PixARMesh: Autoregressive Mesh-Native Single-View Scene Reconstruction

Supplementary Material

1. Additional Implementation Details

1.1. Data Pre-processing

We pre-process meshes from the 3D-FRONT [5] dataset to ensure they are suitable for autoregressive tokenization.

- **Vertex Merging** We merge nearby vertices using a minimum spatial resolution determined by the quantization level $q \in 128, 256, 512, 1024$, where the minimum resolution is $1/q$.
- **Mesh Simplification** We apply planar decimation followed by quadric-error-based edge-collapse decimation with target face counts of 800, 2000, and 4000.
- **Quality Selection** We compute the Hausdorff distance between the simplified meshes and the originals, thresholding with an empirical value $\tau = 0.01$. If the distance is below τ, we choose the mesh with fewer faces; otherwise, we select the quantization level and decimation setting that produce the lowest Hausdorff distance. This procedure balances compactness and geometric fidelity.

The resulting processed dataset contains meshes with an average of 1,809 faces per shape.

1.2. Training Details

To enhance layout prediction accuracy, we employ a two-stage training strategy:

- **Layout Bootstrapping** We first train the model exclusively on the layout prediction prefix tokens. This stage consists of 100k iterations for the EdgeRunner-based variant and 30k iterations for the BPT-based model.
- **Joint Pose-Mesh Training** Building on the bootstrapped weights, we train the model on the full pose and mesh sequence. This final stage requires 30k iterations for EdgeRunner and 25k iterations for BPT.

Both stages utilize the AdamW optimizer with an initial learning rate of 1×10^{-4}, featuring a 500-step linear warmup followed by a cosine decay to 1×10^{-5}. Training is conducted with an effective batch size of 64. We set the optimizer hyperparameters to $\beta_1 = 0.9$ and $\beta_2 = 0.95$, and apply gradient clipping at a threshold of 1.0.

2. Additional Results

2.1. More Qualitative Results on Real Images

We present additional qualitative results on real indoor scenes from Pix3D [8], Matterport3D [1] and ScanNet [3] in Fig. 1. Across diverse environments and lighting conditions, PixARMesh generally achieves stronger layout alignment and produces coherent scene reconstructions, while retaining the advantages of compact, artist-ready mesh outputs.

2.2. Face and Vertex Counts

Method	$\|F\|$	$\|V\|$
InstPIFu [7]	1,942,029	970,951
Uni-3D [11]	141,130	70,844
BUOL [2]	55,493	27,750
Gen3DSR [4]	364,285	216,797
DeepPriorAssembly [13]	250,875	125,420
MIDI [6]	1,936,248	967,886
DepR [12]	319,646	159,905
PixARMesh-EdgeRunner (Ours)	7,110	4,253
PixARMesh-BPT (Ours)	7,506	4,050

Table 1. Face count $|F|$ and vertex count $|V|$ comparisons.

As shown in Tab. 1, we report the average number of faces and vertices per reconstructed scene across different methods. InstPIFu [7] and MIDI [6] produce extremely dense outputs, requiring roughly 2M faces per scene. BUOL [2], Gen3DSR [4], and DeepPriorAssembly [13] reduce this to about 0.3M faces, though still far from mesh-efficient. In contrast, PixARMesh produces dramatically more compact meshes, *i.e.* around 7–8k faces depending on the base model, while maintaining competitive geometric quality. This highlights the practical advantage of generating artist-ready, natively structured meshes rather than relying on iso-surface extraction.

2.3. Object Pose (Layout) Accuracy

Method	GT Depth	Box IoU (%)
MIDI [6]	–	31.07
DepR [12]		33.98
DepR [12]	✓	36.67
PixARMesh-EdgeRunner (Ours)		56.76
PixARMesh-EdgeRunner (Ours)	✓	70.37
PixARMesh-BPT (Ours)		54.54
PixARMesh-BPT (Ours)	✓	65.01

Table 2. Layout accuracy comparisons.

Following MIDI [6], we evaluate scene layout accuracy using the 3D bounding-box IoU between predictions and ground truth. For this setting, we use ground-truth instance masks to isolate layout estimation from segmentation errors. Results are reported in Tab. 2. PixARMesh achieves higher layout accuracy than prior methods that rely on predicted depth, and we observe that the EdgeRunner-based variant consistently produces more accurate layouts than the BPT-based model.

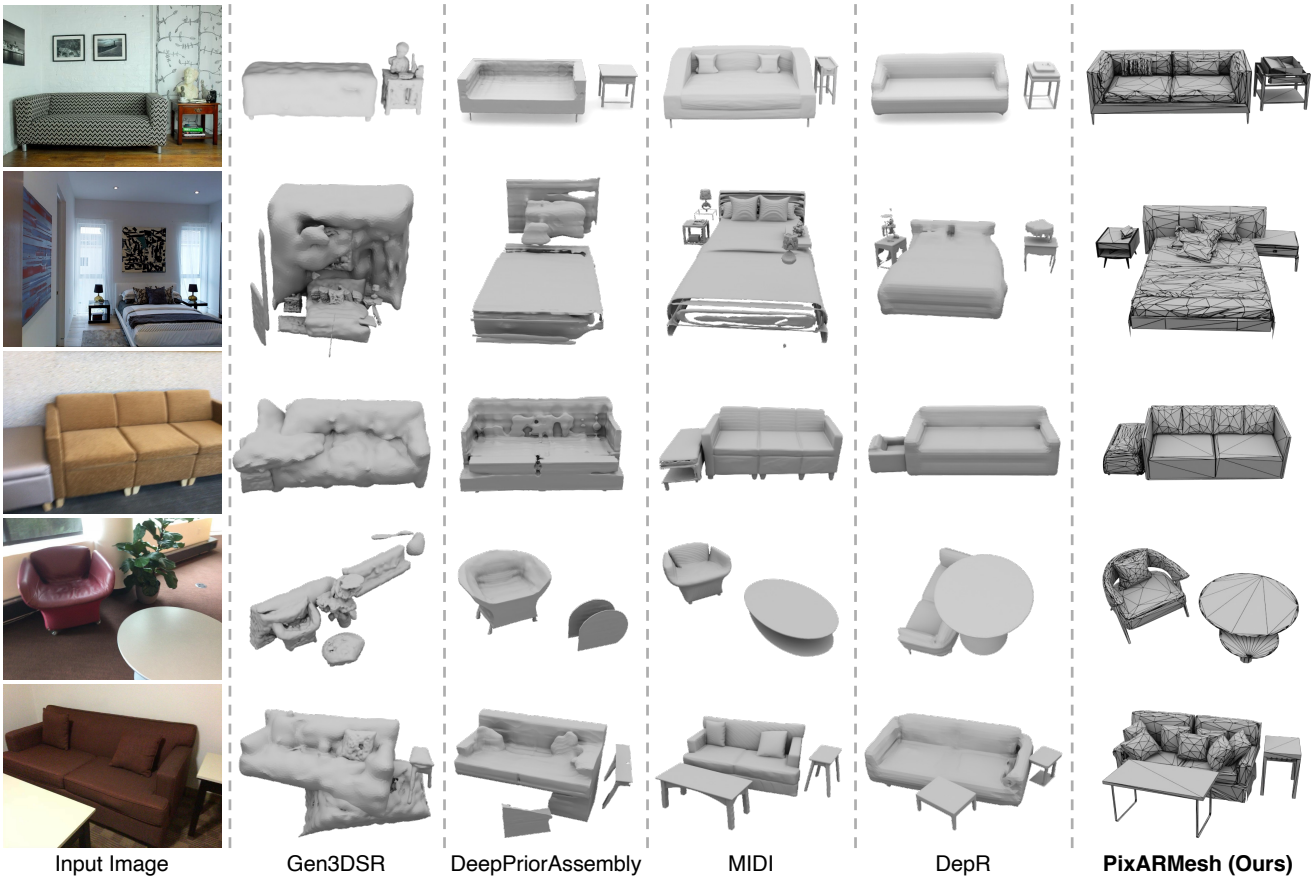

Figure 1. Additional qualitative results on real images from Pix3D [8], Matterport3D [1] and ScanNet [3].

2.4. Runtime Analysis

We measure inference runtime on a single NVIDIA A100 GPU and report the average per-scene latency in Tab. 3. Due to its autoregressive decoding process, PixARMesh is inherently slower than feed-forward approaches (InstPIFu [7], Uni-3D [11], BUOL [2]) as well as the latent-diffusion-based DepR [12]. Nonetheless, PixARMesh remains generally faster than other compositional pipelines such as Gen3DSR [4], while uniquely providing artist-ready native mesh outputs without needing iso-surface extraction.

Method	Runtime
InstPIFu [7]	19.8 s
Uni-3D [11]	3.1 s
BUOL [2]	6.5 s
Gen3DSR [4]	15.5 min
DeepPriorAssembly [13]	6.7 min
MIDI [6]	1.6 min
DepR [12]	1.2 min
PixARMesh-EdgeRunner (Ours)	4.5 min
PixARMesh-BPT (Ours)	6.7 min

Table 3. Inference runtime comparisons.

2.5. BPT-Based Ablation Studies

To further assess the generality of our point-cloud encoder designs, we conduct additional ablations using the BPT-based mesh generative model [10], as summarized in Tab. 4. The trends mirror those observed with EdgeRunner [9]: incorporating scene-level context aggregation without pixel-aligned features slightly reduces object-level fidelity but improves global alignment in the assembled scene. Introducing image features provides a substantial boost in scene-level accuracy, and adding our scene-level context aggregation on top yields further gains. These results confirm that our proposed encoder designs remain effective across different mesh tokenizations and generative backbones.

Img Feat	Ctx Agg	Scene-level CD ($\times 10^{-3}, \downarrow$)	Scene-level CD-S ($\times 10^{-3}, \downarrow$)	Scene-level F-Score (%, $\uparrow$)	Object-level CD ($\times 10^{-3}, \downarrow$)	Object-level F-Score (%, $\uparrow$)
		68.36	25.43	35.96	5.34	76.54
	✓	56.80	23.49	37.98	5.85	76.31
✓		57.03	23.70	41.51	**4.25**	**81.25**
✓	✓	**49.82**	**22.06**	**42.18**	4.57	80.30

Table 4. Ablation studies on our point-cloud encoder design with BPT-based model. *Img Feat*, *Ctx Agg* denote pixel-aligned image features and scene context aggregation, respectively.

References

[1] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niebner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. In *2017 International Conference on 3D Vision*, pages 667–676. IEEE Computer Society, 2017. 1, 2

[2] Tao Chu, Pan Zhang, Qiong Liu, and Jiaqi Wang. Buol: A bottom-up framework with occupancy-aware lifting for panoptic 3d scene reconstruction from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4937–4946, 2023. 1, 2

[3] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5828–5839, 2017. 1, 2

[4] Andreea Dogaru, Mert Özer, and Bernhard Egger. Gen3DSR: Generalizable 3d scene reconstruction via divide and conquer from a single view. In *International Conference on 3D Vision 2025*, 2025. 1, 2

[5] Huan Fu, Bowen Cai, Lin Gao, Ling-Xiao Zhang, Jiaming Wang, Cao Li, Qixun Zeng, Chengyue Sun, Rongfei Jia, Binqiang Zhao, and Hao Zhang. 3d-front: 3d furnished rooms with layouts and semantics. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10933–10942, 2021. 1

[6] Zehuan Huang, Yuan-Chen Guo, Xingqiao An, Yunhan Yang, Yangguang Li, Zi-Xin Zou, Ding Liang, Xihui Liu, Yan-Pei Cao, and Lu Sheng. Midi: Multi-instance diffusion for single image to 3d scene generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23646–23657, 2025. 1, 2

[7] Haolin Liu, Yujian Zheng, Guanying Chen, Shuguang Cui, and Xiaoguang Han. Towards high-fidelity single-view holistic reconstruction of indoor scenes. In *European Conference on Computer Vision*, pages 429–446. Springer, 2022. 1, 2

[8] Xingyuan Sun, Jiajun Wu, Xiuming Zhang, Zhoutong Zhang, Chengkai Zhang, Tianfan Xue, Joshua B Tenenbaum, and William T Freeman. Pix3d: Dataset and methods for single-image 3d shape modeling. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2974–2983, 2018. 1, 2

[9] Jiaxiang Tang, Zhaoshuo Li, Zekun Hao, Xian Liu, Gang Zeng, Ming-Yu Liu, and Qinsheng Zhang. Edgerunner: Auto-regressive auto-encoder for artistic mesh generation. In *The Thirteenth International Conference on Learning Representations*, 2025. 2

[10] Haohan Weng, Zibo Zhao, Biwen Lei, Xianghui Yang, Jian Liu, Zeqiang Lai, Zhuo Chen, Yuhong Liu, Jie Jiang, Chunchao Guo, et al. Scaling mesh generation via compressive tokenization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11093–11103, 2025. 2

[11] Xiang Zhang, Zeyuan Chen, Fangyin Wei, and Zhuowen Tu. Uni-3d: A universal model for panoptic 3d scene reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9256–9266, 2023. 1, 2

[12] Qingcheng Zhao, Xiang Zhang, Haiyang Xu, Zeyuan Chen, Jianwen Xie, Yuan Gao, and Zhuowen Tu. Depr: Depth guided single-view scene reconstruction with instance-level diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5722–5733, 2025. 1, 2

[13] Junsheng Zhou, Yu-Shen Liu, and Zhizhong Han. Zero-shot scene reconstruction from single images with deep prior assembly. *Advances in Neural Information Processing Systems*, 37:39104–39127, 2024. 1, 2