GOV-NeSF: Generalizable Open-Vocabulary Neural Semantic Fields

Supplementary Material

This supplementary material mainly includes: (1) Details about the experimental settings; (2) Further ablation studies; (3) Additional Qualitative Results; (4) Comparison with GNeSF (5) Limitations and Future Work; (6) Potential Social Impacts.

1. Details of experimental settings

For training, we set $\alpha = 1.0$ to balance the loss terms, and train the model end-to-end, using batch size of 1 and initial learning rate of $1e-3$ with cosine decay. We train the models on ScanNet train set for 10 epochs and evaluate them on ScanNet val set using a single NVIDIA L40 GPU.

RDE	DE	NVS	2D *Seg.*		3D *Seg.*	
		PSNR	mIoU	mAcc	mIoU	mAcc
		20.6	51.2	61.2	44.5	58.2
✓		21.1	51.8	61.4	44.7	58.3
✓	✓	**21.5**	**52.2**	**62.2**	**45.7**	**59.1**

Table 1. **Ablation study on ScanNet, evaluating RDE and DE.**

Train	Infer	NVS	2D *Seg.*		3D *Seg.*	
		PSNR	mIoU	mAcc	mIoU	mAcc
		21.5	**52.2**	**62.2**	45.7	59.1
	✓	10.5	46.8	58.3	**53.5**	**64.2**
✓	✓	18.7	50.9	61.8	53.3	64.0

Table 2. **Ablation study evaluating usage of DGM during training and inference.**

2. Further Ablation Studies

Ray Direction Encoders. As presented in Table 1, we evaluate the effectiveness of integrating the ray direction and relative ray directions into the prediction of blending weights. The integration of ray direction serves as the positional encoding for cross-view attention to provide further clues, benefiting both 2D and 3D semantic segmentation performance, while the relative direction encoding mainly improves 2D rendering quality and thus enhancing 2D semantic segmentation through shared density field.

Impact of Using DGM in Training and Inference. An ablation study assessing the effects of employing Depth Guided Masking (DGM) during training and inference phases is detailed in Table 2. The findings reveal that incorporating DGM in the training phase adversely affects both 2D and 3D semantic segmentation performance. Consequently, DGM is utilized only during the inference phase for enhancing 3D semantic segmentation.

$\mathcal{L}_{color}$	$\mathcal{L}_{ov}$	Detach	NVS	2D *Seg.*		3D *Seg.*	
			PSNR	mIoU	mAcc	mIoU	mAcc
✓			**21.9**	48.8	59.4	42.1	56.8
	✓		11.0	51.0	61.4	45.4	58.4
✓	✓		21.3	52.0	**62.2**	45.2	58.7
✓	✓	✓	21.5	**52.2**	**62.2**	**45.7**	**59.1**

Table 3. **Ablation study of learning objectives.**

Analysis of Learning Objectives. Table 3 showcases an evaluation of the impact of employing different loss functions: color loss ($\mathcal{L}_{color}$), OV loss ($\mathcal{L}_{ov}$), and the effect of detaching the gradient from $\mathcal{L}_{ov}$ with respect to density. The absence of $\mathcal{L}_{color}$, where uniform blending weights ($\boldsymbol{w}_s$) are applied, leads to a notable decrease in performance (Row 1). This underscores the importance of learning distinct blending weights for OV features in both 2D and 3D semantic segmentation. Furthermore, Rows 2 through 4 highlight the necessity of applying $\mathcal{L}_{color}$ to accurately learn the density field, thereby benefiting 2D and 3D semantic segmentation performance.

3. Additional Qualitative Results

3D Results with Arbitrary Textual Queries. The visualizations of 3D results on ScanNet when given arbitrary text queries are presented in Figure 1. A notable limitation of OpenScene-3D is its inability to accurately segment smaller objects, such as books, attributed to its exclusive reliance on 3D shapes. In contrast, our approach demonstrates superior performance compared to OpenScene-2D with and without depth maps.

3D Results on the Replica Dataset. Figure 2 illustrates the qualitative 3D results on the Replica dataset. This comparison highlights the limitations in generalizability of OpenScene-3D, which shows poor performance due to the domain gap. Conversely, our proposed approach enhances the results of OpenScene-2D, both in scenarios with and without the guidance of depth maps.

4. Comparison with GNeSF

We additionally provide a thorough comparison with GNeSF [Chen et al, NeurIPS'23]. **1) *Framework Design*:** On high-level design, GNeSF and our method share simi-

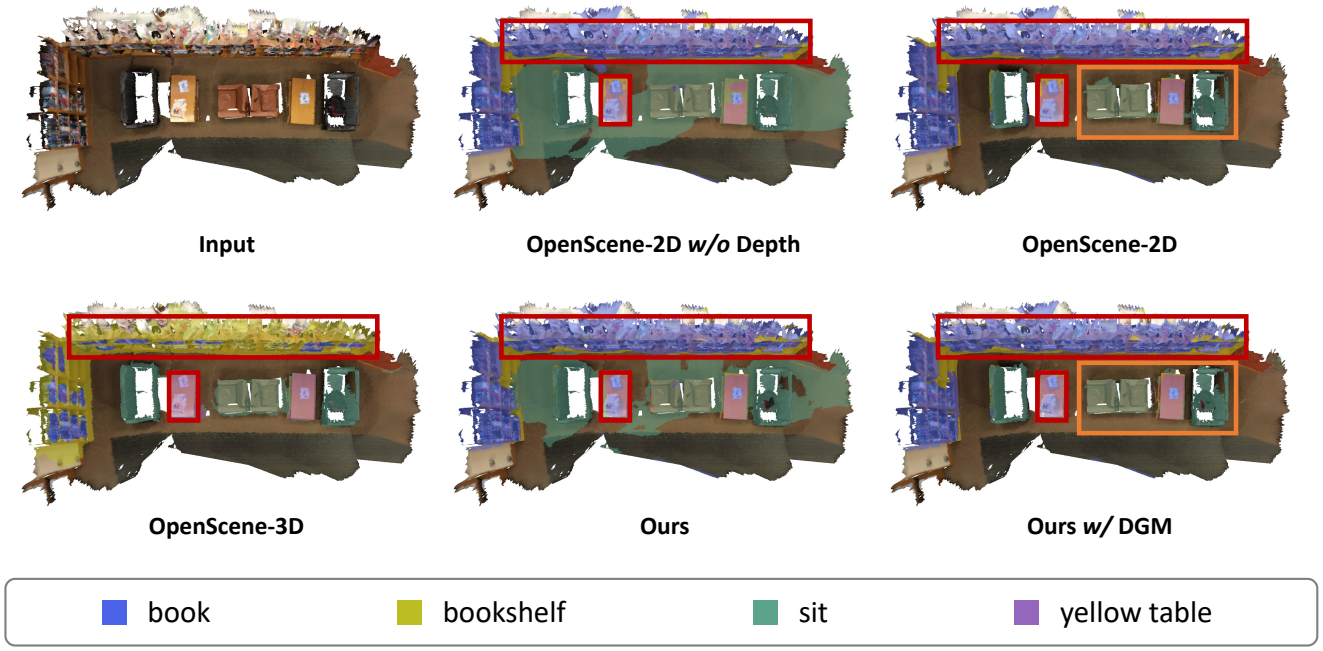
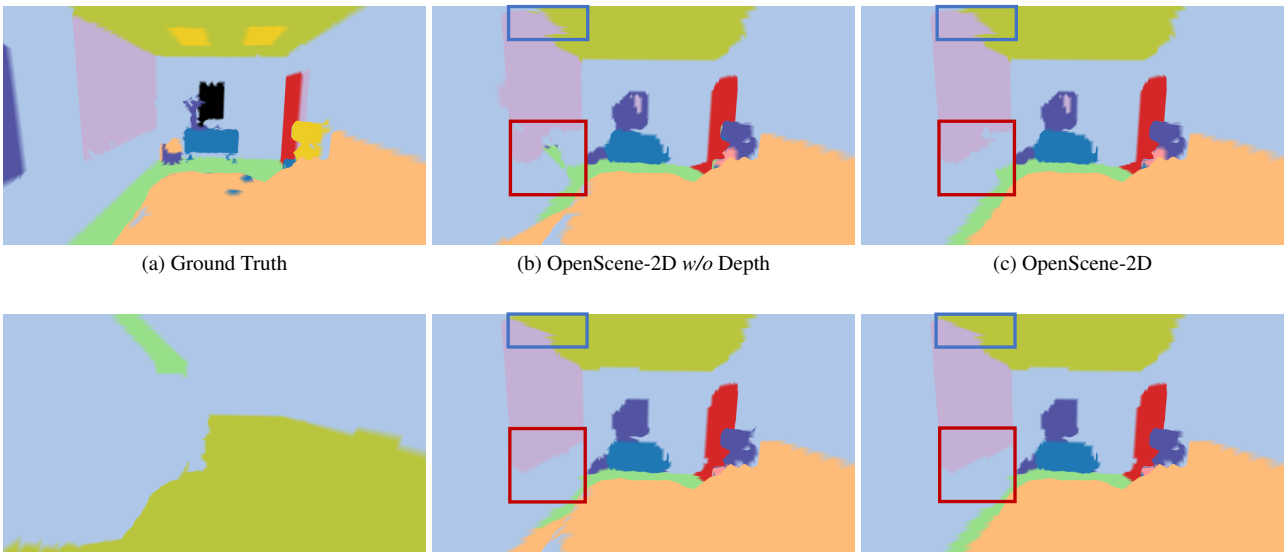


Figure 1. **Visualization of 3D results given arbitrary text queries.**



(a) Ground Truth (b) OpenScene-2D *w/o* Depth (c) OpenScene-2D

(d) OpenScene-3D (e) Ours (f) Ours *w/* DGM

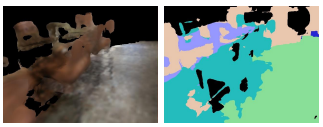
Figure 2. **3D qualitative results on Replica.**

2D Method	mIoU2D	mAcc2D	3D Method	mIoU3D	mAcc3D
GNeSF-2D-OV	47.7	58.0	GNeSF-3D-OV	47.3	60.9
Ours	**52.2**	**62.2**	Ours *w/* DGM*	**50.4**	**62.5**

Table 4. **Comparison with GNeSF.** * denotes we use NeuralRecon predicted mesh and depth maps for a fair comparison.

larities with IBRNet in predicting blending weights to fuse multi-view semantics. However, GNeSF utilizes two separate frameworks for 2D and 3D segmentation. Their 2D version directly adds a semantic MLP to IBRNet and does not aggregate feature volume. Their 3D version uses a NeuralRecon model pretrained using 3D supervision to extract volume features and occupancy which are *kept frozen*

during training, *i.e.*, the density field is already given by NeuralRecon without fine-tuning. This results in holes in the rendered maps as shown in Fig. 3 which prevents it from effective joint 2D and 3D segmentation. In contrast, our model is designed for joint 2D and 3D segmentation through a generalizable NeRF pipeline *without* 3D supervision, and we train VFA, MJF, CVA modules to learn the implicit fields for joint 2D and 3D OV segmentation. **2) *Performance*:** We created GNeSF-based baselines using LSeg feature maps in their public code as shown in Tab. 4, where we outperform by $3.1 - 4.5\%$ mIoU on both settings. Note that GNeSF-3D uses the dense image sequences similarly as NeuralRecon. **3) *Motivation*:** We are targeting the more challenging task of generalizable OV segmentation. Inspired by IBRNet, we blend multi-view OV features into the semantic field for generalizability. Although this is a straightforward design, the experiments demonstrated that the MJF module is crucial specifically for learning generalizable OV semantics.



(a) Color (b) Semantic

Figure 3. **GNeSF-3D rendering.**

5. Limitations and Future Work

One limitation of this study is the necessity for multi-view LSeg prediction during the inference stage. Throughout our experiments, we pre-extracted the LSeg feature maps of the images to reduce the duration of both training and inference. However, in practical applications, the inference process for 5 novel views and 3D point cloud takes approximately 5 minutes in total. Future research could focus on enhancing the efficiency of the cross-view attention module. Specifically, developing a sparse and more efficient variant of this module to speed up the inference process.

6. Potential Social Impacts

Our approach is capable of representing unseen 3D scenes with novel view synthesis, and performing both 2D and 3D semantic segmentation given arbitrary open-vocabulary text queries. As far as we are concerned, this work has the potential of improving open-vocabulary 3D scene understanding, and will not give rise to any social issues.