

Global-Aware Monocular Semantic Scene Completion with State Space Models

Shijie Li $\diamond$ Zhongyao Cheng $\diamond$ Rong Li $\square$ Shuai Li $\triangle$
 Juergen Gall $\triangle \clubsuit$ Xun Xu $\diamond$ Xulei Yang $\diamond, \boxtimes$
 $\diamond$ I²R, A*STAR $\square$ HKUST(GZ) $\triangle$ University of Bonn $\clubsuit$ Lamarr Institute*
 $\boxtimes$ Corresponding author

Abstract

Monocular Semantic Scene Completion (MonoSSC) reconstructs and interprets 3D environments from a single image, enabling diverse real-world applications. However, existing methods are often constrained by the local receptive field of Convolutional Neural Networks (CNNs), making it challenging to handle the non-uniform distribution of projected points and effectively reconstruct missing information caused by the 3D-to-2D projection. In this work, we introduce GA-MonoSSC, a hybrid architecture for MonoSSC that effectively captures global context in both the 2D image domain and 3D space. Specifically, we propose a Dual-Head Multi-Modality Encoder, which leverages a Transformer architecture to capture spatial relationships across all features in the 2D image domain, enabling more comprehensive 2D feature extraction. Additionally, we introduce the Frustum Mamba Decoder, built on the State Space Model (SSM), to efficiently capture long-range dependencies in 3D space. Furthermore, we propose a frustum re-ordering strategy within the Frustum Mamba Decoder to mitigate feature discontinuities in the reordered voxel sequence, ensuring better alignment with the scan mechanism of the State Space Model (SSM) for improved 3D representation learning. We conduct extensive experiments on the widely used Occ-ScanNet and NYUv2 datasets, demonstrating that our proposed method achieves state-of-the-art performance, validating its effectiveness. The code is available [here](#).

1. Introduction

Semantic Scene Completion (SSC) plays a crucial role in various 3D Vision related applications, such as robotic navigation [1–4], augmented reality [5–7], and autonomous driving [8–10]. Its goal is to complete incomplete 3D scenes, typically represented as a voxel grid, while simul-

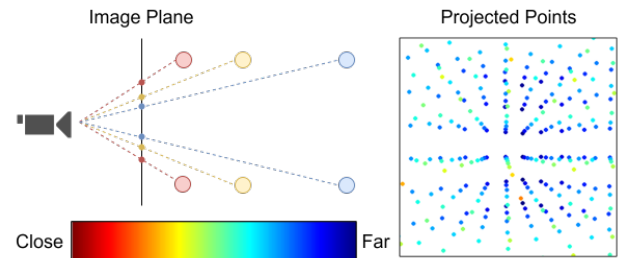


Figure 1. Perspective projection results in an imbalanced distribution of projected points, where nearby points are more widely scattered while distant points are more densely packed. The real projection map is shown on the right.

taneously detecting the relevant semantic information.

This work focuses on Monocular Semantic Scene Completion (MonoSSC) specifically tailored for indoor environments. Many recent SSC studies [11–13] focus on autonomous driving scenarios [14–16], aiming to complete incomplete 3D scenes using multi-view images or 3D LiDAR point clouds as input. However, such rich sensory data is often unavailable in indoor environments where usually only a monocular camera is available. Thus MonoSSC is proposed, which can infer 3D geometry and semantic information from a single image. Compared to SSC in autonomous driving scenarios, indoor MonoSSC is typically more challenging due to less informative sensory data and increased complexity of indoor environments with intricate structures.

The capability of current MonoSSC solutions is often constrained by the inherent limitations of CNN architectures. These methods typically start by extracting 2D features using 2D Convolutional Neural Networks (2D CNNs). Each 3D voxel is then projected onto the 2D feature map to retrieve the corresponding features. Subsequently, 3D Convolutional Neural Networks (3D CNNs) refine these 3D voxel features and generate the final prediction. As shown in Fig. 1, the projected 3D voxel centers are unevenly distributed on the 2D image plane due to perspective projection. Specifically, the projections of 3D points closer to the camera are more widely scattered, while those of farther points are more densely packed. CNN architectures

*Lamarr Institute is Lamarr Institute for Machine Learning and Artificial Intelligence.

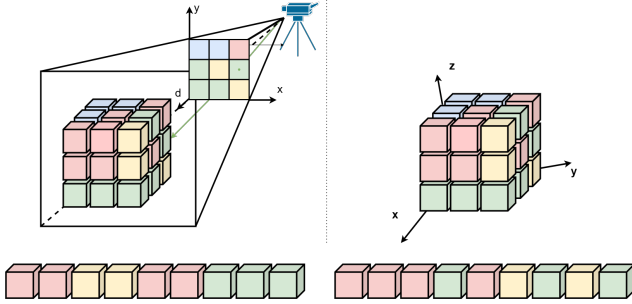


Figure 2. Arranging voxel features into a sequence based on their voxel 3D coordinates introduces discontinuities (right). This issue can be mitigated by sorting voxels within the frustum space, where voxel features are distributed across the image plane and exhibit strong similarity along the depth dimension (left).

adopt a fixed-weight encoding strategy. The resulting feature map usually cannot capture this variance or adapt to the non-uniform distribution, leading to ineffective 2D feature representation and suboptimal spatial reasoning in the MonoSSC task. Furthermore, the 3D-to-2D projection inherently results in information loss along the depth dimension. To maximally recover the lost information, the model needs to capture global dependencies in 3D space. However, traditional CNN architectures are restricted by their local receptive fields, making it challenging to effectively capture long-range spatial relationships in 3D space.

To address these challenges, we propose a MonoSSC method named **GA-MonoSSC**, which possesses global awareness in both 2D image space and 3D space. Specifically, we propose a Dual-Head Multi-Modality Encoder (DM_{enc}), which leverages a Transformer architecture to capture spatial relationships across 2D features. This enables interactions among imbalanced projected points, resulting in a more informative and robust representation. Additionally, multi-modality information, including semantic and geometric features, is learned separately, allowing the model to capture more nuanced representations compared to previous methods that process them jointly.

The 3D model is also expected to exhibit global awareness to effectively reconstruct missing information in 3D space. However, Transformer architectures are impractical for this task due to their quadratic computational complexity, which becomes prohibitive given the typically large input sequence length in 3D space. Instead, we adopt the Selective State Space Model (SSM) [17], specifically Mamba, which efficiently models long-range dependencies with linear complexity. Mamba typically arranges voxel features as sequential data based on their 3D spatial coordinates and processes them accordingly. However, we observe that this approach introduces discontinuities in the reorganized sequential data due to the 2D-to-3D unprojection process, as illustrated in Fig. 2. To address this issue, we propose the Frustum Mamba Decoder (FM_{Dec}), which incorporates a

Frustum Mamba Layer specifically designed to reorder voxels in frustum space. This ensures a more structured and continuous feature flow, better aligning with the underlying data distribution.

Our main contributions can be summarized as follows.

- We propose a hybrid framework, GA-MonoSSC, that integrates Transformer and Mamba architectures to achieve global awareness in both the 2D image domain and 3D space. The proposed method achieves state-of-the-art performance on the Occ-ScanNet and NYUv2 datasets.
- We design a Dual-Head Multi-Modality Encoder that employs a Transformer to capture spatial relationships across 2D features, mitigating the impact of non-uniform distribution of projected points. Additionally, semantic and geometric features are learned separately, enabling model to capture finer details and enhance representation quality.
- We integrate Mamba [17] to capture long-range dependencies in 3D space, enabling maximal recovery of lost information. We also introduce Frustum Mamba Layer, which scans voxels in frustum space, addressing the feature discontinuities caused by 2D-to-3D unprojection.

2. Related Works

2.1. Monocular Semantic Scene Completion

Monocular Semantic Scene Completion (SSC) aims to reconstruct a complete 3D semantic representation from a single 2D image. Some related works attempt to achieve similar effects through Novel View Synthesis [18] or by generating depth maps [19, 20], though these approaches remain within the 2D domain. The first 3D monocular approach was introduced by MonoScene [21], advancing beyond earlier SSC methods [9, 22–34] by relying solely on 2D visual input, thus removing the need for additional 3D data sources. A key limitation in MonoScene [21] is the ambiguity in the unprojection operation, which maps 2D features into 3D space but struggles with depth accuracy. To address this, recent approaches have focused on refining depth information within the unprojection process. NDC-Scene [35] introduced a Depth-Adaptive Dual Decoder that upscales and merges 2D and 3D feature maps, enhancing spatial accuracy. Similarly, ISO [36] proposes directly predicting depth values and incorporating them into the unprojection operation, leading to improved performance.

2.2. State Space Models

State Space Models (SSMs), including Mamba [17], S4 [37], and S4nd [38], have gained considerable attention due to their efficiency in handling long-range dependencies and large-scale data with linear time complexity. Mamba, in particular, integrates selective mechanisms that enhance this capability, making it especially notable. These innovations have led to successful adaptations in computer vision

[39–42]. VMamba [43] incorporates a cross-scan module and Vision Mamba [43] utilizes a bidirectional SSM architecture, both optimized for visual data processing. The architecture has also been extended to process 3D data, such as point clouds, where approaches like [44] use a 3D Hilbert curve to rearrange input patches, and [45] employs an octree-based ordering to efficiently capture spatial relationships.

3. Methodology

In this section, we present a detailed description of **GA-MonoSSC**. We introduce State Space Models and Mamba Module as essential preliminaries in Sec. 3.1. Next, we provide an overview of the overall architecture in Sec 3.2 and delve into specifics of main components in Sec. 3.3 and Sec. 3.4. The employed supervision is shown in Sec. 3.5.

3.1. Preliminaries

State Space Models. To set the stage for the Mamba module, we first revisit the concept of state space models (SSMs). Rooted in control theory, SSMs operate by mapping inputs to outputs via an intermediary hidden state, making them well-suited for processing sequential data. For discrete inputs, SSMs can be formally represented as follows:

$$h_k = \bar{A}h_{k-1} + \bar{B}x_k. \quad (1)$$

$$y_k = \bar{C}h_k. \quad (2)$$

where k is the sequence number, $\bar{A}$, $\bar{B}$, $\bar{C}$ are matrices that represent the discretized parameters of the model, which involve the sampling step Δ . The x_k , y_k , and h_k representing the input, output, and hidden state of the system, respectively. Structured State Space Sequence models (S4) [37] introduce a novel parameterization for SSMs, enabling more efficient computation.

Mamba Module. Mamba introduces an adaptation to S4 models [37], making the matrices $\bar{B}$ and $\bar{C}$, as well as the sampling size Δ , dependent on the input, with this dependency arising from incorporating the sequence length and batch size of the input. This enables dynamic adjustment of the matrices for each input token, allowing $\bar{B}$ and $\bar{C}$ to dynamically adjust state transitions based on the input, thereby enhancing the model’s content awareness.

3.2. Overview

GA-MonoSSC takes a monocular image $\mathcal{X}^{2d} \in \mathcal{R}^{C \times H^{2d} \times W^{2d}}$ as input and outputs a 3D semantic occupancy map $\mathcal{X}^{3d} \in \mathcal{R}^{N_c \times H^{3d} \times W^{3d} \times L^{3d}}$, where N_c is the number of total semantic classes:

$$\mathcal{X}^{3d} = \mathbf{GA-MonoSSC}(\mathcal{X}^{2d}). \quad (3)$$

The input monocular image is first processed by a Dual-head Multi-modality encoder ($\mathbf{DM}_{\text{enc}}$), shown in Fig. 3, which separately extracts global-aware multi-scale semantic and geometric features. These features are then unprojected into 3D voxel space using Feature Line of Sight Projection (**FLoSP**) [21]. In 3D space, multi-scale geometry and semantic features are first fused as a unified representation, which is then processed by the proposed Frustum Mamba Decoder ($\mathbf{FM}_{\text{dec}}$), shown in Fig. 4, to capture long-range dependencies effectively. The 3D predictions will be finally decoded through a prediction head.

3.3. Dual-head Multi-modality Encoder ($\mathbf{DM}_{\text{enc}}$)

We first introduce the Dual-Head Multi-Modality Encoder ($\mathbf{DM}_{\text{enc}}$), as shown in Fig. 3, which utilizes a Transformer architecture to extract multi-scale geometric and semantic features while effectively modeling spatial relationships among 2D features. In previous solutions, CNN architectures have been widely used for 2D feature extraction. However, their fixed-weight encoding strategy and limited local receptive field hinder their ability to capture the non-uniform distribution of projected points. This limitation results in ineffective feature representation and sub-optimal spatial reasoning for the MonoSSC task, as illustrated in Fig. 1. To address this challenge, the model requires the ability to capture long-range dependencies, making the Transformer architecture an ideal choice.

$\mathbf{DM}_{\text{enc}}$ consists of a transformer-based encoder, **Enc**, which comprises N_l transformer encoder blocks and outputs multi-scale 2D tokens. These tokens are then fed into two modality-specific decoders, **Dec_{geo}** and **Dec_{sem}**, to decode geometric and semantic information. **Enc** receives a single-view image $\mathcal{X}^{2d}$ and outputs multi-scale tokens:

$$\mathcal{TK} = \mathbf{Enc}(\mathcal{X}^{2d}). \quad (4)$$

$\mathcal{TK} = \{TK_l \in \mathbb{R}^{N_{tk} \times C} \mid l \in (1, N_l)\}$ represents the set of tokens obtained after each transformer encoder block l . N_{tk} is the number of tokens. To recover spatial information, the tokens TK_l at each scale are reorganized into a spatial feature map:

$$\mathcal{F}_l = \mathbf{Reorganize}(\mathcal{TK}_l). \quad (5)$$

The output feature maps $\mathcal{F} = \{\mathcal{F}_l \in \mathbb{R}^{C \times h \times w} \mid l \in (1, N_l)\}$ where h and w are spatial shape after reorganization, will be fed into the decoder **Dec_{geo}** to decode multi-scale geometric features:

$$\mathcal{F}^{geo} = \mathbf{Dec}_{geo}(\mathcal{F}). \quad (6)$$

where $\mathcal{F}^{geo} = \{\mathcal{F}_l^{geo} \in \mathbb{R}^{C \times H_l^{2d} \times W_l^{2d}} \mid l \in (1, N_l)\}$. Specially, **Dec_{geo}** contains several **UpBlock**, each can be formulated as:

$$\mathcal{F}_l^{geo} = \mathbf{Up2d}(\mathcal{F}_l). \quad (7)$$

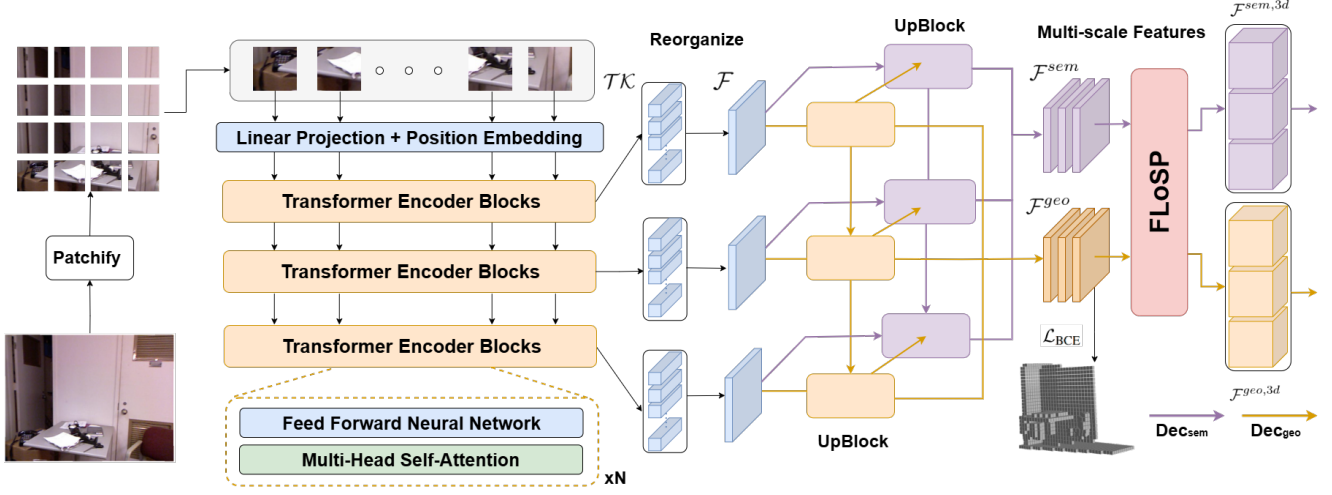


Figure 3. The overview of the proposed Dual-head Multi-modality Encoder (DM_{Enc}) which will extract multi-scale geometry and semantic features separately.

$$\hat{\mathcal{F}}_{l-1}^{\text{geo}} = \text{Up}_{2d}(\text{Conv}_{2d}(\mathcal{F}_{l-1}^{\text{geo}})). \quad (8)$$

$$\mathcal{F}_l^{\text{geo}} = \text{Conv}_{2d}(\text{BN}(\text{ReLU}(\mathcal{F}_l^{\text{geo}} \parallel \hat{\mathcal{F}}_{l-1}^{\text{geo}}))). \quad (9)$$

where Up_{2d} denotes upsample operation, Conv_{2d} , BN and ReLU means convolution layer, batch normalization, and Rectified Linear Unit separately. $\parallel$ denotes the concatenation operation.

The multi-scale semantic features $\mathcal{F}^{\text{sem}} = \{\mathcal{F}_l^{\text{sem}} \in \mathbb{R}^{C \times H_l^{2d} \times W_l^{2d}} | l \in (1, N_l)\}$ are derived similarly with decoder Dec_{sem} which share the same architecture to Dec_{geo} . One thing needs to be noticed is that the geometric features $\mathcal{F}^{\text{geo}}$ will be fed in to introduce more structure information:

$$\mathcal{F}^{\text{sem}} = \text{Dec}_{\text{sem}}(\mathcal{F}^{2d}, \mathcal{F}^{\text{geo}}). \quad (10)$$

Specially, at each scale l , the geometry feature map will be fused with an elementwise addition:

$$\mathcal{F}_l^{\text{sem}} = \mathcal{F}_l^{\text{sem}} + \mathcal{F}_l^{\text{geo}}. \quad (11)$$

Due to the ambiguity introduced by information loss in the 3D-to-2D projection, we enforce 3D supervision on the output of Dec_{geo} to ensure that the extracted geometric features retain awareness of the 3D data distribution. Specifically, we achieve this by predicting depth from Dec_{geo} , which is subsequently converted into a 3D occupancy representation, where the supervision is applied.

3.4. Frustum Mamba Decoder (FM_{Dec})

The extracted multi-scale, geometry and semantic features, $\mathcal{F}^{\text{geo}}$ and $\mathcal{F}^{\text{sem}}$, are unprojected into 3D voxel space using Feature Line of Sight Projection (FLoSP) [21]:

$$\mathcal{F}^{\text{geo},3d} = \text{FLoSP}(\mathcal{F}^{\text{geo}}). \quad (12)$$

$$\mathcal{F}^{\text{sem},3d} = \text{FLoSP}(\mathcal{F}^{\text{sem}}). \quad (13)$$

$$\mathcal{F}^{\text{sem},3d} = \{\mathcal{F}_l^{\text{sem},3d} \in \mathbb{R}^{N_l \times C \times H^{3d} \times W^{3d} \times L^{3d}} | l \in N_l\},$$

$$\mathcal{F}^{\text{geo},3d} = \{\mathcal{F}_l^{\text{geo},3d} \in \mathbb{R}^{N_l \times C \times H^{3d} \times W^{3d} \times L^{3d}} | l \in N_l\}.$$

The resulting 3D features are then processed by the proposed Frustum Mamba Decoder (FM_{Dec}), which effectively captures long-range dependencies and refines the final 3D representation by leveraging structured spatial relationships within the frustum space. In FM_{Dec} , $\mathcal{F}^{\text{sem},3d}$ and $\mathcal{F}^{\text{geo},3d}$ are first fused into a unified 3D feature volume $\mathcal{F}^{3d}$. This volume is further refined by the Frustum Mamba Layer, which enhances spatial reasoning and improves the quality of the 3D representation by modeling long-range dependencies more effectively.

Multi-scale information fusion. To achieve effective multi-scale feature fusion, we incorporate a channel attention mechanism that adaptively reweights features at different scales, enhancing their contribution based on contextual relevance. For $\mathcal{F}^{\text{geo},3d}$, we first apply global average pooling (AvgPool) over spatial dimensions to obtain a compact representation $\mathcal{F}_{\text{avg}}^{\text{geo},3d} \in \mathbb{R}^{N_l \times C}$, which is then mapped into weights $W^{\text{geo}} \in \mathbb{R}^{N_l}$ with a Multilayer perception (MLP):

$$\mathcal{F}_{\text{avg}}^{\text{geo},3d} = \text{AvgPool}(\mathcal{F}^{\text{geo},3d}). \quad (14)$$

$$W^{\text{geo}} = \text{MLP}(\mathcal{F}_{\text{avg}}^{\text{geo},3d}). \quad (15)$$

Finally, the softmax function is applied, and the output weights are used to compute a weighted average of multi-scale 3D feature volumes:

$$\mathcal{F}^{\text{geo},3d} = \sum_{n=1}^{N_l} \mathcal{F}_l^{\text{geo},3d} \cdot w_l, \quad w_l \in W^{\text{geo}}. \quad (16)$$

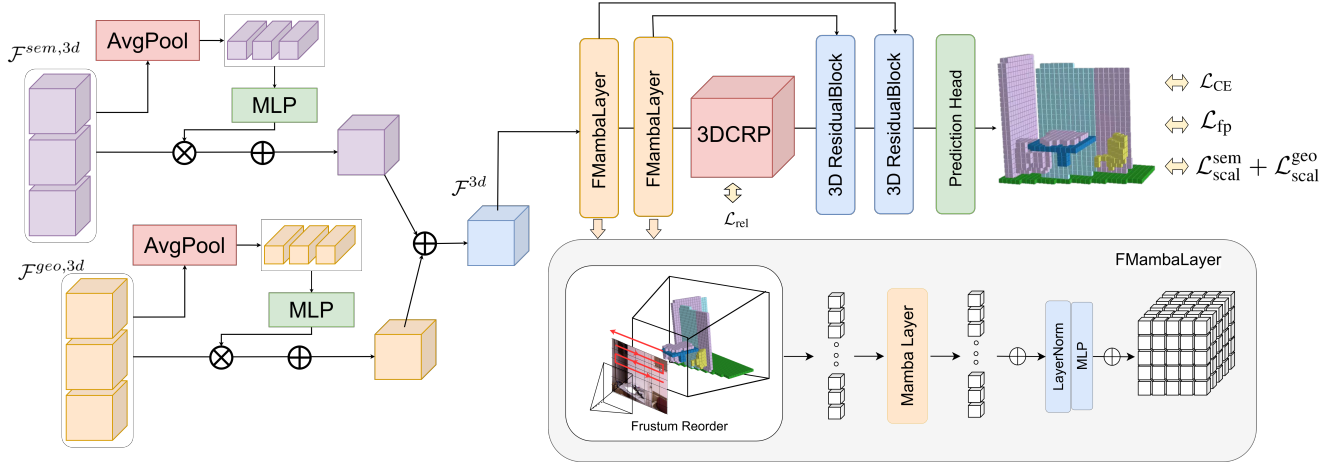


Figure 4. Architecture of the proposed Frustum Mamba Decoder $\mathbf{FM}_{\text{dec}}$. $\mathbf{FM}_{\text{dec}}$ adopts an encoder-decoder architecture, with the proposed Frustum Mamba Layer integrated into each stage of the encoder to capture long-range dependency. The Frustum Reordering Strategy is introduced to arrange voxels into a sequence within frustum space rather than the original 3D space, effectively mitigating feature discontinuities and improving sequential consistency.

$\mathcal{F}^{\text{sem},3d}$ will be gotten in a similar way.

Multi-modality information fusion. The resulting two feature volumes corresponding to semantic and geometry information will be fused by elementwise addition:

$$\mathcal{F}^{3d} = \mathcal{F}^{\text{sem},3d} + \mathcal{F}^{\text{geo},3d}. \quad (17)$$

To mitigate the impact of 3D-to-2D projection ambiguity, which leads to information loss along the depth direction, the 3D model should possess strong global modeling capabilities to effectively capture long-range dependencies, enabling it to seek a globally optimal solution. Unfortunately, the transformer architecture is not suitable for this scenario, as its computational complexity scales quadratically with sequence length. This becomes impractical for 3D applications, where the large number of voxels leads to prohibitive memory and computational costs. Instead, we introduce the State Space Model (SSM) as an alternative solution to capture long-range dependencies in 3D space with linear scalability.

Unlike transformers, Mamba requires reorganizing voxel data into a sequence based on their 3D coordinates before processing them sequentially. As shown in Fig. 2, this strategy introduces discontinuities due to the 2D-to-3D unprojection, which may lead to a performance drop. To mitigate this issue, we propose a Frustum Reordering Strategy that reorganizes voxel data in frustum space. This is motivated by the fact that 3D voxel features are unprojected from 2D image features along rays, meaning they tend to be more structured along the image plane while exhibiting higher similarity along the depth direction.

$\mathbf{FM}_{\text{Dec}}$ adopts a 3D U-Net architecture following an Encoder-Decoder design. The proposed Frustum Mamba Layer, which combines the Frustum Reordering Strategy with a standard Mamba block, is applied in the encoder to

enhance long-range spatial modeling. To ensure each voxel retains awareness of its 3D spatial position, a 3D convolution block is applied at the beginning to inject inductive bias. Specifically, the Frustum Mamba Layer can be formulated as:

$$\mathcal{F}\mathcal{F}^{3d} = \text{Flatten}(\mathcal{F}^{3d}) \quad (18)$$

Flatten is the flatten operation and converts the input 3D feature volume $\mathcal{F}^{3d} \in \mathbb{R}^{C \times H^{3d} \times W^{3d} \times L^{3d}}$ into a voxel list $\mathcal{F}\mathcal{F}^{3d} \in \mathbb{R}^{C \times (H^{3d} \times W^{3d} \times L^{3d})}$:

$$\mathcal{S}\mathcal{F}\mathcal{F}^{3d} = \mathbf{FReorder}(\mathcal{F}\mathcal{F}_i^{3d}). \quad (19)$$

FReorder denotes the proposed Frustum Reordering Strategy, which rearranges voxels in the frustum coordinate system, as illustrated in Fig. 2, and the voxel sequence output from it can be represented as $\mathcal{S}\mathcal{F}\mathcal{F}^{3d}$, which will be processed by the Mamba layer:

$$\mathcal{S}\mathcal{F}\mathcal{F}'^{3d} = \text{MambaLayers}(\mathcal{S}\mathcal{F}\mathcal{F}^{3d}). \quad (20)$$

$$\mathcal{F}'^{3d} = \text{Composit}(\mathcal{S}\mathcal{F}\mathcal{F}'^{3d}). \quad (21)$$

MambaLayers denotes several standard Mamba layers, and **Composit** reconstructs the 3D feature volume $\mathcal{F}'^{3d}$ from the voxel list $\mathcal{S}\mathcal{F}\mathcal{F}'^{3d}$. Finally, the layer normalization LN and skip link will be added:

$$\mathcal{F}^{3d} = \mathbf{LN}(\mathcal{F}'^{3d}) + \mathcal{F}^{3d}. \quad (22)$$

After each Frustum Mamba Layer, a downsampling operation will be applied to enable $\mathbf{FM}_{\text{Dec}}$ encoder to extract more representative multi-scale 3D features.

The multi-scale 3D feature volumes from $\mathbf{FM}_{\text{Dec}}$'s encoder are fed into $\mathbf{FM}_{\text{Dec}}$'s decoder to restore the original

Dataset	Ori-Res	Res	Classes	train / test
NYUv2	240×144×240	60×36×60	13	795 / 654
OSnet	6×60×36	60×60×36	12	45755 / 19764
OSnet-mini	60×60×36	60×60×36	12	4639 / 2007

Table 1. Dataset Summary. OSnet denotes Occ-ScanNet dataset.

volume size. Similar to the encoder, the decoder consists of multiple stages, each containing an upsampling block, 3D residual block and connected to its corresponding encoder block via skip links. To further enhance contextual information modeling, we incorporate the 3D Context Relation Prior (3D CRP) layer from [21] between the encoder and decoder. For additional implementation details, please refer to the supplemental material. Finally, the output feature volume is processed by a prediction head, which generates the final predictions.

3.5. Supervision

The predicted occupancy from $\mathbf{DM}_{\text{enc}}$ is supervised using the ground truth occupancy with the Binary Cross-Entropy (BCE) loss function $\mathcal{L}_{\text{BCE}}$. Additionally, following previous work, we apply cross-entropy loss $\mathcal{L}_{\text{CE}}$ for the final output, along with auxiliary supervision, including Scene-Class Affinity Loss $\mathcal{L}_{\text{scal}}^{\text{sem}}$, $\mathcal{L}_{\text{scal}}^{\text{geo}}$, Frustum Proportion Loss $\mathcal{L}_{\text{fp}}$ and Context Relation Loss $\mathcal{L}_{\text{rel}}$:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{BCE}} + \mathcal{L}_{\text{rel}} + \mathcal{L}_{\text{scal}}^{\text{sem}} + \mathcal{L}_{\text{scal}}^{\text{geo}} + \mathcal{L}_{\text{fp}}. \quad (23)$$

More details can be found in the supplemental materials.

4. Experiments

Datasets. We evaluate the proposed method on three widely used datasets for indoor monocular semantic scene completion: NYUv2 [48], Occ-ScanNet [36], and Occ-ScanNet-mini [36]. Detailed statistics for each dataset are presented in Tab. 1. Notably, Occ-ScanNet-mini is a subset of Occ-ScanNet, and both adopt the original resolution for training and evaluation. For comparison, although a higher resolution is provided ($240 \times 144 \times 240$) for the NYUv2 dataset, the model is trained on a downsampled scene ($60 \times 36 \times 60$) to optimize computational efficiency while preserving essential scene details.

Evaluation Metrics. Consistent with standard practices, we evaluate performance on the Scene Completion (SC) task using Intersection over Union (IoU) for occupied voxels, irrespective of semantic class. For the Semantic Scene Completion (SSC) task, we report the mean IoU (mIoU) across all semantic classes.

4.1. Comparison to state-of-the-art

This section compares the proposed **GA-MonoSSC** with previous methods on the NYUv2 and Occ-ScanNet datasets. The experimental results are presented in Tab. 2 and Tab. 3, respectively. From Tab. 2, we observe that

the proposed MambaSSC achieves state-of-the-art performance in both IoU and mIoU, demonstrating its superior capability in recovering geometric structures while accurately predicting relevant semantic information. Notably, methods explicitly involving 3D data inputs (such as occupancy grids, depth maps, or Truncated Signed Distance Fields (TSDF)) demonstrate comparatively weaker performance. We attribute this limitation to the reduced capacity of their architectures, which may lack the robustness needed to capture the complex spatial relationships in 3D data effectively. We observe a trend of improved overall performance for methods that utilize monocular images as input, likely due to the adoption of more advanced architectures. Compared to the previous state-of-the-art method ISO, the proposed MambaSSC achieves a 3% improvement in mIoU (31.25 vs. 32.32) and a higher IoU (47.11 vs. 47.51), demonstrating its effectiveness in MonoSSC. Furthermore, we observe that this improvement primarily stems from object classes. Specifically, for stuff classes (ceiling, floor, wall), ISO achieves slightly better overall performance. We assume this is because these classes typically have simple structures (e.g., flat planes), which are well-suited for CNN architectures that excel at local area modeling. In comparison, the proposed GA-MonoSSC outperforms ISO across almost all other object classes. This demonstrates that GA-MonoSSC is more effective in recovering and detecting nuanced geometric details and semantic information, thanks to its global awareness ability, which enables it to capture richer contextual information.

For the Occ-ScanNet dataset, we observe a similar trend but with a significantly larger performance gap. Specifically, GA-MonoSSC outperforms ISO by approximately 15% in IoU (42.16 vs. 48.59) and 24% in mIoU (28.71 vs. 35.65). The main reason for the significant performance difference between the NYU and Occ-ScanNet datasets lies in the dataset scale. The Occ-ScanNet dataset contains nearly 60× more samples than the NYU dataset, providing a substantially larger training set for learning richer representations. Furthermore, by equipping GA-MonoSSC with global context modeling in both the 2D and 3D domains through the Transformer and Mamba architectures, the model can efficiently scale to large-scale datasets. In contrast, CNN-based architectures, with their limited local receptive fields and fixed-weight encoding strategies, struggle to scale effectively, constraining their performance on larger datasets.

Finally, qualitative visualizations are presented in Fig. 5 for the NYUv2 dataset. These visualizations demonstrate that the proposed GA-MonoSSC produces more accurate predictions. More qualitative results can be found in the supplemental materials.

Method	Input	IoU	ceiling	floor	wall	window	chair	bed	sofa	table	tv	furniture	objects	mIoU
LMSNet ^{rgb} [46]	$\hat{x}^{occ}$	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 ^{rgb} [22]	$x_{rgb}, \hat{x}^{depth}$	30.03	7.58	82.97	9.15	0.05	6.93	35.87	22.94	11.10	0.71	15.90	6.45	18.15
3DSketch ^{rgb} [47]	$x_{rgb}, \hat{x}^{TSDF}$	38.64	8.53	90.54	9.94	6.75	14.62	49.09	17.73	13.35	8.13	24.68	8.49	22.91
MonoScene [21]	x_{rgb}	42.51	8.89	93.50	12.06	12.57	17.32	48.29	18.45	15.13	15.22	27.96	12.94	26.94
NDC-Scene [35]	x_{rgb}	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
ISO [36]	x_{rgb}	47.11	14.21	93.47	15.89	15.14	18.55	50.01	40.82	18.25	25.9	34.08	17.67	31.25
GA-MonoSSC	x_{rgb}	47.51	12.15	93.39	15.73	17.1	19.23	52.16	44.64	20.63	25.55	35.22	19.71	32.32

Table 2. Quantitative results on the NYUv2 dataset.

Method	Input	IoU	ceiling	floor	wall	window	chair	bed	sofa	table	tv	furniture	objects	mIoU
MonoScene [21]	x_{rgb}	41.60	15.17	44.71	22.41	12.55	26.11	27.03	35.91	28.32	6.57	32.16	19.84	24.62
ISO [36]	x_{rgb}	42.16	19.88	41.88	22.37	16.98	29.09	42.43	42.00	29.60	10.62	36.36	24.61	28.71
GA-MonoSSC	x_{rgb}	48.59	26.01	51.89	30.35	24.58	30.34	48.44	49.92	37.59	22.48	43.13	27.41	35.65

Table 3. Quantitative results on the Occ-ScanNet dataset.

NYU	2D Sup	N.A.	sem→geo	geo→sem	geo↔sem
mIoU	31.65	32.00	31.61	32.32	32.07
IoU	46.91	46.92	47.39	47.51	47.21

Table 4. Ablation Study of $\mathbf{DM}_{\text{Enc}}$ on NYU dataset “2D Sup” means enforce $\mathbf{Dec}_{\text{geo}}$ to predict 2D depth instead 3D occupancy where the supervision is employed. “N.A.” means no interaction between $\mathbf{Dec}_{\text{geo}}$ and $\mathbf{Dec}_{\text{sem}}$. “ $\mathbf{A} \rightarrow \mathbf{B}$ ” denotes the information is injected from $\mathbf{A}$ to $\mathbf{B}$.

4.2. Ablation Study

In this section, we evaluate the effectiveness of the proposed designs for both $\mathbf{DM}_{\text{Enc}}$ and $\mathbf{FM}_{\text{Dec}}$.

4.2.1. Effectiveness of $\mathbf{DM}_{\text{Enc}}$

We first evaluate the effectiveness of the design of $\mathbf{DM}_{\text{Enc}}$ on the NYUv2 dataset, and the experimental results are presented in Tab. 4. In this experiment, we first investigate the necessity of applying 3D supervision (occupancy) to the $\mathbf{Dec}_{\text{geo}}$ output rather than 2D supervision (depth). To enforce $\mathbf{Dec}_{\text{geo}}$ to explicitly learn geometry-related information, a relevant auxiliary supervision signal should be applied to distinguish it from $\mathbf{Dec}_{\text{sem}}$. Since $\mathbf{Dec}_{\text{geo}}$ produces a 2D feature map, a straightforward approach is to align it with the ground-truth 2D depth. However, as indicated by the results (2D Sup), this strategy does not achieve good performance. In contrast, converting predicted depth into 3D space (occupancy) and aligning it with ground-truth occupancy yields better performance. We assume this is because the ultimate goal is to recover geometric structures in 3D space, and the latter approach enables $\mathbf{Dec}_{\text{geo}}$ to learn 3D-aware 2D features, whereas 2D supervision alone fails to provide sufficient 3D spatial awareness.

Next, we investigate the interaction between $\mathbf{Dec}_{\text{geo}}$ and $\mathbf{Dec}_{\text{sem}}$. We observe that the proposed method already achieves good performance even without interaction

between $\mathbf{Dec}_{\text{geo}}$ and $\mathbf{Dec}_{\text{sem}}$ (N.A.). Injecting information from $\mathbf{Dec}_{\text{sem}}$ into $\mathbf{Dec}_{\text{geo}}$ leads to a noticeable increase in IoU, albeit at the cost of a performance drop in mIoU. This is reasonable, as $\mathbf{Dec}_{\text{geo}}$ benefits from high-level semantic information, which aids in recovering accurate geometric structures. However, without explicit geometry information, the semantic-related metric (mIoU) decreases accordingly. Instead, injecting geometry information into $\mathbf{Dec}_{\text{sem}}$ benefits both geometric and semantic performance. This indicates that enhancing semantic-related information contributes to an overall improvement in geometry performance as well. Finally, enabling $\mathbf{Dec}_{\text{geo}}$ and $\mathbf{Dec}_{\text{sem}}$ to interact deeply does not result in improved performance. We hypothesize that a strong interaction between two modalities may lead to a performance drop, similar to previous approaches that jointly learn them. This is also evidenced by the poor performance observed in Tab. 5 (Row 4).

4.2.2. Effectiveness of $\mathbf{FM}_{\text{Dec}}$

Then, we evaluate the effectiveness of the design of $\mathbf{FM}_{\text{Dec}}$ on the NYUv2 and Occ-ScanNet-mini datasets, and the experimental results are presented in Tab. 5.

We first verify the necessity of utilizing the Mamba architecture, which effectively and efficiently captures long-range dependencies in 3D space. Its effectiveness is assessed by replacing all Mamba blocks with 3D convolution-based blocks, widely adopted in previous methods [21, 35, 36], which consist of several bottleneck blocks. We observe that this replacement leads to a significant performance drop (Row 1 vs. Row 6), demonstrating the necessity of capturing global context in 3D space.

Next, we evaluate the effectiveness of each component in $\mathbf{FM}_{\text{Dec}}$. We begin by verifying the necessity of multi-modality and multi-scale feature learning. We observe that

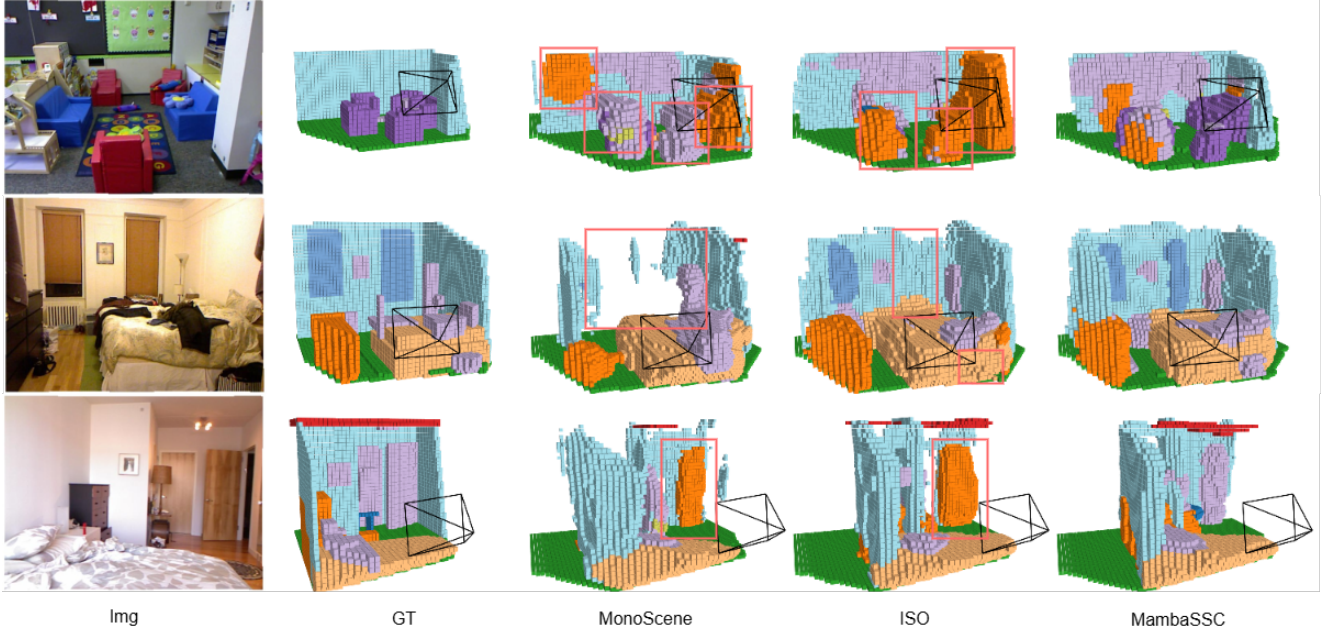


Figure 5. Qualitative results on the NYU dataset. The red rectangle indicates incorrect semantic predictions. Compared to previous methods, the proposed MambaSSC produces significantly fewer incorrect predictions.

on the Occ-ScanNet-mini dataset, omitting multi-scale features while jointly learning multi-modality features (instead of learning them separately) leads to a significant performance drop (Row 2 vs. Row 6). Learning multi-scale features primarily improves IoU (Row 2 vs. Row 4), likely because it enables the extraction of more detailed geometric features at different scales. Meanwhile, learning multi-modality features separately enhances both mIoU and IoU (Row 2 vs. Row 3), demonstrating that separately extracting these features allows the model to learn more nuanced information. Learning multi-scale features while separately learning multi-modality information yields the best overall performance, as shown in Row 6. Similar conclusion can be drawn from the NYU dataset, whereas the performance improvement is even larger. We assume that this is due to the limited amount of data in the NYU dataset, and that both strategies contribute to learning meaningful information in this scenario. Finally, we highlight the importance of re-ordering voxels as sequential data in frustum space rather than in the original 3D space. We observe that applying the Mamba block to voxel sequences reordered in frustum space achieves better performance across all metrics on both datasets (Row 5 vs. Row 6). This demonstrates that ensuring feature continuity in the resulting voxel sequence, as achieved by the Frustum Reorder Strategy, is crucial for Mamba, which processes sequential data from start to end.

5. Conclusion

In this work, we present GA-MonoSSC, a hybrid architecture for the MonoSSC task. Compared to previous CNN-based methods, which are limited by their local receptive

Row	Arch	MM	MS	FS	NYUv2		OSnet-mini	
					IoU	mIoU	IoU	mIoU
0	ISO	-	-	-	47.11	31.25	51.03	39.08
1	Conv	✓	✓	-	46.99	31.60	57.33	46.83
2	MB			✓	46.89	31.80	56.91	46.07
3		✓		✓	47.02	32.00	58.42	47.39
4			✓	✓	46.96	31.81	57.59	46.20
5		✓	✓		47.10	31.90	57.51	46.66
6		✓	✓	✓	47.51	32.32	58.97	48.19

Table 5. Ablation Study of $\mathbf{FM}_{Dec}$ on NYU and Occ-ScanNet-mini (**OSnet-mini**) datasets. **MB** denotes Mamba Block while **Conv** means replace Mamba Block with convolution block. **FS** denotes Frustum Reordering Strategy, **MM** denotes learning multi-modality features separately, **MS** denotes learning multi-scale features. We present ISO results as a reference, which show significantly lower performance compared to the proposed method.

fields and fixed-weight encoding strategies, GA-MonoSSC effectively captures global context in both the 2D image domain and 3D space. Specifically, it employs a Dual-Head Multi-Modality Encoder, based on a Transformer architecture, to capture spatial relationships among all features in the 2D image domain. The Frustum Mamba Decoder then processes 3D features using the State Space Model (SSM), which enables the efficient capture of long-range dependencies with linear computational complexity. Furthermore, we introduce a frustum reordering strategy within the Frustum Mamba Decoder to reduce feature discontinuities in the reordered voxel sequence, making it more compatible with the scan mechanism of the State Space Model (SSM). We conduct extensive experiments on the Occ-ScanNet and NYUv2 datasets, achieving state-of-the-art performance, which validates the effectiveness of our design.

Acknowledgments

This work is supported by the Agency for Science, Technology and Research (A*STAR) under its Career Development Fund (Project No. C243512026) and is also supported by the Agency for Science, Technology and Research (A*STAR) under its MTC Programmatic Funds (Grant No. M23L7b0021). Juergen Gall is supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) 1927/5-2 (FOR 2535 Anticipating Human Behavior).

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

References

- [1] Guilherme N DeSouza and Avinash C Kak. Vision for mobile robot navigation: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(2):237–267, 2002. [1](#)
- [2] Yanying Zhou, Shijie Li, and Jochen Garcke. Foresight social-aware reinforcement learning for robot navigation. In *Chinese Control and Decision Conf.*, pages 3501–3507. IEEE, 2023.
- [3] Jonathan Tremblay, Stephen Tyree, Terry Mosier, and Stan Birchfield. Indirect object-to-robot pose estimation from an external monocular rgb camera. In *IEEE/RSJ Int. Conf. Intelligent Robots and Systems*, pages 4227–4234. IEEE, 2020.
- [4] Rong Li, Shijie Li, Lingdong Kong, Xulei Yang, and Junwei Liang. Seeground: See and ground for zero-shot open-vocabulary 3d visual grounding. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3707–3717, 2025. [1](#)
- [5] Shi-Jie Li, Bo Ren, Yun Liu, Ming-Ming Cheng, Duncan Frost, and Victor Adrian Prisacariu. Direct line guidance odometry. In *IEEE Int. Conf. Robotics and Automation*, pages 5137–5143. IEEE, 2018. [1](#)
- [6] Ronald T Azuma. A survey of augmented reality. *Presence: Teleoperators and Virtual Environments/MIT press*, 1997.
- [7] Shi-Jie Li, Ming-Ming Cheng, Yun Liu, Shao-Ping Lu, Yahui Wang, and Victor Adrian Prisacariu. Structured skip list: A compact data structure for 3d reconstruction. In *IEEE/RSJ Int. Conf. Intelligent Robots and Systems*, pages 1–7. IEEE, 2018. [1](#)
- [8] Shijie Li, Xieyuanli Chen, Yun Liu, Dengxin Dai, Cyrill Stachniss, and Juergen Gall. Multi-scale interaction for real-time lidar data segmentation on an embedded platform. *IEEE Robot. Autom. Lett.*, 7(2):738–745, 2021. [1](#)
- [9] 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 *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 20258–20267, 2024. [2](#)
- [10] Shijie Li, Yun Liu, and Juergen Gall. Rethinking 3-d lidar point cloud segmentation. *IEEE Trans. Neural Netw. Learn. Syst.*, 2021. [1](#)
- [11] Xiaoyu Tian, Tao Jiang, Longfei Yun, Yucheng Mao, Huitong Yang, Yue Wang, Yilun Wang, and Hang Zhao. Occ3d: A large-scale 3d occupancy prediction benchmark for autonomous driving. *Adv. Neural Inform. Process. Syst.*, 36, 2024. [1](#)
- [12] Zhiqi Li, Zhiding Yu, David Austin, Mingsheng Fang, Shiyi Lan, Jan Kautz, and Jose M Alvarez. Fb-occ: 3d occupancy prediction based on forward-backward view transformation. *arXiv preprint arXiv:2307.01492*, 2023.
- [13] Lizi Wang, Hongkai Ye, Qianhao Wang, Yuman Gao, Chao Xu, and Fei Gao. Learning-based 3d occupancy prediction for autonomous navigation in occluded environments. In *IEEE/RSJ Int. Conf. Intelligent Robots and Systems*, pages 4509–4516. IEEE, 2021. [1](#)
- [14] Rong Li, Shijie Li, Xieyuanli Chen, Teli Ma, Juergen Gall, and Junwei Liang. Tfnet: Exploiting temporal cues for fast and accurate lidar semantic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4547–4556, 2024. [1](#)
- [15] Yahui Wang, Shaojun Cai, Shi-Jie Li, Yun Liu, Yangyan Guo, Tao Li, and Ming-Ming Cheng. Cubemapslam: A piecewise-pinhole monocular fisheye slam system. In *ACCV*, pages 34–49. Springer, 2018.
- [16] Xieyuanli Chen, Shijie Li, Benedikt Mersch, Louis Wiesmann, Jürgen Gall, Jens Behley, and Cyrill Stachniss. Moving object segmentation in 3d lidar data: A learning-based approach exploiting sequential data. *IEEE Robot. Autom. Lett.*, 6(4):6529–6536, 2021. [1](#)
- [17] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023. [2](#)
- [18] Shijie Li, Farhad G Zanjani, Haitam Ben Yahia, Yuki Asano, Juergen Gall, and Amirhossein Habibian. Valid: variable-length input diffusion for novel view synthesis. In *IEEE/CVF Winter Conf. Appl. Comput. Vis.*, pages 2240–2249. IEEE, 2025. [2](#)
- [19] Shijie Li, Rong Li, and Juergen Gall. Semantic rgb-d image synthesis. In *Int. Conf. Comput. Vis.*, pages 944–952, 2023. [2](#)
- [20] Shijie Li, Ming-Ming Cheng, and Juergen Gall. Dual pyramid generative adversarial networks for semantic image synthesis. *arXiv preprint arXiv:2210.04085*, 2022. [2](#)
- [21] Anh-Quan Cao and Raoul De Charette. Monoscene: Monocular 3d semantic scene completion. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3991–4001, 2022. [2](#), [3](#), [4](#), [6](#), [7](#)
- [22] Jie Li, Kai Han, Peng Wang, Yu Liu, and Xia Yuan. Anisotropic convolutional networks for 3d semantic scene completion. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3351–3359, 2020. [2](#), [7](#)
- [23] Shice Liu, Yu Hu, Yiming Zeng, Qiankun Tang, Beibei Jin, Yinhe Han, and Xiaowei Li. See and think: Disentangling semantic scene completion. *Adv. Neural Inform. Process. Syst.*, 31, 2018.
- [24] Shuran Song, Fisher Yu, Andy Zeng, Angel X Chang, Manolis Savva, and Thomas Funkhouser. Semantic scene completion from a single depth image. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1746–1754, 2017.
- [25] Jiahui Zhang, Hao Zhao, Anbang Yao, Yurong Chen, Li Zhang, and Hongen Liao. Efficient semantic scene completion network with spatial group convolution. In *Eur. Conf. Comput. Vis.*, pages 733–749, 2018.

- [26] Min Zhong and Gang Zeng. Semantic point completion network for 3d semantic scene completion. In *Eur. Conf. Artif. Intell.*, pages 2824–2831. IOS Press, 2020.
- [27] Yingjie Cai, Xuesong Chen, Chao Zhang, Kwan-Yee Lin, Xiaogang Wang, and Hongsheng Li. Semantic scene completion via integrating instances and scene in-the-loop. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 324–333, 2021.
- [28] Jiayang Tang, Xiaokang Chen, Jingbo Wang, and Gang Zeng. Not all voxels are equal: Semantic scene completion from the point-voxel perspective. In *AAAI*, volume 36, pages 2352–2360, 2022.
- [29] Pingping Zhang, Wei Liu, Yinjie Lei, Huchuan Lu, and Xiaoyun Yang. Cascaded context pyramid for full-resolution 3d semantic scene completion. In *Int. Conf. Comput. Vis.*, pages 7801–7810, 2019.
- [30] Yunpeng Zhang, Zheng Zhu, and Dalong Du. Occformer: Dual-path transformer for vision-based 3d semantic occupancy prediction. In *Int. Conf. Comput. Vis.*, pages 9433–9443, 2023.
- [31] Zhu Yu, Runmin Zhang, Jiacheng Ying, Junchen Yu, Xiaohai Hu, Lun Luo, Si-Yuan Cao, and Hui-Liang Shen. Context and geometry aware voxel transformer for semantic scene completion. *arXiv preprint arXiv:2405.13675*, 2024.
- [32] Junming Wang, Wei Yin, Xiaoxiao Long, Xingyu Zhang, Zebin Xing, Xiaoyang Guo, and Qian Zhang. Occrkv: Rethinking efficient 3d semantic occupancy prediction with linear complexity. *arXiv preprint arXiv:2409.19987*, 2024.
- [33] Zhu Yu, Bowen Pang, Lizhe Liu, Runmin Zhang, Qihao Peng, Maochun Luo, Sheng Yang, Mingxia Chen, Si-Yuan Cao, and Hui-Liang Shen. Language driven occupancy prediction. *arXiv preprint arXiv:2411.16072*, 2024.
- [34] Li Liang, Naveed Akhtar, Jordan Vice, Xiangrui Kong, and Ajmal Saeed Mian. Skip mamba diffusion for monocular 3d semantic scene completion. *arXiv preprint arXiv:2501.07260*, 2025. 2
- [35] 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 *Int. Conf. Comput. Vis.*, pages 9421–9431. IEEE Computer Society, 2023. 2, 7
- [36] Hongxiao Yu, Yuqi Wang, Yuntao Chen, and Zhaoxiang Zhang. Monocular occupancy prediction for scalable indoor scenes. In *Eur. Conf. Comput. Vis.*, 2024. 2, 6, 7
- [37] Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. *arXiv preprint arXiv:2111.00396*, 2021. 2, 3
- [38] Eric Nguyen, Karan Goel, Albert Gu, Gordon Downs, Preey Shah, Tri Dao, Stephen Baccus, and Christopher Ré. S4nd: Modeling images and videos as multidimensional signals with state spaces. *Adv. Neural Inform. Process. Syst.*, 35: 2846–2861, 2022. 2
- [39] Ali Hatamizadeh and Jan Kautz. Mambavision: A hybrid mamba-transformer vision backbone. *arXiv preprint arXiv:2407.08083*, 2024. 3
- [40] Md Maklachur Rahman, Abdullah Aman Tutul, Ankur Nath, Lamyamba Laishram, Soon Ki Jung, and Tracy Hammond. Mamba in vision: A comprehensive survey of techniques and applications. *arXiv preprint arXiv:2410.03105*, 2024.
- [41] Badri N Patro and Vijay S Agneeswaran. Simba: Simplified mamba-based architecture for vision and multivariate time series. *arXiv preprint arXiv:2403.15360*, 2024.
- [42] Yijun Yang, Zhaohu Xing, Lequan Yu, Chunwang Huang, Huazhu Fu, and Lei Zhu. Vivim: A video vision mamba for medical video segmentation. *arXiv preprint arXiv:2401.14168*, 2024. 3
- [43] Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. Vision mamba: Efficient visual representation learning with bidirectional state space model. *arXiv preprint arXiv:2401.09417*, 2024. 3
- [44] Dingkang Liang, Xin Zhou, Wei Xu, Xingkui Zhu, Zhikang Zou, Xiaoqing Ye, Xiao Tan, and Xiang Bai. Pointmamba: A simple state space model for point cloud analysis. *arXiv preprint arXiv:2402.10739*, 2024. 3
- [45] Jiuming Liu, Ruiji Yu, Yian Wang, Yu Zheng, Tianchen Deng, Weicai Ye, and Hesheng Wang. Point mamba: A novel point cloud backbone based on state space model with octree-based ordering strategy. *arXiv preprint arXiv:2403.06467*, 2024. 3
- [46] Luis Roldao, Raoul de Charette, and Anne Verroust-Blondet. Lmscnet: Lightweight multiscale 3d semantic completion. In *Int. Conf. 3D Vision*, pages 111–119. IEEE, 2020. 7
- [47] Xiaokang Chen, Kwan-Yee Lin, Chen Qian, Gang Zeng, and Hongsheng Li. 3d sketch-aware semantic scene completion via semi-supervised structure prior. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4193–4202, 2020. 7
- [48] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *Eur. Conf. Comput. Vis.*, pages 746–760. Springer, 2012. 6