Global-Aware Monocular Semantic Scene Completion with State Space Models

Supplementary Material

Table of Contents

A. Implementation Details

In this section, we provide additional details to facilitate the implementation and reproducibility of the proposed GA-MonoSSC.

A.1. Experiment Details

We use the AdamW optimizer with an initial learning rate of 2×10^{-4} for the NYUv2 dataset (30 epochs) and 1×10^{-4} for the Occ-ScanNet-mini (60 epochs) and Occ-ScanNet (10 epochs) datasets. A learning rate decay is applied during training to ensure effective convergence. The model is trained on 2 NVIDIA A40 GPUs for the NYUv2 and Occ-ScanNet-mini datasets, while Occ-ScanNet, due to its significantly larger scale, is trained on 8 NVIDIA A40 GPUs. The batch size is set to 2 per GPU.

A.2. Architecture Details

We present the architecture details of the proposed method, which consists of Dual-head Multi-modality Encoder ($\mathbf{DM}_{Enc}$), Features Line of Sight Projection ($\mathbf{FLoSP}$) and Frustum Mamba Decoder ($\mathbf{FM}_{Dec}$).

Dual-head Multi-modality Encoder ($\mathbf{DM_{Enc}}$) is employed to extract 2D features. It consists of a vision transformer encoder $\mathbf{Enc}$ and two modality-specific decoders, $\mathbf{Dec_{sem}}$ and $\mathbf{Dec_{geo}}$, which separately decode semantic and geometric features. Specifically, $\mathbf{Enc}$ consists of four transformer encoder blocks, each composed of multiple transformer encoder layers, and is initialized with Dinov2 [2] pre-trained weights. The output tokens from each stage are fed into the modality-specific decoders to generate the final multi-scale feature maps.

Features Line of Sight Projection (FLoSP) was introduced in [1] and is designed to unproject 2D features into 3D voxel space. The camera intrinsic parameters are assumed to be known. Each 3D voxel centroid (x^c) is projected onto the 2D image plane, where the corresponding 2D features from $\mathbf{DM}_{\text{Enc}}$ are sampled:

$$\mathcal{F}^{X,3D} = \Phi_{\rho(x^c)}(\mathcal{F}^X), \tag{1}$$

where X represents either semantic information (sem) or geometric information (geo). $\Phi_a(b)$ denotes the sampling of b at coordinates a, while $\rho(\cdot)$ represents the perspective projection. Voxels projected outside the image are assigned a feature vector of 0. The resulting feature map $\mathcal{F}^{X,3D}$ serves as input to the $\mathbf{FM_{Dec}}$.

Frustum Mamba Decoder ($\mathbf{FM_{Dec}}$) is designed to capture long-range dependencies in 3D space for accurate scene completion and semantic information inference. It follows a 3D U-Net-like architecture, consisting of an encoder and a decoder, each with two stages. A skip connection is introduced between each encoder stage and its corresponding decoder stage. The proposed Frustum Mamba Layers is primarily applied in the encoder. The input multi-scale semantic features are first fused into a unified feature volume. A 3D convolution block is then applied to introduce inductive bias, followed by two Frustum Mamba Layers, which process the features while progressively reducing resolution. Next, the 3D Context Relation Prior (3D CRP) proposed in [1] is applied to model voxel relationships. Finally, the decoder employs 3D Residual blocks and progressive upsampling to restore the original resolution and generate the final prediction.

A.3. Supervision Details

In this section, we provide a detailed description of the training loss, which is primarily based on [1].

Scene-Class Affinity Loss

This loss aims to enable the model to be aware of the global SSC performance by directly optimizing scene-level and class-level metrics, including (P)recision, (R)ecall, and (S)pecificity. Among these metrics, P_c and R_c evaluate the performance of voxels belonging to class c, while S_c assesses the performance of voxels that do not belong to class c:

$$P_c(\hat{p}, p) = \log \frac{\sum_i \hat{p}_{i,c} \mathbb{I}[p_i = c]}{\sum_i \hat{p}_{i,c}}, \tag{2}$$

$$R_c(\hat{p}, p) = \log \frac{\sum_i \hat{p}_{i,c} \mathbb{I}[p_i = c]}{\sum_i \mathbb{I}[p_i = c]}, \tag{3}$$

$$S_c(\hat{p}, p) = \log \frac{\sum_i (1 - \hat{p}_{i,c})(1 - \mathbb{I}[p_i = c])}{\sum_i (1 - \mathbb{I}[p_i = c])}, \quad (4)$$

where p_i is the ground truth class of voxel i, and $\hat{p}_{i,c}$ is its predicted probability of being class c. $\mathbb{I}[\cdot]$ represents the Iverson brackets. The Scene-Class Affinity Loss optimizes both semantic $\mathcal{L}_{\text{scal}}^{\text{sem}} = \mathcal{L}_{\text{scal}}(\hat{y}, y)$ and geometric $\mathcal{L}_{\text{scal}}^{\text{geo}} = \mathcal{L}_{\text{scal}}(\hat{y}^{\text{geo}}, y^{\text{geo}})$. $\{y, y^{\text{geo}}\}$ are semantic and geometric labels with respective predictions $\{\hat{y}, \hat{y}^{\text{geo}}\}$. $\mathcal{L}_{\text{scal}}$ builds on the affinity loss in [3]:

$$\mathcal{L}_{\text{scal}}(\hat{p}, p) = -\frac{1}{C} \sum_{c=1}^{C} (P_c(\hat{p}, p) + R_c(\hat{p}, p) + S_c(\hat{p}, p)). \quad (5)$$

Frustum Proportion Loss

This loss captures the impact of occlusion by partitioning the 3D space into multiple small, non-overlapping frustums and explicitly enforcing consistency in the class distribution within each frustum k using the Kullback-Leibler (KL) divergence:

$$\mathcal{L}_{\text{fp}} = \sum_{k=1}^{\ell^2} D_{\text{KL}}(P_k \| \hat{P}_k) = \sum_{k=1}^{\ell^2} \sum_{c \in C_k} P_k(c) \log \frac{P_k(c)}{\hat{P}_k(c)}. \quad (6)$$

Here, P_k represents the ground truth class distribution of voxels within frustum k, and $P_{k,c}$ denotes the proportion of class c in k. $\hat{P}_k$ and $\hat{P}_{k,c}$ are their corresponding soft predicted distributions, obtained by summing per-class predicted probabilities.

Context Relation Loss

The relation matrices $\hat{A}^m$ are inferred from the 3D Context Relation Prior (3D CRP), where each matrix encodes a unique relation $m \in \mathcal{M}$. These matrices are supervised using a relation loss defined as a weighted multi-label binary cross-entropy loss:

$$\mathcal{L}_{\text{rel}} = -\sum_{m \in \mathcal{M}, i} \left[(1 - A_i^m) \log(1 - \hat{A}_i^m) + w_m A_i^m \log \hat{A}_i^m \right], \quad (7)$$

where i loops through all elements of the relation matrix, A^m represents the ground truth, and the weight term w_m is defined as:

$$w_m = \frac{\sum_i (1 - A_i^m)}{\sum_i A_i^m}. \quad (8)$$

The total training loss with be:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{BCE}} + \mathcal{L}_{\text{rel}} + \mathcal{L}_{\text{scal}}^{\text{sem}} + \mathcal{L}_{\text{scal}}^{\text{geo}} + \mathcal{L}_{\text{fp}}. \quad (9)$$

where $\mathcal{L}_{\text{CE}}$ represents the final class-weighted cross-entropy loss, and $\mathcal{L}_{\text{BCE}}$ denotes the binary cross-entropy loss applied to the output of $\mathbf{Dec}_{\text{geo}}$.

B. Additional Quantitative Results

In this section, to pursue a more comprehensive comparison, we provide additional quantitative results of the proposed GA-MonoSSC.

2D Enc	3D Dec	NYUv2		Occ-ScanNet-mini	
		IoU	**mIoU**	**IoU**	**mIoU**
ISO	ISO	47.11	31.25	51.03	39.08
ISO	Ours	46.80	31.09	55.61	45.10
Ours	ISO	47.20	31.68	57.04	45.56
Ours	Ours	47.51	32.32	58.97	48.19

Table A. Component Interchange with ISO.

B.1. Component Interchange with ISO

To validate the effectiveness of our design, we interchange the 2D encoder and 3D decoder between the proposed GA-MonoSSC and the previous state-of-the-art method, ISO, and evaluate their performance on the NYUv2 and Occ-ScanNet-mini datasets. The experimental results, presented in Table A, show that the proposed method significantly outperforms ISO on both datasets, with an especially large margin on the large-scale Occ-ScanNet-mini dataset. This demonstrates the model's strong scalability to large datasets, attributed to its global modeling capability in both the 2D image domain and 3D space. Replacing either the 2D encoder or 3D decoder in the proposed method results in a noticeable performance drop, highlighting the effectiveness of our design. Specifically, integrating our 2D encoder with the ISO 3D decoder outperforms the original ISO model on both datasets. In contrast, replacing our 2D encoder with the ISO 2D encoder while retaining our 3D decoder results in a slight performance drop on the NYUv2 dataset compared to the original ISO. We attribute this to the small scale of the NYUv2 dataset, as the Mamba-based architecture excels at modeling global context and benefits from larger training data. This is further validated by the results on the large-scale Occ-ScanNet-mini dataset, where this architecture significantly outperforms the original ISO.

C. Broader Impact and Limitations

C.1. Broader Impact

Our model enables more accurate Monocular Semantic Scene Completion by incorporating global awareness in both 2D and 3D ccomponents. By introducing a transformer architecture into the 2D feature extraction process, the irregular distribution of projected 3D points is mitigated, allowing for the extraction of more representative information. Furthermore, the State-Space Model enables the 3D model to capture long-range dependencies, which was challenging in previous methods due to high computational costs.

C.2. Potential Limitations

Despite the advancements of the proposed method, certain limitations remain. There is still considerable room for performance improvement, particularly in handling complex scenarios where the method may struggle to recover detailed 3D structures and accurately detect semantic information. Additionally, model efficiency can be further optimized, which is crucial for real-world applications but has not been a primary focus in current works. In future work, we aim to enhance both model efficiency and performance.

D. Public Resource Used

In this section, we acknowledge the use of the following public resources, during this work:

- Pytorch [1] ..... Pytorch License
- ISO [2] ..... Apache License 2.0
- MonoScene [3] ..... Apache License 2.0
- NDCScene[4] ..... Apache License 2.0
- ScanNet [5] ..... Apache 2.0 License
- NYUv2[6] ..... non-commercial license

E. More Qualitative Visualization.

More qualitative visualizations on the OccScanNet dataset are presented in Fig. A. These results further demonstrate that the proposed method can infer more detailed 3D structures and accurately detect relevant semantic information.

References

[1] Anh-Quan Cao and Raoul De Charette. Monoscene: Monocular 3d semantic scene completion. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3991–4001, 2022. 1

[2] Maksym Oquab, Théo Darcet, Thibaut Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, and et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 1

[3] Changqian Yu, Jingbo Wang, Changxin Gao, Gang Yu, Chunhua Shen, and Nong Sang. Context prior for scene segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 12416–12425, 2020. 2

[1] https://github.com/pytorch/pytorch

[2] https://github.com/hongxiaoy/ISO

[3] https://github.com/astra-vision/MonoScene

[4] https://github.com/Jiawei-Yao0812/NDCScene/tree/main

[5] http://www.scan-net.org/

[6] https://cs.nyu.edu/~fergus/datasets/nyu_depth_v2.html

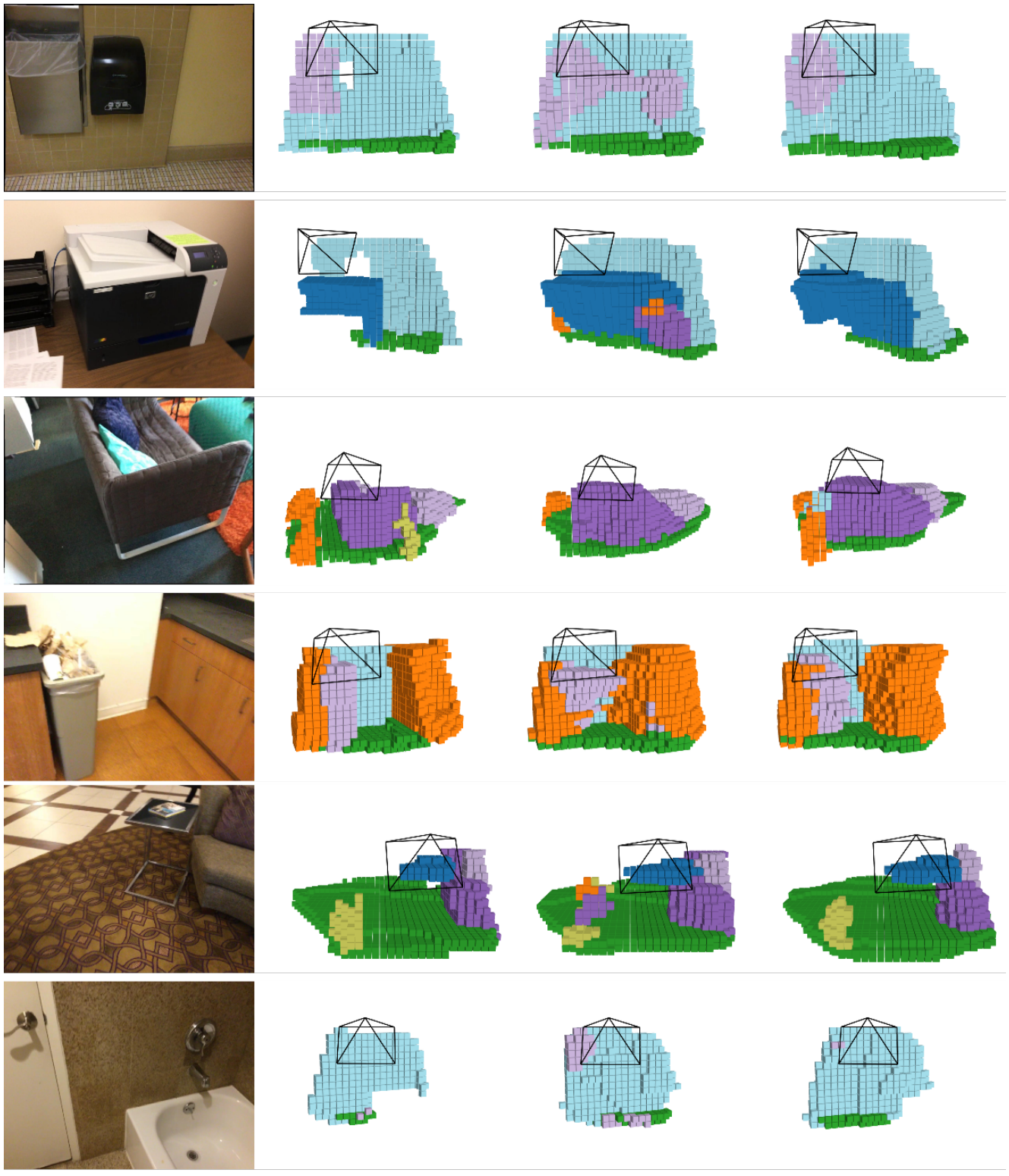

Figure A. Qualitative Results on OccScanNet dataset.