Know Your Neighbors: Improving Single-View Reconstruction via Spatial Vision-Language Reasoning

Supplementary Material

Method	O_{acc} ↑	IE_{acc} ↑	IE_{rec} ↑
Monodepth2 [1]	**0.94**	n/a	n/a
Monodepth2 [1] + $4m$	0.91	0.63	0.22
PixelNeRF [3]	0.92	0.63	0.43
BTS [2]	**0.94**	**0.77**	0.43
Ours	**0.94**	0.76	**0.55**

Table 1. **Comparison on KITTI-360 using GT accumulated by 20 LiDAR frames**. Our method achieves competitive or better results.

F_{app}	F_{fused}	VL-Mod.	Attn.	VL-Attn.	Scene Recon.			Object Recon.		
					O_{acc}	IE_{acc}	IE_{rec}	O_{acc}	IE_{acc}	IE_{rec}
✓					0.84	0.60	0.53	0.72	0.61	0.48
✓			✓		0.85	0.61	0.60	0.73	0.61	0.56
	✓		✓		0.85	0.61	0.67	0.72	0.60	0.62
	✓			✓	0.85	0.60	0.66	0.73	0.62	0.61
	✓	✓	✓		**0.86**	**0.63**	**0.75**	0.74	0.62	**0.73**
	✓	✓		✓	**0.86**	**0.63**	0.73	**0.75**	**0.63**	0.68

Table 2. **Ablations on the of the spatial attention.** We report the performance in the $[4, 50]$ meters range. The spatial attention improves in each variant by enabling the 3D context awareness. Combining the vision-language (VL) features with spatial attention yields the best performance.

1. Object-level Annotation

As illustrated in Sec. 4.2 of the main paper, to evaluate the object-level reconstruction on KITTI-360, we annotate the object area from the GT occupancy maps accumulated by 300 LiDAR frames. Specifically, for each GT occupancy slice parallel to the ground plane, we 1) manually annotate the object area by selecting the connected region of each object, and 2) enlarge the object area into a bounding box with $2m$ margins on the front and back border, and $0.5m$ margins on the left and right sides. We then compute the metrics within the bounding box. Notably, with the enlarged evaluation area, we can evaluate not only the occupied area but also the empty space, facilitating a comprehensive evaluation of the object's shape. Fig. 1 demonstrates this process.

2. Results with 20-Frame Accumulated GT

In addition to the evaluation using high-quality GT accumulated by 300 frames, we also conduct comparisons using GT accumulated by 20 frames to align with previous work [2]. The results are shown in Tab. 1, where we compare with previous methods within the $[4, 20]$ meters range following [2]. Our method achieves competitive or better performance.

3. Spatial Attention v.s. VL Spatial Attention

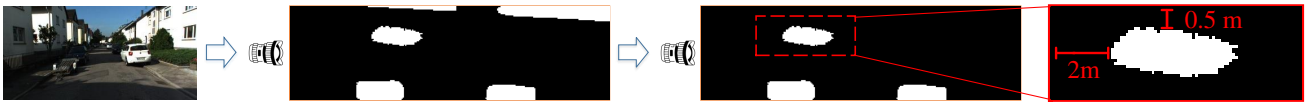
In Sec. 4.6.2 of the main paper, we investigate different variants of spatial attention. We show the duplicated results of using generic spatial attention and the vision-language (VL) spatial attention in Tab. 2. As shown in the last two rows, the two variants show close quantitative performances under current evaluation metrics. However, the visualization shows that using VL spatial attention yields better reconstruction details. As shown in Fig. 2, compared with the generic spatial attention, the VL attention produces sharper object borders that better separate the objects from the background (row 2, 3, 4) and completes object shapes that better describe the true geometry (row 1, 3, 5).

4. Ablations of Text Embeddings and Varying Segmentation Categories

Intuitively, the effectiveness of text embeddings correlates with the category types used for segmentation. As described in Sec. 3.1 of the main paper, we utilize category names commonly encountered in outdoor scenes as the prompts to the text encoder. The categories are *road, person, rider, car, truck, bus, train, bicycle, sidewalk, ground, parking, rail track, building, wall, fence, bridge, pole, traffic light, traffic sign, vegetation, others*. These 21 categories include most scene types present in outdoor scenes. To further investigate the effectiveness of text embeddings and the influence of category types, we ablate our method by feeding random text embeddings and using different category types. As shown in Tab. 3, random text embeddings (row 1) lead to inferior performance, manifesting the importance of informative text embeddings. Meanwhile, introducing more categories from 5 to 21 or 150 increases the performance, showing that our method can benefit from text embeddings retrieved from fine-grained segmentation. We observe the best accuracy when on the 21 street scene-specific categories.

5. Additional Qualitative Results

We provide more scene reconstruction results represented as occupancy grids. As shown in Fig. 3 and 4, our method consistently outperforms previous work and recovers accurate object shapes. In particular, it estimates plausible object

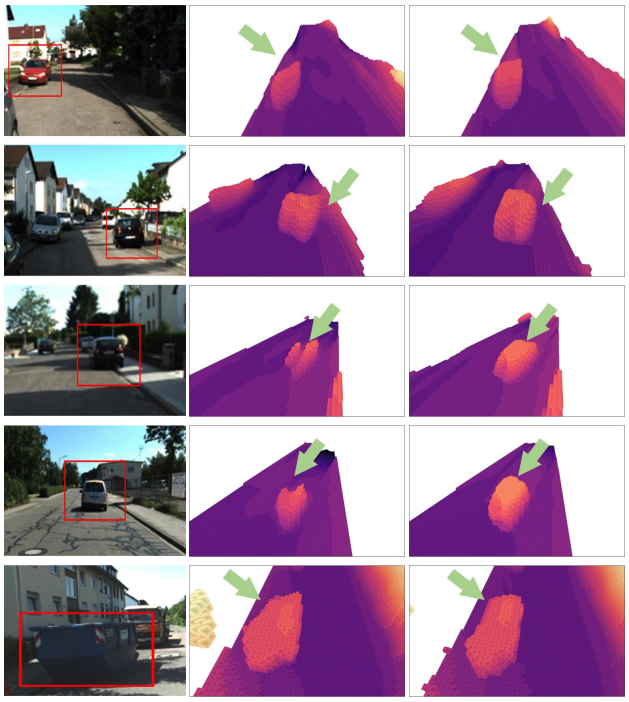


(a) Input (b) GT Occupancy (c) Object connected region (d) Object evaluation area

Figure 1. **Object annotation process.** Given an (a) input image, we show the object GT annotation process, where the camera icon indicates the direction of the observing view. For (b) a slice of the GT occupancy map parallel to the ground plane, we first identify the objects by (c) annotating their connected regions. For each object, we compute metrics within a (d) bounding box with $2m$ margin of its front and back border, as well as $0.5m$ margin of its left and right sides.

Method	Scene Recon.			Object Recon.		
	O_{acc}	IE_{acc}	IE_{rec}	O_{acc}	IE_{acc}	IE_{rec}
Ramdom text embeddings	0.84	0.61	0.55	0.69	0.56	0.52
5-merged street categories	0.84	0.60	0.61	0.7	0.57	0.57
150-general categories	0.85	0.61	**0.77**	**0.75**	0.63	**0.71**
21-street categories (Ours)	**0.86**	**0.63**	0.73	**0.75**	**0.64**	0.68

Table 3. Ablation of text embeddings and category numbers. Adopting random text embeddings leads to inferior performance. Meanwhile, incorporating more categories from 5 to 21 or 150 increases the performances.



(a) Input (b) Spatial Attention (c) VL Spatial Attention

Figure 2. **Qualitative comparison between generic and VL spatial attentions.** Compared to spatial attention, the VL attention mechanism produces fine-grained reconstructions with reasonable object shapes and sharp borders.

shapes even in the absence of direct visual cues.

References

[1] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3828–3838, 2019. 1, 3

[2] Felix Wimbauer, Nan Yang, Christian Rupprecht, and Daniel Cremers. Behind the scenes: Density fields for single view reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9076–9086, 2023. 1, 3

[3] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4578–4587, 2021. 1, 3

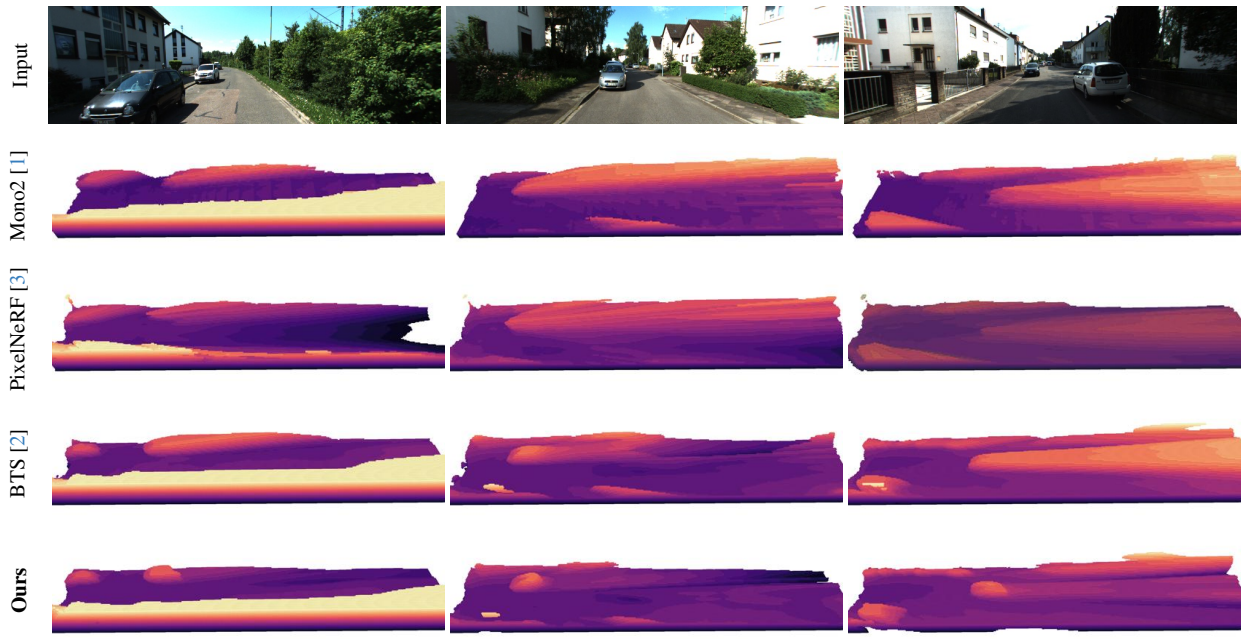


Figure 3. **Qualitative comparisons on KITTI-360 dataset (part-1).** The scene reconstructions are represented as occupancy grids, where the camera is on the left side and points to the right along the z-axis.

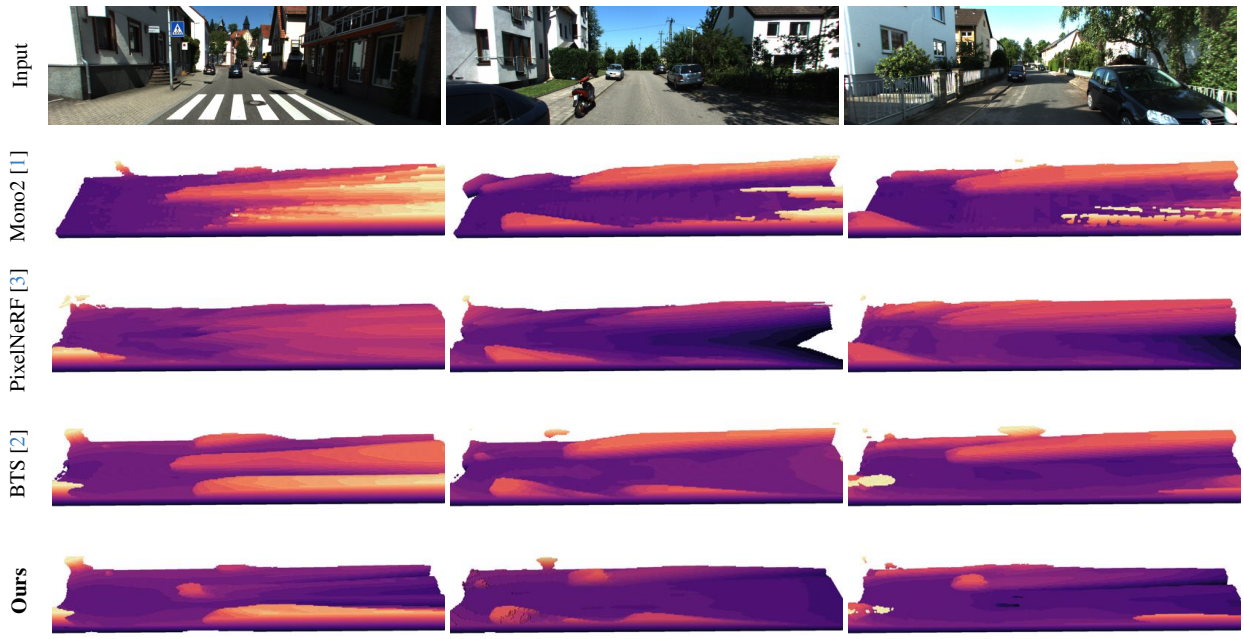


Figure 4. **Qualitative comparisons on KITTI-360 dataset (part-2).** The scene reconstructions are represented as occupancy grids, where the camera is on the left side and points to the right along the z-axis.