MVBoost: Boost 3D Reconstruction with Multi-View Refinement

Xiangyu Liu[1,2*]   Xiaomei Zhang[1,2*]   Zhiyuan Ma[3,4]   Xiangyu Zhu[1,2 †]   Zhen Lei[1,2,4,5]

[1]MAIS, Institute of Automation, Chinese Academy of Sciences, Beijing, China
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
[3]The Hong Kong Polytechnic University, Hong Kong, China   [4]CAIR, HKISI, CAS
[5]School of Computer Science and Engineering, FIE, M.U.S.T

https://github.com/Piggy-ch/MVBoost

Figure 1. Given a single image as input, our MVBoost can generate a high-quality 3D asset.

Abstract

Recent advancements in 3D object reconstruction have been remarkable, yet most current 3D models rely heavily on existing 3D datasets. The scarcity of diverse 3D datasets results in limited generalization capabilities of 3D reconstruction models. In this paper, we propose a novel framework for boosting 3D reconstruction with multi-view refinement (MVBoost) by generating pseudo-GT data. The key of MVBoost is combining the advantages of the high accuracy of the multi-view generation model and the consis-

---

*Equal contribution.
†Corresponding Authors.

tency of the 3D reconstruction model to create a reliable data source. Specifically, given a single-view input image, we employ a multi-view diffusion model to generate multiple views, followed by a large 3D reconstruction model to produce consistent 3D data. MVBoost then adaptively refines these multi-view images, rendered from the consistent 3D data, to build a large-scale multi-view dataset for training a feed-forward 3D reconstruction model. Additionally, the input view optimization is designed to optimize the corresponding viewpoints based on the user's input image, ensuring that the most important viewpoint is accurately tailored to the user's needs. Extensive evaluations demonstrate that our method achieves superior reconstruction results and robust generalization compared to prior works.

1. Introduction

Generating 3D assets from a single-view image is a critical task in 3D computer vision [11, 18, 19, 21, 26, 31, 35], offering a broad range of applications such as video games [9], virtual reality [23], 3D content creation [1], and animation [16]. High-fidelity 3D reconstruction models greatly reduce the labor involved in creating 3D digital assets. However, generating high-fidelity 3D assets from a single-view image while preserving consistent surface details is a challenge, especially for complex objects. A core difficulty in developing these models is the limited availability of high-quality 3D data. Creation of such datasets is a complex task that often requires specialized equipment, advanced capture techniques, or intricate 3D modeling processes. Current publicly available 3D asset datasets [2–4] lack high-quality textures and contain significant repetitiveness, making them insufficient for training 3D generative models effectively.

Several methods leverage 2D diffusion models to enhance 3D generative models due to the widespread success of diffusion models in image generation. DreamFusion proposes Score Distillation Sampling (SDS) [26], which distills 3D knowledge from 2D diffusions and inspires the advancement of SDS-based 2D lifting methods [22, 26, 34]. Although SDS-based methods can produce highly realistic visual effects, their optimization-based approaches require several hours to generate a refined 3D asset, making them impractical for 3D content creators. Additionally, SDS-based methods often suffer from poor geometry and inconsistencies, such as the "Janus" problem. To address these issues, various feed-forward methods [11, 14, 18, 31] have developed that use four views of ground truth as input during training, generating 3D assets through a robust 3D reconstruction network. However, these methods rely on multi-view diffusion models [20, 33] during inference, which may yield inconsistent multi-view outputs. In practical application, the reconstructed model will perform poorly

if the training set lacks examples for the current scene. Collecting additional data for new scenarios is costly, making reconstruction very inflexible to apply.

In this work, we propose a novel framework for boosting 3D reconstruction with multi-view refinement (MVBoost) from a single-view image. The method consists of a multi-view refinement strategy and a boosting reconstruction model. The multi-view refinement strategy combines the advantages of the high accuracy of the multi-view generation model and the consistency of the 3D reconstruction model, as our data source. While, the multi-view generation model excels in accuracy but lacks consistency across views, whereas the 3D reconstruction model provides consistent multi-view data, but with lower accuracy. Specifically, given a single-view input image, the multi-view refinement strategy uses a multi-view diffusion model to generate high accuracy multi-views. Then, these multi-view images are sent into a large 3D reconstruction model to produce consistent original 3D representation. By refining multi-view images rendered from original 3D data, the multi-view refinement strategy generates a large-scale multi-view dataset to train a feed-forward 3D reconstruction model. The boosting reconstruction model using the recently proposed Large Multi-View Gaussian Model as our starting point, we introduce LoRA [12] to stabilize the training process. In practical applications, the alignment of the reconstructed 3D model with the input image is one of the key criteria for assessing the quality of the reconstruction. Therefore, we provide an input view optimization process to optimize the corresponding viewpoint image.

We conduct extensive experiments on the GSO [6] dataset. Both qualitative and quantitative results demonstrate that our proposed method outperforms related reconstruction methods. In summary, our contributions are:

- We present a novel framework for constructing pseudo-ground truth, which harnesses the advantages of high-accuracy multi-view generation to improve overall reconstruction quality.
- We propose a sophisticated framework designed to integrate diverse single-view datasets into the 3D reconstruction training. This framework is further refined based on user-provided input images, culminating in a significant enhancement of the reconstruction results.
- Results validate that our method could reconstruct high-fidelity 3D reconstruction and achieve new state-of-the-art results.

2. Related Work

3D Generation through Score Distillation. Learning to generate 3D assets from single images presents a formidable challenge due to the limited availability of 3D-image paired data. DreamFusion [26] introduces Score Distillation Sampling (SDS), along with the concurrent effort

Score Jacobian Chaining [32], which incrementally integrates scores from various perspectives to form a 3D representation. Building on SDS, numerous subsequent studies have emerged. Notably, ProlificDreamer [34] aims to minimize the noise prediction error to ensure that the model's output closely aligns with the desired distribution of rendered images. Taking it a step further, ScaleDreamer [22] presents Asynchronous Score Distillation (ASD), which seeks to minimize the noise prediction error without altering the weights of the pre-trained diffusion network. Nevertheless, due to recent limitations on the number of perspectives, SDS-based optimization methods often encounter the "Janus" problem, a common issue where multiple conflicting views lead to inconsistencies in the generated 3D asset.

Multi-view Diffusion Models. Due to the limited availability of 3D datasets, many approaches adapt 2D diffusion models for 3D reconstruction. For example, Zero123 [19] demonstrates that Stable Diffusion can be fine-tuned to generate novel views by conditioning on relative camera poses. To address inconsistencies among the generated views in Zero123, several studies [15, 17, 20, 21, 29, 33, 35] have focused on fine-tuning 2D diffusion models to simultaneously generate multiple consistent views of the same object. Wonder3D [21] uses Cross-Domain Diffusion to produce more consistent RGB and normal images. CRM [35] incorporates the Canonical Coordinates Map (CCM) into multi-view generation to enhance the geometric features of 3D assets. Building on Wonder3D, Era3D [15] employs a regression and conditioning scheme to unify arbitrary camera inputs into canonical camera outputs. VIAT [27] and MVTN [7] focus on view selection for 3D object understanding. VTR [39] enhances view transformation robustness by generating new views.

Multi-view to 3D Reconstruction. In the field of 3D reconstruction and generation, a significant trend is to directly learn a model that reconstructs 3D representations from multi-view based on 3D datasets. LRM [11] utilizes a transformer backbone to map multi-view image tokens to an implicit 3D triplane NeRF [25]. The CRM [35] replaces the transformer backbone with a convolutional U-Net backbone and employs a differentiable Flexicubes [28] for direct optimization on the grid. Taking advantage of the speed and quality improvements offered by 3D Gaussian Splatting, both LGM [31] and GRM [37] substitute triplane NeRF with 3D Gaussian Splatting [13]. The VFusion3D [8] fine-tuning video diffusion model generates multi-view data for the refinement of LRM [11]. Several other works have also made commendable progress in this feed-forward approach. However, training predominantly on the publicly available 3D dataset Objaverse [3], and employing input views with domain inconsistencies during training and inference, have resulted in limited practicality and generalization in real-world applications. In contrast, our research incorporates a

diverse range of 2D datasets with the multi-view refinement strategy into the model training process, ensuring domain-consistent input views throughout both training and inference to improve the generalization and robustness of the feed-forward model.

3. Method

The overall framework of our network, MVBoost, is illustrated in Figure 2. We first provide an overview of the methods and processes behind diffusion models and 3D generation in Section 3.1. Next, we present our multi-view refinement strategy (Section 3.2), which addresses inconsistencies across views generated by the diffusion model. In Section 3.3, we describe our boosting reconstruction model, which leverages a refined 2D multi-view dataset to achieve high-fidelity 3D reconstruction without relying on any 3D datasets. Finally, in Section 3.4, we explain how we optimize the alignment of the 3D Gaussian Splatting with the corresponding input viewpoint for improved 3D representation.

3.1. Preliminaries

Diffusion Model. Diffusion Model [5, 10, 30] is a generative model that involves forward and reverse diffusion processes. In forward diffusion, data x_0 is incrementally corrupted with Gaussian noise ($x_t = \alpha_t x + \sigma_t \epsilon, t \in \{1, ..., T\}$) to approach a noise distribution. The reverse process, which is the generative phase, uses a neural network to predict and apply denoising steps, progressively recovering the original data as it transitions from $t = T$ back to $t = 0$.

3D Generation. Given a single-view image c, we utilize the Multi-View Diffusion Model $\mathcal{G}$ to generate multiple views $C^\pi = \{c^{\pi_i}\}_{i=1}^n$. This process can be represented as

$$C^\pi = \{c^{\pi_i}\}_{i=1}^n = \mathcal{G}(c_T^{\pi_1}, \cdots, c_T^{\pi_n}; c, T), \qquad (1)$$

where $c_T^{\pi_1}, \cdots, c_T^{\pi_n}$ are initialized as Gaussian noise, T represents the timestep, and π_i corresponds to the camera pose for each view.

Using the multi-view data C^π as input, we apply a large 3D reconstruction model $\mathcal{R}_\phi$ to generate the 3D representation θ:

$$\theta = \mathcal{R}_\phi(C^\pi). \qquad (2)$$

The 3D representation θ can be either a NeRF [24] or 3D Gaussian Splatting [13]. In our specific implementation, we utilize 3D Gaussian Splatting. We first generate an original 3D Gaussian Splatting denoted as θ using a standard 3D asset generation process. Then, we render θ from specific viewpoints to obtain multi-view images, represented as x^π. The rendering process can be be represented as
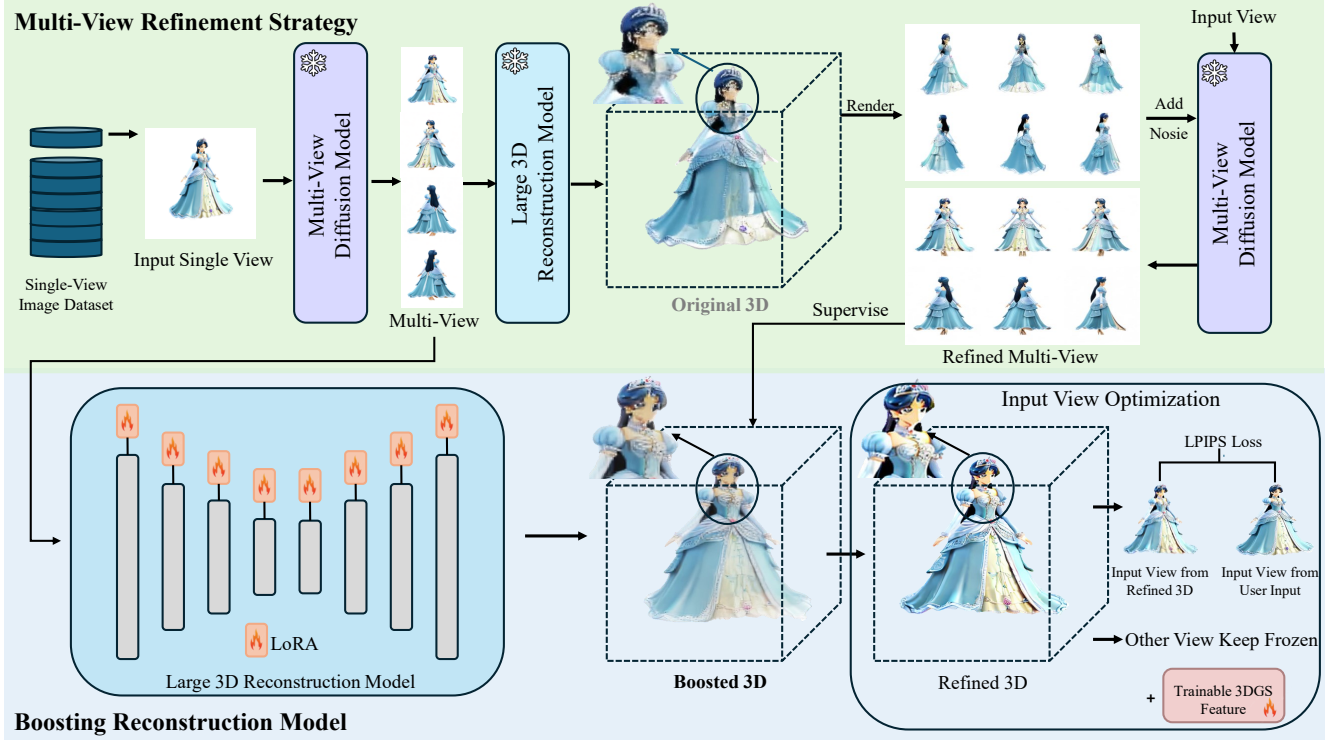
$$x^\pi = g(\theta, \pi). \qquad (3)$$

Figure 2. The overview of our MVBoost framework. Given a single-view image dataset, we first employ a multi-view diffusion model to generate the original multi-view dataset. Then the original multi-view is sent into a large 3D reconstruction model to produce the 3D Gaussian Splatting. Several views are rendered from this 3D Gaussian Splatting, and refined by the diffusion model to produce the refined multi-view dataset. During training, the refined multi-view dataset is used to supervise the 3D reconstruction model with LoRA. Finally, the generated 3D assets are optimized to align with the specific input viewpoint, obtaining high-fidelity reconstruction results.

3.2. Multi-View Refinement Strategy

The single-view image dataset can be enhanced using Multi-View Refinement to generate a large, refined multi-view dataset.

Previous methods often relied on limited 3D data as ground truth, which constrained their performance. While VFusion3D [8] leverages video diffusion models for multi-view generation, it lacks explicit view consistency constraints, relying only on implicitly learned multi-view relationships, which can result in geometric inconsistencies. We propose a method that creates pseudo-ground truth by combining 3D multi-view consistency with the high accuracy of the diffusion model. The process is as follows:

In our forward refinement diffusion, the Gaussian Splatting rendered view x^π is progressively corrupted by Gaussian noise, approaching a noise distribution:

$$x_t^\pi = \alpha_t x^\pi + \sigma_t \epsilon. \tag{4}$$

The reverse refinement diffusion process predicts the denoising operations using $X_t^\pi = \{x_t^\pi\}_{i=1}^n$ and the input view c, resulting in the refined multi-view set $C_\uparrow^\pi = \{c_\uparrow^\pi\}_{i=1}^n$:

$$C_\uparrow^\pi = \mathcal{G}(X_t^\pi; c, t), \tag{5}$$

where $t = sT$ and T represents the maximum timestep. The parameter s controls the strength of the noise added during the process, ranging from 0 to 1.

3.3. Boosting Reconstruction Model

As illustrated in Figure 2, we use the refined multi-view dataset generated from Section 3.2 to boost the reconstruction model eliminating the need for 3D datasets.

We input the original multi-view data $C'^\pi = c^{\pi_i}{}_{i=1}^n$ into our boosted model to ensure consistency between the training and inference inputs. In our model, we apply LoRA exclusively to the original 3D reconstruction model, $\mathcal{R}\phi$, specifically within its Cross-view Self-Attention component, resulting in the boosted model $\mathcal{R}\phi^*$. By training $\mathcal{R}_{\phi^*}$ with the refined multi-view dataset $C_\uparrow^\pi$, we obtain the boosted 3D Gaussian Splatting, $\theta^* = \mathcal{R}_{\phi^*}(C^\pi)$.

In summary, we use a multi-view diffusion model along with a refinement strategy to convert a single-view image into a multi-view tuple $(C^\pi, C_\uparrow^\pi)$. The training process is defined as follows:

$$\phi^* = \arg\min_\phi \mathbb{E}_{(C^\pi, C_\uparrow^\pi), i} \left[\mathcal{L}(g(\mathcal{R}_{\phi^*}(C^\pi), \pi), C_\uparrow^\pi) \right], \tag{6}$$

where $\mathcal{L}$ is the loss function that supervises the reconstruction model by comparing the render views $g(\theta^*, \pi)$ against

the refined multi-view $C_\uparrow^\pi$. This ensures consistent improvement across different viewpoints and adaptability to various 3D representations. The loss function $\mathcal{L}$ is a combination of Mean Squared Error (MSE) and LPIPS loss.

The training framework, including the LoRA-augmented Self-Attention, is compatible with various reconstruction models, supporting different types of 3D representations. Detailed algorithmic illustrations can be found in the Supplementary Material.

Next, we integrate a set of learnable parameters into the generated 3D Gaussian Splatting. This process involves aligning the user-provided input view with the corresponding view in the 3D Gaussian Splatting through Input View Optimization.

3.4. Input View Optimization

To improve the alignment between the generated 3D assets and the input view, we search over the possible camera poses of the 3D assets to find the pose that minimizes the LPIPS loss. Let π represent the set of all possible camera poses. The optimal camera pose π_{opt} is defined as:

$$\pi_{opt} = \arg\min_\pi \text{LPIPS}(g(\theta^*, \pi), c), \qquad (7)$$

where $g(\theta^*, \pi)$ denotes the rendered view from the original 3D Gaussian Splatting θ^* at camera pose π, and c is the input view.

Subsequently, we enhance the 3D Gaussian Splatting θ^* by adding a learnable matrix W:

$$\theta_\uparrow = \theta^* + W. \qquad (8)$$

This produces the refined 3D Gaussian Splatting $\theta_\uparrow$. The matrix W is optimized by minimizing the LPIPS loss between the rendered view at the optimal camera pose π_{opt} and the input view c:

$$W = \arg\min_W \text{LPIPS}(g(\theta_\uparrow, \pi_{opt}), c). \qquad (9)$$

During this optimization, only the view corresponding to π_{opt} is optimized, while other views remain frozen. This ensures that the refined 3D Gaussian Splatting $\theta_\uparrow$ aligns more closely with the input view, without affecting the fidelity of the other perspectives.

4. Experiment

4.1. Implementation Details

Datasets. We employ ChatGPT-4o-mini to generate over $100k$ prompts for distinct objects in bulk, and utilize a text-to-image model (LayerDiffuse [38]) to transform the prompts into finely crafted object images. We do not rely on any existing image datasets for training, which renders our dataset acquisition nearly cost-free, allowing users to

expand the image dataset according to their specific needs. Our SOTA experiments are conducted on the entire Google Scanned Objects (GSO [6]) dataset, while our ablation studies are carried out on a randomly selected subset of 300 assets from the GSO dataset.

Baselines. In our research, the Multi-View Diffusion Model utilizes Era3D [15], while the large 3D reconstruction model employs LGM [31] which uses Gaussian Splatting for 3D representation. Our pipeline theoretically does not impose any restrictions on the specific types of multi-view diffusion models, reconstruction models, or 3D representations. Among the models we compare, TriplaneGaussian and LGM both use 3D Gaussian Splatting, while OpenLRM [11] and VFusion3D [8] employ NeRF, and InstantMesh [36] and CRM [35] utilize mesh. To ensure fairness, in the main results (Section 4.2), the comparative experiments of our method against other models are conducted without utilizing Input View Optimization.

Training. We train our model using 8 NVIDIA A100 (80G) GPUs for about a day, which, to our knowledge, represents the lowest training cost and shortest training time compared to similar models [8]. Similar to the training of LGM, we resize the images to 256×256 for LPIPS loss to save memory. Both our training and inference input views come from the unrefined multi-view diffusion. All rendered images and supervision images were set against a white background with a resolution of 512×512.

4.2. Main Results

Quantitative Results. We compare MVBoost with recent open-world feed-forward single/sparse-view to 3D methods, including TriplaneGaussian [40], LGM [31], OpenLRM [11], CRM [35], and InstantMesh [36]. For the comparative methods, we utilize their official implementations. Since input settings differ among the baselines, we evaluate all methods in a unified single-view to 3D setting. For the GSO dataset, we utilize the first thumbnail image as the single-view input. Then, we use the official multi-view diffusion models of each respective model to generate multi-view inputs, thereby simulating the real user environment. We conducted experiments on the Omni3D dataset, with results in the Supplementary Material.

Table 1. Quantitative comparison for the visual quality on Google Scanned Objects (GSO) orbiting views between our method and baselines for single image to 3D methods.

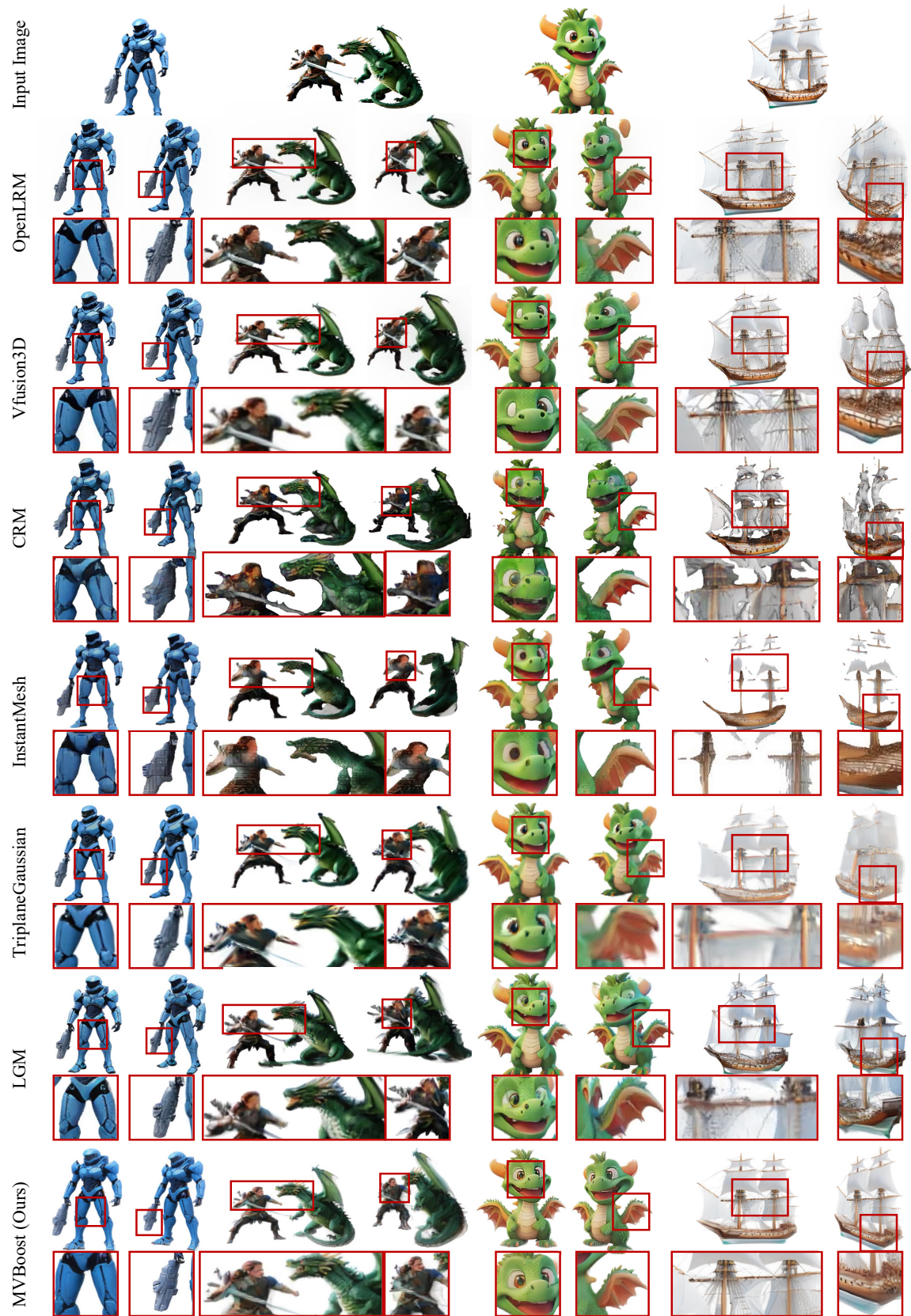| Method	PSNR↑	SSIM↑	LPIPS↓
OpenLRM [11]	16.728	0.785	0.208
VFsuion3D [8]	17.416	0.846	0.155
CRM [35]	17.435	0.800	0.195
InstantMesh [36]	16.796	0.786	0.207
TriplaneGaussian [40]	13.702	0.805	0.243
LGM [31]	17.148	0.776	0.220
MVBoost (Ours)	**18.561**	**0.859**	**0.131**

Figure 3. Qualitative comparison of image-to-3D methods. Our approach demonstrates superior 3D generation across challenging images.

Table 2. Quantitative comparison for the geometry quality on Google Scanned Objects (GSO) between our method and baselines for single image to 3D methods.

Method	CD↓	F-Score↑
OpenLRM [11]	0.14786	0.6562
VFsuion3D [8]	0.16118	0.6365
CRM [35]	0.12379	0.7306
InstantMesh [36]	0.12334	0.7369
TriplaneGaussian [40]	0.12132	0.7598
LGM [31]	0.17032	0.6349
MVBoost (Ours)	**0.10110**	**0.7977**

The quantitative experiments on the quality of 2D images are illustrated in Table 1. In this study, our method, TriplaneGaussian, and LGM produce 3D representations using 3D Gaussian Splatting, while OpenLRM and VFusion3D use NeRF, and InstantMesh and CRM employ meshes as their 3D representation. We render the corresponding orbiting views of the 3D Gaussian Splatting, NeRF, and mesh outputs, and compare them against the Ground Truth for evaluation metrics. Our method achieves state-of-the-art performance across various metrics of 2D novel view synthesis quality.

The quantitative experiments on the quality of 3D geometry are illustrated in Table 2. In this study, our method, TriplaneGaussian, and LGM convert 3D Gaussian Splatting into a mesh following the mesh extraction framework established by LGM. Similarly, OpenLRM and VFusion3D export their NeRF representations as meshes according to their official implementations. Subsequently, all methods are uniformly rescaled to operate at the same mesh scale for the experiments. Even though we choose 3D Gaussian Splatting as our 3D representation, we still achieve state-of-the-art performance across all metrics of geometric quality.

Qualitative Results. Figure 3 visualizes the qualitative results. Compared with other methods, our boosted results show very high fidelity for diverse inputs, including Robots, cartoon characters, and boat images of various subjects. Our MVBoost correctly models complex geometry (e.g., dinosaur and warrior) and generates the details, such as the ship's sail and halyard, reflecting the great generalization ability of our model. Figure 4 shows the pseudo GT multi-view data generated by our refinement strategy.

Figure 4. Multi-view data generated by our refinement strategy.

4.3. Ablation Study and Disscussion

To demonstrate the effectiveness of the multi-view refinement strategy and the input view optimization model, we conduct several ablation studies, which include: (1) qualita-

Table 3. The quantitative ablation experiments of the multi-view refinement strategy conducted on the GSO dataset demonstrate that a noise intensity of 0.95 is the most suitable.

Refine Strength	PSNR↑	SSIM↑	LPIPS↓
original	17.811	0.815	0.118
0.10	17.798	0.818	0.114
0.50	17.76	0.813	0.113
0.70	17.80	0.817	0.111
0.90	18.27	0.821	0.106
0.95	**19.132**	**0.827**	**0.100**
1.00	18.583	0.825	0.104

tive and quantitative comparative experiments on 2D multi-view data before and after refinement; (2) quantitative experiments employing different intensities of refined multi-view supervision; (3) comparative experiments evaluating the Input View Optimization module before and after its implementation; (4) the experimental performance of our model as the dataset grows.

Figure 5 illustrates the qualitative experiments conducted with our multi-view refinement strategy. The first row shows the original multi-view images, while the second row presents the refined multi-view images. The restoration of the red sports car's side door highlights the module's ability to enhance texture details. The chandelier and telescope examples demonstrate its effectiveness in preserving geometric consistency. In the keyboard example, the side view, which initially appeared unrealistic, is corrected to a more plausible representation. In the train example (Figure 6), significant viewpoint errors in the original multi-view images are successfully corrected by our method to achieve accurate alignment.

Table 4. The quantitative ablation experiments on the GSO dataset using a reconstruction model supervised with different noise strengths.

Refine Strength	PSNR↑	SSIM↑	LPIPS↓
original	17.851	0.837	0.147
0.10	17.624	0.839	0.141
0.50	17.764	0.840	0.140
0.70	17.904	0.840	0.139
0.90	18.021	0.841	0.139
0.95	**18.093**	**0.842**	**0.138**
1.00	18.053	0.840	0.141

In the selection of noise intensity within the multi-view refinement strategy, we conduct quantitative experiments on the GSO dataset, as illustrated in Table 3. The results indicate that the optimal refinement effect is achieved at a strength of 0.95. To further illustrate the efficacy of a noise intensity of 0.95, we employ a reconstruction model supervised with varying noise intensities, subsequently conducting quantitative experiments on image quality using the GSO dataset. The experimental results are presented in the Table 4.
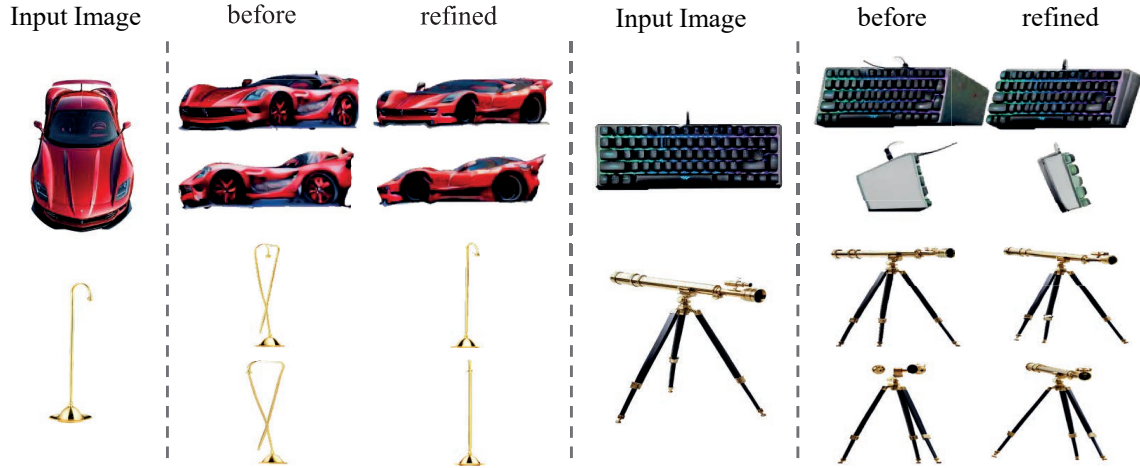
Table 5 show the scaling law of model performance

Figure 5. Qualitative comparison of 2D multi-view data before and after refinement. Our refined multi-view shows enhanced geometric correction and superior fidelity in detail reproduction compared to the original.
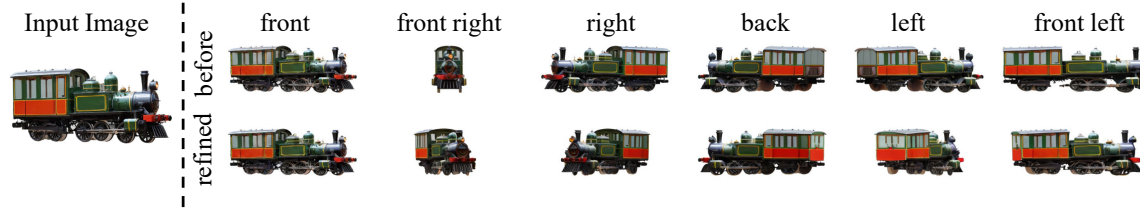

Figure 6. Our multi-view refinement strategy effectively corrects substantial viewpoint errors in the original views.

Figure 7. Post-processing with Input View Optimization aligns the 3DGS rendered views with the user's input view.

given the number of new data. As the dataset grows, we observe a significant improvement in model performance.

Table 5. Improved performance with increased data.

Num of Data	PSNR↑	SSIM↑	LPIPS↓	CD↓	F-Score↑
10K	17.212	0.783	0.223	0.175	0.633
50K	18.098	0.790	0.187	0.155	0.693
100K	**18.561**	**0.859**	**0.131**	**0.101**	**0.798**

The input view optimization is a post-processing method. All previous quantitative and qualitative experiments are conducted without Input View Optimization. The experimental results with Input View Optimization are shown in Figure 7. On the GSO dataset, the LPIPS loss of the input view decreased from an average of 0.108 to 0.002.

5. Conclusion

This paper designs a novel framework MVBoost to boost 3D reconstruction. MVBoost combines the advantages of the high accuracy of the multi-view generation model and the consistency of the 3D reconstruction model to generate pseudo-ground truth, as the data source. A feed-forward 3D reconstruction model is trained by synthetic data, showing superior performance in the reconstruction of 3D assets. Followed by a post-processing method, input-view optimization further optimizes the corresponding viewpoints based on the input image, ensuring that the most important viewpoint is accurately tailored to the user's needs. Extensive evaluations on benchmark tasks demonstrate that MVBoost achieves state-of-the-art performance in 3D reconstruction. Our approach provides a new pipeline for 3D reconstruction, and we hope to contribute to the development of the 3D reconstruction community.

Acknowledgement

References

[1] Bharat Lal Bhatnagar, Garvita Tiwari, Christian Theobalt, and Gerard Pons-Moll. Multi-garment net: Learning to dress 3d people from images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5420–5430, 2019. 2

[2] Angel X. Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. ShapeNet: An Information-Rich 3D Model Repository. Technical Report arXiv:1512.03012 [cs.GR], Stanford University — Princeton University — Toyota Technological Institute at Chicago, 2015. 2

[3] 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. 3

[4] Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, et al. Objaverse-xl: A universe of 10m+ 3d objects. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[5] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 3

[6] 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, 2022. 2, 5

[7] Abdullah Hamdi, Silvio Giancola, and Bernard Ghanem. Mvtn: Multi-view transformation network for 3d shape recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1–11, 2021. 3

[8] Junlin Han, Filippos Kokkinos, and Philip Torr. Vfusion3d: Learning scalable 3d generative models from video diffusion models. *European Conference on Computer Vision (ECCV)*, 2024. 3, 4, 5, 7

[9] Zekun Hao, Arun Mallya, Serge Belongie, and Ming-Yu Liu. Gancraft: Unsupervised 3d neural rendering of minecraft worlds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14072–14082, 2021. 2

[10] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 3

[11] 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. In *The Twelfth International Conference on Learning Representations*, 2024. 2, 3, 5, 7

[12] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. 2

[13] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42 (4), 2023. 3

[14] Jiahao Li, Hao Tan, Kai Zhang, Zexiang Xu, Fujun Luan, Yinghao Xu, Yicong Hong, Kalyan Sunkavalli, Greg Shakhnarovich, and Sai Bi. Instant3d: Fast text-to-3d with sparse-view generation and large reconstruction model. In *The Twelfth International Conference on Learning Representations*, 2024. 2

[15] Peng Li, Yuan Liu, Xiaoxiao Long, Feihu Zhang, Cheng Lin, Mengfei Li, Xingqun Qi, Shanghang Zhang, Wenhan Luo, Ping Tan, et al. Era3d: High-resolution multiview diffusion using efficient row-wise attention. *arXiv preprint arXiv:2405.11616*, 2024. 3, 5

[16] Tingting Liao, Xiaomei Zhang, Yuliang Xiu, Hongwei Yi, Xudong Liu, Guo-Jun Qi, Yong Zhang, Xuan Wang, Xiangyu Zhu, and Zhen Lei. High-fidelity clothed avatar reconstruction from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8662–8672, 2023. 2

[17] Minghua Liu, Ruoxi Shi, Linghao Chen, Zhuoyang Zhang, Chao Xu, Xinyue Wei, Hansheng Chen, Chong Zeng, Jiayuan Gu, and Hao Su. One-2-3-45++: Fast single image to 3d objects with consistent multi-view generation and 3d diffusion. *arXiv preprint arXiv:2311.07885*, 2023. 3

[18] Minghua Liu, Chao Xu, Haian Jin, Linghao Chen, Mukund Varma T, Zexiang Xu, and Hao Su. One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[19] 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. 2, 3

[20] 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. In *The Twelfth International Conference on Learning Representations*, 2024. 2, 3

[21] Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, and Wenping Wang. Wonder3d: Single image to 3d using cross-domain diffusion. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9970–9980, 2024. 2, 3

[22] Zhiyuan Ma, Yuxiang Wei, Yabin Zhang, Xiangyu Zhu, Zhen Lei, and Lei Zhang. Scaledreamer: Scalable text-to-3d synthesis with asynchronous score distillation. In *European Conference on Computer Vision*, pages 1–19. Springer, 2025. 2, 3

[23] Oier Mees, Maxim Tatarchenko, Thomas Brox, and Wolfram Burgard. Self-supervised 3d shape and viewpoint estimation from single images for robotics. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 6083–6089. IEEE, 2019. 2

[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 synthesis. *Commun. ACM*, 65(1):99–106, 2021. 3

[25] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 3

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

[27] Shouwei Ruan, Yinpeng Dong, Hang Su, Jianteng Peng, Ning Chen, and Xingxing Wei. Towards viewpoint-invariant visual recognition via adversarial training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4709–4719, 2023. 3

[28] Tianchang Shen, Jacob Munkberg, Jon Hasselgren, Kangxue Yin, Zian Wang, Wenzheng Chen, Zan Gojcic, Sanja Fidler, Nicholas Sharp, and Jun Gao. Flexible isosurface extraction for gradient-based mesh optimization. *ACM Trans. Graph.*, 42(4), 2023. 3

[29] 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, 2023. 3

[30] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015. 3

[31] 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, 2025. 2, 3, 5, 7

[32] Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A. Yeh, and Greg Shakhnarovich. Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. *arXiv preprint arXiv:2212.00774*, 2022. 3

[33] Peng Wang and Yichun Shi. Imagedream: Image-prompt multi-view diffusion for 3d generation, 2023. 2, 3

[34] Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 3

[35] Zhengyi Wang, Yikai Wang, Yifei Chen, Chendong Xiang, Shuo Chen, Dajiang Yu, Chongxuan Li, Hang Su, and Jun Zhu. Crm: Single image to 3d textured mesh with convolutional reconstruction model. *arXiv preprint arXiv:2403.05034*, 2024. 2, 3, 5, 7

[36] 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. 5, 7

[37] Yinghao Xu, Zifan Shi, Wang Yifan, Sida Peng, Ceyuan Yang, Yujun Shen, and Wetzstein Gordon. Grm: Large gaussian reconstruction model for efficient 3d reconstruction and generation. *arxiv: 2403.14621*, 2024. 3

[38] Lvmin Zhang and Maneesh Agrawala. Transparent image layer diffusion using latent transparency, 2024. 5

[39] Qi Zhang, Zhouhang Luo, Tao Yu, and Hui Huang. View transformation robustness for multi-view 3d object reconstruction with reconstruction error-guided view selection, 2024. 3

[40] Zi-Xin Zou, Zhipeng Yu, Yuan-Chen Guo, Yangguang Li, Ding Liang, Yan-Pei Cao, and Song-Hai Zhang. Triplane meets gaussian splatting: Fast and generalizable single-view 3d reconstruction with transformers. *arXiv preprint arXiv:2312.09147*, 2023. 5, 7