EscherNet++: Simultaneous Amodal Completion and Scalable View Synthesis through Masked Fine-Tuning and Enhanced Feed-Forward 3D Reconstruction

Xinan Zhang[1] Muhammad Zubair Irshad[2] Anthony Yezzi[1] Yi-Chang Tsai[1] Zsolt Kira[1]
Georgia Institute of Technology[1] Toyota Research Institute[2]

Abstract

We propose EscherNet++, a masked fine-tuned diffusion model that can synthesize novel views of objects in a zero-shot manner with amodal completion ability. Existing approaches utilize multiple stages and complex pipelines to first hallucinate missing parts of the image and then perform novel view synthesis, which fail to consider cross-view dependencies and require redundant storage and computing for separate stages. Instead, we apply masked fine-tuning including input-level and feature-level masking to enable an end-to-end model with the improved ability to synthesize novel views and conduct amodal completion. In addition, we empirically integrate our model with other feed-forward image-to-mesh models without extra training and achieve competitive results with reconstruction time decreased by 95%, thanks to its ability to synthesize arbitrary query views. Our method's scalable nature further enhances fast 3D reconstruction. Despite fine-tuning on a smaller dataset and batch size, our method achieves state-of-the-art results, improving PSNR by 3.9 and Volume IoU by 0.28 on occluded tasks in 10-input settings, while also generalizing to real-world occluded reconstruction.

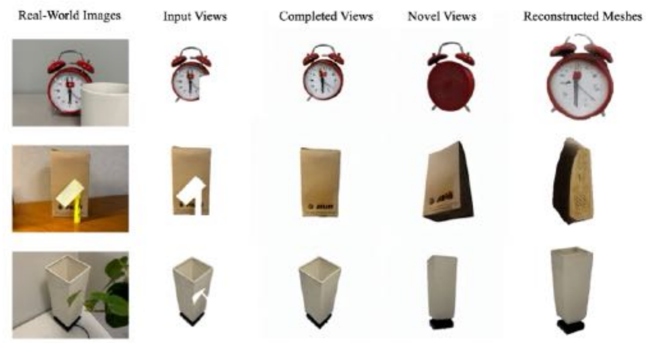

Figure 1. Given occluded input views of any number, our unified model EscherNet++ is able to complete occluded views and synthesize novel views simultaneously, without need for multiple specialized models. Synthesized views can be queried from any viewpoints, which allows instant integration with other feed-forward 3D reconstruction models [39]. It can be generalizable to unseen data such as real-world captures.

1. Introduction

Novel view synthesis (NVS) of objects is an important topic in computer vision due to its wide range of applications, including virtual and augmented reality [18, 26], computer graphics [6], robotics [13, 42] and 3D reconstruction [11, 19, 23] . It involves generating new images of an object from viewpoints that were not observed during data capture, enabling more immersive and interactive experiences. Recent progresses represented by nueral radiance field (NeRF) [24], have achieved high-quality results by modeling the scene as a continuous volumetric function using a neural network. However NeRF and its following works come with several limitations that hinder their practical application, including 1) slow training/rendering speeds, 2) limited extrapolation/few-shot/generalization ability and 3) inability to handle occlusion well.

Various methods have been proposed to alleviate these problems while maintaining the quality of the synthesis, such as grid-based methods [25], point-based methods [16], incorporation of learned prior knowledge [12, 40]. In addition, Diffusion methods [8, 29, 31], which are a group of generative models previously used in content generation, began to gain popularity in NVS [10, 15, 17, 20–22, 30, 34, 38, 41]. Among these methods, EscherNet [17] stands out for its ability to generate high-quality consistent views and support multiple inputs as the condition. Besides, diffusion models have also been used in amodal completion to deal with occlusion [1, 3, 27]; however, current amodal completion models typically serve as a stand-alone model and primarily focus on single-view context.

Departing from existing approaches that often treat these tasks separately [3, 27], we ask *"Can these two problems be solved with a more integrated solution?"* Such a solution should be able to 1) leverage a shared understanding of object semantics and geometry from the input views with possible occlusions and 2) possess the ability to be optimized collectively for both tasks. These requirements moti-

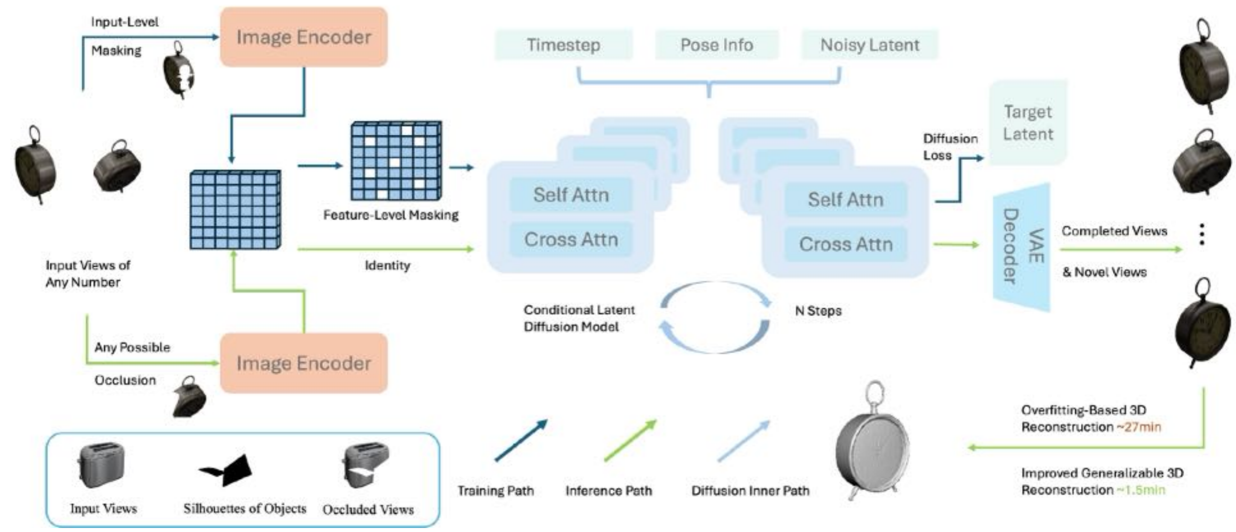

Figure 2. **The pipeline of EscherNet++**. Our unified model enables simultaneous novel view synthesis and amodal completion. During training, hierarchical masking—at both input and feature levels—helps the model learn complete geometry from occluded views while improving robustness. During inference, our model not only supports commonly used overfitting approaches—such as NueS [36], which iteratively refines geometry—but also seamlessly integrates with pre-trained feed-forward models like InstantMesh [39]. We empirically find that this integration achieves competitive performance while significantly reducing computational time. Bottom right corner shows input-level masking is applied. Silhouettes are extracted from rendered objects and overlayed on complete input views to get occluded views paired with groundtruth.

vate the development of our propose method **EscherNet++** as follows:

- We propose a unified diffusion-based network **EscherNet++** as shown in Fig. 2, designed for **occlusion-aware novel view synthesis**. It flexibly adapts to varying numbers of input and output views, extending the original task for **multi-view amodal completion**—a challenging yet underexplored task.
- Introduce an effective approach to **enhance fast 3D reconstruction using pre-trained feed-forward models**, leveraging the scalability and consistency of our synthesized novel views without requiring additional fine-tuning
- Our proposed work excels in extensive experiments on NVS and 3D reconstruction, particularly **under occlusions**, outperforming prior work by an average of 3.9 PSNR in occluded NVS tests and 0.28 Volume IoU in occluded 3D reconstruction tests with 10-input settings.

2. Methodology

We introduce **EscherNet++**, detailed in this section. We first introduce our masked fine-tuning approach in Sec. 2.1, and our view-to-3D reconstruction method in Sec. 2.2. An overview of the pipeline is illustrated in Fig. 2.

2.1. Masked Fine-Tuning

Built upon EscherNet, we aim to achieve an end-to-end model that can synthesize novel views and complete the occluded regions in input views simultaneously. There are two key aspects to consider when tackling the compound problem, 1) dataset acquisition and 2) training method.

Curated Dataset: A well-structured dataset is crucial for training a model to handle the problem effectively. The requirements on the dataset lead us to create a paired dataset curated from Objaverse-1.0 [2].We employ silhouettes of objects as masks to randomly overlay occlusions onto objects in the dataset, as shown in Fig. 2. Specifically, we sampled single objects from then rendered Objaverse dataset to extract their silhouettes. Then we group, rescale, shift them to create various occlusions.

Input-Level & Feature-Level Masking: We fine-tune the model using two techniques, input-level masking and feature-level masking. Input level masking can be achieved with the above curated dataset naturally. Similar to the original training of Eschernet, we randomly choose three input views with 50 percent chance of being partially occluded, the model learns to synthesize novel three other complete views. In addition, inspired by previous works [5, 7, 14, 37], we propose to further randomly mask the encoded input feature maps to further improve the performance, as shown in Sec. 3 by strengthening model's ability in overall comprehension of the object in semantics and intricate structure details. We empirically found that 25 percent is a suitable choice for feature-level masking probability as shown in App. B. That is, around 1/4 of input data

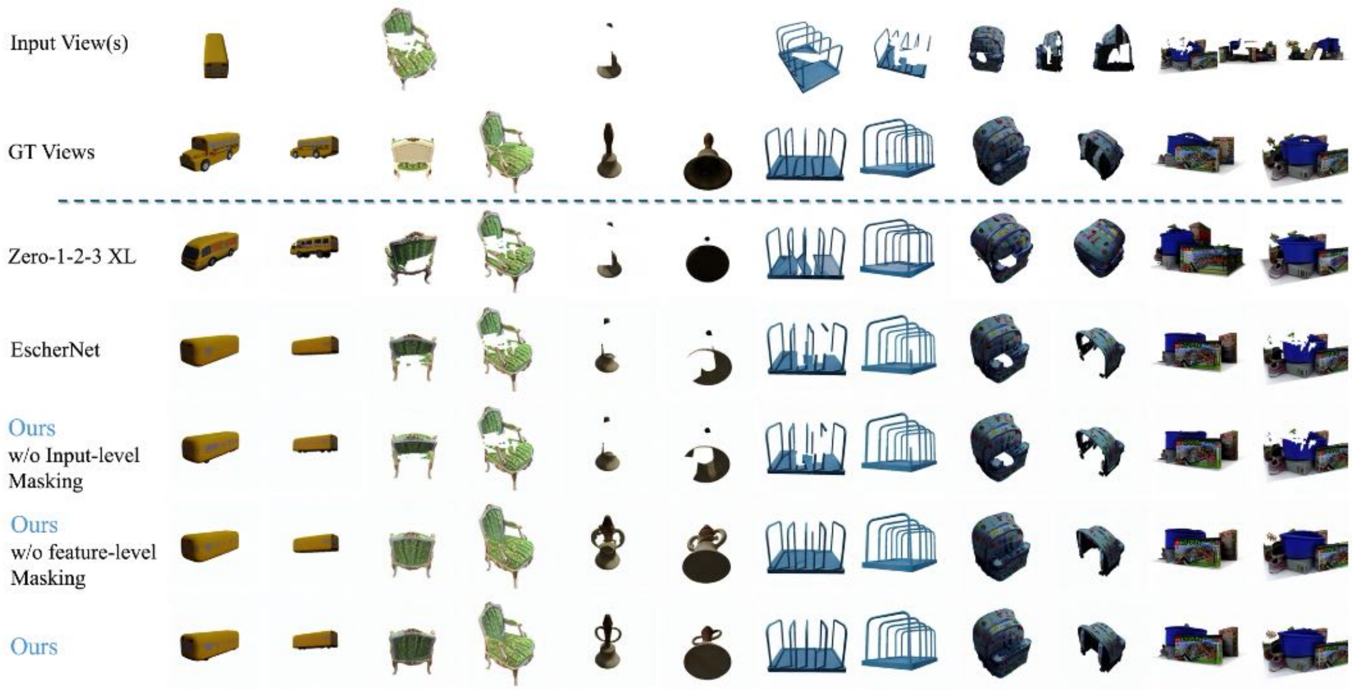

Figure 3. Visualization of synthesized views from different models with our OccNVS benchmark.

will be processed by random feature-level masking during training.

2.2. Novel View Synthesis to 3D reconstruction

Reconstructing objects from synthesized novel views is a crucial downstream task. Broadly, two main approaches exist: 1) Overfitting methods, where a model is trained per object, and 2) Generalizable models, which learn a universal 3D representation applicable across objects with minimal adaptation. We experiment with both methods and propose a simple yet effective way to enhance a feed-forward generalizable model in a training-free manner.

2.2.1. Overfitting Method

The prior work EscherNet opts to train separate NeuS [36] models for each object, which is able to memorize the details of a particular object by overfitting, leading to highly accurate and detailed reconstruction. Such overfitting method can yield high-quality reconstruction however they usually involve extensive per-object training as shown in Sec. 3.

2.2.2. Generalizable Method

There have been several feed-forward generalizable reconstruction models available in recent years [9, 28, 32, 33, 39], designed to quickly infer 3D representations from sparse inputs such as single view or a few views. We pick one generalizable model, InstantMesh [39] for case study in this paper. It is found InstantMesh performs worse when given inputs from poses other than those used in their paper, although their model design supports any input poses.

Target View Synthesis: Luckily, we can take advantage of our model that can generate any view from any query pose to generate preferred views for generalizable reconstruction models. We further find that performance can be elevated if more generated views can be provided to the reconstruction model. No additional training or extra inference time is introduced as we show in Sec. 3 and App. C.

3. Experiments

Experiments are conducted to compare our proposed method EscherNet++ with other state-of-the-art methods.

Training & Test Settings: Objaverse-1.0 is used to train our models. Specifically, a subset of 300K objects is sampled from Objaverse-1.0 for faster fine-tuning and data-efficient purposes. A small learning rate of $1 \cdot 10^{-5}$ is used for fine-tuning weights from the public checkpoint of EscherNet. A batch size of 48 is adopted on each of 8 A40 GPUs, it takes around 3 days to complete 28K iterations. 4DoF object-centric setting is set for all experiments. We evaluate all the models with two settings, one with complete input views and one with randomly occluded views with a new set of masks to simulate any possible occlusions from query viewpoints. We term the occluded benchmark **OccNVS**, including complete/occluded views from Google Scanned Objects dataset (GSO) [4], RTMV and NeRF Synthetic [24]. The structure of EscherNet++ and other settings

Figure 4. Amodal complation results by different models on OccNVS.

are kept the same as EscherNet.

We conduct three sets of experiments with OccNVS in this section, including NVS, amodal completion and 3D reconstruction. Quantitative results can be found in App. A.

Results on Novel View Synthesis: Experiments in Tab. 1 in occluded tests show that our model successfully achieves the intended goal of synthesizing complete novel views even with occlusion in input views, with semantic accuracy and geometric consistency well maintained. our model in tasks with occlusion significantly outperforms baselines by improvement of at least 5 in PSNR for GSO in all settings over EscherNet.

Results on Amodel Completion: We also compare the amodel completion performance of our model with two other recent models specifically designed for this task. As shown in Fig. 4 and Tab. 2, our model stands out for its distinct ability to consider multi-view reference in amodel completion.

Results on 3D Reconstruction: We evaluate 3D reconstruction quality across various models, normalizing mesh outputs for comparison, with two image-to-3D reconstruction approaches: an overfitting method (NeuS [36]) and a feed-forward model [39]. Qualitative and quantitative results are presented in Fig.5 and Tab.3. For our model, 36 synthesized views serve as inputs for NeuS-based reconstruction, while an additional 6 views are used for InstantMesh, totaling 42 views for enhanced reconstruction.

By synthesizing more consistent and precise views, our model and EscherNet outperform prior methods when

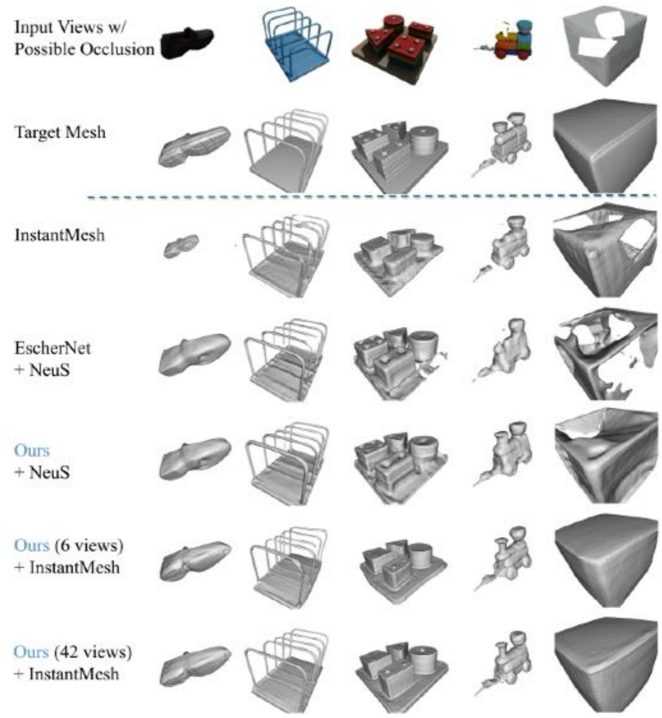

Figure 5. Rendered meshes from 3D reconstruction by different models on OccNVS benchmark. Note that a floater occurs in the first example with InstantMesh.

paired with NeuS under both settings (Fig.5, Tab.3). Further, it enables seamless integration with pre-trained feed-forward 3D reconstruction models. We validate this by integrating InstantMesh, achieving over a 10% increase in volume IoU by providing more accurate views at the same viewpoints at occluded settings, with reconstruction time reduced by 95% while maintaining competitive performance.

4. Conclusion

In this paper, we propose EscherNet++, a masked fine-tuned diffusion model that can synthesize novel views of objects in a zero-shot way with amodal completion ability. We find that properly masked input images and input feature maps can contribute to better performance of the model. In addition, it can be seamlessly integrated with other fast feed-forward image-to-mesh models because of its flexible feature to synthesize any query views without the need for extra training, and the fast 3D reconstruction performance can be further boosted by its scalable nature. Limitations of the current work as well as future work can be found in App. D.

References

[1] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402, 2023. 1, 3

[2] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13142–13153, 2023. 2

[3] Andreea Dogaru, Mert Özer, and Bernhard Egger. Generalizable 3d scene reconstruction via divide and conquer from a single view. *arXiv:2404.03421*, 2024. 1, 3

[4] Laura Downs, Anthony Francis, Nate Koenig, Brandon Kinman, Ryan Hickman, Krista Reymann, Thomas B McHugh, and Vincent Vanhoucke. Google scanned objects: A high-quality dataset of 3d scanned household items. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 2553–2560. IEEE, 2022. 3

[5] Shanghua Gao, Pan Zhou, Ming-Ming Cheng, and Shuicheng Yan. Mdtv2: Masked diffusion transformer is a strong image synthesizer. *arXiv preprint arXiv:2303.14389*, 2023. 2

[6] Ayaan Haque, Matthew Tancik, Alexei Efros, Aleksander Holynski, and Angjoo Kanazawa. Instruct-nerf2nerf: Editing 3d scenes with instructions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 1

[7] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 2

[8] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 1

[9] Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. Lrm: Large reconstruction model for single image to 3d. *arXiv preprint arXiv:2311.04400*, 2023. 3

[10] Zehuan Huang, Hao Wen, Junting Dong, Yaohui Wang, Yangguang Li, Xinyuan Chen, Yan-Pei Cao, Ding Liang, Yu Qiao, Bo Dai, et al. Epidiff: Enhancing multi-view synthesis via localized epipolar-constrained diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9784–9794, 2024. 1

[11] Muhammad Zubair Irshad, Sergey Zakharov, Rares Ambrus, Thomas Kollar, Zsolt Kira, and Adrien Gaidon. Shapo: Implicit representations for multi-object shape appearance and pose optimization. 2022. 1

[12] Muhammad Zubair Irshad, Sergey Zakharov, Katherine Liu, Vitor Guizilini, Thomas Kollar, Adrien Gaidon, Zsolt Kira, and Rares Ambrus. Neo 360: Neural fields for sparse view synthesis of outdoor scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9187–9198, 2023. 1

[13] Muhammad Zubair Irshad, Mauro Comi, Yen-Chen Lin, Nick Heppert, Abhinav Valada, Rares Ambrus, Zsolt Kira, and Jonathan Tremblay. Neural fields in robotics: A survey, 2024. 1

[14] Muhammad Zubair Irshad, Sergey Zakharov, Vitor Guizilini, Adrien Gaidon, Zsolt Kira, and Rares Ambrus. Nerf-mae: Masked autoencoders for self-supervised 3d representation learning for neural radiance fields. In *European Conference on Computer Vision (ECCV)*, 2024. 2

[15] Yash Kant, Aliaksandr Siarohin, Ziyi Wu, Michael Vasilkovsky, Guocheng Qian, Jian Ren, Riza Alp Guler, Bernard Ghanem, Sergey Tulyakov, and Igor Gilitschenski. Spad: Spatially aware multi-view diffusers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10026–10038, 2024. 1

[16] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023. 1

[17] Xin Kong, Shikun Liu, Xiaoyang Lyu, Marwan Taher, Xiaojuan Qi, and Andrew J Davison. Eschernet: A generative model for scalable view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9503–9513, 2024. 1, 2, 3

[18] Zhengfei Kuang, Kyle Olszewski, Menglei Chai, Zeng Huang, Panos Achlioptas, and Sergey Tulyakov. Neroic: Neural rendering of objects from online image collections. *ACM Transactions on Graphics (TOG)*, 41(4):1–12, 2022. 1

[19] Zhaoshuo Li, Thomas Müller, Alex Evans, Russell H Taylor, Mathias Unberath, Ming-Yu Liu, and Chen-Hsuan Lin. Neuralangelo: High-fidelity neural surface reconstruction. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 1

[20] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9298–9309, 2023. 1, 2

[21] Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. Syncdreamer: Generating multiview-consistent images from a single-view image. *arXiv preprint arXiv:2309.03453*, 2023. 1, 3

[22] Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, et al. Wonder3d: Single image to 3d using cross-domain diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9970–9980, 2024. 1

[23] Mayank Lunayach, Sergey Zakharov, Dian Chen, Rares Ambrus, Zsolt Kira, and Muhammad Zubair Irshad. Fsd: Fast self-supervised single rgb-d to categorical 3d objects. In *Int. Conf. on Robotics and Automation*. IEEE, 2024. 1

[24] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view syn-

thesis. *Communications of the ACM*, 65(1):99–106, 2021. 1, 3

[25] Thomas Muller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM transactions on graphics (TOG)*, 41(4):1–15, 2022. 1

[26] Thang-Anh-Quan Nguyen, Amine Bourki, Matyas Macudzinski, Anthony Brunel, and Mohammed Bennamoun. Semantically-aware neural radiance fields for visual scene understanding: A comprehensive review. *arXiv preprint arXiv:2402.11141*, 2024. 1

[27] Ege Ozguroglu, Ruoshi Liu, Dídac Surís, Dian Chen, Achal Dave, Pavel Tokmakov, and Carl Vondrick. pix2gestalt: Amodal segmentation by synthesizing wholes. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3931–3940. IEEE Computer Society, 2024. 1, 3

[28] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 3

[29] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1

[30] Ruoxi Shi, Hansheng Chen, Zhuoyang Zhang, Minghua Liu, Chao Xu, Xinyue Wei, Linghao Chen, Chong Zeng, and Hao Su. Zero123++: a single image to consistent multi-view diffusion base model. *arXiv preprint arXiv:2310.15110*, 2023. 1, 3

[31] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 1

[32] Jiaxiang Tang, Jiawei Ren, Hang Zhou, Ziwei Liu, and Gang Zeng. Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. *arXiv preprint arXiv:2309.16653*, 2023. 3, 1

[33] Jiaxiang Tang, Zhaoxi Chen, Xiaokang Chen, Tengfei Wang, Gang Zeng, and Ziwei Liu. Lgm: Large multi-view gaussian model for high-resolution 3d content creation. In *European Conference on Computer Vision*, pages 1–18. Springer, 2024. 3, 1

[34] Vikram Voleti, Chun-Han Yao, Mark Boss, Adam Letts, David Pankratz, Dmitry Tochilkin, Christian Laforte, Robin Rombach, and Varun Jampani. Sv3d: Novel multi-view synthesis and 3d generation from a single image using latent video diffusion. In *European Conference on Computer Vision*, pages 439–457. Springer, 2024. 1

[35] Peng Wang and Yichun Shi. Imagedream: Image-prompt multi-view diffusion for 3d generation. *arXiv preprint arXiv:2312.02201*, 2023. 1, 3

[36] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *arXiv preprint arXiv:2106.10689*, 2021. 2, 3, 4, 1

[37] Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. Convnext v2: Co-designing and scaling convnets with masked autoencoders. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16133–16142, 2023. 2

[38] Chao Xu, Ang Li, Linghao Chen, Yulin Liu, Ruoxi Shi, Hao Su, and Minghua Liu. Sparp: Fast 3d object reconstruction and pose estimation from sparse views. In *European Conference on Computer Vision*, pages 143–163. Springer, 2024. 1

[39] Jiale Xu, Weihao Cheng, Yiming Gao, Xintao Wang, Shenghua Gao, and Ying Shan. Instantmesh: Efficient 3d mesh generation from a single image with sparse-view large reconstruction models. *arXiv preprint arXiv:2404.07191*, 2024. 1, 2, 3, 4

[40] 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

[41] Chuanxia Zheng and Andrea Vedaldi. Free3d: Consistent novel view synthesis without 3d representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9720–9731, 2024. 1

[42] Siting Zhu, Guangming Wang, Xin Kong, Dezhi Kong, and Hesheng Wang. 3d gaussian splatting in robotics: A survey, 2024. 1

EscherNet++: Simultaneous Amodal Completion and Scalable View Synthesis through Masked Fine-Tuning and Enhanced Feed-Forward 3D Reconstruction

Supplementary Material

A. Quantitative Results in Experiment

This section presents the quantitative results in experiments conducted in Sec. 3, with Tab. 1, Tab. 2, Tab. 3 summarizing results on novel view synthesis, amodel completion and 3D reconstruction accordingly.

B. Ablation Study on Feature-Level Masking

In experiment, we empirically find the proper ratio for feature-level masking. Consider a batch of feature maps from the image encoder, its tenser shape is $[b * t, l, c]$, in which b is the batch size of samples, t is number of input views in each sample, l is the feature map area (number of feature vectors associated with each input view) and c is the feature dimension.

We start by masking all (100% of $b * t$ dimension) feature maps by half feature map area (50% of l) randomly and the performance is sub-optimal. Then we gradually decrease the ratio on the second dimension by 25% (in $b * t$ dimension), and finally found that 25 % is a proper ratio for feature-level masking. That is, we report performance of the model with 25% masked in $b * t$ dimension and 50% masked in l dimension in training, as the representative results of feature-level masking.

We also attach the full tables for evaluating models with **OccNVS** in the ablation study on feature-level masking. It is found that feature-level masking with proper ratio can improve overall performance including better understanding of semantics from input views, better capture of intricate structures. However, it will lead to sub-optimal performance is too large ratio is picked, as shown is Fig. 6, Tab. 4, Tab. 5, Tab. 6.

C. Implementation Details of Models in Comparison

We compare our model with several recent SoTA models: Zero-1-2-3, Zero-1-2-3 XL [20] and EscherNet [17] for comparison in NVS tasks; DreamGaussian [32], Large Multi-View Gaussian Model(LGM) [33], SyncDreamer [21], InstantMesh [39] and EscherNet [17] for mesh quality comparison in 3D reconstruction tasks. OccNVS is used for comparison. For 3D reconstruction tasks, raw meshes from the models are normalized first and then compared with ground truth as in [17, 21].

Zero-1-2-3 & Zero-1-2-3 XL It is the first work in diffusion-based NVS for objects. In its model design, one input view can be referenced at a time and one target view can be synthesized afterwards. As a result, Zero-1-2-3 and its XL version are only adopted for one-input settings.

EscherNet Our model shares the same model structure with EscherNet. As the result, EscherNet can be used for direct comparison in all tasks and settings in this paper, including NVS and 3D reconstruction. For NVS, EscherNet is able to synthesize multiple novels view from any query viewpoints. For 3D reconstruction, 36 fixed view are synthesized, with the azimuth from 0°to 360°with a rendering every 30°at a set of elevations (-30°, 0°, 30°) for reconstruction with NeuS, the same setting as reconstruction with our model.

We fine-tune our model based on public weights shared by authors of Eschernet, and we have confirmed with them about the performance of EscherNet in the experiments.

DreamGaussian It is a two-stage model, which uses the first stage for reconstruction conditioned on a single input view and second image for texture refinement. Hence, there are no novel views required before reconstruction. Rotation is conducted for evaluation as in EscherNet. It is worth noting that DreamGaussian and LGM are the fastest methods for reconstruction in our experiment.

LGM As a two-stage method, LGM [33] depends on four views from fixed viewpoints synthesized by ImageDream [35] conditioned on one input view to reconstruct 3D. It is also a fast pipeline, however, it is found to struggle with significant elevation and azimuth angles in input views. Therefore, it does not perform well in our tests. The fundamental reason is that ImageDream may not be able to provide consistent and reasonable novel views when conditioned on inputs with significant angles, as shown in Fig.7. The same rotation mechanism is conducted as with DreamGaussain.

SyncDreamer 16 fixes views are synthesized conditioned on one input view and then given to NeuS [36] by SyncDreamer [21]. Compared with reconstruction time which usually takes near 30 minutes, the time spent on synthesis is almost insignificant. That is, the time used to reconstruct an object from one input view to a complete mesh is largely dependent on the reconstruction method, which shares a similar case with reconstitution based on our model with overfitting methods like NeuS.

InstantMesh In the original pipeline, Zero123++ Shi et al. [30] is used for NVS at the first stage and InstantMesh [39] construct the mesh based on novel views. Zero123++ is designed to generate 6 fixed views of an object with relative azimuth rotations and absolute elevations. The 6 in-

Table 1. Performance comparison on GSO-30, RTMV, NeRF Synthetic datasets and occluded counterparts (OccNVS). The **best number** is highlighted in bold, and the second best is underlined.

Method	# Ref. Views	GSO-30			Occluded GSO-30			RTMV			Occluded RTMV			NeRF			Occluded NeRF		
		PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
Zero-1-to-3 [20]	1	18.55	0.86	0.122	14.5	0.83	0.192	10.27	0.514	0.409	9.33	0.505	0.428	12.61	0.639	0.31	11.95	0.634	0.338
Zero-1-to-3 XL [20]	1	18.74	0.855	0.124	14.55	0.823	0.198	10.47	0.516	0.402	9.38	0.503	0.429	12.62	0.637	0.309	11.65	0.625	0.346
EscherNet[17]	1	20.05	0.883	0.096	15.64	0.852	0.161	10.43	0.520	0.411	9.63	0.511	0.432	13.35	0.658	0.293	12.55	0.654	0.317
	2	**22.85**	**0.908**	0.063	15.82	0.865	0.145	12.55	0.581	0.306	10.92	0.566	0.344	14.93	0.699	0.21	13.39	0.685	0.253
	3	23.87	**0.918**	0.052	16.32	0.874	0.130	13.58	**0.611**	0.259	11.68	0.594	0.295	16.19	**0.729**	**0.161**	14.57	0.716	0.119
	5	24.91	0.926	0.044	16.67	0.883	0.118	**14.48**	**0.633**	**0.222**	12.28	0.611	0.264	17.11	0.748	0.128	15.28	0.731	0.167
	10	25.65	0.933	0.037	16.92	0.889	0.111	**15.44**	0.657	0.186	13.00	0.634	0.230	**17.72**	**0.76**	**0.115**	15.80	0.746	0.150
Ours	1	20.11	0.883	0.094	**19.72**	**0.879**	**0.103**	10.5	0.523	0.408	10.34	0.52	0.416	13.35	**0.661**	0.29	13.51	**0.666**	**0.29**
	2	22.83	**0.908**	**0.062**	**21.86**	**0.902**	**0.07**	12.57	0.583	0.303	**12.32**	**0.577**	0.316	14.96	0.698	0.21	**14.74**	**0.692**	**0.221**
	3	**24.02**	**0.918**	**0.051**	**23.22**	**0.913**	**0.056**	13.45	0.608	0.262	13.29	0.603	0.269	16.14	0.727	0.164	15.85	**0.721**	**0.174**
	5	**25.15**	0.926	**0.043**	**24.22**	**0.921**	**0.047**	14.38	0.631	0.223	**14.16**	**0.627**	**0.232**	16.97	0.745	0.132	16.79	0.74	**0.138**
	10	**25.98**	**0.934**	**0.036**	**25.06**	**0.929**	**0.04**	15.42	**0.658**	0.186	**15.13**	**0.652**	0.196	**17.72**	0.759	**0.115**	**17.49**	**0.755**	**0.121**
Ours w/o Input-Level Masking	1	**20.33**	**0.886**	**0.091**	15.78	0.856	0.158	10.59	**0.531**	0.399	9.64	0.519	0.42	13.35	0.657	0.292	12.8	0.659	0.309
	2	22.7	0.907	0.063	15.87	0.866	0.145	**12.66**	**0.585**	**0.299**	10.99	0.57	0.336	14.97	**0.7**	**0.209**	13.47	0.688	0.251
	3	23.92	**0.918**	**0.051**	16.35	0.875	0.129	**13.59**	**0.611**	**0.258**	11.62	0.595	0.294	16.16	0.728	0.165	14.53	0.714	0.203
	5	25.00	**0.927**	**0.043**	16.66	0.883	0.118	14.41	0.632	0.223	12.21	0.612	0.266	17.0	0.745	0.131	15.24	0.73	0.169
	10	25.91	**0.934**	**0.036**	17.02	0.891	0.11	15.3	0.655	0.189	12.88	0.632	0.234	17.53	0.756	0.119	15.76	0.744	0.152
Ours w/o Feature-Level Masking	1	19.95	0.88	0.1	19.31	0.875	0.109	**10.78**	0.53	**0.391**	**10.57**	**0.526**	**0.405**	**13.47**	0.658	**0.289**	**13.57**	0.66	0.295
	2	22.72	0.907	0.064	21.65	0.9	0.073	12.57	0.582	0.301	12.26	0.575	**0.315**	**14.98**	0.697	0.211	14.69	0.691	0.226
	3	23.93	0.917	0.052	22.97	0.91	0.059	13.5	0.609	0.259	**13.31**	**0.609**	**0.259**	**16.25**	**0.729**	0.163	15.91	**0.721**	0.175
	5	25.05	0.926	**0.043**	23.98	0.919	0.049	14.37	0.63	0.223	14.12	0.624	0.233	**17.22**	**0.749**	**0.128**	**16.86**	**0.742**	**0.138**
	10	25.85	**0.934**	0.037	24.77	0.927	0.042	15.38	**0.658**	**0.185**	15.08	0.65	**0.195**	17.7	**0.76**	0.116	17.43	0.754	0.123

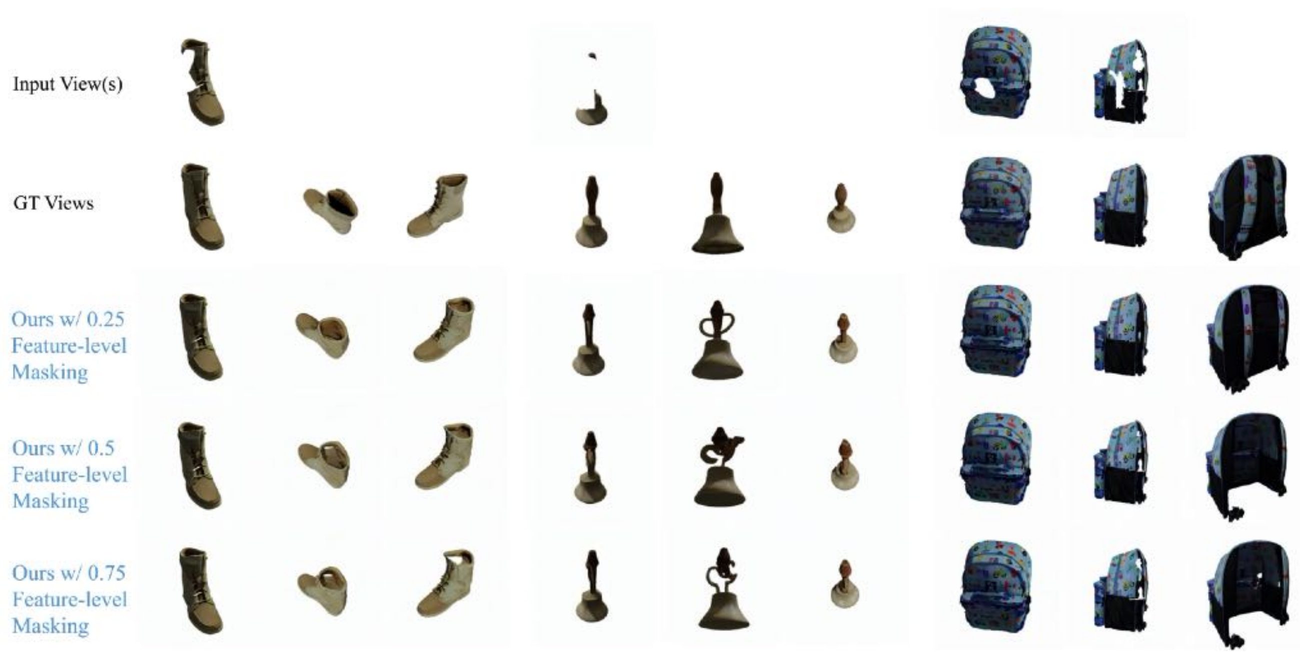

Figure 6. Qualitative results with different ratios for feature-level masking.

put images have poses with alternating absolute elevations of 20°and -10°, and their azimuths are defined relative to the query image, beginning at 30°and increased by 60°for subsequent poses. However, it sometimes generate meshes with floaters around the object, which leads to erroneous scale in normalization, as shown in Fig. 5. It is found that we can make use of our model to generate more consistent novel views at the preferred viewpoints for InstantMesh so that the performance can be improved significantly without floaters in the final meshes. The performance can be further enhanced by providing more novel views covering more viewpoints to InstantMesh. We provide one example comparing novel views from Zero123++ and our method in Fig. 8. No extra training or extra reference time is induced in this whole process.

Although it is able to provide views from any viewpoints, we find that the six viewpoints used in the original pipeline and their absolute values are necessary to the network. Therefore, we define that the input views are at 0°azimuth angle and we rotate the meshes back before evaluation.

As noticed by authors of InstantMesh, InstantMesh is able to take in various numbers of input views because of its transformer-based structure. However, in contrast to their

Table 2. Performance comparison on amodel completion on Occluded GSO-30, RTMV, and NeRF Synthetic datasets (OccNVS).

Method	# Ref. / NoI. Views	Occluded GSO-30			Occluded RTMV			Occluded NeRF		
		PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
InstructPix2Pix [1, 3]	1	18.08	0.92	0.098	15.19	0.829	0.142	16.77	0.843	0.144
	2	17.84	0.917	0.107	15.14	0.837	0.141	17.49	0.859	0.123
	3	17.86	0.918	0.11	15.03	0.837	0.141	18.23	0.868	0.117
	5	17.88	0.92	0.108	15.03	0.824	0.149	18.71	0.869	0.113
	10	17.51	0.918	0.114	15.38	0.828	0.145	18.31	0.872	0.111
Pix2gestalt [27]	1	20.71	0.942	0.072	16.22	0.85	0.109	16.98	0.849	0.123
	2	19.87	0.937	0.082	16.52	0.859	0.11	17.45	0.854	0.117
	3	20.2	0.938	0.08	16.06	0.859	0.112	18.02	0.863	0.115
	5	20.38	0.939	0.079	15.96	0.852	0.114	18.43	0.863	0.11
	10	19.94	0.937	0.084	16.08	0.85	0.115	18.2	0.866	0.108
Ours	1	28.42	0.952	0.029	19.99	0.832	0.09	21.24	0.841	0.071
	2	28.62	0.954	0.027	20.93	0.845	0.078	21.59	0.852	0.065
	3	29.29	0.956	0.025	22.28	0.848	0.072	22.22	0.863	0.06
	5	29.33	0.957	0.024	22.26	0.832	0.075	22.12	0.858	0.06
	10	28.34	0.95	0.027	20.81	0.799	0.088	21.47	0.843	0.062

Table 3. 3D reconstruction comparison on GSO3D and Occluded GSO3D datasets. Time is measured from when input views are given to networks to when the reconstructed meshes are ready in the batch inference mode.

Method	# Ref. Views	# NoI. Views	GSO3D		Occluded GSO3D		Time
			Chamfer Dist. ↓	Volume IoU ↑	Chamfer Dist. ↓	Volume IoU ↑	Minutes ↓
Dream Gaussian[32]	1	-	0.0543	0.4515	0.0611	0.3448	1.5
ImageDream[35]+LGM[33]	1	4	0.0877	0.2521	0.1787	0.095	1.5
SyncDreamer[21]+NeuS[36]	1	16	0.0427	0.5191	0.0624	0.2784	27
Zero123++[30]+InstantMesh[39]	1	6	0.0608	0.4557	0.0655	0.2478	1.6
EscherNet [17] + NeuS[36]	1	36	0.0312	0.5941	0.0477	0.3736	27
	2	36	0.0217	0.6878	0.0671	0.286	
	3	36	0.0186	0.7117	0.0346	0.3853	
	5	36	0.0177	0.7377	0.0351	0.3976	
	10	36	0.0169	0.7442	0.0312	0.4498	
Ours + NeuS	1	36	0.0305	0.6018	0.0376	0.5602	27
	2	36	0.0214	0.6921	0.0249	0.664	
	3	36	0.0185	0.7277	0.0197	0.7139	
	5	36	0.0182	0.7294	0.0189	0.7221	
	10	36	0.0168	0.7437	0.0176	0.7352	
Ours + InstantMesh	1	6	0.0304	0.5912	0.0392	0.5405	1.3
	2	6	0.0259	0.633	0.0301	0.5954	
	3	6	0.0251	0.6491	0.0257	0.6413	
	5	6	0.0238	0.6667	0.0291	0.6376	
	10	6	0.0275	0.6472	0.0282	0.6414	
Ours + InstantMesh	1	42	0.0278	0.6244	0.04	0.5501	1.3
	2	42	0.0224	0.6803	0.0311	0.6118	
	3	42	0.0265	0.6744	0.0277	0.6605	
	5	42	0.0253	0.6857	0.024	0.6886	
	10	42	0.0179	0.7295	0.0233	0.6987	

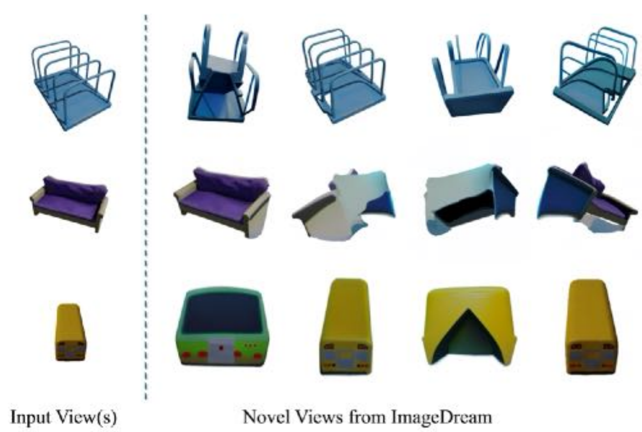

Figure 7. Examples of novel views generated by ImageDream. It struggles with significant elevations and azimuths. Therefore, the challenge is propagated to the reconstruction pipeline of LGM.

finding that decrease the number of input views can boost the performance in some hard cases, we found with our model, simply increasing the number of input views can further improve the overall reconstruction performance without extra overheads, thanks to the ability to synthesize high-quality views from any query viewpoints from our model.

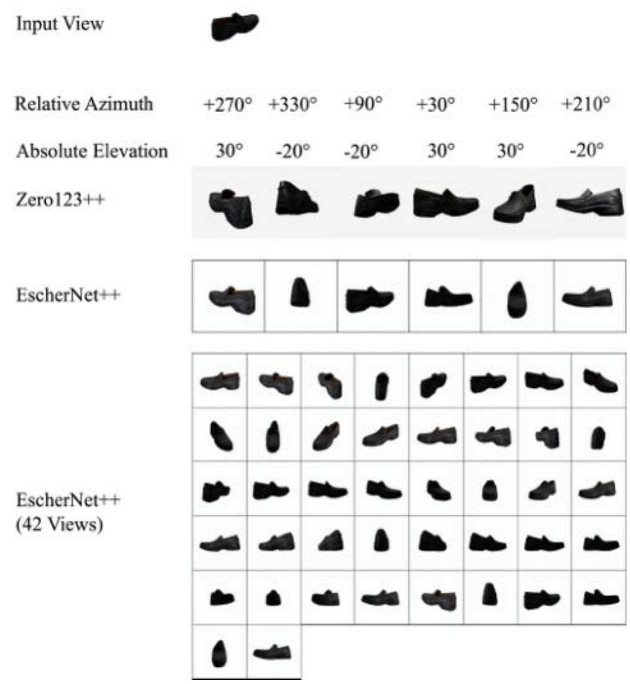

Figure 8. Examples of novel views generated by Zero123++ and EscherNet++ for reconstruction by InstantMesh. The last row contains all 42 views by our model. The scale and pose of the object in novel views by Zero123++ are not consistent sometimes, which can lead to confusion for InstantMesh.

D. Conclusion, Limitations & Future Work

In this paper, we propose EscherNet++, a masked fine-tuned diffusion model that can synthesize novel views of objects in a zero-shot way with amodal completion ability. We find that properly masked input images and input feature maps can contribute to better performance of the model. In addition, it can be seamlessly integrated with other fast feed-forward image-to-mesh models because of its flexible feature to synthesize any query views without the need for extra training, and the fast 3D reconstruction performance can be further boosted by its scalable nature.

During experiments, we found there are several aspects in which our model still falls short, including 1) degraded performance with data incorporating intricate details and complex layouts, 2) hallucination especially with occluded inputs. Future work can explore robust architecture designs with more diverse datasets, more explicit guidance with multi-modal inputs. Feed-forward 3D reconstruction methods also have the potential to be improved in terms of how to increase robustness to inconsistency in inputs views and utilize increasing number of views more efficiently. Last, a comprehensive framework is necessary to make our work more accessible in applications that includes object segmen-

Table 4. Performance comparison on GSO-30 and Occluded GSO-30 datasets with different ratios for feature-level masking.

Base Method	Input-Level Masking Ratio	Feature-Level Masking Ratio	# Ref. Views	GSO-30			Occluded GSO-30		
				PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
EscherNet (ckpt)	0.5	1.0	1	19.62	0.879	0.1	19.11	0.874	0.11
EscherNet (ckpt)	0.5	1.0	2	22.21	0.903	0.067	21.36	0.897	0.076
EscherNet (ckpt)	0.5	1.0	3	23.54	0.915	0.054	22.58	0.908	0.061
EscherNet (ckpt)	0.5	1.0	5	24.51	0.922	0.046	23.81	0.917	0.051
EscherNet (ckpt)	0.5	1.0	10	25.41	0.93	0.039	24.68	0.926	0.043
EscherNet (ckpt)	0.5	0.75	1	19.68	0.879	0.099	19.21	0.875	0.108
EscherNet (ckpt)	0.5	0.75	2	22.4	0.905	0.066	21.47	0.898	0.074
EscherNet (ckpt)	0.5	0.75	3	23.78	0.916	0.053	22.65	0.908	0.061
EscherNet (ckpt)	0.5	0.75	5	24.82	0.924	0.044	23.89	0.918	0.05
EscherNet (ckpt)	0.5	0.75	10	25.71	0.933	0.038	24.84	0.927	0.042
EscherNet (ckpt)	0.5	0.5	1	19.93	0.883	0.095	19.27	0.877	0.107
EscherNet (ckpt)	0.5	0.5	2	22.72	0.907	0.063	21.76	0.9	0.072
EscherNet (ckpt)	0.5	0.5	3	23.87	0.917	0.051	22.97	0.91	0.059
EscherNet (ckpt)	0.5	0.5	5	24.93	0.925	0.043	24.04	0.919	0.049
EscherNet (ckpt)	0.5	0.5	10	25.88	0.933	0.037	24.95	0.927	0.041
EscherNet (ckpt)	0.5	0.25	1	20.11	0.883	0.094	19.72	0.879	0.103
EscherNet (ckpt)	0.5	0.25	2	22.83	0.908	0.062	21.86	0.902	0.07
EscherNet (ckpt)	0.5	0.25	3	24.02	0.918	0.051	23.22	0.913	0.056
EscherNet (ckpt)	0.5	0.25	5	25.15	0.926	0.043	24.22	0.921	0.047
EscherNet (ckpt)	0.5	0.25	10	25.98	0.934	0.036	25.06	0.929	0.04
EscherNet (ckpt)	0.5	0	1	19.95	0.88	0.1	19.31	0.875	0.109
EscherNet (ckpt)	0.5	0	2	22.72	0.907	0.064	21.65	0.9	0.073
EscherNet (ckpt)	0.5	0	3	23.93	0.917	0.052	22.97	0.91	0.059
EscherNet (ckpt)	0.5	0	5	25.05	0.926	0.043	23.98	0.919	0.049
EscherNet (ckpt)	0.5	0	10	25.85	0.934	0.037	24.77	0.927	0.042

Table 5. Performance comparison on RTMV and Occluded RTMV datasets with different ratios for feature-level masking.

Base Method	Input-Level Masking Ratio	Feature-Level Masking Ratio	# Ref. Views	RTMV			Occluded RTMV		
				PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
EscherNet (ckpt)	0.5	1.0	1	10.62	0.532	0.401	10.37	0.525	0.414
EscherNet (ckpt)	0.5	1.0	2	12.38	0.58	0.31	12.14	0.574	0.322
EscherNet (ckpt)	0.5	1.0	3	13.23	0.606	0.267	13.02	0.6	0.279
EscherNet (ckpt)	0.5	1.0	5	14.23	0.628	0.232	13.94	0.62	0.243
EscherNet (ckpt)	0.5	1.0	10	15.2	0.654	0.192	14.96	0.648	0.201
EscherNet (ckpt)	0.5	0.75	1	10.29	0.522	0.418	10.12	0.518	0.428
EscherNet (ckpt)	0.5	0.75	2	12.3	0.577	0.316	12.17	0.576	0.32
EscherNet (ckpt)	0.5	0.75	3	13.3	0.606	0.267	13.1	0.6	0.278
EscherNet (ckpt)	0.5	0.75	5	14.3	0.63	0.227	14.01	0.623	0.239
EscherNet (ckpt)	0.5	0.75	10	15.17	0.652	0.193	14.9	0.647	0.203
EscherNet (ckpt)	0.5	0.5	1	10.37	0.521	0.415	10.23	0.518	0.42
EscherNet (ckpt)	0.5	0.5	2	12.3	0.575	0.318	12.08	0.571	0.327
EscherNet (ckpt)	0.5	0.5	3	13.23	0.604	0.272	13.1	0.599	0.28
EscherNet (ckpt)	0.5	0.5	5	14.26	0.628	0.229	14.02	0.622	0.24
EscherNet (ckpt)	0.5	0.5	10	15.21	0.652	0.192	14.93	0.645	0.202
EscherNet (ckpt)	0.5	0.25	1	10.5	0.523	0.408	10.34	0.52	0.416
EscherNet (ckpt)	0.5	0.25	2	12.57	0.583	0.303	12.32	0.577	0.316
EscherNet (ckpt)	0.5	0.25	3	13.45	0.608	0.262	13.29	0.603	0.269
EscherNet (ckpt)	0.5	0.25	5	14.38	0.631	0.223	14.16	0.627	0.232
EscherNet (ckpt)	0.5	0.25	10	15.42	0.658	0.186	15.13	0.652	0.196
EscherNet (ckpt)	0.5	0	1	10.78	0.53	0.391	10.57	0.526	0.405
EscherNet (ckpt)	0.5	0	2	12.57	0.582	0.301	12.26	0.575	0.315
EscherNet (ckpt)	0.5	0	3	13.5	0.609	0.259	13.31	0.609	0.259
EscherNet (ckpt)	0.5	0	5	14.37	0.63	0.223	14.12	0.624	0.233
EscherNet (ckpt)	0.5	0	10	15.38	0.658	0.185	15.08	0.65	0.195

tation, pose estimation, etc, combined as integrated modules or a single unified model.

Table 6. Performance comparison on NeRF and Occluded NeRF datasets with different ratios for feature-level masking.

Base Method	Input-Level Masking Ratio	Feature-Level Masking Ratio	# Ref. Views	NeRF			Occluded NeRF		
				PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
EscherNet (ckpt)	0.5	1.0	1	13.43	0.657	0.292	13.5	0.661	0.295
EscherNet (ckpt)	0.5	1.0	2	14.99	0.696	0.214	14.72	0.688	0.229
EscherNet (ckpt)	0.5	1.0	3	16.19	0.728	0.166	15.87	0.722	0.178
EscherNet (ckpt)	0.5	1.0	5	17.01	0.744	0.133	16.71	0.738	0.143
EscherNet (ckpt)	0.5	1.0	10	17.46	0.754	0.121	17.19	0.749	0.128
EscherNet (ckpt)	0.5	0.75	1	13.37	0.659	0.3	13.9	0.671	0.282
EscherNet (ckpt)	0.5	0.75	2	14.93	0.695	0.214	14.66	0.688	0.229
EscherNet (ckpt)	0.5	0.75	3	16.19	0.727	0.166	15.87	0.721	0.177
EscherNet (ckpt)	0.5	0.75	5	17.12	0.747	0.13	16.74	0.739	0.141
EscherNet (ckpt)	0.5	0.75	10	17.53	0.756	0.119	17.26	0.751	0.126
EscherNet (ckpt)	0.5	0.5	1	13.43	0.659	0.295	13.47	0.659	0.3
EscherNet (ckpt)	0.5	0.5	2	14.85	0.695	0.212	14.66	0.689	0.224
EscherNet (ckpt)	0.5	0.5	3	16.14	0.727	0.164	15.84	0.721	0.176
EscherNet (ckpt)	0.5	0.5	5	16.97	0.745	0.132	16.69	0.738	0.142
EscherNet (ckpt)	0.5	0.5	10	17.4	0.754	0.121	17.16	0.749	0.128
EscherNet (ckpt)	0.5	0.25	1	13.35	0.661	0.29	13.51	0.666	0.29
EscherNet (ckpt)	0.5	0.25	2	14.96	0.698	0.21	14.74	0.692	0.221
EscherNet (ckpt)	0.5	0.25	3	16.14	0.727	0.164	15.85	0.721	0.174
EscherNet (ckpt)	0.5	0.25	5	16.97	0.745	0.132	16.79	0.74	0.138
EscherNet (ckpt)	0.5	0.25	10	17.72	0.759	0.115	17.49	0.755	0.121
EscherNet (ckpt)	0.5	0	1	13.47	0.658	0.289	13.57	0.66	0.295
EscherNet (ckpt)	0.5	0	2	14.98	0.697	0.211	14.69	0.691	0.226
EscherNet (ckpt)	0.5	0	3	16.25	0.729	0.163	15.91	0.721	0.175
EscherNet (ckpt)	0.5	0	5	17.22	0.749	0.128	16.86	0.742	0.138
EscherNet (ckpt)	0.5	0	10	17.7	0.76	0.116	17.43	0.754	0.123