Amodal Depth Anything: Amodal Depth Estimation in the Wild

Zhenyu Li[1], Mykola Lavreniuk[2], Jian Shi[1], Shariq Farooq Bhat[1], Peter Wonka[1]
[1]KAUST, [2]Space Research Institute NASU-SSAU

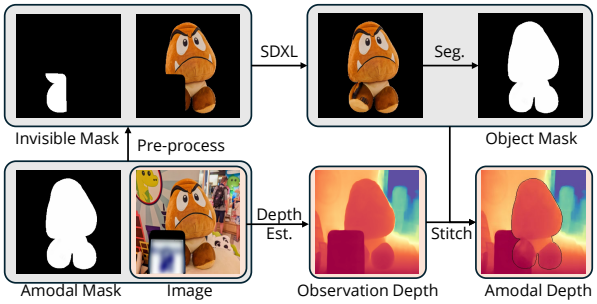

Figure 1. **Invisible Stitch for Amodal Depth.**

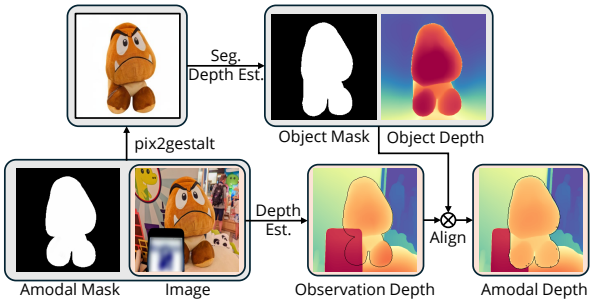

Figure 2. **Pix2gestalt Stitch for Amodal Depth.**

1. Baseline

In this section, we describe how two inpainting-based methods, Invisible Stitch [1] and pix2gestalt [3], are adapted as potential solutions for the amodal depth estimation task.

1.1. Invisible Stitch

As illustrated in Fig. 1, given an input image and the target amodal mask, we first use SD-XL [4] to inpaint the invisible parts of the target object. Note that generating satisfactory inpainting results requires an accurate textual description. For example, we use the textual prompt *"a stuffed toy of an angry-looking character on display"* for the case shown in Fig. 1. Next, RMBG-1.4 is applied to create the object mask. Finally, the Invisible Stitch model [1] estimates the observation depth based on the input image and generates the corresponding amodal depth map.

Figure 3. **Reconstructed 3D Mesh for Occluded Object.** Blue arrows indicate the target object and red arrows highlight the reconstructed meshes for occluded parts of objects, respectively. Left: Input image. Middle: Mesh from general depth. Right: Reconstructed mesh combining general depth and amodal depth.

1.2. pix2gestalt

The pipeline for adapting pix2gestalt for amodal depth estimation is shown in Fig. 2. Given the input image and the target amodal mask, pix2gestalt first inpaints the invisible parts of the target object. Before processing, the amodal mask is preprocessed into the visible mask for the target object. Subsequently, Depth-Anything V2 ViT-G [5] is used to estimate the observation and amodal depth for both the original image and the inpainting result. Finally, the depth maps are aligned using shared visible regions, producing the final amodal depth prediction.

2. Amodal 3D Reconstruction

Using a depth map and the corresponding camera intrinsic parameters, we can reconstruct the 3D point cloud and

Method	RMSE↓
Jo *et al*. [2]	0.1534
Amodal DAV2-L	0.1436

Table 1. **Comparisons on Front-3D.** Amodal DAV2-L achieves a better performance even in a zero-shot manner, indicating the effectiveness and generalization of our method.

Method	RMSE↓	$\delta(\%)\uparrow$
Jo *et al*. [2]	0.1744	96.759
Amodal DAV2-L	0.1339	98.234

Table 2. **Comparisons on Amodal NYU-V2.** We evaluate both our Amodal DAV2-L and Jo *et al*.'s model on Amodal NYU-V2 in the same zero-shot manner.

convert it into a 3D mesh. As shown in Fig. 3, general depth estimation methods only account for visible pixels, leaving holes in areas corresponding to occluded regions. Our amodal depth estimation provides a simple yet effective solution by predicting reasonable geometry for the invisible parts of objects. This approach can serve as a valuable prior for 3D reconstruction tasks and generative applications, such as novel view synthesis and inpainting.

3. More Experiments

Comparison on Amodal-Front3D: In this experiment, we assess the performance of our Amodal-DAV2-L on the Amodal-Front3D dataset [2] in a zero-shot setting and compare it to the results reported by Jo *et al*. [2]. We align the output of our model with the metric ground-truth depth by the scale-and-shift alignment. As indicated in Tab. 1, despite the fact that their model is specifically trained on the Amodal-Front3D dataset (thus benefiting from an in-domain setting), our model achieves a 6.3% reduction in RMSE, consistently outperforming theirs. This significant improvement highlights the superior effectiveness and generalization capability of our method.

Comparison on Amodal NYU-V2: In this experiment, we generate amodal pairs from the original NYU-V2 dataset, creating the Amodal NYU-V2. We then compare our Amodal-DAV2-L to Jo *et al*.'s officially released model [2] using the same zero-shot inference approach. Both methods adopt the scale-and-shift alignment with the ground-truth depth to ensure a fair comparison. As shown in Tab. 2, our model achieves a 23.2% reduction in RMSE, underscoring the effectiveness of our approach.

4. Limitation and Future Work

While our methods demonstrate promising potential for amodal depth estimation, several limitations remain to be addressed in future work. First, our model's reliance on the input amodal mask, where inaccurate or ambiguous masks can propagate errors and lead to cascading failures in the amodal depth estimation. Additionally, our framework requires an observed depth map as guidance. Consequently, a forward pass through a depth model is necessary before applying our framework to obtain the amodal depth, which introduces additional computational costs.

Future directions could extend the single-frame framework to handle videos and develop a unified framework capable of predicting amodal segmentation, RGB, depth, surface normals, and more. These advancements could further enhance the scope and utility of amodal depth estimation.

References

[1] Paul Engstler, Andrea Vedaldi, Iro Laina, and Christian Rupprecht. Invisible stitch: Generating smooth 3d scenes with depth inpainting. *arXiv preprint arXiv:2404.19758*, 2024. 1

[2] Seong-Uk Jo, Du Yeol Lee, and Chae Eun Rhee. Occlusion-aware amodal depth estimation for enhancing 3d reconstruction from a single image. *IEEE Access*, 2024. 2

[3] Ege Ozguroglu, Ruoshi Liu, Dídac Surís, Dian Chen, Achal Dave, Pavel Tokmakov, and Carl Vondrick. pix2gestalt: Amodal segmentation by synthesizing wholes. In *CVPR*, pages 3931–3940. IEEE Computer Society, 2024. 1

[4] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 1

[5] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *arXiv preprint arXiv:2406.09414*, 2024. 1