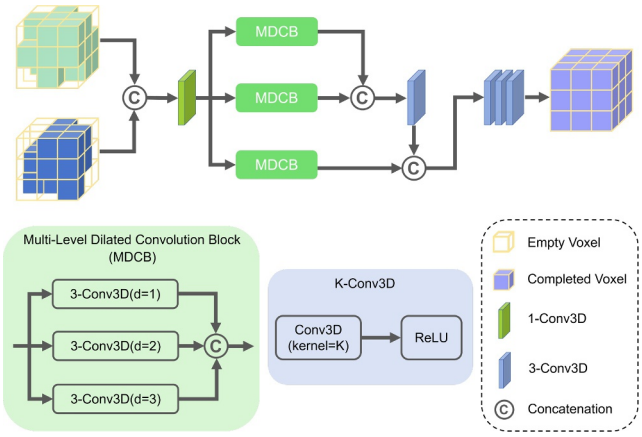


FIGURE 4 | The structure of the voxel feature complete module. We adopt an hierarchical architecture to aggregate occupied voxels and complement empty voxels, which improves the completeness of voxel representation.

During the feature optimisation process, we found that preserving the intra-voxel structures and inter-voxel relations of voxel features is crucial, which, however, is ignored by previous camera-based methods. As shown in Figure 5, the intra-voxel structures represent the geometric shape of objects within each voxel, whereas the inter-voxel relations denote the spatial relations between objects in different voxels. Note that the range of points varies significantly across different voxels. However, the normalisation layers, such as batch normalisation, will rescale the distribution of voxel features and thus disrupt the spatial relations between voxel features during the optimisation of voxel features. Furthermore, the bias term in 3D convolution is prone to altering the original structures of voxel features. We verify these findings in our ablation experiment. To address these issues, in our VFC module, we remove all the normalisation layers and disable the bias of all 3D convolutions to preserve the intra-voxel structures and inter-voxel relations of the voxel features. Besides, we do not use any upsampling or downsampling operations to reduce the structural information loss of the scene.

As illustrated in Figure 4, we first use a 3D convolution with a small kernel to extract fine-grained local 3D details, and then employ three hierarchical multi-level dilated convolution blocks (MDCB) to aggregate rich semantic information and diffuse them into the surrounding unoccupied voxels in a progressive way. Each MDCB block contains three 3D convolutions with different dilation rates to aggregate features from different receptive fields. Subsequently, after optimisation with three consecutive blocks consisting of 3D convolution and ReLU, we can obtain the dense and completed voxel features. Our sparse-to-dense voxel feature completion module effectively fills in the missing semantic information in unoccupied voxels, especially in occluded regions, enabling our model to reconstruct a more complete scene.

Furthermore, our method differs from SCPNet [41] in two key aspects: First, SCPNet [41] introduces a 3D completion module consisting of three multi-path blocks with large receptive fields to aggregate features of occupied voxels and complete unoccupied ones. In contrast, we propose three hierarchical MDCB blocks with smaller receptive fields to enhance completion performance. Second, our completion mechanism focuses on the fusion of two distinct voxel feature streams. Notably, we disable convolution biases and batch normalisation in our completion module, which is not validated in SCPNet [41].

Finally, the completed voxel features through our completion module are fed to a segmentation sub-network from Cylinder3D [22] for predicting the voxel-wise semantic classes.

3.5 | Training Loss

Our method is trained end-to-end from scratch by optimising a combination of four loss functions, that is, the cross-entropy loss, the Lovasz–Softmax loss [51], the scene-class affinity loss [15] and our 3D occupancy prediction loss.

The cross-entropy loss is defined as

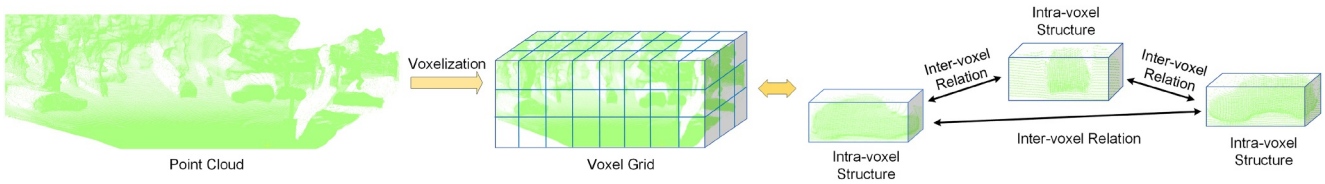


FIGURE 5 | Visualisation of the intra-voxel structures and the inter-voxel relations in our voxelised point clouds. The intra-voxel structures describe the geometric shape of objects within each voxel, whereas the inter-voxel relations characterise the spatial relations between objects in different voxels.

$$\mathcal{L}_{ce} = -\sum_{c=1}^{M} w_c \hat{y}_{i,c} \log\left(\frac{e^{y_{i,c}}}{\sum_{c'}^{M} e^{y_{i,c'}}}\right) \tag{6}$$

where y is the model prediction, $\hat{y}_{i,c}$ is a one-hot vector of ground truth (i.e., $\hat{y}_{i,c} = 1$ if voxel i is the semantic class c, otherwise $\hat{y}_{i,c} = 0$) and w_c is the weight of class c, which is computed by the inverse of the class-frequency f_c as in ref. [52] (i.e., $w_c = \frac{1}{\log(f_c + \epsilon)}$ with $\epsilon \ll 1$).

The Lovasz–Softmax loss is defined as

$$\mathcal{L}_{lovasz} = \frac{1}{|C|} \sum_{c \in \mathcal{C}} \overline{\Delta}_{J_c}(m(c)) \tag{7}$$

where $m(c) \in [0,1]^p$ is the vector of pixel p errors for class c [51], and $\overline{\Delta}_{J_c}$ is the tight convex closure of Jaccard loss Δ_{J_c}.

The scene-class affinity loss is defined as

$$\mathcal{L}_{scal} = -\frac{1}{M} \sum_{c=1}^{M} (\mathtt{P}_c(\hat{y}, y) + \mathtt{R}_c(\hat{y}, y) + \mathtt{S}_c(\hat{y}, y)) \tag{8}$$

where y is the predicted class, $\hat{y}$ is the ground truth class and $\mathtt{P}, \mathtt{R}, \mathtt{S}$ are the class-wise derivable precision, recall and specificity, respectively.

Thus, the total loss is computed by

$$\mathcal{L} = \mathcal{L}_{ce} + \mathcal{L}_{lovasz} + \mathcal{L}_{scal} + \lambda \mathcal{L}_{occ} \tag{9}$$

where λ is a weighting factor to balance the effect of the 3D occupancy prediction loss.

4 | Experiments

4.1 | Experimental Setting

Datasets: SemanticKITTI [23] provides dense outdoor LiDAR scans from KITTI Odometry Benchmark [53], which are voxelised as 256 × 256 × 32 voxel grids with a size of 0.2 m. It comprises 22 sequences of outdoor driving scenarios and labelled with 20 semantic classes (19 semantics, 1 free), in which 10 sequences are for training, 1 sequence for validation and 11 sequences for testing. Besides, SemanticKITTI provides LiDAR sweep or RGB image as input, and we use cropped monocular image of size 1220 × 370 as input in this work.

SSCBench-KITTI-360 [54] consists of over 320k images in the suburban area with rich 360° sensory ground truth, which are voxelised as 256 × 256 × 32. It is split as 7/1/1 sequences with 8487/1812/2566 frames for training, validation and testing, respectively. Each voxel is labelled as one of the 19 semantic classes. In this work, we only use RGB images of resolution 1408 × 376 as input.

Implementation details: We train our network for 11 epochs with a batch size of 1 on the PyTorch platform using two NVIDIA Tesla A40 GPUs. In our method, we employ MobileStereoNet [33] as the depth predictor to estimate the depth maps. During training, we employ the Adam optimiser with a learning rate of 1e-4 and a beta value of 0.9. The parameter λ is 1.0 in all the experiments.

Evaluation Metrics: Following the official evaluation, we report two widely used metrics: Intersection over Union (IoU) and mean IoU (mIoU). The IoU measures occupancy regardless of semantic classes, which is computed as

$$IoU = \frac{\mathtt{TP}_i}{\mathtt{TP}_i + \mathtt{FP}_i + \mathtt{FN}_i}, \tag{10}$$

where $\mathtt{TP}_i$, $\mathtt{FP}_i$ and $\mathtt{FN}_i$ denote the true positive, false positive and false negative of class i, respectively. The mIoU is the arithmetic average of the IoUs across all classes:

$$mIoU = \frac{1}{M} \sum_{i=1}^{M} \frac{\mathtt{TP}_i}{\mathtt{TP}_i + \mathtt{FP}_i + \mathtt{FN}_i}, \tag{11}$$

where M is the total number of classes. Note that IoU and mIoU have a strong interaction that a high mIoU can be achieved by the low IoU. Therefore, the ideal model should achieve excellent performance by high mIoU and IoU metrics.

However, the SSC data often involve class imbalances, with some more common classes in scenes, such as vegetation, road, building, terrain and car, whereas others, such as bicycle, motorcycle, person and motorcyclist, appear infrequently. This imbalance presents challenges in evaluating model performance, as traditional metrics like mIoU treat all classes equally, irrespective of their occurrence in the dataset.

The arithmetic mean used in mIoU fails to account for the varying importance of different classes based on their frequency [55]. Specifically, it disproportionately amplifies the impact of rare classes while diminishing the contribution of frequent classes. For example, strong performance on a rare class

(e.g., bicycle) can significantly inflate the overall mIoU score, even if the model performs poorly on a highly frequent class (e.g., road). Conversely, poor performance on a frequent class is diluted when averaged across all classes, leading to an overly optimistic evaluation of the model's capabilities. As shown in Figure 6, the predictions of the terrain and vegetation by symphonies [18] are highly inaccurate, resulting in an incomplete scene. In contrast, our method provides more complete and accurate predictions for the terrain and vegetation. However, the mIoU of symphonies [18] is higher than that of our method.

To address this issue, a weighted average is a more appropriate approach for aggregating IoU scores. By assigning weights proportional to the frequency of each class in the dataset, a weighted average ensures that frequent classes have a greater impact on the overall metric, whereas rare classes contribute less. To this end, we introduce a weighted IoU (wIoU), which aligns the evaluation metric with the practical significance of each class, providing a more balanced and fair assessment of model performance. The wIoU is defined as follows:

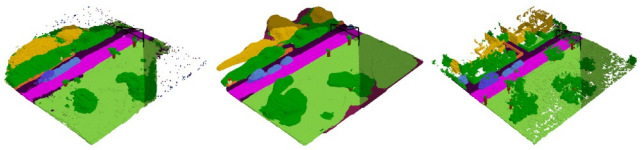
$$wIoU = \sum_{i=1}^{M} w(i) * \frac{\mathrm{TP}_i}{\mathrm{TP}_i + \mathrm{FP}_i + \mathrm{FN}_i}, \tag{12}$$

where $w(i)$ is the proportion of the i class. As illustrated in Figure 6, by introducing wIoU to assess the quality of the reconstructed scene, we establish a reasonable quantitative metric that effectively measures the structural completeness of the reconstructed scene.

Baselines: We compare CDFNet against current state-of-the-art baseline methods: (1) the camera-based methods, for example, MonoScene [15], TPVFormer [44], VoxelFromer [17], OccFormer [46], H2GFormer [47], symphonise [18], SGN [19], MonoMRN [56], SG-SSC [57], GaussianFormer2 [58] and SDFormer [59]. (2) The RGB-inferred baselines: LMSCNet [12], AICNet [25] and JS3CNet [42].

4.2 | Comparison Against Camera-Based Methods

As shown in Table 1, CDFNet achieves state-of-the-art performance in the validation set of the SemanticKITTI dataset [23], and it outperforms MonoScene [15] and the transformer-based method VoxelFormer [17] by a large margin in terms of mIoU (+3.22 and +1.37) and wIoU (+7.35, +3.67), respectively. Although our method yields a slightly lower mIoU (by 0.17) than the latest method, symphonies [18], it significantly outperforms symphonies in terms of IoU (by 3.86) and wIoU (by 3.06) and achieves the best results across most semantic classes in outdoor scenarios, particularly for road, sidewalk, building, car, vegetation, terrain, fence, pole and traffic sign.



(a) Symphonies [18] (b) CDFNet (c) Ground Truth

FIGURE 6 | Visual results of symphonies [18] and CDFNet. The mIoU and wIoU for (a) are 15.20, 36.07, respectively, whereas for (b) are 14.61 and 36.23.

Table 2 shows the results on the test set of SemanticKITTI dataset [23]. It is evident that our method also yields the best wIoU, leading the second place by 0.8. Although our method slightly falls behind symphonies [18] in terms of IoU and mIoU, it demonstrates significant improvements in key semantic classes such as road, car, sidewalk, vegetation, trunk, terrain, pole and traffic sign. This highlights our method's effectiveness in reconstructing and segmenting these key categories, which are critical for applications in autonomous driving and urban scene analysis.

To visually validate the performance of our method, we show some qualitative results from the validation set in Figure 7. VoxFormer [17] generates numerous incorrect scenes when dealing with complex scenarios, whereas symphonies [18] produces scenes lacking details and missing some corner areas. In contrast, our method can generate more complete scenes with accurate objects and clear details, especially in complex scenarios such as crossroads and rows of parked vehicles.

Moreover, we perform a comparative evaluation against representative classical baselines, including Voxformer [17], symphonies [18] and SGN [19]. Specifically, the average inference time of Voxformer [17], symphonies [18] and SGN [19] is 0.57, 0.49 and 0.65 s, respectively, whereas our method achieves a latency of 0.42 s. This result shows that our approach outperforms these baselines by a margin, which validates the superior computational efficiency of our model.

We also conduct experiments on SSCBench-KITTI-360 [56]. As shown in Table 3, our method achieves excellent performance with an IoU of 47.45, mIoU of 17.22 and wIoU of 37.74. Moreover, this analysis further demonstrates the effectiveness of our method. Moreover, our method outperforms VLScene [61] on the SSC-Bench-KITTI360 dataset [56]. Although SOAP [62] uses five consecutive frames as input to gain more temporal context, our method still achieves comparable and competitive IoU scores.

4.3 | Ablation Studies

We conduct extensive ablation studies on the validation set of SemanticKITTI to evaluate the contribution of each component in CDFNet.

Effect of each component in CDFNet: We first conduct ablation experiments on the three key components of our method, that is, the voxel feature filter, the voxel feature completion module and the point cloud branch. As demonstrated in Table 4, our method achieves the highest IoU, mIoU and wIoU when all three components are presented. When the point cloud is absent, the model shows a significant decline across all three metrics. This demonstrates that the 3D geometric structural features provided by our proposed point cloud greatly benefit 3D scene reconstruction. Additionally, the other two modules also positively contribute to the model's performance, especially the VFF module, whose absence leads to a notable reduction in IoU by 4.21.

TABLE 1 | Quantitative comparisons against the camera-based baseline methods on the SemanticKITTI validation set.

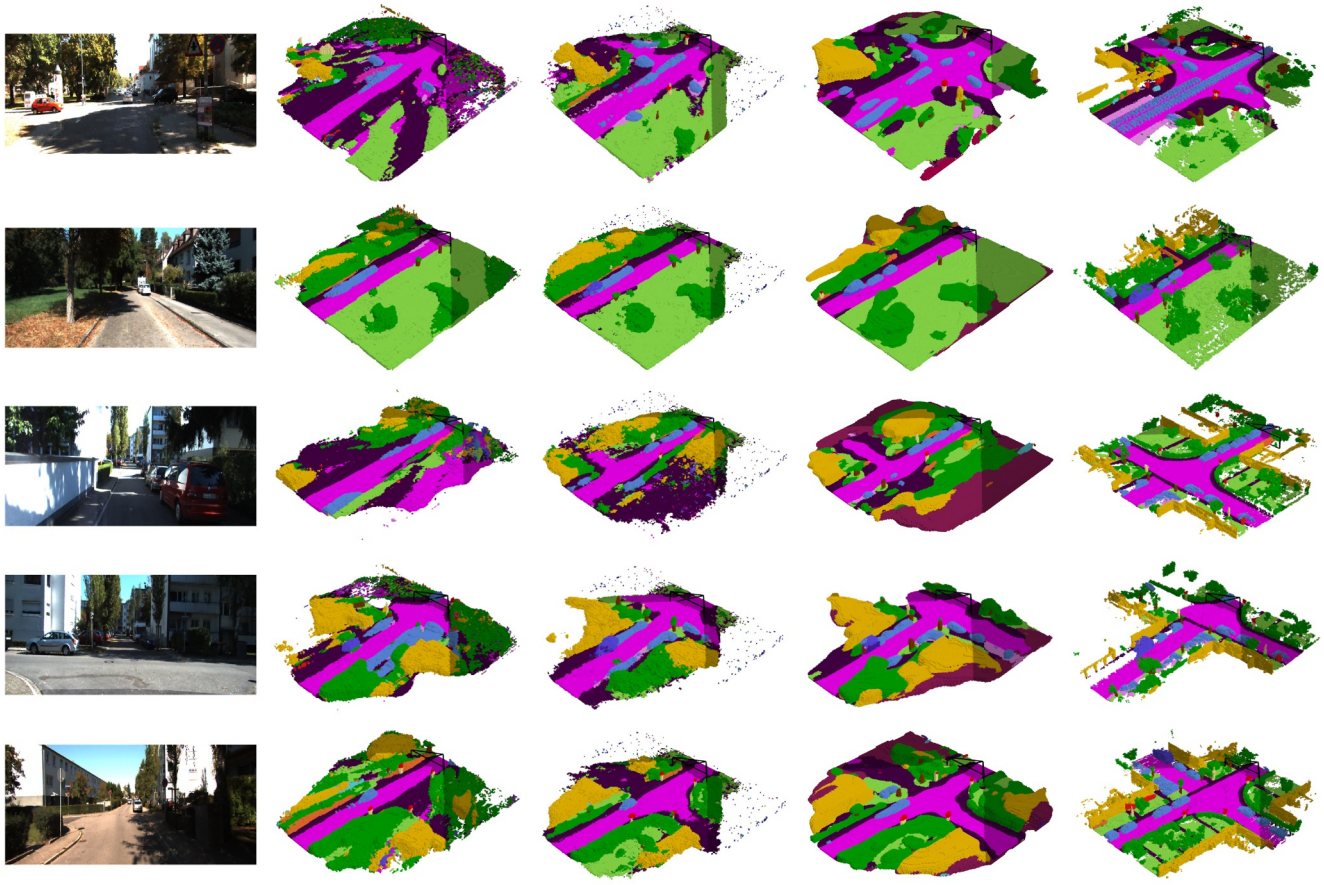
Method	Pub	IoU	mIoU	wIoU	Road (15.30%)	Sidewalk (11.13%)	Parking (1.12%)	Other ground (0.56%)	Building (14.1%)	Car (3.92%)	Truck (0.16%)	Bicycle (0.03%)	Motorcycle (0.03%)	Other-vehicle (0.20%)	Vegetation (39.3%)	Trunk (0.51%)	Terrain (9.17%)	Person (0.07%)	Bicyclist (0.07%)	Motorcyclist (0.05%)	Fence (3.90%)	Pole (0.29%)	Traffic-sign (0.08%)
LMSCNet* [12]	2020	28.61	6.70	17.77	40.68	18.22	4.38	0.00	10.31	18.33	0.00	0.00	0.00	0.00	13.66	0.02	20.54	0.00	0.00	0.00	1.21	0.00	0.00
AICNet* [25]	2020	29.59	8.31	20.27	43.55	20.55	11.97	0.07	12.94	14.71	4.53	0.00	0.00	0.00	15.37	2.90	28.71	0.00	0.00	0.00	2.52	0.06	0.00
JS3C-net* [42]	2021	38.98	10.31	23.37	50.49	23.74	11.94	0.07	15.03	24.65	4.41	0.00	0.00	6.15	18.11	4.33	26.86	0.67	0.27	0.00	3.94	3.77	1.45
MonoScene [15]	2022	37.12	11.50	25.15	57.47	27.05	15.72	0.87	14.24	23.55	7.83	0.20	0.77	3.59	18.12	2.57	30.76	1.79	1.03	0.00	6.39	4.11	2.48
TPVFormer [44]	2023	35.61	11.36	24.36	56.50	25.87	20.60	0.85	13.88	23.81	8.08	0.36	0.05	4.35	16.92	2.26	30.38	0.51	0.89	0.00	5.94	3.14	1.52
VoxFormer-S [17]	2023	44.02	12.35	27.68	54.76	26.35	15.50	0.70	17.65	25.79	5.63	0.59	0.51	3.77	24.39	5.08	29.96	1.78	**3.32**	0.00	7.64	7.11	4.18
VoxFormer-T [17]	2023	44.15	13.35	28.83	53.57	26.52	19.69	0.42	19.54	26.54	7.26	1.28	0.56	7.81	26.10	6.10	33.06	1.93	1.97	0.00	7.31	9.15	4.94
OccFormer [46]	2023	36.50	13.46	26.25	58.85	26.88	19.61	0.31	14.40	25.09	**25.53**	0.81	1.19	8.52	19.63	3.93	32.62	2.78	2.82	0.00	5.61	4.26	2.86
H2GFormer-S [47]	2024	44.57	13.73	29.73	56.08	29.12	17.83	0.45	19.74	27.60	10.00	0.50	0.47	7.39	26.25	6.80	34.42	1.54	2.88	0.00	7.24	7.88	4.68
H2GFormer-T [47]	2024	44.69	14.29	30.75	57.00	29.37	**21.74**	0.34	20.51	28.21	6.80	0.95	0.91	9.32	27.44	**7.80**	36.26	1.15	0.10	0.00	7.98	9.88	5.81
Symphonies [18]	2024	41.91	**14.89**	29.44	56.36	27.59	15.35	**0.93**	21.63	28.68	20.42	**2.53**	**2.80**	**13.95**	25.73	6.61	30.87	**3.50**	2.24	0.00	8.41	9.58	5.76
SGN [19]	2024	43.60	14.55	30.80	59.32	30.51	18.46	0.42	21.43	31.88	13.18	0.58	0.17	5.68	25.98	7.43	34.42	1.28	1.49	0.00	9.66	9.83	4.71
CDFNet (Ours)	—	**45.77**	14.72	**32.50**	**60.75**	**31.99**	11.87	0.18	**24.54**	**33.00**	8.51	0.13	1.48	5.15	**27.79**	7.31	**36.47**	1.03	0.80	0.00	**10.21**	**12.03**	**6.38**

Note: The bold values indicate the best results, the underlined values denote the second-best results, * denotes the results reported in MonoScene and '-T' denotes the results with temporal inputs.

TABLE 2 | Quantitative comparisons against the camera-based baseline methods on the SemanticKITTI hidden test set.

Method	Pub	IoU	mIoU	wIoU	Road (15.30%)	Sidewalk (11.13%)	Parking (1.12%)	Other ground (0.56%)	Building (14.1%)	Car (3.92%)	Truck (0.16%)	Bicycle (0.03%)	Motorcycle (0.03%)	Other-vehicle (0.20%)	Vegetation (39.3%)	Trunk (0.51%)	Terrain (9.17%)	Person (0.07%)	Bicyclist (0.07%)	Motorcyclist (0.05%)	Fence (3.90%)	Pole (0.29%)	Traffic-sign (0.08%)
LMSCNet* [12]	2020	31.38	7.07	16.90	46.70	19.50	13.50	3.10	10.30	14.30	0.30	0.00	0.00	0.00	10.80	0.00	10.40	0.00	0.00	0.00	5.40	0.00	0.00
AICNet* [25]	2020	23.93	7.09	15.45	39.30	18.30	19.80	1.60	9.60	15.30	0.70	0.00	0.00	0.00	9.60	1.90	13.50	0.00	0.00	0.00	5.00	0.10	0.00
JS3C-net* [42]	2021	34.00	8.97	19.55	47.30	21.70	19.90	2.80	12.70	20.10	0.80	0.00	0.00	4.10	14.20	3.10	12.40	0.00	0.20	0.20	8.70	1.90	0.30
MonoScene [15]	2022	34.16	11.08	22.57	54.70	27.10	24.80	5.70	14.40	18.80	3.30	0.50	0.70	4.40	14.90	2.40	19.50	1.00	1.40	0.40	11.10	3.30	2.10
TPVFormer [44]	2023	34.25	11.26	22.43	55.10	27.20	27.40	6.50	14.80	19.20	3.70	1.00	0.50	2.30	13.90	2.60	20.40	1.10	2.40	0.30	11.00	2.90	1.50
VoxFormer-S [17]	2023	42.95	12.20	26.20	53.90	25.30	21.10	5.60	19.80	20.80	3.50	1.00	0.70	3.70	22.40	7.50	21.30	1.40	**2.60**	0.20	11.10	5.10	4.90
VoxFormer-T [17]	2023	43.21	13.41	28.15	54.10	26.90	25.10	7.30	23.50	21.70	3.60	1.90	1.60	4.10	24.40	8.10	24.20	1.60	1.10	0.00	13.10	6.60	5.70
OccFormer [46]	2023	34.53	12.32	24.43	55.90	30.30	**31.50**	6.50	15.70	21.60	1.20	1.50	1.70	3.20	16.80	3.90	21.30	2.20	1.10	0.20	11.90	3.80	3.70
H2GFormer-S [47]	2024	**44.20**	13.72	28.72	56.40	28.60	26.50	4.90	22.80	23.40	4.80	0.80	0.90	4.10	24.60	9.10	23.80	1.20	2.50	0.90	13.30	6.40	6.30
H2GFormer-T [47]	2024	43.52	14.60	29.87	57.90	**30.40**	30.00	6.90	24.00	23.70	**5.20**	0.60	1.20	**5.00**	25.20	10.70	25.80	1.10	0.10	1.20	14.60	7.50	7.10
Symphonies [18]	2024	42.00	**14.80**	29.25	58.10	29.10	27.40	**9.70**	**24.10**	23.00	3.20	**3.60**	**2.90**	4.80	24.70	9.70	22.40	**2.60**	1.80	**2.60**	**15.60**	7.60	**7.70**
SGN [19]	2024	41.88	14.01	29.30	57.80	29.20	27.70	5.20	23.90	24.90	2.70	0.40	0.30	4.00	24.20	10.00	25.80	1.10	2.50	0.30	14.20	7.40	4.40
SG-SSC [53]	2025	41.91	12.90	26.97	55.30	26.00	18.70	7.60	22.30	21.50	3.10	0.80	1.00	2.60	22.20	9.20	21.80	1.70	1.10	0.40	14.50	8.30	7.30
MonoMRN [54]	2025	42.00	13.80	28.78	**62.10**	28.70	22.50	5.60	21.30	23.60	2.60	0.80	0.70	4.40	22.40	5.10	26.00	2.40	1.80	0.40	16.00	7.90	7.10
CDFNet (ours)	—	40.70	14.20	**30.67**	61.50	30.30	22.40	4.60	23.00	**25.40**	3.00	0.30	0.70	2.80	**26.10**	**11.10**	**27.00**	1.10	0.80	0.00	14.00	**9.30**	7.10

Note: The bold values indicate the best results, the underlined values denote the second-best results, * denotes the results reported in MonoScene and '-T' denotes the results with temporal inputs.



(a) Input (b) VoxFormer [17] (c) Symphonies [18] (d) CDFNet (e) Ground Truth

FIGURE 7 | Qualitative results of our method and state-of-the-art methods (VoxFormer [17], symphonies [18]). Our method (CDFNet) shows more accurate objects and clear details of the 3D scene when dealing with complex scenarios.

To more rigorously observe the effect of each module, we display the visual results both after progressively adding each module (in Figure 8) and selectively removing specific modules from our complete model (in Figure 9). As shown in Figure 8, we can see that the voxel feature filter is able to remove some redundant scenes while preserving important details, and the voxel feature completion module can augment the scene with more key instances or regions. Furthermore, our method can generate clearer details and more realistic scenes by adding the point cloud branch, which validates the effectiveness of 3D geometric features. Figure 9 demonstrates that our model produces a larger amount of redundant and false voxels without the voxel feature filter. Without the voxel feature completion module, our method cannot effectively aggregate the occupied voxel features and adequately complete the unoccupied voxels in the complex scene, which results in inaccurate and incomplete scene representations.

Moreover, since the point cloud can provide detailed structural and spatial information of complex scenes, our model without the point cloud exhibits poorer performance in the overall structure of the generated scenes.

Ablation on the components in the CSF module: The CSF module incorporates multiple sub-modules, including the SELayer, MPAC and DSE, to effectively extract and fuse 2D textural and spatial features. We conduct experiments to validate the individual contributions of each sub-module to the overall performance of the framework. Note that we remove the VFF, VFC and the point cloud branch from our model to eliminate their influence. As demonstrated in Table 5, introducing the SELayer results in a slight decrease in IoU, but mIoU and wIoU are slightly improved to 11.69 and 27.34, respectively. The combination of SELayer and MPAC improves the IoU to 41.01. Finally, the inclusion of the DSE module further boosts the performance, achieving the highest IoU of 42.64, with mIoU and wIoU also reaching 11.89 and 28.57, respectively. These results demonstrate the significant contribution of the three components, especially the DSE unit, in enhancing the overall performance of the CSF module.

Since the DSE unit in our CSF module plays an important role, to comprehensively validate the effect of the DSE unit, we further conduct experiments to test the impact of the DSE unit on our complete model. The results presented in Table 6 show that the DSE unit significantly improves the model's performance across all three metrics, demonstrating its effectiveness and robustness. As visualised in Figure 10, our method obviously boosts the depth perception performance of the model through the DSE unit. We believe that the significant effect of

TABLE 3 | Quantitative comparisons against the camera-based baseline methods on the SSCBench-KITTI360 test set.

Method	Pub	IoU	mIoU	wIoU	Car (2.85%)	Bicycle (0.01%)	Motorcycle (0.01%)	Truck (0.16%)	Other-veh. (5.75%)	Person (0.02%)	Road (14.98%)	Parking (2.31%)	Sidewalk (6.43%)	Other-grnd. (2.05%)	Building (15.67%)	Fence (0.96%)	Vegetation (41.99%)	Terrain (7.10%)	Pole (0.22%)	Traf.-sign (0.06%)	Other-struct. (4.33%)	Other-obj. (0.28%)
MonoScene [15]	2022	37.87	12.31	27.63	19.34	0.43	0.58	8.02	2.03	0.86	48.35	11.38	28.13	3.32	32.89	3.53	26.15	16.75	6.92	5.67	4.20	3.09
TPVFormer [44]	2023	40.22	13.64	30.70	21.56	1.09	1.37	8.06	2.57	2.38	52.99	11.99	31.07	3.78	34.83	4.80	30.08	17.52	7.46	5.86	5.48	2.70
VoxFormer [17]	2023	38.76	11.81	28.04	17.84	1.16	0.89	4.56	2.06	1.63	47.01	9.67	27.21	2.89	31.18	4.97	28.99	14.69	6.51	6.92	3.79	2.43
OccFormer [46]	2023	40.27	13.81	31.91	22.58	0.66	0.26	9.89	3.82	2.77	54.30	13.44	31.53	3.55	36.42	4.80	31.00	19.51	7.77	8.51	6.95	4.60
DepthSSC [60]	2023	40.85	14.28	30.94	21.90	2.36	4.30	11.51	4.56	2.92	50.88	12.89	30.27	2.49	37.33	5.22	29.61	21.59	5.97	7.71	5.24	3.51
Symphonies [18]	2024	44.12	18.58	35.57	30.02	1.85	5.90	**25.07**	**12.06**	**8.20**	54.94	13.83	32.76	**6.93**	35.11	8.58	38.33	11.52	14.01	9.57	**14.44**	**11.28**
SGN [19]	2024	46.22	17.71	36.44	28.20	2.09	3.02	11.95	3.68	4.20	59.49	14.50	36.53	4.24	39.79	7.14	36.61	23.10	14.86	16.14	8.24	4.95
GaussianFormer2 [58]	2025	38.37	13.90	30.18	21.08	2.55	4.21	12.41	5.73	1.59	54.12	11.04	32.31	3.34	32.01	4.98	28.94	17.33	3.57	5.48	5.88	3.54
VLScene [61]	2025	46.08	19.10	37.62	29.00	**4.70**	**7.70**	18.30	7.60	7.40	60.10	**17.40**	**39.00**	6.00	42.10	**9.60**	36.50	**24.80**	**17.00**	**18.80**	10.50	6.50
SDFormer [59]	2025	46.99	**20.59**	**38.72**	**30.78**	2.85	3.65	13.58	6.77	5.96	61.83	16.37	38.88	5.43	**43.16**	8.86	**38.48**	23.48	15.86	17.35	10.33	6.51
CDFNet (Ours)	—	**47.45**	17.22	37.74	28.49	0.53	0.43	1.45	4.93	2.25	**62.62**	14.62	38.13	5.06	39.65	5.88	38.21	22.64	14.03	17.15	9.02	4.84

Note: The bold values indicate the best results, and the underlined values denote the second-best results.

TABLE 4 | Ablation studies on three key components in CDFNet.

Voxel feature filter	Voxel feature completion	Point cloud	IoU	mIoU	wIoU	Params (M)	Memory (MB)
×	×	×	42.64	11.89	28.57	61.4	245.655
✓	×	×	43.25	13.01	29.52	61.5	246.113
✓	✓	×	43.71	13.36	30.66	66.2	264.656
✓	×	✓	45.01	14.56	31.85	61.6	246.371
×	✓	✓	41.56	14.51	32.43	66.1	264.455
✓	✓	✓	**45.77**	**14.72**	**32.50**	66.2	264.913

Note: The best results are in bold.

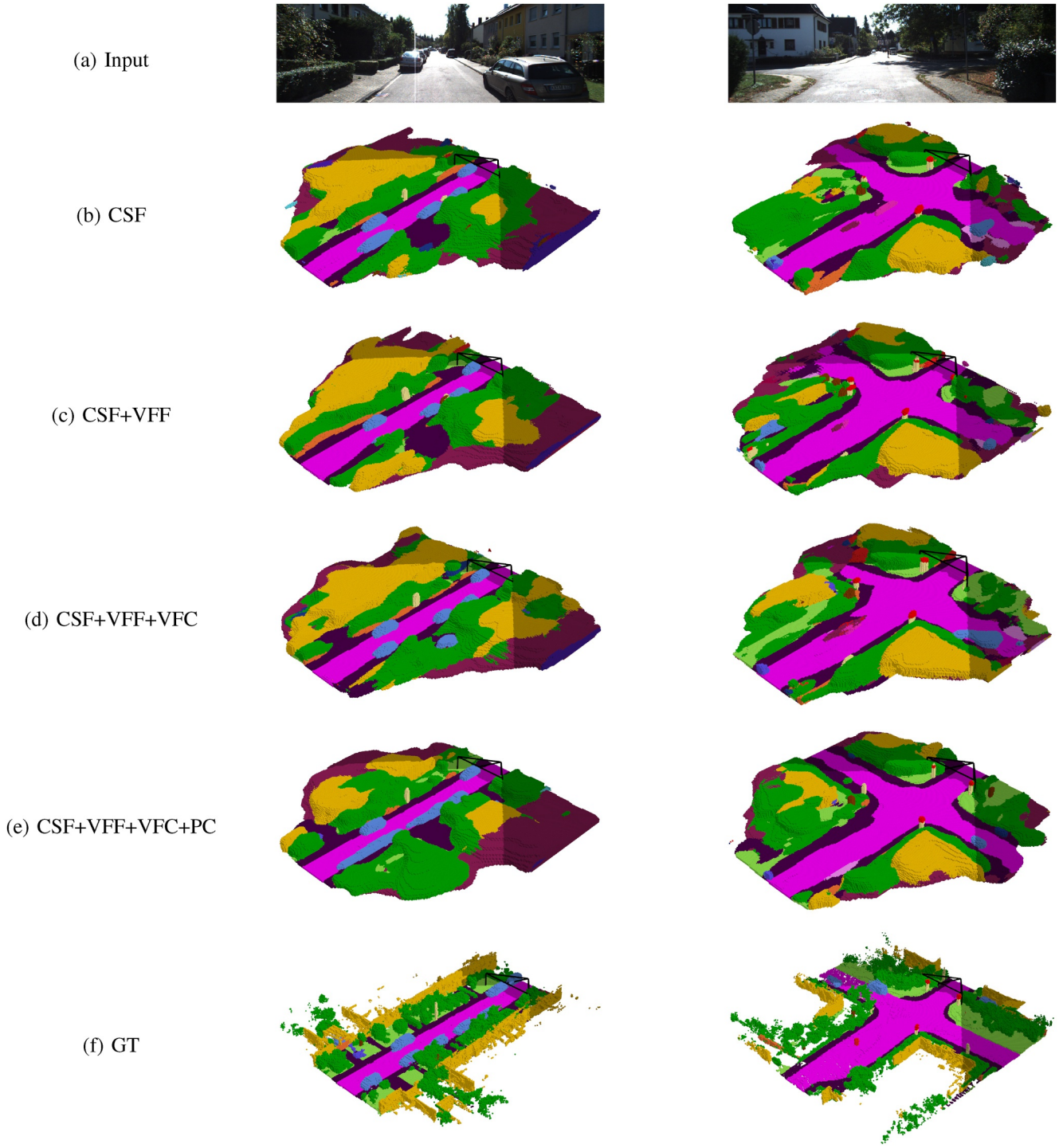


FIGURE 8 | Visual results of progressively adding three architectural components. CDFNet generates clearer and more realistic scenes by progressively adding the voxel feature filter (VFF), voxel feature completion (VFC) and point cloud branch (PC) based on the cross-modal semantic fusion module (CSF).

DSE is due to its ability to effectively enrich the information of 2D features, thereby enhancing the model's comprehensive and accurate perception of the 3D scene.

Ablation on efficient of depth estimator: To more fairly evaluate the fusion efficiency of our architecture, we adopt the state-of-the-art stereo depth estimator (DEFOMStereo) [63] to replace our default depth estimator (MobileStereoNet) [33]. DEFOM-Stereo outperforms MobileStereoNet by a significant margin in terms of depth estimation quality across multiple datasets, including KITTI 2012, KITTI 2015, Middlebury and ETH3D. As demonstrated in Table 7, our method still achieves competitive performance even when integrated with DEFOMStereo [63]. This demonstrates the effectiveness of the depth estimation strategy we employed (Table 8).

Ablation on normalisation layers and Bias: CDFNet discards the normalisation layers and bias in the voxel feature filter and

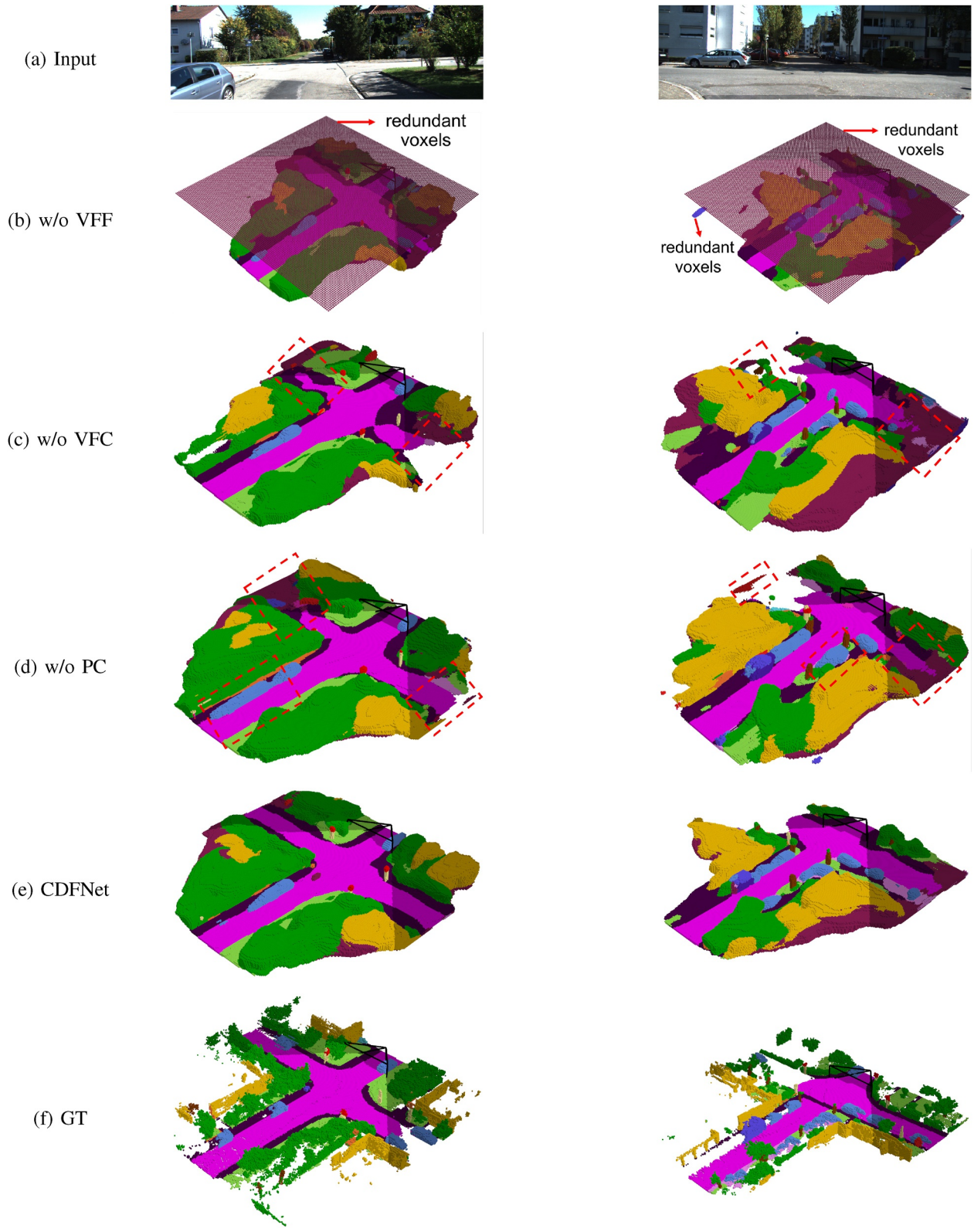


FIGURE 9 | Visual results of selectively removing the key components, including voxel feature filter (VFF), voxel feature completion (VFC) and point cloud branch (PC).

completion module to preserve the intra-voxel structure and inter-voxel relation voxel features during optimisation. In this section, we conduct the ablation experiments to validate the impact of the normalisation layers and/or bias on the completion results. As shown in Table 8, after adding normalisation layers and bias, the model's performance on IoU, mIoU, and wIoU decreased by 0.45, 0.15 and 0.25, respectively. Moreover, regardless of adding bias or normalisation layers, they all degrade the performance of the proposed method. These findings indicate that the normalisation layer and bias have a

TABLE 5 | Ablation studies on the key components in the CSF module.

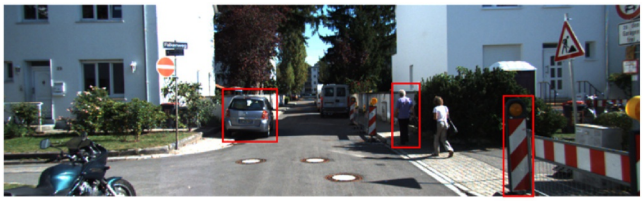
Model	IoU	mIoU	wIoU
Baseline	36.56	11.28	26.53
w/SELayer	35.13	11.69	27.34
w/SELayer + MPAC	41.01	11.25	27.03
w/SELayer + MPAC + DSE	**42.64**	**11.89**	**28.57**

Note: The 'Baseline' indicates the model without SELayer, MPAC and DSE. The bold values represent the best results.

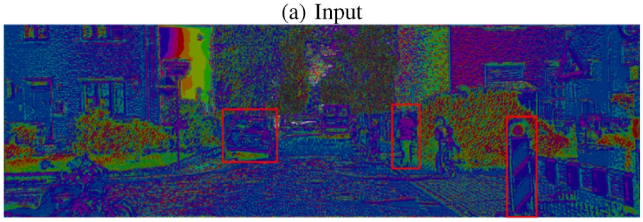
TABLE 6 | Effect of the depth semantic enhancement module.

Model	IoU	mIoU	wIoU
CSF w/o DSE	41.01	11.25	27.03
CSF w/DSE	**42.64**	**11.89**	**28.57**
Ours w/o DSE	45.59	14.26	32.06
Ours w/DSE	**45.77**	**14.72**	**32.50**

Note: The best results are marked in bold.



(a) Input



(b) Heatmap of feature perception enhanced by DSE Unit

FIGURE 10 | Visualisation results of the heatmap enhanced by the DSE unit, our method achieves an accurate scene perception, such as car, person and traffic sign.

TABLE 7 | Effect of the depth estimator.

Component	IoU	mIoU	wIoU
MobileStereoNet [33]	**45.77**	14.72	**32.51**
DEFOMStereo [63]	45.36	**14.76**	32.42

Note: The best results are marked in bold.

TABLE 8 | Impact of the normalisation layers and bias.

Component	IoU	mIoU	wIoU
w/Bias	44.38	14.27	31.80
w/Normalisation	45.33	14.59	32.23
w/Normalisation + bias	45.32	14.57	32.25
w/o normalisation + bias	**45.77**	**14.72**	**32.50**

Note: The best results are marked in bold.

negative impact on completion results. Although removing these operations from the model is not innovative, our work is the first to identify and analyse the negative impact of normalisation and bias on 3D semantic scene completion, which provides valuable insights for future 3D scene analysis and reconstruction tasks.

Ablation on multi-scale feature encoder: In many vision-based tasks, there are two commonly used modules, that is, ResNet and U-Net, to extract the multi-scale features. In this work, we explore the results of employing these two modules separately in our CSF module. As shown in Table 9, the performance of the ResNet50-based encoder is inferior to that of the U-Net encoder used by our model. This indicates that our U-Net architecture has a stronger capability for extracting multi-scale features than ResNet.

Moreover, our feature encoder generates three levels of feature maps for lifting 3D features, which is different from MonoScene [15], VoxelFromer [17] and H2GFormer [47] that generate four levels of features. The detailed quantitative analysis is shown in Table 10. It can be seen from Table 10 that extracting three-layer feature maps achieves the best performance, whereas four-layer feature maps actually degrade the performance while increasing computational overhead. This suggests that more layers are not necessarily better for the U-Net-based feature encoder in the SSC task. In our framework, three layers are sufficient to achieve the best performance.

Ablation on the λ coefficient: As mentioned, the voxel feature filter is a critical component of our approach. Therefore, we conduct experiments to explore the choice of the coefficient λ in the joint loss function Equation (9). As demonstrated in Table 11, although $\lambda = 0.5$ yields the best mIoU, the best IoU and wIoU are achieved at $\lambda = 1.0$. Since IoU and wIoU are more accurate and reliable than mIoU, we thus set $\lambda = 1.0$ in our experiment.

TABLE 9 | Comparison of the feature encoders in the cross-modal semantic fusion module.

Component	IoU	mIoU	wIoU
ResNet50	44.37	12.99	29.18
Multi-scale feature encoder	**45.77**	**14.72**	**32.50**

Note: The best results are marked in bold.

TABLE 10 | Effect of the number of multi-level features in the feature encoder.

Layers	IoU	mIoU	wIoU
1	45.57	14.32	32.25
2	44.70	14.34	32.11
3	**45.77**	**14.72**	**32.50**
4	45.02	14.21	31.90

Note: The best results are marked in bold.

TABLE 11 | Analysis of coefficient λ in the loss function.

λ	IoU	mIoU	wIoU
0.5	45.26	**14.86**	32.46
0.8	45.21	14.54	31.80
1.0	**45.77**	14.72	**32.50**

Note: The best results are marked in bold.

5 | Conclusion

In this paper, we present CDFNet, a cross-modal 3D semantic scene completion framework. Our model integrates 2D textural, 2D spatial and 3D geometric information from RGB images, depth maps and point clouds to reconstruct complete scenes with detailed semantics. Moreover, we finely and delicately process cross-modal features by designing a voxel feature filter to remove redundant features, a depth semantic enhancement module to enhance multi-dimension information for 2D features and a voxel feature completion module to fill in voxel features in sparse and occluded regions. Additionally, we are the first to examine the impact of normalisation and bias in camera-based semantic scene completion tasks and we propose a new evaluation metric to address the issue of class imbalance in this task. We think that our work provides valuable ideas and clues for future 3D semantic scene completion and reconstruction, as well as 3D scene understanding tasks.

Funding

This work was supported in part by the National Nature Science Foundation of China under Grants 62466005, 62572149 and 82360445.

Conflicts of Interest

Yong Xu is a Deputy Editors-in-Chief for the journal, and was not involved in peer review process or the decision to publish this article. The authors declare that they have no conflict of interest.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

References

1. X. Cheng, L. Wu, J. Hou, Z. Hu, J. Wen, and Y. Xu, "Arbitrary-Scale Point Cloud Upsampling via Enhanced Geometric Spatial Consistency," *CAAI Transactions on Intelligence Technology* (2025).

2. Y. Zhao, W. Nie, J. Nie, Y. Zhang, and B. Wang, "Rjan: Region-Based Joint Attention Network for 3d Shape Recognition," *CAAI Transactions on Intelligence Technology* 10, no. 2 (2025): 460–473, https://doi.org/10.1049/cit2.12388.

3. R. Chang, Y. Ma, T. Hao, W. Wang, and W. Nie, "3d Shape Knowledge Graph for Cross-Domain 3d Shape Retrieval," *CAAI Transactions on Intelligence Technology* 9, no. 5 (2024): 1199–1216, https://doi.org/10.1049/cit2.12326.

4. L. Hong, M. H. Modirrousta, M. Hossein Nasirpour, et al., "Gan-lstm-3d: An Efficient Method for Lung Tumour 3d Reconstruction Enhanced by Attention-Based Lstm," *CAAI Transactions on Intelligence Technology* (2023).

5. N. Yu, H. Li, Q. Xu, O. Sie, and E. Firdaous, "3d Reconstruction and Defect Pattern Recognition of Bonding Wire Based on Stereo Vision," *CAAI Transactions on Intelligence Technology* 9, no. 2 (2024): 348–364, https://doi.org/10.1049/cit2.12240.

6. B. Wen, H. Xie, Z. Chen, F. Hong, and Z. Liu, "3d Scene Generation: A Survey," *arXiv preprint arXiv:2505.05474* (2025).

7. L. Roldao, R. De Charette, and A. Verroust-Blondet, "3d Semantic Scene Completion: A Survey," *International Journal of Computer Vision* 130, no. 8 (2022): 1978–2005, https://doi.org/10.1007/s11263-021-01504-5.

8. L. Wu, X. Cheng, Y. Xu, H. Zeng, and J. Hou, "Unsupervised 3d Point Cloud Completion Via Multi-View Adversarial Learning," *IEEE Transactions on Visualization and Computer Graphics* 31, no. 10 (2025): 7890–7905, https://doi.org/10.1109/tvcg.2025.3559340.

9. X. Cheng, L. Wu, Z. Wang, J. Hou, J. Wen, and Y. Xu, "Pvnet: Point-Voxel Interaction LiDAR Scene Upsampling via Diffusion Models," *IEEE Transactions on Image Processing* 34 (2025): 6895–6910, https://doi.org/10.1109/tip.2025.3620666.

10. L. Wu, J. Hou, L. Song, and Y. Xu, "3d Shape Completion on Unseen Categories: A Weakly-Supervised Approach," *IEEE Transactions on Visualization and Computer Graphics* 31, no. 9 (2025): 5547–5561, https://doi.org/10.1109/tvcg.2024.3459018.

11. 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 IEEE Conference Computer Vision and Pattern Recognition*, (2017), 1746–1754.

12. L. Roldao, R. de Charette, and A. Verroust-Blondet, "Lmscnet: Lightweight Multiscale 3d Semantic Completion," in *Proceedings of International Conference on 3D Vision* (IEEE, 2020), 111–119.

13. T. Kühner and J. Kümmerle, "Large-Scale Volumetric Scene Reconstruction Using Lidar," in *2020 IEEE International Conference on Robotics and Automation* (IEEE, 2020), 6261–6267.

14. L. Wu, Q. Zhang, J. Hou, and Y. Xu, "Leveraging Single-View Images for Unsupervised 3d Point Cloud Completion," *IEEE Transactions on Multimedia* 27 (2025): 940–953, https://doi.org/10.1109/tmm.2023.3340892.

15. A.-Q. Cao and R. De Charette, "Monoscene: Monocular 3d Semantic Scene Completion," in *Proceedings of IEEE Conference Computer Vision and Pattern Recognition*, (2022), 3991–4001.

16. S. Popov, P. Bauszat, and V. Ferrari, "Corenet: Coherent 3d Scene Reconstruction From a Single Rgb Image," in *Proceedings of European Conference on Computer Vision* (Springer, 2020), 366–383.

17. Y. Li, Z. Yu, C. Choy, et al., "Voxformer: Sparse Voxel Transformer for Camera-based 3d Semantic Scene Completion," in *Proceedings of IEEE Conference Computer Vision and Pattern Recognition*, (2023), 9087–9098.

18. H. Jiang, T. Cheng, N. Gao, et al., "Symphonize 3d Semantic Scene Completion With Contextual Instance Queries," in *Proceedings of IEEE Conference Computer Vision and Pattern Recognition*, (2024), 20258–20267.

19. J. Mei, Y. Yang, M. Wang, et al., "Camera-Based 3d Semantic Scene Completion With Sparse Guidance Network," *IEEE Transactions on Image Processing* (2024).

20. Z. Yu, R. Zhang, J. Ying, et al., "Context and Geometry Aware Voxel Transformer for Semantic Scene Completion," *arXiv preprint arXiv:2405.13675* (2024).

21. J. Hu, L. Shen, and G. Sun, "Squeeze-And-Excitation Networks," in *Proceedings of IEEE Conference Computer Vision and Pattern Recognition*, (2018), 7132–7141.

22. X. Zhu, H. Zhou, T. Wang, et al., "Cylindrical and Asymmetrical 3d Convolution Networks for LiDAR Segmentation," in *Proceedings of IEEE Conference Computer Vision and Pattern Recognition*, (2021), 9939–9948.

23. J. Behley, M. Garbade, A. Milioto, et al., "SemanticKITTI: A Dataset for Semantic Scene Understanding of LiDAR Sequences," in *Proceedings of IEEE International Conference on Computer Vision*, (2019).

24. S. Li, C. Zou, Y. Li, X. Zhao, and Y. Gao, "Attention-Based Multi-Modal Fusion Network for Semantic Scene Completion," *Proceedings of the AAAI Conference on Artificial Intelligence* 34, no. 7 (2020): 11402–11409, https://doi.org/10.1609/aaai.v34i07.6803.

25. J. Li, K. Han, P. Wang, Y. Liu, and X. Yuan, "Anisotropic Convolutional Networks for 3d Semantic Scene Completion," in *Proceedings of IEEE Conference Computer Vision and Pattern Recognition*, (2020), 3351–3359.

26. S. F. Bhat, R. Birkl, D. Wofk, P. Wonka, and M. Müller, "Zoedepth: Zero-Shot Transfer by Combining Relative and Metric Depth," *arXiv preprint arXiv:2302.12288* (2023).

27. L. Yang, B. Kang, Z. Huang, X. Xu, J. Feng, and H. Zhao, "Depth Anything: Unleashing the Power of Large-Scale Unlabeled Data," in *Proceedings of IEEE Conference Computer Vision and Pattern Recognition*, (2024), 10371–10381.

28. S. Patni, A. Agarwal, and C. Arora, "Ecodepth: Effective Conditioning of Diffusion Models for Monocular Depth Estimation," in *Proceedings of IEEE Conference Computer Vision and Pattern Recognition*, (2024), 28285–28295.

29. J. Song, C. Meng, and S. Ermon, "Denoising Diffusion Implicit Models," *arXiv preprint arXiv:2010.02502* (2020).

30. J. Ho, A. Jain, and P. Abbeel, "Denoising Diffusion Probabilistic Models," in *Proceedings of Advanced Neural Information Processing System*, Vol. 33, (2020), 6840–6851.

31. F. Zhang, V. Prisacariu, R. Yang, and P. H. Torr, "Ga-Net: Guided Aggregation Net for End-to-End Stereo Matching," in *Proceedings of IEEE Conference Computer Vision and Pattern Recognition*, (2019), 185–194.

32. H. Xu and J. Zhang, "Aanet: Adaptive Aggregation Network for Efficient Stereo Matching," in *Proceedings of IEEE Conference Computer Vision and Pattern Recognition*, (2020), 1959–1968.

33. F. Shamsafar, S. Woerz, R. Rahim, and A. Zell, "Mobilestereonet: Towards Lightweight Deep Networks for Stereo Matching," in *Proceedings of IEEE/CVF Winter Conference on Applications of Computer Vision*, (2022), 2417–2426.

34. L. Lipson, Z. Teed, and J. Deng, "Raft-Stereo: Multilevel Recurrent Field Transforms for Stereo Matching," in *Proceedings of International Conference on 3D Vision* (IEEE, 2021), 218–227.

35. K. Li, L. Wang, Y. Zhang, K. Xue, S. Zhou, and Y. Guo, "Los: Local Structure-Guided Stereo Matching," in *Proceedings of IEEE Conference Computer Vision and Pattern Recognition*, (2024), 19746–19756.

36. X. Wang, G. Xu, H. Jia, and X. Yang, "Selective-Stereo: Adaptive Frequency Information Selection for Stereo Matching," in *Proceedings of IEEE Conference Computer Vision and Pattern Recognition*, (2024), 19701–19710.

37. 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 European Conference on Computer Vision*, (2018), 733–749.

38. X. Yang, H. Zou, X. Kong, et al., "Semantic Segmentation-Assisted Scene Completion for LiDAR Point Clouds," in *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems* (IEEE, 2021), 3555–3562.

39. Y. Wang, V. C. Guizilini, T. Zhang, Y. Wang, H. Zhao, and J. Solomon, "Detr3d: 3d Object Detection From Multi-View Images Via 3d-to-2d Queries," in *Conference on Robot Learning* (PMLR, 2022), 180–191.

40. Z. Chen and H. Zhang, "Learning Implicit Fields for Generative Shape Modeling," in *Proceedings of IEEE Conference Computer Vision and Pattern Recognition*, (2019), 5939–5948.

41. Z. Xia, Y. Liu, X. Li, et al., "Scpnet: Semantic Scene Completion on Point Cloud," in *Proceedings of IEEE Conference Computer Vision and Pattern Recognition*, (2023), 17642–17651.

42. X. Yan, J. Gao, J. Li, et al., "Sparse Single Sweep LiDAR Point Cloud Segmentation via Learning Contextual Shape Priors From Scene Completion," *Proceedings of the AAAI Conference on Artificial Intelligence* 35, no. 4 (2021): 3101–3109, https://doi.org/10.1609/aaai.v35i4.16419.

43. Z. Li, W. Wang, H. Li, et al., "Bevformer: Learning Bird's-Eye-View Representation From Multi-Camera Images Via Spatiotemporal Transformers," in *Proceedings of European Conference on Computer Vision* (Springer, 2022), 1–18.

44. Y. Huang, W. Zheng, Y. Zhang, J. Zhou, and J. Lu, "Tri-Perspective View for Vision-Based 3d Semantic Occupancy Prediction," in *Proceedings of IEEE Conference Computer Vision and Pattern Recognition*, (2023), 9223–9232.

45. Y. Wang, I. Skorokhodov, and P. Wonka, "Pet-Neus: Positional Encoding Tri-Planes for Neural Surfaces," in *Proceedings of IEEE Conference Computer Vision and Pattern Recognition*, (2023), 12598–12607.

46. Y. Zhang, Z. Zhu, and D. Du, "Occformer: Dual-Path Transformer for Vision-Based 3d Semantic Occupancy Prediction," in *Proceedings of IEEE International Conference on Computer Vision*, (2023), 9433–9443.

47. Y. Wang and C. Tong, "H2gformer: Horizontal-To-Global Voxel Transformer for 3d Semantic Scene Completion," *Proceedings of the AAAI Conference on Artificial Intelligence* 38, no. 6 (2024): 5722–5730, https://doi.org/10.1609/aaai.v38i6.28384.

48. M. Tan and Q. Le, "Efficientnet: Rethinking Model Scaling for Convolutional Neural Networks," in *Proceedings of International Conference on Machine Learning* (PMLR, 2019), 6105–6114.

49. J. Philion and S. Fidler, "Lift, Splat, Shoot: Encoding Images From Arbitrary Camera Rigs by Implicitly Unprojecting to 3d," in *Proceedings of European Conference on Computer Vision* (Springer, 2020), 194–210.

50. M. Ye, R. Wan, S. Xu, T. Cao, and Q. Chen, "Efficient Point Cloud Segmentation With Geometry-Aware Sparse Networks," in *European Conference on Computer Vision* (Springer, 2022), 196–212.

51. M. Berman, A. R. 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 IEEE Conference Computer Vision and Pattern Recognition*, (2018), 4413–4421.

52. A. Milioto, I. Vizzo, J. Behley, and C. Stachniss, "Rangenet++: Fast and Accurate LiDAR Semantic Segmentation," in *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems* (IEEE, 2019), 4213–4220.

53. M. Menze and A. Geiger, "Object Scene Flow for Autonomous Vehicles," in *Proceedings of IEEE Conference Computer Vision and Pattern Recognition*, (2015).

54. Y. Liao, J. Xie, and A. Geiger, "Kitti-360: A Novel Dataset and Benchmarks for Urban Scene Understanding in 2d and 3d," *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2022).

55. J. M. Johnson and T. M. Khoshgoftaar, "Survey on Deep Learning With Class Imbalance," *Journal of Big Data* 6, no. 1 (2019): 1–54, https://doi.org/10.1186/s40537-019-0192-5.

56. X. Wang, X. Wu, S. Wang, L. Kong, and Z. Zhao, "Monocular Semantic Scene Completion via Masked Recurrent Networks," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, (2025), 24811–24822.

57. X. Liu, H. Xie, S. Zhang, et al., "2d Semantic-Guided Semantic Scene Completion," *International Journal of Computer Vision* 133, no. 3 (2025): 1306–1325, https://doi.org/10.1007/s11263-024-02244-y.

58. Y. Huang, A. Thammatadatrakoon, W. Zheng, Y. Zhang, D. Du, and J. Lu, "Gaussianformer-2: Probabilistic Gaussian Superposition for Efficient 3d Occupancy Prediction," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, (2025), 27477–27486.

59. Y. Xue, H. Pi, J. Zhang, et al., "Sdformer: Vision-Based 3d Semantic Scene Completion via Sam-Assisted Dual-Channel Voxel Transformer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, (2025), 26837–26847.

60. J. Yao and J. Zhang, "Depthssc: Depth-Spatial Alignment and Dynamic Voxel Resolution for Monocular 3d Semantic Scene Completion," *arXiv preprint arXiv:2311.17084* 7 (2023): 16, https://doi.org/10.48550/arXiv.2311.17084.

61. M. Wang, H. Pi, R. Li, Y. Qin, Z. Tang, and K. Li, "Vlscene: Vision-Language Guidance Distillation for Camera-Based 3d Semantic Scene Completion," *Proceedings of the AAAI Conference on Artificial Intelligence* 39, no. 8 (2025): 7808–7816, https://doi.org/10.1609/aaai.v39i8.32841.

62. H.-J. Lee, Y. J. Koh, H. Kim, H. Kim, Y. Lee, and J. Lee, "Soap: Vision-Centric 3d Semantic Scene Completion With Scene-Adaptive Decoder and Occluded Region-Aware View Projection," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, (2025), 17145–17154.

63. H. Jiang, Z. Lou, L. Ding, et al., "Defom-Stereo: Depth Foundation Model Based Stereo Matching," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, (2025), 21857–21867.