Monocular Semantic Scene Completion via Masked Recurrent Networks

Xuzhi Wang[1] Xinran Wu[1] Song Wang[2] Lingdong Kong[3*] Ziping Zhao[1*]
[1]Tianjin Normal University [2]Zhejiang University [3]National University of Singapore

Abstract

Monocular Semantic Scene Completion (MSSC) aims to predict the voxel-wise occupancy and semantic category from a single-view RGB image. Existing methods adopt a single-stage framework that aims to simultaneously achieve visible region segmentation and occluded region hallucination, while also being affected by inaccurate depth estimation. Such methods often achieve suboptimal performance, especially in complex scenes. We propose a novel two-stage framework that decomposes MSSC into coarse MSSC followed by the Masked Recurrent Network. Specifically, we propose the Masked Sparse Gated Recurrent Unit (MS-GRU) which concentrates on the occupied regions by the proposed mask updating mechanism, and a sparse GRU design is proposed to reduce the computation cost. Additionally, we propose the distance attention projection to reduce projection errors by assigning different attention scores according to the distance to the observed surface. Experimental results demonstrate that our proposed unified framework, MonoMRN, effectively supports both indoor and outdoor scenes and achieves state-of-the-art performance on the NYUv2 and SemanticKITTI datasets. Furthermore, we conduct robustness analysis under various disturbances, highlighting the role of the Masked Recurrent Network in enhancing the model's resilience to such challenges. The source code is publicly available at: https://github.com/alanWXZ/MonoMRN.

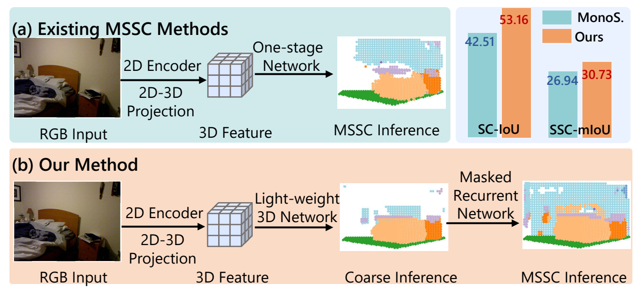

Figure 1. The framework of our method and existing methods. Our method decomposes MSSC into coarse SSC and Masked Recurrent Network. MonoS. represents the work of MonoScene [7].

1. Introduction

Semantic scene completion is an essential task in 3D scene understanding, which aims to jointly infer the holistic semantics and geometry of the 3D scene from partial observation [9, 23, 43, 53]. The task is motivated by the fact that the occupancy patterns and semantic labels of the scene are mutually reinforcing [5, 8, 44, 47, 70]. Researchers gradually discover the important applications of semantic scene completion in indoor scene analysis, robotics, virtual reality, and many downstream tasks in autonomous driving.

Recently, the completion of monocular 3D semantic scenes has become a hot topic of research. Compared with methods relying on 3D sensors like LiDAR or depth sensors and stereo vision using RGB stereo cameras, monocular-based approaches offer advantages such as lower cost, simpler configuration, and greater portability [7, 36, 50, 56, 84]. Additionally, compared with stereo and surround-view setups, a monocular configuration offers greater adaptability to both indoor and outdoor environments, as indoor scenarios particularly emphasize portability. A monocular setup facilitates seamless cross-scenario generalization, making it a more versatile choice for a wide range of applications.

Most existing Monocular Semantic Scene Completion (MSSC) models [7, 45] attempt to simultaneously complete the coupled 3D segmentation and completion tasks in a single stage. However, this framework presents significant challenges, particularly in complex scenes, where performance tends to be poor. When performing scene completion, it is necessary not only to restore the geometric structure but also to ensure the consistency of semantic labels. At the same time, during scene segmentation, the accuracy and completeness of the completion results directly affect the quality of the segmentation [7, 12, 13, 60, 68, 93]. Moreover, most existing methods [45] rely on estimated depth maps to provide geometric information. The noise introduced by inaccurate depth estimations further intensifies the complexity and challenges associated with the problem.

In contrast to adopting a one-stage framework, it is natural to decompose MSSC into several stages. Therefore, we propose a new two-stage framework, named MonoMRN,

*Corresponding authors.

that decomposes MSSC into initial estimation and a novel Masked Recurrent Network for iterative refinement. The GRU (Gated Recurrent Unit) is also used to update partial features based on the previous stage, aiming to reduce task complexity and improve performance.

Due to the standard GRU processing of all positions of the input feature, it results in significant computational overhead. We propose the MS-GRU (Masked Sparse Gated Recurrent Unit), which enhances feature sparsity by applying a dynamic mask that focuses on the occupancy region and designs a sparse version of the GRU to leverage this sparsity for acceleration. Moreover, motivated by the fact that completion and segmentation tasks mutually reinforce each other, we designed a dynamic mask module, which updates the mask at each iteration stage. Additionally, to efficiently update the MS-GRU in each iteration, we propose the distance attention projection, which effectively reduces projection error by reasonably distributing feature weights.

The overall framework is shown in Fig. 2. Our method consists of two stages: Coarse MSSC and Masked Recurrent Network. For the initial estimation stage, the single RGB input is fed into the encoder to extract 2D features. The 2D features are projected to 3D space according to the estimated depth. Subsequently, the projected features are fed into the 3D network to obtain a coarse MSSC prediction. For the recurrent refinement stage, the MS-GRU module takes the distance attention projection x_0, initial estimation feature h_0, and initial mask m_0 as input to recurrently update the semantic scene completion results. The mask is updated in each iteration by the mask updating module.

The key contributions can be summarized as follows:

- We propose a novel semantic scene completion framework, MonoMRN, that decomposes semantic scene completion to coarse MSSC and Masked Recurrent Network for iterative refinement.
- We design a spectrum of useful components in our framework. The MS-GRU is proposed to focus on the occupied region to enhance performance, while also accelerating the process by increasing sparsity. We propose a mask update mechanism to update the mask during each iteration. We propose the distance attention projection to reduce the projection error.
- Extensive experiments validate the effectiveness of our proposed approach. We achieve state-of-the-art performance on the indoor NYUv2 and outdoor SemanticKITTI datasets. We further validate the robustness against various disturbances and reveal the role of the Masked Recurrent Network in the network optimization process.

2. Related Work

3D from a Single Image. The exploration of estimating object or scene geometry from a single-view RGB image has been a longstanding area of study. Early researchers focus on single/multi 3D objects reconstruction by explicit representation [1, 6, 18, 22, 28, 37, 55, 61, 64] or implicit representation [17, 41, 42, 54, 57, 58, 65, 67, 71, 77, 85]. Recently, some researchers have focused on scene-level 3D reconstruction from a single image. The pioneer work [20] unifies 3D reconstruction, 3D semantic segmentation, and 3D instance segmentation into a task known as panoptic 3D scene reconstruction [28]. BUOL [19] proposes a bottom-up framework with occupancy-aware lifting to address instance-channel ambiguity and voxel-reconstruction ambiguity. Recently, there has been a surge of interest in monocular semantic scene completion. The seminal work MonoScene [7] conducts monocular semantic scene completion by line of sight projection and proposes a 3D context relation prior to enforce spatial-semantic consistency.

Semantic Scene Completion from 3D. The initial SSC methods [70, 91] rely on depth data and are designed for indoor scenarios. To further enhance the performance of SSC, researchers [48, 48, 48, 72, 78] have explored the integration of complementary RGB-D data, leveraging both geometric depth information and rich color cues from RGB images. In the past few years, outdoor SSC [15, 63, 81, 88] has gained much attention because it can aid in holistic scene understanding. JS3CNet [88] explores and enhances the interaction between LiDAR segmentation [2, 16, 31, 34, 37, 49, 51, 59, 73, 74, 79, 86, 89], and Semantic Scene Completion (SSC) through a point-voxel interaction module. SCPNet [81] redesigns the completion sub-network using dense-to-sparse knowledge distillation.

Semantic Scene Completion from RGB. The RGB-based SSC can be categorized into two types based on application scenarios: 1) unified methods [7, 90] applicable to both indoor and outdoor scenes, and 2) methods specifically designed for outdoor scenes [11, 29, 30, 39, 45, 52, 66, 75, 80, 87, 92]. The first category is unified methods for both indoor and outdoor scenes. MonoScene [7] first proposes a monocular semantic scene completion method, which is a unified method for both indoor and outdoor scenes. It introduces a line of sight projection and a 3D context relation prior to enforcing spatio-semantic consistency. NDCScene [90] identifies several critical issues in monocular semantic scene completion, including the feature ambiguity of size, depth, and pose. The second category focuses on outdoor scenes [3, 23]. VoxFormer [45] proposes a class-agnostic query proposal and class-specific segmentation for SSC. HASSC [75] introduces a hardness-aware approach that dynamically selects hard voxels based on global hardness and incorporates a self-distillation strategy to ensure stable and consistent training.

Existing methods adopt a single-stage framework for MSSC prediction but struggle to perform well in complex scenes. To address the above issue, we propose a novel framework to decompose MSSC into coarse MSSC and

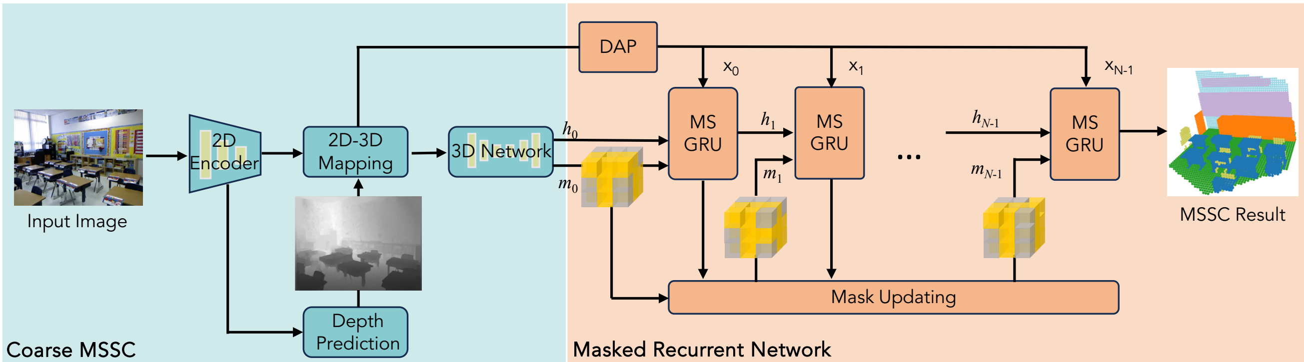

Figure 2. The overall framework of MonoMRN. We first employ the base network to produce the coarse MSSC h_0 and the mask m_0. Subsequently, a series of MS-GRUs operate in a recurrent manner to update MSSC estimation. For each iteration, the mask is updated by the proposed Mask Updating Module. 'DAP' represents the Distance Attention Projection.

Masked Recurrent Network stages.

3. Methodology

In this section, we first provide an overview of the problem formulation and our method (Sec. 3.1). Next, we give a brief introduction to the coarse MSSC stage (Sec. 3.2). Following this, we delve into the Masked Recurrent Network (Sec. 3.3), which is the core of our approach. The masked recurrent network consists of an MS-GRU module (Sec. 3.3.1), a Distance Attention Projection (DAP) mechanism (Sec. 3.3.2), and a Mask Updating Module (Sec. 3.3.3). Lastly, we introduce the loss function for model training (Sec. 3.4).

3.1. Overview

Problem Formulation. Given a single-view RGB image I_{RGB}, which is a partial observation of a 3D scene, monocular semantic scene completion aims to infer volumetric occupancy of the 3D scene, and each voxel is assigned with a semantic label C_i, where $i \in [0, 1, ..., N]$. N is the number of semantic categories and C_0 represents the empty space. Our objective is to train a two-stage network $\{F_1, F_2\}$ with the learnable parameters θ_1 and θ_2. For the first stage, $h_0, m_0 = F_1(I_{\text{RGB}}; \theta_1)$ generate the coarse estimation feature h_0 and initial mask m_0. For the second stage, $\hat{Y} = F_2(h_0, m_0, x_0; \theta_2)$ takes initial estimation feature h_0, initial mask m_0, the feature of distance attention projection x_0 as input and generates the final completion results.

Overall Architecture. As shown in Fig. 2, our method consists of two stages: coarse MSSC and Masked Recurrent Network. The coarse estimation stage takes RGB images as input, extracts 2D features through an encoder, and projects the 2D features into 3D space using the estimated depth values. Then, a 3D network performs the coarse estimation. The Masked Recurrent Network iteratively updates the coarse estimation, which consists of a set of MS-GRU, Mask Updating Module, and Distance Attention Projection. In each iteration, the MS-GRU updates the hidden state h_t by processing the previous hidden state h_{t-1}, the previous mask m_{t-1}, and the distance attention projection feature.

3.2. Coarse MSSC

In this stage, we obtain the coarse MSSC. Given a single RGB input, we extract 2D features and estimate the depth map. The 2D features are projected into 3D space according to the depth map. Subsequently, the 3D features are fed into the 3D network to obtain the coarse MSSC results.

2D Feature Learning. We use the pre-trained ResNet-50 [27] as our backbone to extract RGB features. Then we leverage the off-the-shelf depth prediction method [4] to predict the depth of each image pixel.

2D-3D Projection. To alleviate the gap between 2D and 3D, we apply surface projection [48] to project the 2D features to corresponding 3D positions. Given the estimated depth image I_d, the intrinsic camera matrix $K \in \mathbb{R}^{3\times3}$ and extrinsic camera matrix $[R|t] \in \mathbb{R}^{3\times4}$, we map the 2D features to 3D positions according to equation $p_{u,v} = K[R|t]p_{x,y,z}$.

3D Network for Coarse MSSC. The projected 3D features are fed into the 3D network that is a stack of AIC [40] modules, and we obtain the initial coarse MSSC results. The details of the 3D Coarse MSSC network are provided in the supplementary material.

3.3. Masked Recurrent Network (MRN)

Masked Recurrent Network estimates a sequence of MSSC results $\{\hat{y}^1, ...\hat{y}^N\}$ and masks $\{\hat{m}^1, ...\hat{m}^{N-1}\}$ from the feature of initial coarse MSSC h^0, mask m_0 and the feature of distance attention projection x_0. With each iteration, the information of each MSSC prediction state is selectively memorized or forgotten by the proposed MS-GRU. MS-

GRU is able to concentrate on occupied regions to boost performance by masks and exploit the sparsity of 3D scenes to alleviate the computation cost. We design the mask updating module to adjust the mask in each iteration. The proposed distance attention projection assigns different attention according to the distance to the observed surface, which alleviates the projection error and serves as the input for MS-GRU.

3.3.1. Masked Sparse Gated Recurrent Unit

Standard GRU operates on every position of the input features with dense convolutions, which is computationally inefficient and leads to high memory requirements, especially for processing 3D data. To address this concern, we propose the Masked Sparse Gated Recurrent Unit (MS-GRU), which enhances feature sparsity by applying a dynamic mask that focuses on the occupancy region and designs a sparse version of the GRU to leverage the sparsity for acceleration. Given the input feature x_t of distance attention projection at the current time step, the hidden state h_t is updated as:

$$z_t = \sigma(\mathrm{SubConv}([m_{t-1} \cdot h_{t-1}, m_{t-1} \cdot x_t], W_z)), \quad (1)$$

$$r_t = \sigma(\mathrm{SubConv}([m_{t-1} \cdot h_{t-1}, m_{t-1} \cdot x_t], W_r)), \quad (2)$$

$$h_t^{'} = \tanh(\mathrm{SConv}([r_t \odot h_{t-1}, m_{t-1} \cdot x_t], W_h)), \quad (3)$$

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot h_t^{'}, \quad (4)$$

where m_{t-1} represents the predicted binary mask of previous state, z_t is the update gate, r_t denotes the reset gate, σ indicates the sigmoid activation function, $\tanh$ is the hyperbolic tangent activation function. Moreover, $\mathrm{SubConv}$ and SConv represent the submanifold convolution and sparse convolution, respectively. The proposed MS-GRU differs from the standard GRU in its mask and sparse design, with the specific details as follows.

Mask Design. In contrast to standard GRU, the mask m_t is applied to the input x_t and h_{t-1}, directing the GRU to focus only on the information within the mask. The mask m_t indicates that if the location at (x, y, z) is occupied. It can be defined as:

$$m(x,y,z) = \begin{cases} 1, & \text{if voxel(x, y, z) is occupied ,} \\ 0, & \text{if voxel(x, y, z) is empty .} \end{cases} \quad (5)$$

The mask m_t is updated by the Mask Updating Module (detailed in Sec. 3.3.3) in each iteration and supervised by the Sequential Mask Loss (detailed in Sec. 3.4).

Sparse Design. We leverage the mixture of submanifold sparse convolution [24] and sparse convolution [21] to design the MS-GRU block. Submanifold convolution and sparse convolution operate only on non-zero elements of the input tensor and their neighborhoods. Submanifold convolution preserves the sparsity in the output, while sparse convolutions can increase the sparsity of the output. When

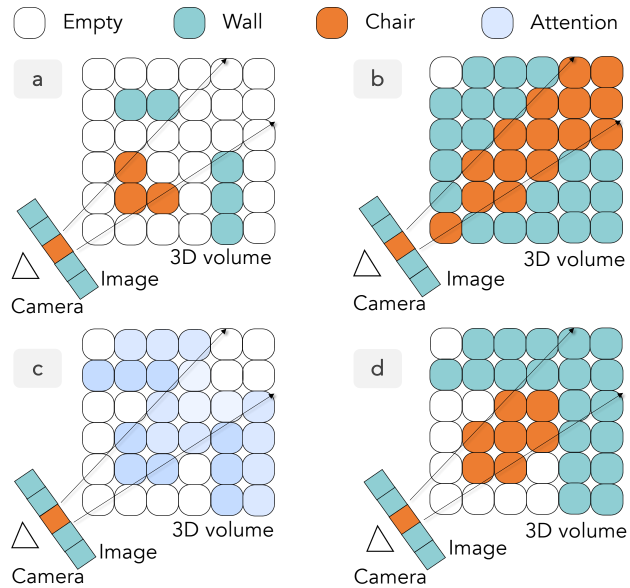

Figure 3. In (a) Surface Projection, voxels behind the surface are not assigned features. In (b) Sight Projection, although occluded regions are assigned features, a large number of incorrect features are also introduced. Distance Attention Projection determines the attention weight (c) based on the distance to the estimated surface and then weights the sight-projected features, allowing feature propagation to occluded regions while reducing the impact of introduced inaccurate features.

calculating the reset gate z_t and update gate r_t, we apply submanifold convolution to ensure sparsity. In updating the hidden state $h_t^{'}$, we use sparse convolution to appropriately update the non-zero positions (occupied regions).

3.3.2. Distance Attention Projection

At each iteration, x_t updates the hidden state h_t. Ideally, x_t should be informationally complete, encapsulating essential data for updating h_t, while maintaining high relevance and minimizing noise to ensure effective state transitions. We initially project the 2D RGB features into 3D space to mitigate the dimensional gap, and subsequently obtain x_t through processing via convolutional layers. However, existing projection methods present several drawbacks as shown in Fig. 3. Surface projection [45, 48] projects the 2D features onto a surface in 3D space based on the estimated depth. Consequently, voxels behind the surface are not assigned features, resulting in insufficient information to update the GRU in the occluded regions. Sight projection [7], on the other hand, maps 2D features to their corresponding line of sight without requiring depth data, allowing for rapid propagation of 2D features to distant occluded regions. However, it also introduces a significant number of incorrect features into these occluded regions.

To overcome these limitations, we introduce the distance attention projection, which assigns soft features to the 3D

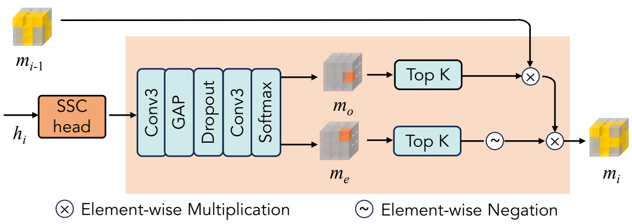

Figure 4. The architecture of Mask Updating Module.

scene. This approach enables rapid feature propagation to occluded regions while allocating soft labels based on the correlation between 2D features and 3D regions. Consequently, it integrates the advantages of both surface and sight projections. The distance attention projection determines the attention weight w_d based on the spatial occlusion relationship and the distance from different 3D locations to the estimated observation surface. The basic idea is that the further a point in space is from the estimated observation surface, the lower its relevance to the 2D image features, and thus the lower the attention weight w_d.

We allocate attention weights based on the line of sight. In a 2D image, each pixel is part of a line of sight. The distribution of attention along the line of sight is described by the following formula:

$$w_d = \begin{cases} 1 / \left(d - d' + 1\right), & d > d' \\ 1, & d = d' \\ 0.5, & \delta < d < d' \end{cases} \tag{6}$$

where d' represents the distance from the camera to the observed surface, $d > d'$ denotes the occluded region in the line-of-sight, $d < d'$ indicates the region before the observed surface, and δ is defined according to the RMS of predicted depth. The distance attention projection feature is obtained by element-wise multiplying the line-of-sight-projection with the attention matrix w_d. Subsequently, the feature of distance attention projection is fed into the AIC [40] module to extract the 3D context feature x_t.

3.3.3. Mask Updating Module

The strong coupling between environmental occupancy patterns and object semantics has been established [7, 70]. We hypothesize that 3D scene occupancy information aids in voxel-wise semantic estimation. To leverage this, we integrate a mask into the standard GRU to emphasize occupied regions and introduce a mask updating module that iteratively enhances occupancy information. The mask m indicates whether a voxel is occupied or empty, supervised by the mask loss (see Sec. 3.4). The ground truth for the mask is derived from the summation of all non-empty classes.

Mask Initialization. The mask m is initialized during the Coarse MSSC stage and iteratively updated at each recurrent step. The initial mask m_0 is determined based on the occupancy score:

$$s_{x,y,z} = 1 - P_{x,y,z}(y = 0 | I_{\text{RGB}}; \theta_1), \tag{7}$$

where θ_1 denotes the Coarse MSSC network, $P(y = 0)$ represents the probability to be empty and $1 - P(y = 0)$ indicates the probability to be occupied. In practice, a predefined threshold t is used: if $s_{x,y,z} \geq t$, then $m_0(x, y, z)$ is set to 1.

Mask Updating Module. We propose the mask updating module to sequentially update the mask m as shown in Fig. 4. The mask update module takes the output of the MSSC head as input, instead of the hidden state h_t from the MS-GRU, in order to reduce the interference between the mask prediction and MSSC prediction tasks. The features are successively fed into a 3×3 convolution, GAP (Global Average Pooling) operation, dropout layer, 3×3 convolution layer, and softmax layer to obtain the probability of occupancy. The global average pooling operation is utilized to downsample the feature map two times. We use dropout to prevent overfitting by randomly setting 0.1 of the input units to zero during training. After the softmax function, m_o indicates the probability of occupancy for each voxel, and m_e indicates the probability of emptiness for each voxel. Subsequently, we select the top K voxels in m_o and m_e. We add the top K voxels in m_o to m_{i-1} and delete the top K voxels in m_{i-1} for updating.

3.4. Objective & Optimization

Sequential MSSC Loss. We first apply a sequential cross-entropy loss to both the initial coarse estimation $\hat{y}^0$ and each output state of recurrent refinement $\{\hat{y}^1, \hat{y}^2, ..., \hat{y}^N\}$, formulated as follows:

$$L_{\text{MSSC}} = \sum_{i=0}^{N} \gamma^i L_{\text{ce}}(y^{\text{gt}}, \hat{y}^i), \tag{8}$$

where L_{ce} indicates the cross-entropy loss and $\gamma = 0.8$ in our experiments.

Sequential Mask Loss. The predicted masks from previous steps, *i.e.*, $\{\hat{m}^1, \hat{m}^2, ..., \hat{m}^N\}$, are supervised using a sequential binary weighted cross-entropy loss, defined as:

$$L_{\text{mask}} = \sum_{i=1}^{N-1} \gamma^{N-i} L_{\text{wbce}}(m_{\text{gt}}, m_i), \tag{9}$$

where L_{wbce} represents the binary weighted cross entropy loss and $\gamma = 0.6$ in our experiments. Note that the ground truth for the mask is obtained by summing over all non-empty classes.

Overall, the total loss function can be formulated as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{MSSC}} + \mathcal{L}_{\text{mask}} + \mathcal{L}_{\text{scal}}, \tag{10}$$

where L_{scal} denotes the Scene-Class Affinity Loss proposed by MonoScene [7].

Table 1. **Comparisons among state-of-the-art MSSC methods** on the NYUv2 dataset. The symbol † denotes the results presented by [7]. The Best and 2nd Best scores from each metric are highlighted in Orange and Teal, respectively.

Method	Venue	SC IoU	Ceiling	Floor	Wall	window	chair	bed	sofa	table	tvs	furniture	objects	SSC mIoU	FPS
LMSCNet† [63]	3DV'20	33.93	4.49	88.41	4.63	0.25	3.94	32.03	15.44	6.57	0.02	14.51	4.39	15.88	-
AICNet† [40]	CVPR'20	30.03	7.58	82.97	9.15	0.05	6.93	35.87	22.92	11.11	0.71	15.90	6.45	18.15	3.68
3DSketch† [14]	CVPR'20	38.64	8.53	90.45	9.94	5.67	10.64	42.29	29.21	13.88	9.38	23.83	8.19	22.91	3.12
MonoScene [7]	CVPR'22	42.51	8.89	93.50	12.06	12.57	13.72	48.19	36.11	15.13	15.22	27.96	12.94	26.94	1.96
NDC-Scene [90]	ICCV'23	44.17	12.02	93.51	13.11	13.77	15.83	49.57	39.87	17.17	24.57	31.00	14.96	29.03	-
MonoMRN	**Ours**	53.16	26.80	92.02	19.39	18.50	17.66	44.60	31.02	19.60	17.22	32.90	18.31	30.73	2.56

Table 2. **Comparisons among state-of-the-art SSC methods** on the test set of SemanticKITTI [3]. We include stereo (**S**), monocular (**M**) and temporal (**T**) methods for fair comparisons. '**I**', '**O**', and '**I/O**' denote methods that applied to indoor, outdoor, and both scenes. Symbol † denotes the results presented in [7]. The Best and 2nd Best scores from each metric are highlighted in Orange and Teal, respectively.

Method	Input	App. Scens.	SC IoU	Car	Bicycle	Motorcycle	Truck	Other-Vehicle	Person	Bicyclist	Motorcyclist	Road	Parking	Sidewalk	Other-Ground	Building	Fence	Vegetation	Trunk	Terrain	Pole	Traffic Sign	SSC mIoU
VoxFormer-S [45]	S	O	44.0	25.8	0.6	0.5	5.6	3.8	1.8	3.3	0.0	54.8	15.5	26.4	0.7	17.7	7.6	24.4	5.1	30.0	7.1	4.2	12.4
HASSC-S [75]	S	O	44.8	27.2	0.9	0.9	9.9	5.6	2.8	4.7	0.0	57.1	15.9	28.3	1.0	19.1	6.6	25.5	6.2	33.0	7.7	4.1	13.5
H2GFormer-S [80]	S	O	44.2	23.4	0.8	0.9	4.8	4.1	1.2	2.5	0.1	56.4	26.5	28.6	4.9	22.8	13.3	24.6	9.1	23.8	6.4	6.3	14.6
MonoOcc-S [93]	M&T	O	–	23.2	2.2	1.5	5.2	5.4	1.7	2.0	0.2	55.2	25.1	27.8	9.7	21.4	13.4	24.0	8.7	23.0	5.8	6.4	13.8
HTCL-M [38]	S&T	O	44.2	27.2	1.8	2.2	5.7	5.4	1.1	3.1	0.9	64.4	33.8	34.8	12.4	25.9	21.1	25.3	10.8	31.2	9.0	8.3	17.1
LMSCNet† [63]	M	O	28.6	18.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	40.7	4.4	18.2	0.0	10.3	1.2	13.7	0.0	20.5	0.0	0.0	6.7
3DSketch† [14]	M	I	33.3	18.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	41.3	0.0	21.6	0.0	14.8	0.7	19.1	0.0	26.4	0.0	0.0	7.5
AICNet† [40]	M	I	29.6	14.7	0.0	0.0	2.9	0.0	0.0	0.0	0.0	43.6	12.0	20.6	0.1	12.9	2.5	15.4	2.9	28.7	0.1	0.0	8.3
MonoScene [7]	M	I/O	34.2	18.8	0.5	0.7	3.3	4.4	1.0	1.4	0.4	54.7	24.8	27.1	5.7	14.4	11.1	14.9	2.4	19.5	3.3	2.1	11.1
VoxFormer-S [45]	M	O	38.7	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	10.7
TPVFormer [29]	M	O	34.3	19.2	1.0	0.5	3.7	2.3	1.1	2.4	0.3	55.1	27.4	27.2	6.5	14.8	11.0	12.9	3.7	20.4	2.9	1.5	11.3
OccFormer [92]	M	O	34.5	21.6	1.5	1.7	1.2	3.2	2.2	1.1	0.2	55.9	31.5	33.3	6.5	15.7	11.9	16.8	3.9	21.3	3.8	3.7	12.3
NDC-Scene [90]	M	I/O	37.2	26.3	1.7	2.4	14.8	7.7	3.6	2.7	0.0	59.2	21.4	28.2	1.7	14.9	6.7	19.1	3.5	31.0	4.5	2.7	12.7
MonoMRN	M	I/O	42.0	23.6	0.8	0.7	2.6	4.4	2.4	1.8	0.4	62.1	22.5	28.7	5.6	21.3	16.0	22.4	5.1	26.0	7.9	7.1	13.8

4. Experiments

We conduct an evaluation of our method using widely recognized real-world MSSC datasets, specifically the indoor NYUv2 [69] and the outdoor SemanticKITTI [3]. We compare our method with state-of-the-art MSSC methods. Then, we provide a detailed analysis of our evaluation results and the ablation studies. The qualitative performance of SemanticKITTI and more ablation studies are shown in the supplementary material.

4.1. Experimental Setup

Datasets. The NYUv2 [69] dataset includes $1,449$ depth images captured by the Kinect from a diverse set of indoor scenes, partitioned into 795 for training and 654 for testing. Following SSCNet [70], we use the 3D labels annotated by [62] and the mapping object categories based on [25]. SemanticKITTI [3] is a large-scale autonomous driving dataset, which consists of 22 point cloud sequences. Sequences 00 to 10, 08, and 11 to 21 are used as training, validation, and testing, respectively. The LiDAR scans are represented as $256 \times 256 \times 32$ voxel grids of length 0.2 meters. We use the RGB image of camera-2 and crop the image into size 1220×370.

Evaluation Metrics. We follow the SSCNet [70] protocol to ensure fair evaluations. For SSC, we use the intersection over union (IoU) between the predicted voxel labels and ground-truth labels for all object classes. The overall performance is evaluated by computing the mean intersection over union (mIoU) over all classes. To evaluate the scene completion performance, all voxels in the scene are categorized as empty or occupied. We compute IoU, precision, and recall for scene completion.

Implementation Details. We use the PyTorch framework to implement our approach. Our method adopts the SGD optimizer. We begin with an initial learning rate of 0.1 and adjust it using the polynomial learning rate policy. We train the model for 200 epochs on NYUv2 and for 30 epochs on the SemanticKITTI dataset. The output 3D feature maps for NYU and SemanticKITTI are $60 \times 36 \times 60$ and $128 \times 128 \times 16$, respectively. For SemanticKITTI, the generated features are upsampled to $256 \times 256 \times 32$ to match the resolution of the GT for evaluation. MS-GRU is iterated 2 times. The mask initialization threshold is set as 0.6. We set the mask

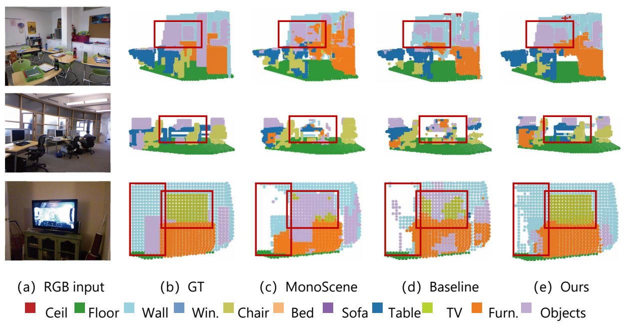

Figure 5. **Qualitative comparisons** of semantic scene completion results with different methods on the NYUv2 dataset. From left to right: (a) Single RGB input; (b) Ground truth; (c) MonoScene [7]; (d) Our baseline method; and (e) Our approach.

updating parameter K to 5.

4.2. Comparisons with State of the Arts

We compare our method with state-of-the-art methods, including those designed specifically for indoor scenes [14, 40], outdoor scenes [45, 63], and those applicable to both indoor and outdoor scenes [7, 90]. For outdoor MSSC, we also list the methods that use stereo images as input [45, 75, 80]. Stereo-based methods can provide more accurate geometric information. Taking VoxFormer [45] as an example, it can achieve significant improvement when using stereo input as shown in Tab. 2.

Quantitative Comparisons. Tab. 1 and Table 2 report the performance of the proposed MonoMRN and other comparison methods on the indoor NYUv2 dataset (test set) and outdoor SemanticKITTI dataset (hidden test set). As shown in Tab. 1, we significantly outperform the state-of-the-art methods on the NYUv2 dataset, achieving an 8.99% IoU increase in SC and a 1.60% mIoU improvement in SSC. As shown in Table 2, our method outperforms state-of-the-art methods by 3.35% in SC and by 1.07% in SSC on the SemanticKITTI dataset. Note that MonoScene and NDC-Scene adopt different 2D-to-3D projection strategies without using the estimated depth, which leads to significantly lower performance on the SC IoU compared with our methods. We additionally include comparison methods that utilize stereo and/or temporal images as input. While these methods leverage extra information to enhance performance, our approach still achieves competitive results without relying on such auxiliary data.

Qualitative Assessments. Fig. 5 presents a comparison of the qualitative results on the NYUv2 dataset between our method, our baseline method, and MonoScene. We can observe that our method achieves the best qualitative results by restoring more precise details. By the proposed decomposed SSC framework, our method achieves noticeable improvement. As NDC-Scene [90] does not release the complete code, we can not compare our visualization results with it. Additionally, we visualized the outputs of MonoMRN at different stages. Fig. 6 indicates that the Masked Recurrent Network can progressively refine the coarse MSSC results.

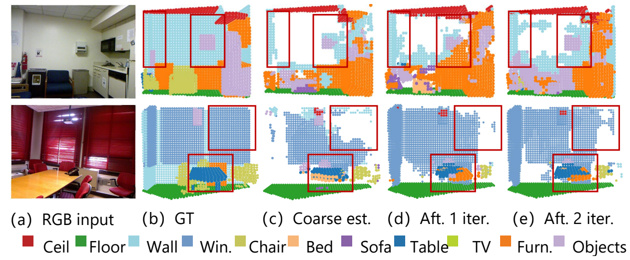

Figure 6. **Visual outputs** of MonoMRN at different stages. From left to right: (a) Single RGB input; (b) Ground truth; (c) Coarse estimation; (d) After one iteration; and (e) After two iterations.

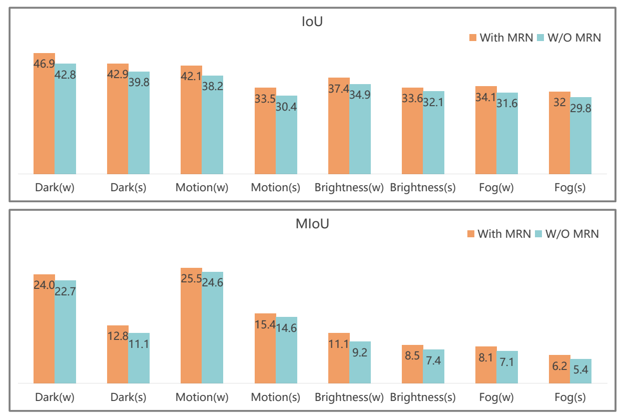

Figure 7. The robustness analysis with and w/o MRN Network. For indoor scenarios, we analyze robustness under dark and motion conditions, while for outdoor scenarios, we examine performance under brightness and fog conditions.

4.3. Robustness Analysis

In real-world scenarios, environmental disturbances are unavoidable. For instance, indoor MSSC systems often suffer from low lighting and motion blur, whereas outdoor MSSC systems are more vulnerable to adverse conditions such as fog and intense lighting. A model with high adaptability to such disturbances is crucial for practical applications [10, 32, 33, 35, 76, 83]. Therefore, in this section, we evaluate the improvement in model robustness achieved by the decoupled framework. Enhanced robustness not only enables the model to better handle complex environments but also reflects its improved capacity for information recovery [26, 33, 46, 82], highlighting the role of the Masked Recurrent Network in the optimization process.

We simulate four potential out-of-distribution scenarios during inference: darkness and motion blur for indoor envi-

Table 3. Ablation study on the core components in MonoMRN.

Method	SC-IoU (%) ↑	SSC-mIoU (%) ↑
Baseline	48.23	27.47
+ MS-GRU	50.61 (+2.63)	29.16 (+1.69)
+ Distance Attention Projection	51.86 (+1.25)	30.11 (+0.95)
+ Masking Updating	**53.16** (+1.30)	**30.73** (+0.62)

Table 4. Ablation study on the Distance Attention Projection.

Projection Method	SC-IoU (%) ↑	SSC-mIoU (%) ↑
Surface Projection	51.23	29.36
Sight Projection	52.56	30.12
Distance Attention Projection	**53.16**	**30.73**

Table 5. Ablation study on different design choices of MS-GRU.

Design Choice	SC-IoU (%)	SSC-mIoU (%)	Params (M)	MACs (G)
GRU	50.48	29.67	1.33	171.99
MS-GRU	**53.16**	**30.73**	1.33	**52.44**
Submanifold Conv	52.66	30.12	1.33	**50.06**
Sparse Conv	**53.16**	**30.73**	1.33	52.44
Untied Weights	52.85	30.34	2.66	**104.88**
Tied Weights	**53.16**	**30.73**	1.33	**104.88**
1× Iteration	53.01	30.11	1.33	**52.44**
2× Iteration	53.16	30.73	1.33	104.88
3× Iteration	**53.51**	**30.86**	1.33	157.32
4× Iteration	53.51	30.17	1.33	209.76

ronments, and brightness and fog for outdoor environments. Each out-of-distribution scenario includes both weak (w) and strong (s) perturbation levels, providing a comprehensive evaluation of the model's resilience to varying environmental disturbances. As shown in Fig. 7, the model exhibits enhanced robustness after processing through the Masked Recurrent Network under four types of disturbances, demonstrating stronger information recovery and completion capabilities.

4.4. Ablation Study

We present ablation results on the NYUv2 dataset to validate the effectiveness of our method from several perspectives: core components, Distance Attention Projection, the design choice of MS-GRU, and the design choice of mask updating. Due to space limits, more details and visualizations are placed in the Appendix.

The Effectiveness of Core Components. We evaluate the effectiveness of the core components of our method, including MS-GRU, Distance Attention Projection, and Mask Updating, as shown in Tab. 3. The vanilla baseline is the Coarse MSSC stage. The recurrent refinement by the proposed MS-GRU brings an improvement by 1.69% mIoU. The proposed Distance Attention Projection boosts the performance by 0.95% mIoU. The proposed mask updating mechanism enhances the performance by 0.62% mIoU.

The Effectiveness of Distance Attention Projection. To validate the effectiveness of the proposed Distance Attention Projection, we replace it with Surface Projection and Sight Projection. As shown in Tab. 4, our method achieves the best performance among all comparisons. Compared with the other two methods, our approach efficiently assigns 2D features to occluded regions while reducing the impact of inaccurate assignments.

Different Design Choices of MS-GRU. We evaluate the effectiveness of different design choices of MS-GRU as shown in Table 5. We experiment with replacing the proposed MS-GRU with a standard GRU. We achieve better performance by using MS-GRU. We design a mixture of submanifold sparse convolution and sparse convolution in MS-GRU. Here, we test a version that only uses the submanifold sparse convolution. We achieve an improvement through the mixture design of sparse convolution. We ablate a version of each MS-GRU to learn a separate set of weights. We can observe that when weights are tied, the performance is better and the parameter count is significantly lower. We conducted ablation experiments with different numbers of iterations. It can be observed that the performance increases with more iterations at first, then gets stable around 2 and 3. Due to the marginal improvement in the third iteration, we opt for two iterations in our method. Further increasing the number of iterations will degrade the performance.

4.5. Limitations & Discussions

The predicted depth is far from the demand for semantic scene completion. For indoor scene completion, the depth estimation method [4] obtains 0.364 m RMS on NYUv2, while the voxel size of our method is 0.08 m. This indicates that the depth estimation algorithm still falls short of the requirements for MSSC and has a significant impact on the results. In the future, we will explore adding depth correction to our framework to further boost the performance.

5. Conclusion

In this paper, we proposed a novel decomposed framework that conducts Coarse MSSC followed by Masked Recurrent Networks. To better leverage the sparsity of 3D scenes and focus on the occupied 3D regions in each iteration, we introduced MS-GRU and the mask updating mechanism. Additionally, we proposed the Distance Attention Projection to mitigate projection errors. As a result, our approach achieves the state-of-the-art performance on the NYUv2 and SemanticKITTI datasets. We expect that the MonoMRN framework could inspire future research directions and drive more advancements in 3D perception tasks.

Acknowledgments

This work was partially supported by the National Natural Science Foundation of China under Grant Nos. 62071330, 61831022, and U21B2020.

The authors would like to sincerely thank the Program Chairs, Area Chairs, and Reviewers for the time and effort devoted during the review process.

References

[1] Panos Achlioptas, Olga Diamanti, Ioannis Mitliagkas, and Leonidas Guibas. Learning representations and generative models for 3d point clouds. In *International Conference on Machine Learning*, pages 40–49. PMLR, 2018.

[2] Angelika Ando, Spyros Gidaris, Andrei Bursuc, Gilles Puy, Alexandre Boulch, and Renaud Marlet. Rangevit: Towards vision transformers for 3d semantic segmentation in autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5240–5250, 2023.

[3] Jens Behley, Martin Garbade, Andres Milioto, Jan Quenzel, Sven Behnke, Cyrill Stachniss, and Jurgen Gall. Semantickitti: A dataset for semantic scene understanding of lidar sequences. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9297–9307, 2019.

[4] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. Adabins: Depth estimation using adaptive bins. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4009–4018, 2021.

[5] Hengwei Bian, Lingdong Kong, Haozhe Xie, Liang Pan, Yu Qiao, and Ziwei Liu. Dynamiccity: Large-scale 4d occupancy generation from dynamic scenes. In *International Conference on Learning Representations*, 2025.

[6] Alexandre Boulch, Corentin Sautier, Björn Michele, Gilles Puy, and Renaud Marlet. Also: Automotive lidar self-supervision by occupancy estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13455–13465, 2023.

[7] Anh-Quan Cao and Raoul De Charette. Monoscene: Monocular 3d semantic scene completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3991–4001, 2022.

[8] Anh-Quan Cao and Raoul De Charette. Scenerf: Self-supervised monocular 3d scene reconstruction with radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9387–9398, 2023.

[9] Anh-Quan Cao, Gilles Puy, Alexandre Boulch, and Renaud Marlet. Pcam: Product of cross-attention matrices for rigid registration of point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13229–13238, 2021.

[10] Anh-Quan Cao, Angela Dai, and Raoul de Charette. Pasco: Urban 3d panoptic scene completion with uncertainty awareness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14554–14564, 2024.

[11] Haoming Chen, Zhizhong Zhang, Yanyun Qu, Ruixin Zhang, Xin Tan, and Yuan Xie. Building a strong pre-training baseline for universal 3d large-scale perception. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19925–19935, 2024.

[12] Runnan Chen, Youquan Liu, Lingdong Kong, Nenglun Chen, Xinge Zhu, Yuexin Ma, Tongliang Liu, and Wenping Wang. Towards label-free scene understanding by vision foundation models. In *Advances in Neural Information Processing Systems*, pages 75896–75910, 2023.

[13] Runnan Chen, Youquan Liu, Lingdong Kong, Xinge Zhu, Yuexin Ma, Yikang Li, Yuenan Hou, Yu Qiao, and Wenping Wang. Clip2scene: Towards label-efficient 3d scene understanding by clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7020–7030, 2023.

[14] Xiaokang Chen, Kwan-Yee Lin, Chen Qian, Gang Zeng, and Hongsheng Li. 3d sketch-aware semantic scene completion via semi-supervised structure prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4193–4202, 2020.

[15] Ran Cheng, Christopher Agia, Yuan Ren, Xinhai Li, and Liu Bingbing. S3cnet: A sparse semantic scene completion network for lidar point cloud. In *Conference on Robot Learning*, pages 2148–2161, 2021.

[16] Ran Cheng, Ryan Razani, Ehsan Taghavi, Enxu Li, and Bingbing Liu. Af2-s3net: Attentive feature fusion with adaptive feature selection for sparse semantic segmentation network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12547–12556, 2021.

[17] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3075–3084, 2019.

[18] Christopher B Choy, Danfei Xu, JunYoung Gwak, Kevin Chen, and Silvio Savarese. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In *European Conference on Computer Vision*, pages 628–644. Springer, 2016.

[19] Tao Chu, Pan Zhang, Qiong Liu, and Jiaqi Wang. Buol: A bottom-up framework with occupancy-aware lifting for panoptic 3d scene reconstruction from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4937–4946, 2023.

[20] Manuel Dahnert, Ji Hou, Matthias Nießner, and Angela Dai. Panoptic 3d scene reconstruction from a single rgb image. *Advances in Neural Information Processing Systems*, 34: 8282–8293, 2021.

[21] Martin Engelcke, Dushyant Rao, Dominic Zeng Wang, Chi Hay Tong, and Ingmar Posner. Vote3deep: Fast object detection in 3d point clouds using efficient convolutional neural networks. In *IEEE International Conference on Robotics and Automation*, pages 1355–1361, 2017.

[22] Haoqiang Fan, Hao Su, and Leonidas J Guibas. A point set generation network for 3d object reconstruction from a single image. In *Proceedings of the IEEE/CVF Conference on*

Computer Vision and Pattern Recognition, pages 605–613, 2017.

[23] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3354–3361, 2012.

[24] Benjamin Graham and Laurens Van der Maaten. Submanifold sparse convolutional networks. *arXiv preprint arXiv:1706.01307*, 2017.

[25] Ankur Handa, Viorica Patraucean, Vijay Badrinarayanan, Simon Stent, and Roberto Cipolla. Scenenet: Understanding real world indoor scenes with synthetic data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4077–4085, 2016.

[26] Xiaoshuai Hao, Mengchuan Wei, Yifan Yang, Haimei Zhao, Hui Zhang, Yi Zhou, Qiang Wang, Weiming Li, Lingdong Kong, and Jing Zhang. Is your hd map constructor reliable under sensor corruptions? In *Advances in Neural Information Processing Systems*, pages 22441–22482, 2024.

[27] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[28] Fangzhou Hong, Lingdong Kong, Hui Zhou, Xinge Zhu, Hongsheng Li, and Ziwei Liu. Unified 3d and 4d panoptic segmentation via dynamic shifting networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(5): 3480–3495, 2024.

[29] Yuanhui Huang, Wenzhao Zheng, Yunpeng Zhang, Jie Zhou, and Jiwen Lu. Tri-perspective view for vision-based 3d semantic occupancy prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9223–9232, 2023.

[30] Haoyi Jiang, Tianheng Cheng, Naiyu Gao, Haoyang Zhang, Tianwei Lin, Wenyu Liu, and Xinggang Wang. Symphonize 3d semantic scene completion with contextual instance queries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20258–20267, 2024.

[31] Lingdong Kong, Youquan Liu, Runnan Chen, Yuexin Ma, Xinge Zhu, Yikang Li, Yuenan Hou, Yu Qiao, and Ziwei Liu. Rethinking range view representation for lidar segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 228–240, 2023.

[32] Lingdong Kong, Youquan Liu, Xin Li, Runnan Chen, Wenwei Zhang, Jiawei Ren, Liang Pan, Kai Chen, and Ziwei Liu. Benchmarking 3d perception robustness to common corruptions and sensor failure. In *International Conference on Learning Representations 2023 Workshop on Scene Representations for Autonomous Driving*, 2023.

[33] Lingdong Kong, Youquan Liu, Xin Li, Runnan Chen, Wenwei Zhang, Jiawei Ren, Liang Pan, Kai Chen, and Ziwei Liu. Robo3d: Towards robust and reliable 3d perception against corruptions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19994–20006, 2023.

[34] Lingdong Kong, Jiawei Ren, Liang Pan, and Ziwei Liu. Lasermix for semi-supervised lidar semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21705–21715, 2023.

[35] Lingdong Kong, Shaoyuan Xie, Hanjiang Hu, Lai Xing Ng, Benoit R. Cottereau, and Wei Tsang Ooi. Robodepth: Robust out-of-distribution depth estimation under corruptions. In *Advances in Neural Information Processing Systems*, pages 21298–21342, 2023.

[36] Lingdong Kong, Xiang Xu, Youquan Liu, Jun Cen, Runnan Chen, Wenwei Zhang, Liang Pan, Kai Chen, and Ziwei Liu. Largead: Large-scale cross-sensor data pretraining for autonomous driving. *arXiv preprint arXiv:2501.04005*, 2025.

[37] Lingdong Kong, Xiang Xu, Jiawei Ren, Wenwei Zhang, Liang Pan, Kai Chen, Wei Tsang Ooi, and Ziwei Liu. Multimodal data-efficient 3d scene understanding for autonomous driving. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(5):3748–3765, 2025.

[38] Bohan Li, Jiajun Deng, Wenyao Zhang, Zhujin Liang, Dalong Du, Xin Jin, and Wenjun Zeng. Hierarchical temporal context learning for camera-based semantic scene completion. In *European Conference on Computer Vision*, pages 131–148. Springer, 2024.

[39] Bohan Li, Yasheng Sun, Zhujin Liang, Dalong Du, Zhuanghui Zhang, Xiaofeng Wang, Yunnan Wang, Xin Jin, and Wenjun Zeng. Brgscene: Bridging stereo geometry and bev representation with reliable mutual interaction for semantic scene completion. In *Proceedings of the International Joint Conference on Artificial Intelligence*, 2024.

[40] Jie Li, Kai Han, Peng Wang, Yu Liu, and Xia Yuan. Anisotropic convolutional networks for 3d semantic scene completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3351–3359, 2020.

[41] Li Li, Hubert PH Shum, and Toby P Breckon. Less is more: Reducing task and model complexity for 3d point cloud semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9361–9371, 2023.

[42] Li Li, Hubert PH Shum, and Toby P. Breckon. Rapid-seg: Range-aware pointwise distance distribution networks for 3d lidar segmentation. In *European Conference on Computer Vision*, pages 222–241. Springer, 2024.

[43] Rong Li, Anh-Quan Cao, and Raoul de Charette. Coarse3d: Class-prototypes for contrastive learning in weakly-supervised 3d point cloud segmentation. *arXiv preprint arXiv:2210.01784*, 2022.

[44] Rong Li, Shijie Li, Lingdong Kong, Xulei Yang, and Junwei Liang. Seeground: See and ground for zero-shot open-vocabulary 3d visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3707–3717, 2025.

[45] Yiming Li, Zhiding Yu, Christopher Choy, Chaowei Xiao, Jose M Alvarez, Sanja Fidler, Chen Feng, and Anima Anandkumar. Voxformer: Sparse voxel transformer for camera-based 3d semantic scene completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9087–9098, 2023.

[46] Ye Li, Lingdong Kong, Hanjiang Hu, Xiaohao Xu, and Xiaonan Huang. Is your lidar placement optimized for 3d scene

understanding? In *Advances in Neural Information Processing Systems*, pages 34980–35017, 2024.

[47] Xuewu Lin, Tianwei Lin, Zixiang Pei, Lichao Huang, and Zhizhong Su. Sparse4d: Multi-view 3d object detection with sparse spatial-temporal fusion. *arXiv preprint arXiv:2211.10581*, 2022.

[48] Shice Liu, Yu Hu, Yiming Zeng, Qiankun Tang, Beibei Jin, Yinhe Han, and Xiaowei Li. See and think: Disentangling semantic scene completion. In *Proceedings of the Advances in Neural Information Processing Systems*, pages 261–272, 2018.

[49] Youquan Liu, Runnan Chen, Xin Li, Lingdong Kong, Yuchen Yang, Zhaoyang Xia, Yeqi Bai, Xinge Zhu, Yuexin Ma, Yikang Li, et al. Uniseg: A unified multi-modal lidar segmentation network and the openpcseg codebase. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21662–21673, 2023.

[50] Youquan Liu, Lingdong Kong, Jun Cen, Runnan Chen, Wenwei Zhang, Liang Pan, Kai Chen, and Ziwei Liu. Segment any point cloud sequences by distilling vision foundation models. In *Advances in Neural Information Processing Systems*, pages 37193–37229, 2023.

[51] Youquan Liu, Lingdong Kong, Xiaoyang Wu, Runnan Chen, Xin Li, Liang Pan, Ziwei Liu, and Yuexin Ma. Multi-space alignments towards universal lidar segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14648–14661, 2024.

[52] Anas Mahmoud, Jordan SK Hu, Tianshu Kuai, Ali Harakeh, Liam Paull, and Steven L Waslander. Self-supervised image-to-point distillation via semantically tolerant contrastive loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7102–7110, 2023.

[53] Tetiana Martyniuk, Gilles Puy, Alexandre Boulch, Renaud Marlet, and Raoul de Charette. Lidpm: Rethinking point diffusion for lidar scene completion. *arXiv preprint arXiv:2504.17791*, 2025.

[54] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4460–4470, 2019.

[55] Björn Michele, Alexandre Boulch, Tuan-Hung Vu, Gilles Puy, Renaud Marlet, and Nicolas Courty. Train till you drop: Towards stable and robust source-free unsupervised 3d domain adaptation. In *European Conference on Computer Vision*, pages 1–19. Springer, 2024.

[56] Lucas Nunes, Louis Wiesmann, Rodrigo Marcuzzi, Xieyuanli Chen, Jens Behley, and Cyrill Stachniss. Temporal consistent 3d lidar representation learning for semantic perception in autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5217–5228, 2023.

[57] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 165–174, 2019.

[58] Songyou Peng, Michael Niemeyer, Lars Mescheder, Marc Pollefeys, and Andreas Geiger. Convolutional occupancy networks. In *European Conference on Computer Vision*, pages 523–540. Springer, 2020.

[59] Xidong Peng, Runnan Chen, Feng Qiao, Lingdong Kong, Youquan Liu, Yujing Sun, Tai Wang, Xinge Zhu, and Yuexin Ma. Learning to adapt sam for segmenting cross-domain point clouds. In *European Conference on Computer Vision*, pages 54–71. Springer, 2024.

[60] Gilles Puy, Alexandre Boulch, and Renaud Marlet. Using a waffle iron for automotive point cloud semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3379–3389, 2023.

[61] Gilles Puy, Spyros Gidaris, Alexandre Boulch, Oriane Siméoni, Corentin Sautier, Patrick Pérez, Andrei Bursuc, and Renaud Marlet. Three pillars improving vision foundation model distillation for lidar. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21519–21529, 2024.

[62] Jason Rock, Tanmay Gupta, Justin Thorsen, JunYoung Gwak, Daeyun Shin, and Derek Hoiem. Completing 3d object shape from one depth image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2484–2493, 2015.

[63] Luis Roldao et al. Lmscnet: Lightweight multiscale 3d semantic completion. In *International Conference on 3D Vision*, pages 111–119, 2020.

[64] Nermin Samet, Oriane Siméoni, Gilles Puy, Georgy Ponimatkin, Renaud Marlet, and Vincent Lepetit. You never get a second chance to make a good first impression: Seeding active learning for 3d semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18445–18457, 2023.

[65] Nermin Samet, Gilles Puy, Oriane Siméoni, and Renaud Marlet. Milan: Milli-annotations for lidar semantic segmentation. *arXiv preprint arXiv:2407.15797*, 2024.

[66] Corentin Sautier, Gilles Puy, Spyros Gidaris, Alexandre Boulch, Andrei Bursuc, and Renaud Marlet. Image-to-lidar self-supervised distillation for autonomous driving data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9891–9901, 2022.

[67] Corentin Sautier, Gilles Puy, Alexandre Boulch, Renaud Marlet, and Vincent Lepetit. Bevcontrast: Self-supervision in bev space for automotive lidar point clouds. In *International Conference on 3D Vision*, pages 559–568, 2024.

[68] Corentin Sautier, Gilles Puy, Alexandre Boulch, Renaud Marlet, and Vincent Lepetit. Clustering is back: Reaching state-of-the-art lidar instance segmentation without training. *arXiv preprint arXiv:2503.13203*, 2025.

[69] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *European Conference on Computer Vision*, pages 746–760. Springer, 2012.

[70] Shuran Song, Fisher Yu, Andy Zeng, Angel X. Chang, Manolis Savva, and Thomas Funkhouser. Semantic scene completion from a single depth image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1746–1754, 2017.

[71] Haotian Tang, Zhijian Liu, Shengyu Zhao, Yujun Lin, Ji Lin, Hanrui Wang, and Song Han. Searching efficient 3d architectures with sparse point-voxel convolution. In *European Conference on Computer Vision*, pages 685–702. Springer, 2020.

[72] Jiaxiang Tang, Xiaokang Chen, Jingbo Wang, and Gang Zeng. Semantic scene completion via integrating instances and scene in-the-loop. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 324–333, 2021.

[73] Song Wang, Jianke Zhu, and Ruixiang Zhang. Meta-rangeseg: Lidar sequence semantic segmentation using multiple feature aggregation. *IEEE Robotics and Automation Letters*, 7(4):9739–9746, 2022.

[74] Song Wang, Wentong Li, Wenyu Liu, Xiaolu Liu, and Jianke Zhu. Lidar2map: In defense of lidar-based semantic map construction using online camera distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5186–5195, 2023.

[75] Song Wang, Yu Jiawei, Li Wentong, Liu Wenyu, Liu Xiaolu, Chen Junbo, and Zhu Jianke. Not all voxels are equal: Hardness-aware semantic scene completion with self-distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14792–14801, 2024.

[76] Song Wang, Zhongdao Wang, Jiawei Yu, Wentong Li, Bailan Feng, Junbo Chen, and Jianke Zhu. Reliocc: Towards reliable semantic occupancy prediction via uncertainty learning. In *arXiv:2409.18026*, 2024.

[77] Song Wang, Xiaolu Liu, Lingdong Kong, Jianyun Xu, Chunyong Hu, Gongfan Fang, Wentong Li, Jianke Zhu, and Xinchao Wang. Pointlora: Low-rank adaptation with token selection for point cloud learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6605–6615, 2025.

[78] Xuzhi Wang, Di Lin, and Liang Wan. Ffnet: Frequency fusion network for semantic scene completion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2550–2557, 2022.

[79] Xuzhi Wang, Wei Feng, Lingdong Kong, and Liang Wan. Nuc-net: Non-uniform cylindrical partition network for efficient lidar semantic segmentation. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.

[80] Yu Wang and Chao Tong. H2gformer: Horizontal-to-global voxel transformer for 3d semantic scene completion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5722–5730, 2024.

[81] Zhaoyang Xia, Youquan Liu, Xin Li, Xinge Xu, Yuexin Ma, Yikang Li, Yuenan Hou, and Yu Qiao. Scpnet: Semantic scene completion on point cloud. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17642–17651, 2023.

[82] Shaoyuan Xie, Lingdong Kong, Yuhao Dong, Chonghao Sima, Wenwei Zhang, Qi Alfred Chen, Ziwei Liu, and Liang Pan. Are vlms ready for autonomous driving? an empirical study from the reliability, data, and metric perspectives. *arXiv preprint arXiv:2501.04003*, 2025.

[83] Shaoyuan Xie, Lingdong Kong, Wenwei Zhang, Jiawei Ren, Liang Pan, Kai Chen, and Ziwei Liu. Benchmarking and improving bird's eye view perception robustness in autonomous driving. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(5):3878–3894, 2025.

[84] Xiang Xu, Lingdong Kong, Hui Shuai, Wenwei Zhang, Liang Pan, Kai Chen, Ziwei Liu, and Qingshan Liu. 4d contrastive superflows are dense 3d representation learners. In *European Conference on Computer Vision*, pages 58–80. Springer, 2024.

[85] Xiang Xu, Lingdong Kong, Hui Shuai, and Qingshan Liu. Frnet: Frustum-range networks for scalable lidar segmentation. *IEEE Transactions on Image Processing*, 34:2173–2186, 2025.

[86] Xiang Xu, Lingdong Kong, Hui Shuai, Liang Pan, Ziwei Liu, and Qingshan Liu. Limoe: Mixture of lidar representation learners from automotive scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27368–27379, 2025.

[87] Yujie Xue, Ruihui Li, Fan Wu, Zhuo Tang, Kenli Li, and Mingxing Duan. Bi-ssc: Geometric-semantic bidirectional fusion for camera-based 3d semantic scene completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20124–20134, 2024.

[88] Xu Yan, Jiantao Gao, Jie Li, Ruimao Zhang, Zhen Li, Rui Huang, and Shuguang Cui. Sparse single sweep lidar point cloud segmentation via learning contextual shape priors from scene completion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3101–3109, 2021.

[89] Xu Yan, Jiantao Gao, Chaoda Zheng, Chao Zheng, Ruimao Zhang, Shuguang Cui, and Zhen Li. 2dpass: 2d priors assisted semantic segmentation on lidar point clouds. In *European Conference on Computer Vision*, pages 677–695. Springer, 2022.

[90] Jiawei Yao, Chuming Li, Keqiang Sun, Yingjie Cai, Hao Li, Wanli Ouyang, and Hongsheng Li. Ndc-scene: Boost monocular 3d semantic scene completion in normalized device coordinates space. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9421–9431, 2023.

[91] Jiahui Zhang, Hao Zhao, Anbang Yao, Yurong Chen, Li Zhang, and Hongen Liao. Efficient semantic scene completion network with spatial group convolution. In *European Conference on Computer Vision*, pages 733–749. Springer, 2018.

[92] Yunpeng Zhang, Zheng Zhu, and Dalong Du. Occformer: Dual-path transformer for vision-based 3d semantic occupancy prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9433–9443, 2023.

[93] Yupeng Zheng, Xiang Li, Pengfei Li, Yuhang Zheng, Bu Jin, Chengliang Zhong, Xiaoxiao Long, Hao Zhao, and Qichao Zhang. Monoocc: Digging into monocular semantic occupancy prediction. In *IEEE International Conference on Robotics and Automation*, pages 18398–18405, 2024.