WildSeg3D: Segment Any 3D Objects in the Wild from 2D Images

Supplementary Material

6. Implementation Details

In the DGA module, the dynamic adjustment function incorporates two hyperparameters α_p and α_n, which are set to 0.5 and 0.9 in both quantitative and qualitative experiments. Our approach constructs 3D scenes using sparse views, and for the training phase of the dynamic aligning loss, we employ 500 iterations. For each scene, the entire workflow, from pre-processing to global alignment reconstruction and user-prompted 3D segmentation, runs on a single NVIDIA RTX 3090 GPU in approximately 30 seconds.

7. Limitations and Further Direction

SAM2's 2D segmentation has limitations in tracking small targets and objects with similar appearances, and our method consequently faces these challenges. To address this, semantic embedding of point clouds in 3D may be a future direction. Although we successfully integrate 3D reconstruction and segmentation into a single pipeline, it is not strictly end-to-end differentiable. Therefore, feed-forward semantic reconstruction is also a topic worth exploring.

8. Experimental Details

8.1. More Qualitative Evaluation

We implemented real-time intuitive segmentation based on mask cache, supporting user interaction for multi-round segmentation. Furthermore, leveraging SAM2, we demonstrate that users can add further interactions to refine the segmentation mask. Visualization results based on the LERF [18] dataset are shown in Figure 6.

To validate the effectiveness of our WildSeg3D, we conducted visualization-based ablation studies on segmentation tasks. As outlined in Sec. 4.4, DGA demonstrates its ability to effectively reduce 3D alignment errors by mitigating the influence of misaligned points and redundant background information. This improvement significantly enhances the clarity of object boundaries while minimizing artifacts such as blurred details and background confusion. The efficiency of DGA has previously been demonstrated. Furthermore, Figures 7, 8, 9, and 10 present visualization experiments conducted on the NVOS [44], SPIn-NeRF [35], Mip-NeRF360 [1], and T&T [20] datasets, respectively.

Without DGA, aligning pointmaps from multiple views to a unified coordinate system is significantly affected by the presence of redundant background points, which adversely affect the global aligning accuracy. This issue becomes particularly pronounced in complex scenes. Moreover, directly using confidence scores as aligning weights for 3D points across different views, which vary in matching difficulty, can lead to the accumulation of 3D alignment errors. Notably, the incorporation of DGA enables our method to accurately demarcate object boundaries while suppressing the influence of background pixels and dynamically adjusting aligning weights.

Scene			α_p		
	-0.9	-0.5	0	0.5	0.9
fern	93.9%	93.9%	94.2%	94.2%	94.1%
flower	91.0%	91.1%	91.2%	91.1%	91.0%
fortress	97.8%	97.8%	97.7%	97.7%	97.6%
horns (left)	95.2%	95.1%	95.2%	95.2%	95.1%
horns (center)	94.1%	93.9%	93.5%	93.5%	93.2%
leaves	90.9%	91.3%	92.2%	94.0%	94.1%
orchids	89.4%	90.1%	90.0%	91.0%	90.6%
trex	39.7%	79.9%	84.5%	86.4%	86.3%
average	86.5%	91.6%	92.3%	**92.9%**	92.8%

Table 5. Ablation experiment on the NVOS dataset for the adjustment factor α_p.

Scene			α_n		
	-0.9	-0.5	0	0.5	0.9
fern	73.2%	75.3%	94.2%	94.0%	94.2%
flower	84.0%	40.4%	91.2%	89.4%	89.5%
fortress	98.2%	78.6%	97.7%	97.6%	97.7%
horns (left)	96.3%	96.4%	95.2%	95.1%	95.0%
horns (center)	35.5%	32.2%	93.5%	92.9%	93.0%
leaves	93.8%	66.8%	92.2%	96.3%	96.6%
orchids	92.3%	91.2%	90.0%	90.1%	90.3%
trex	82.0%	78.6%	84.5%	86.3%	86.4%
average	81.9%	69.9%	92.3%	92.7%	**92.8%**

Table 6. Ablation experiment on the NVOS dataset for the adjustment factor α_n.

8.2. Evaluation on Adjustment Factor α

Tables 5 and 6 present the results of our ablation experiments conducted on the NVOS dataset. For each scene, we select five views as training views and use the mask segmented by SAM2 from one reference view as the ground truth for evaluation. With the Adjustment Factor α ranging from $[-1, 1]$, we project the 3D segmentation results from all scenes onto the reference views and compute the mIoU with their ground truth masks. The Adjustment Factor involves two hyperparameters, α_p and α_n, representing matching and non-matching points, respectively. To evaluate the impact of each hyperparameter on segmentation performance, we systematically investigate their individual contributions. For example, to analyze the influence of α_p, we fix α_n at 0, and vice versa. Among all values of α_p and

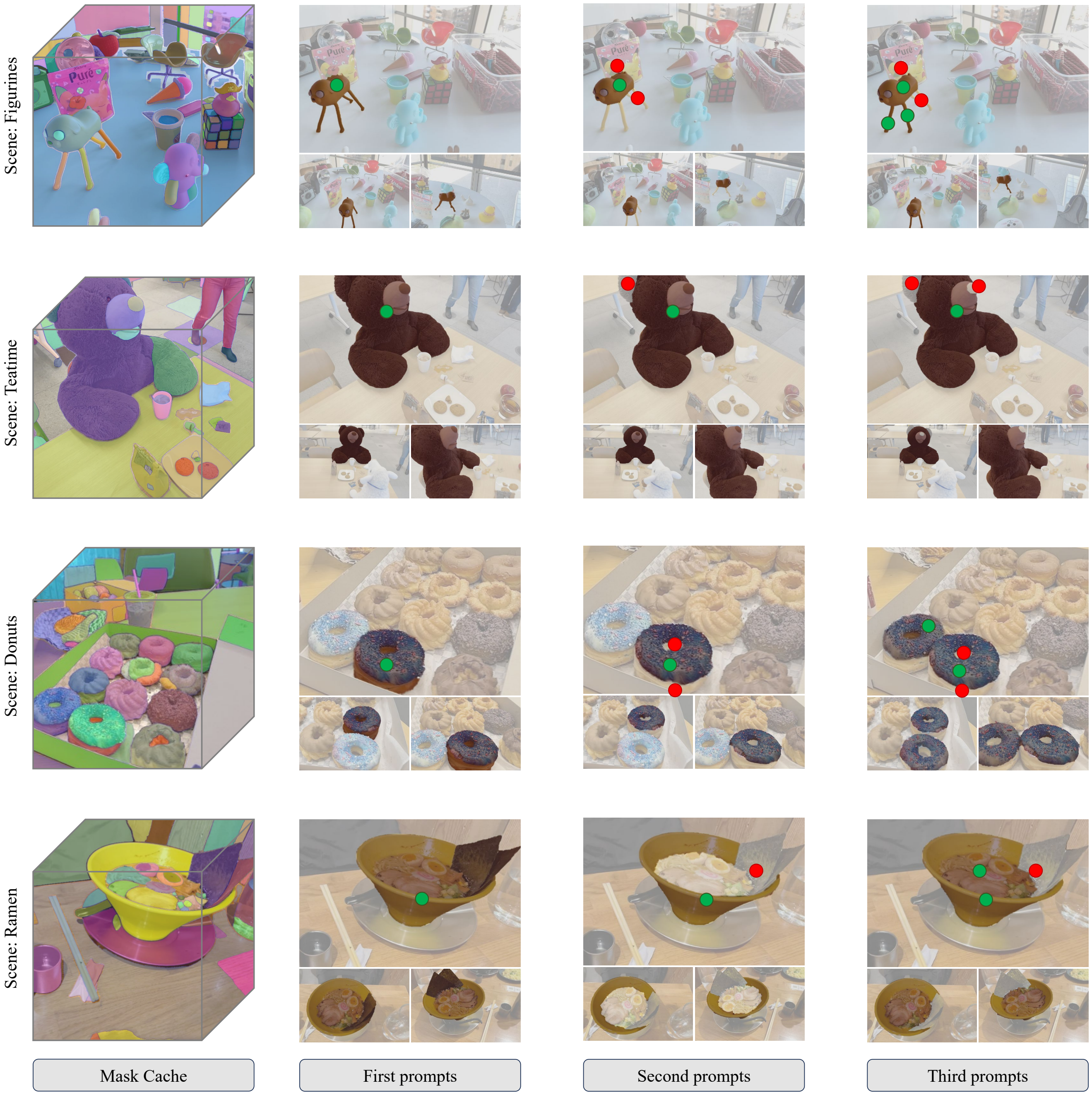

Figure 6. **Visualization of mask cache and interactive segmentation on the LERF dataset.** In each row, the segmentation masks in the first column represent the stored masks in our mask cache. Each scene undergoes three interactive refinement steps, with three distinct viewpoints displayed per prompt. Green and red points denote user-provided positive and negative prompts, respectively.

α_n, the settings of $\alpha_p = 0.5$ and $\alpha_n = 0.9$ achieve the optimal mean IoU of 92.9% and 92.8%, respectively.

9. Segmentation in the Wild

Our method demonstrates the capability of real-time segmentation in arbitrary scenes. As shown in Figure 3, we perform segmentation on both indoor and outdoor scenes. Notably, the "Eiffel Tower" and "Big Ben" scenes are sourced from outdoor aerial videos, where a selection of frames is extracted and used as images for 3D segmentation. To showcase WildSeg3D's capability in segmenting arbitrary objects, we evaluate it on additional scenes, as shown in

Figure 11. Our method achieves robust performance on diverse real-world scenes with highly sparse views, completing reconstruction and interactive segmentation within 10 seconds in unconstrained environments.

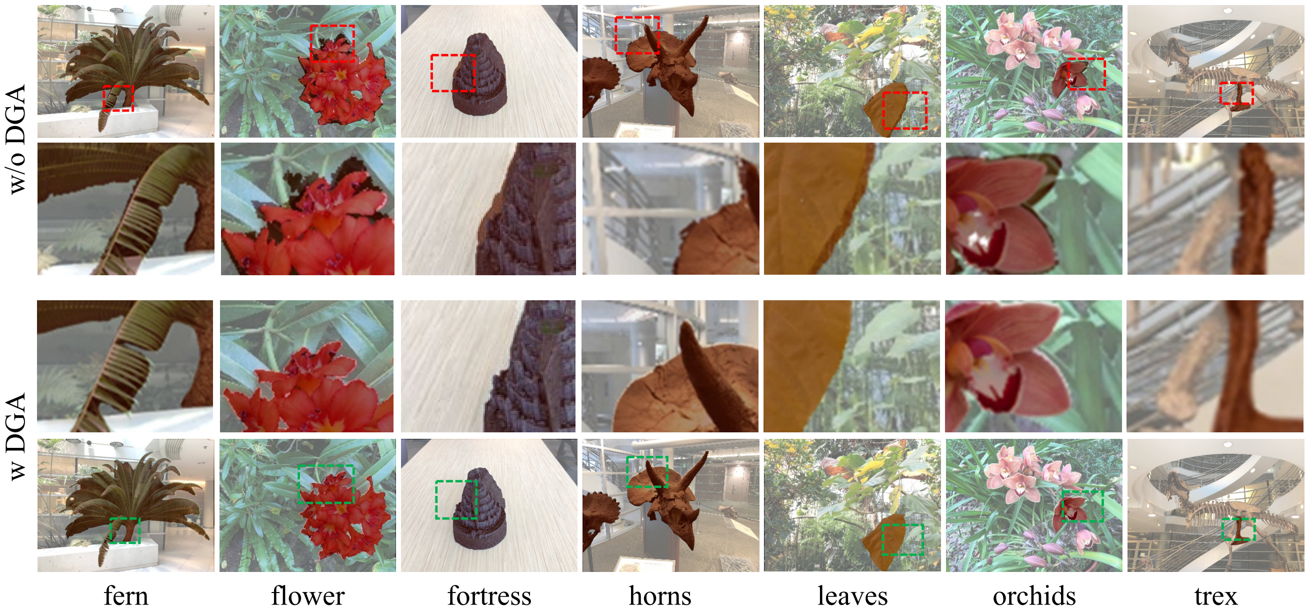

Figure 7. **Visualization of ablation experiment on the NVOS dataset.** In each column, the images depicted in the top and bottom rows illustrate the segmentation results without and with DGA, respectively. The second and third rows highlight zoomed-in views of the areas within the red and green dashed boxes.

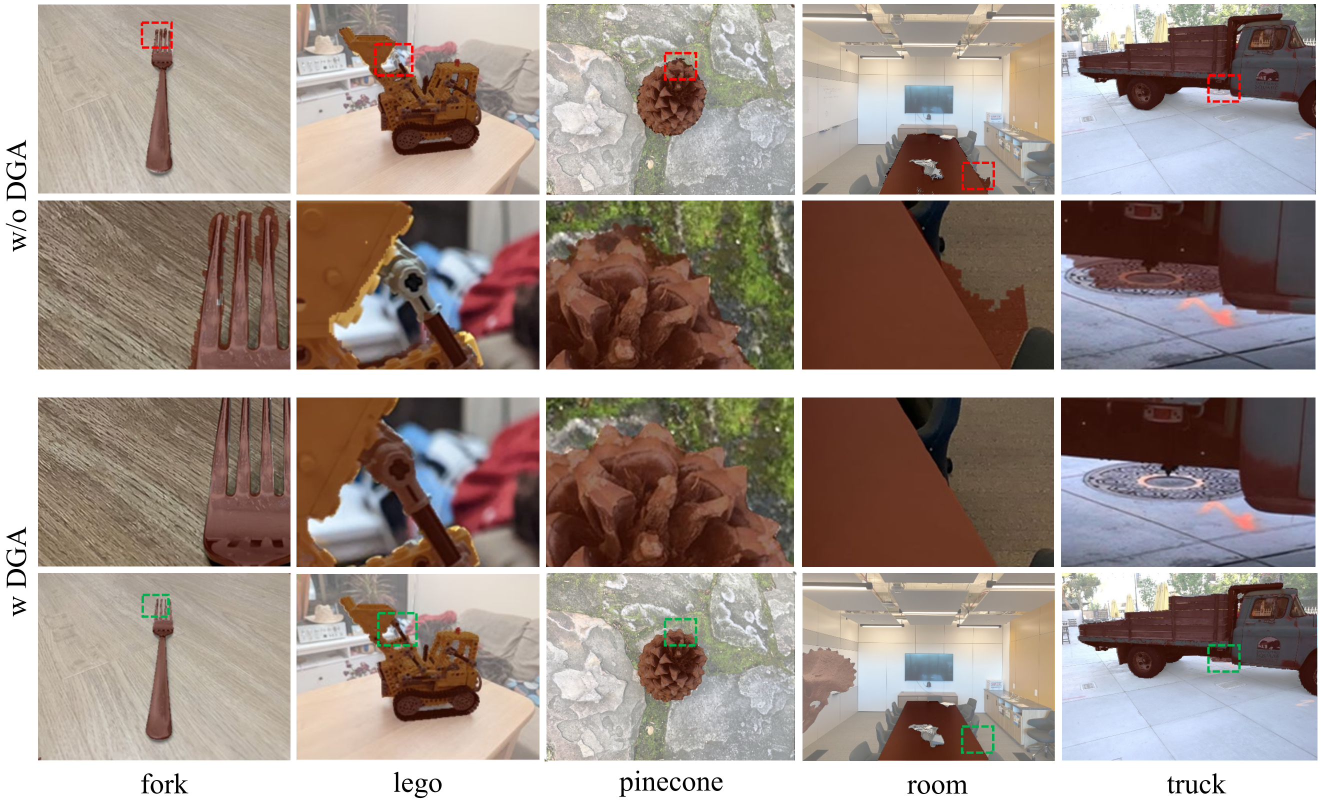

Figure 8. **Visualization of ablation experiment on the SPIn-NeRF dataset.**

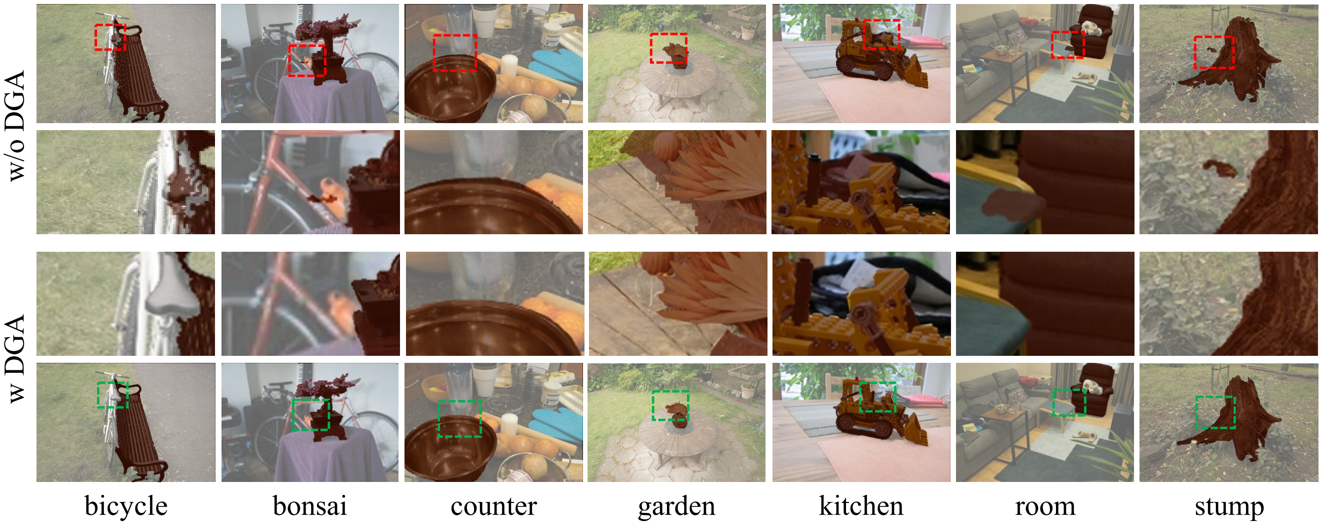

Figure 9. **Visualization of ablation experiment on the Mip-NeRF360 dataset.**

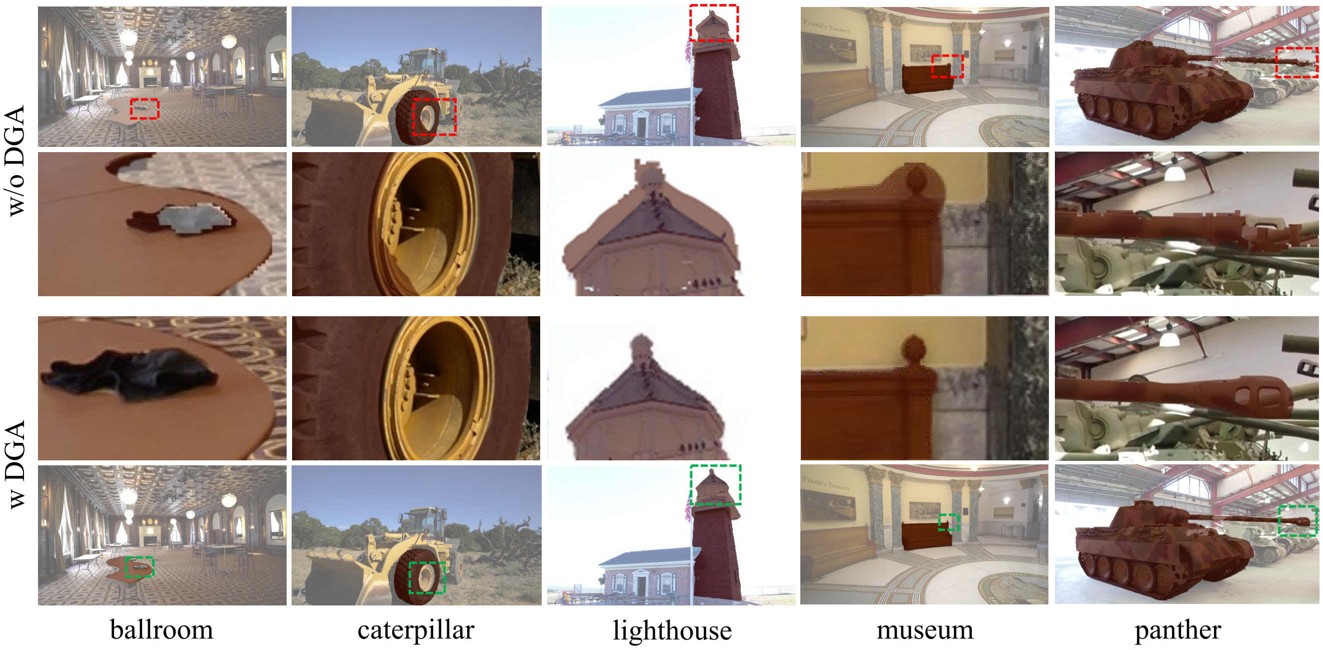

Figure 10. **Visualization of ablation experiment on the T&T dataset.**

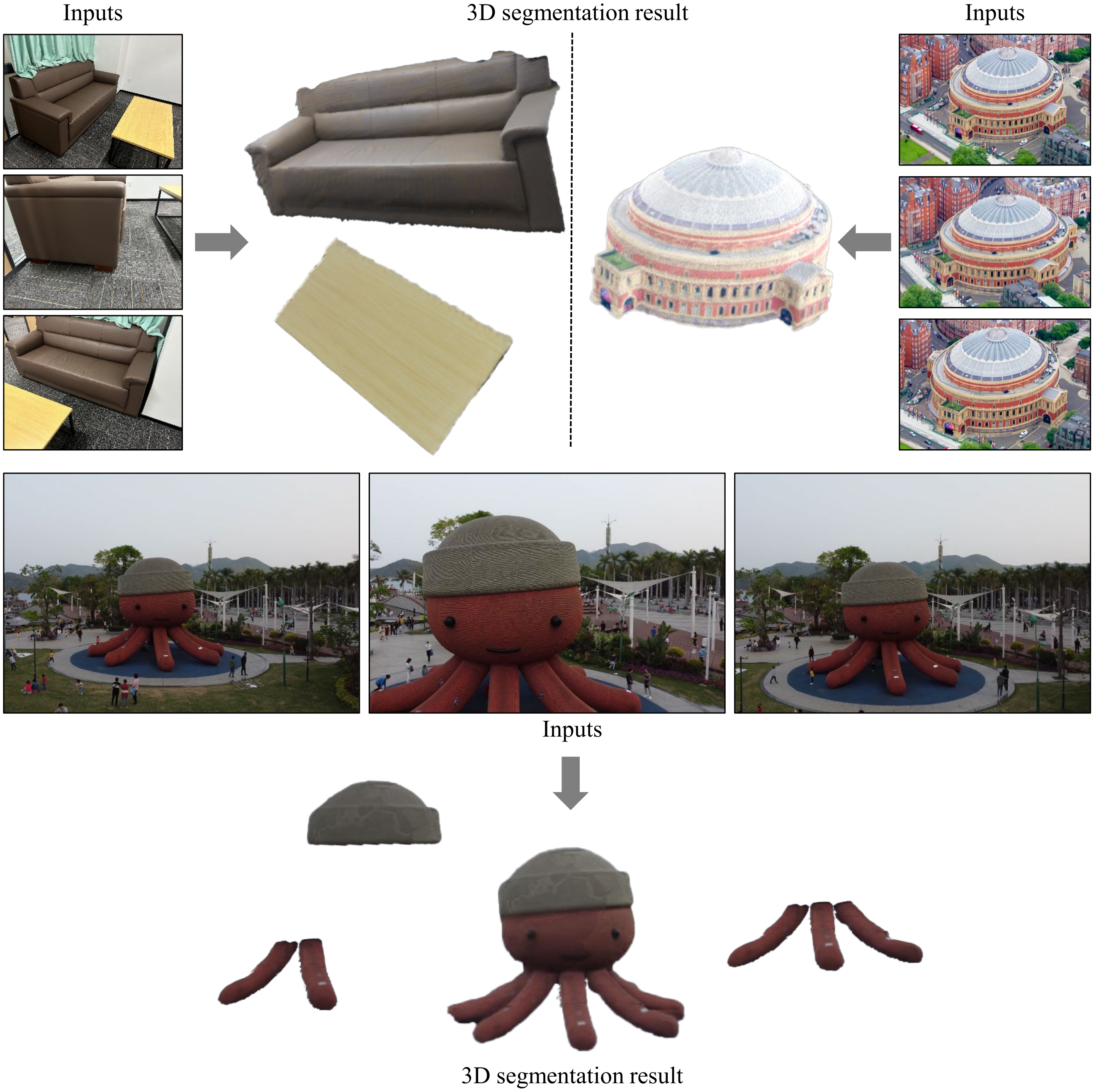

Figure 11. Visualization of WildSeg3D's segmentation results on scenes in the wild with highly sparse views.