
Era3D: High-Resolution Multiview Diffusion using Efficient Row-wise Attention

Peng Li^{1*} Yuan Liu^{2,3*} Xiaoxiao Long^{1,2†} Feihu Zhang³ Cheng Lin² Mengfei Li¹

Xingqun Qi¹ Shanghang Zhang⁴ Wenhan Luo¹ Ping Tan^{1,5} Wenping Wang²

Qifeng Liu¹ Yike Guo^{1†}

*Equal contribution †Corresponding author

¹HKUST ²HKU ³DreamTech ⁴PKU ⁵LightIllusion

Abstract

In this paper, we introduce **Era3D**, a novel multiview diffusion method that generates high-resolution multiview images from a single-view image. Despite significant advancements in multiview generation, existing methods still suffer from camera prior mismatch, inefficacy, and low resolution, resulting in poor-quality multiview images. Specifically, these methods assume that the input images should comply with a predefined camera type, e.g. a perspective camera with a fixed focal length, leading to distorted shapes when the assumption fails. Moreover, the full-image or dense multiview attention they employ leads to an exponential explosion of computational complexity as image resolution increases, resulting in prohibitively expensive training costs. To bridge the gap between assumption and reality, Era3D first proposes a diffusion-based camera prediction module to estimate the focal length and elevation of the input image, which allows our method to generate images without shape distortions. Furthermore, a simple but efficient attention layer, named row-wise attention, is used to enforce epipolar priors in the multiview diffusion, facilitating efficient cross-view information fusion. Consequently, compared with state-of-the-art methods, Era3D generates high-quality multiview images with up to a 512×512 resolution while reducing computation complexity by 12x times. Comprehensive experiments demonstrate that Era3D can reconstruct high-quality and detailed 3D meshes from diverse single-view input images, significantly outperforming baseline multiview diffusion methods. Project page: <https://penghtyx.github.io/Era3D/>.

1 Introduction

3D reconstruction from single-view images is an essential task in computer vision and graphics due to its potential applications in game design, virtual reality, and robotics. Early research [68, 70, 79, 118] mainly relies on direct 3D regression on voxels [64, 88, 16], which often leads to oversmoothed results and has difficulty in generalizing to real-world unseen objects due to limited 3D training data [7]. Recently, diffusion models (DMs) [27, 75] show strong generation ability on image or video synthesis by training on extremely large-scale datasets. These diffusion models are promising tools for single-view 3D reconstruction because it is possible to generate novel-view images from the given image to enable 3D reconstruction.



Figure 1: Given single-view image with arbitrary intrinsic and viewpoints, Era3D can generate high-quality multiview images with a resolution of 512×512 on the orthogonal camera setting, which can be used in mesh reconstruction by NeuS [102].

To utilize image DMs for single-view 3D reconstruction, a pioneer work DreamFusion [72] tries to distill a 3D representation like NeRF [59] or Gaussian Splatting [92] from a 2D image diffusion by a Score Distillation Sampling (SDS) loss and many follow-up works improve the distillation-based methods in quality [106] and efficiency [92]. However, these methods suffer from unstable convergence and degenerated quality. Alternatively, recent works such as MVDream [84], Sync-Dreamer [53], Wonder3D [55] and Zero123++ [83] explicitly generate multiview images by multiview diffusion [95, 53] and then reconstruct 3D models from the generated images by neural reconstruction methods [102] or large reconstruction models (LRMs) [30, 45]. Explicitly generating multiview images makes these methods more controllable and efficient than SDS methods and thus is more popular in the single-view 3D reconstruction task.

Despite impressive advancements in multiview diffusion methods [84, 55, 53, 80, 52, 51, 103, 95, 95, 94, 112, 45], efficiently generating novel-view images for high-quality 3D reconstruction remains an open challenge. There are three noticeable challenges in the current multiview diffusion methods. (1) **Inconsistent predefined camera type.** Most multiview diffusion methods assume that the input images are captured by a camera with a predefined focal length. This leads to unwanted distortions when input images are captured by cameras with different camera types or intrinsics, as exemplified in Fig. 3 (e.g., Wonder3D’s assumption of an orthogonal camera leads to distorted meshes when the input image is captured by a perspective camera with a small focal length). (2) **Inefficiency of multiview diffusion.** Multiview diffusion methods usually rely on multiview attention layers to exchange information among different views to generate multiview-consistent images. However, these multiview attention layers are usually implemented by extending the self-attention in Stable Diffusion [78] to all multiview images, which is called dense multiview attention in Fig. 2(a) and results in a significant increase in computation complexity and memory consumption. (3) **Low resolution of generated images.** The limitation above restricts most existing multiview diffusion

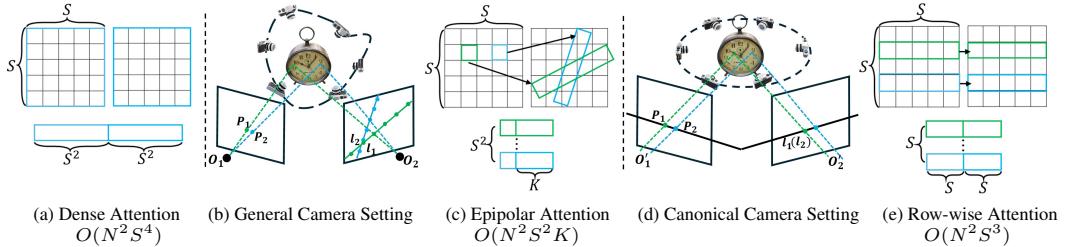


Figure 2: Different types of multiview attention layers. **(a)** In a dense multiview attention layer, all feature vectors of multiview images are fed into an attention block. For a general camera setting **(b)** with arbitrary viewpoints and intrinsics, utilizing epipolar constraint to construct an epipolar attention **(c)** needs to correlate the features on the epipolar line. This means that we need to sample K points along each epipolar line to compute such an attention layer. In our canonical camera setting **(d)** with orthogonal cameras and viewpoints on an elevation of 0° , epipolar lines align with the row of the images across different views **(e)**, which eliminates the need to resample epipolar line to compute epipolar attention. We assume the latent feature map has a resolution of $H \times W$ and $H = W = S$. In such a N -view camera system, row-wise attention reduces the computational complexity to $O(N^2 S^3)$.

models to resolutions of 256×256 , preventing them from reconstructing detailed meshes. Addressing these above challenges is crucial for developing practical and scalable multiview diffusion methods.

In this paper, we introduce Era3D, a novel multiview diffusion method that efficiently generates high-resolution (512×512) consistent-multiview images for single-view 3D reconstruction. Unlike existing methods, Era3D allows images of commonly used camera types as inputs while mitigating the unwanted distortion brought by different camera models. To this end, we employ a unique approach: using different camera models for input images and the generated ones for training, meaning that the input images are allowed to have arbitrary focal lengths and elevations while generated images are with orthogonal cameras and fixed viewpoints of 0° elevations. However, this requires DMs to implicitly infer and rectify the focal lengths and viewpoints of input images in the generation process, which is a challenging task and degenerates the generation quality. To overcome this challenge and improve generation quality, we propose a novel regression and condition scheme and utilize the low-level feature maps of UNet at each denoising step to predict camera information of input images. We find that such a regression and condition scheme facilitates much more accurate camera pose prediction than existing methods [51] and leads to more details in the generation. As shown in Fig. 3, Era3D successfully avoids the above distortion problem brought by the different camera types and focal lengths.

Moreover, drawing inspiration from epipolar attention [98], Era3D enables efficient training for high-resolution multiview generation by introducing a novel row-wise multiview attention. Epipolar constraint can be utilized to constrain the attention regions across views and thus improve attention efficiency. However, directly applying such epipolar attention [98] for a general camera setting (Fig. 2(b)) is still memory and computationally inefficient because we have to sample multiple points on epipolar lines for attention. This requires us to construct a 3D grid of features in the view frustums for multiview images, which is too slow and memory-consuming. In contrast, since Era3D generates images with orthogonal cameras on viewpoints of 0° , we find that epipolar lines in our camera setting are aligned with rows of images across different views (Fig. 2(d)), which enables us to propose an efficient row-wise attention layer. Compared with dense multiview attention, row-wise attention significantly reduces memory consumption (35.32GB v.s. 1.66GB), and the computation complexity (220.41ms v.s. 2.23ms) of multiview attention (Fig. 2(e)). Even with Xformers [44],

an accelerating library for attention, the efficiency of row-wise attention still outperforms existing methods by approximately twelve-fold as evident in Tab. 3. Consequently, the proposed row-wise attention allows us to easily scale Era3D to a high resolution of 512×512 to reconstruct more detailed 3D meshes.

Overall, our main contributions are summarized as follows: (1) Era3D is the first method that tries to solve the distortion artifacts brought by the inconsistent camera intrinsic in 3D generation; (2) we design a novel regression and condition scheme to enable diffusion models to take images of arbitrary cameras as inputs while outputting the orthogonal images on the canonical camera setting; (3) we propose row-wise multiview attention, an efficient attention layer for high-resolution multiview image generation; (4) our method achieve state-of-the-art performance for single-view 3D generation [18].

2 Related Works

Our study is primarily centered on the domain of image-to-3D. Unlike early works [10, 73, 92, 54, 106, 81, 82, 48, 97, 31, 13, 80, 32, 25, 2, 37, 105, 14, 113, 46, 17, 47, 49, 74, 114], which concentrate on per-scene optimization based on Score Distillation Sampling [72, 100], we emphasize feed-forward 3D generation.

Image to 3D. Generating 3D assets from images has been extensively researched, paralleling the development of GAN [23, 38, 40, 41, 39, 61, 63, 19, 86, 5, 22, 35] and diffusion models (DMs) [87, 28, 34, 62, 65, 101, 20, 76, 29]. A stream of these works [33, 119, 8, 30, 112], directly produce 3D representations, like SDF [69, 16, 67], NeRF [59, 24, 115, 9, 56, 116, 3], Triplane [6, 26, 85], Gaussian [42, 89, 1, 109] or 3D volume [90]. Zero-1-to-3 [52] and subsequent works [83, 51, 12] represent the scene as a diffusion model conditioned on reference image and camera pose. LRM-based methods [30, 45, 104, 110–112] employ a large transformer architecture to train a triplane representation with a data-driven approach. Another technical line involves generating consistent multiview images first [84, 103, 83, 93, 96, 107], and then robustly reconstructing 3D shapes with NeuS [53, 55], Gaussian Splatting [91, 30, 111] or LRM [108]. Despite great advancement, challenges remain in reconstruction quality, training resolution, or efficiency.

Multiview Diffusion. Cross-view consistency is critical in 3D reconstruction and generation, relying on multiview feature correspondence to estimate 3D structures. MVffusion [95] first proposes generating multiview images in parallel with correspondence-aware attention, facilitating cross-view information interaction, and applies to texture scene meshes. [98, 36] introduces epipolar features into DM to enhance fusion between viewpoints. Zero123++ [83] tiles multi-views into a single image and performs a single pass for multiview generation, also used in Direct2.5 and Instant3D. MVDream [84] and Wonder3D [55] also design multiview self-attention to improve multiview consistency. Syncdreamer [53] composes multiview features into 3D volumes, conducting 3D-aware fusion in 3D noise space. All of the aforementioned methods share the same idea: modeling 3D generation with multiview joint probability distribution. Other works [15, 57, 99] explore the priors from video diffusion model to achieve consistent multiview generation. Following widely used multiview self-attention, we propose more efficient row-wise attention to reduce computation workloads, but without loss of multiview feature interaction.

Camera pose. For 3D generation, early works [52, 84] are trained with fixed focal lens. When inference, one also needs to provide elevation of input for better performance. Further research seeks to mitigate this issue by leveraging fixed poses [55, 83] or incorporating additional elevation prediction modules [50]. LEAP [33] uses pixel-wise similarity, rather than estimated poses, to aggregate multiview features. However, none of these methods consider the distortion error caused by cameras, which can severely affect the reconstruction of real-world data. To overcome this, we opt to generate images at fixed views in the canonical orthogonal space, with simultaneous predictions of elevation and focal distortion.

3 Methods

Era3D is proposed to generate a 3D mesh from a single-view image. The overview is shown in Fig. 4, consisting of three key components. Given an input image with commonly used focal length and arbitrary viewpoint, Era3D generates multiview images in a canonical camera setting as introduced in Sec. 3.1. To improve the generation quality, in Sec. 3.2, we propose a regression and condition

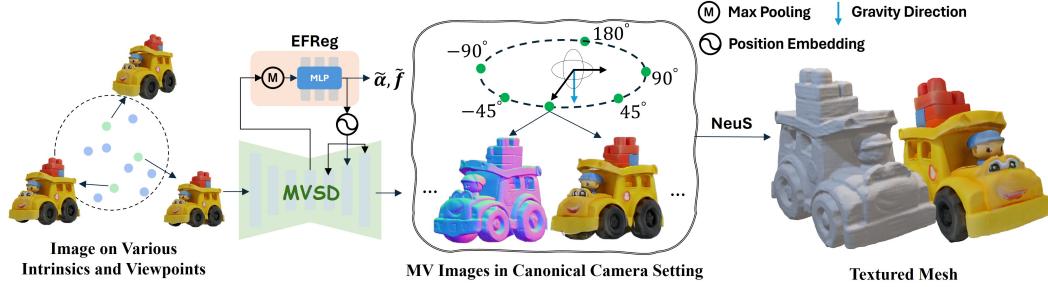


Figure 4: **Overview.** Given a single-view image as input, Era3D applies multiview diffusion to generate multiview consistent images and normal maps in the canonical camera setting, which enables us to reconstruct 3D meshes using NeuS [102, 60].

scheme, enabling diffusion models to predict accurate camera pose and focal lengths and guiding the denoising process. Finally, we considerably reduce memory consumption and improve computation efficiency by proposing row-wise multiview attention (Sec. 3.3), which exchanges information among multiview images to maintain multiview consistency. Finally, we reconstruct the 3D mesh from the generated images and normal maps using neural reconstruction methods like NeuS [102].

3.1 Camera Canonicalization

Perspective distortion problem. Existing multiview diffusion methods [53, 55, 83, 52] assume that the input image and the generated images share the same fixed intrinsic parameters. However, this assumption is often violated in practice, as input images may be captured by arbitrary cameras with varying focal lengths. When the input image has a different intrinsic matrix from the assumed one, the generated multiview images and reconstructed 3D meshes will exhibit perspective distortion, as illustrated in Fig. 3. This issue arises because these models are trained on renderings of the Objaverse [18] dataset with a fixed intrinsic and thus these models are severely biased towards the geometry patterns present in this fixed intrinsic matrix.

Canonical camera setting. To address this problem, Era3D uses different intrinsic parameters for input and generated images. Regardless of the focal length and pose of the input image, we consistently generate orthogonal images with an elevation of 0° . For example, when processing an input image captured at elevation α and azimuth β , Era3D produces a set of multiview images at an azimuth of $\{\beta, \beta + 45^\circ, \beta + 90^\circ, \beta - 45^\circ, \beta - 90^\circ, \beta + 180^\circ\}$ and an elevation of 0° . We refer to the setup of these output images as a *Canonical* camera setting.

3.2 Regression and Condition Scheme

Given an image with an arbitrary viewpoint and focal length, generating novel-view images in the canonical camera setting is challenging because this implicitly puts an additional task on the diffusion model to infer the focal lengths and elevation of the input image. To make it easier, previous methods [52, 53, 4] rely on additional elevation inputs or predictions as the conditions for the diffusion model. However, pose estimation from a single image is inherently ill-posed due to the lack of geometry information. Even though estimating a rough elevation is possible, it is almost impossible for users to estimate the focal length of the input image.

To address this problem, we propose incorporating an Elevation and Focal length Regression module (**EFReg**) into the diffusion model. We use the feature maps of UNet to predict the camera pose in the diffusion process. Our motivation stems from the fact that the feature maps of UNet not only contain the input images but also include the current generation results which provide richer and more informative features for predicting the camera pose.

Specifically, within the middle-level transformer block of the UNet, we apply global average pooling to the hidden feature map \mathbf{H} , yielding a feature vector that is subsequently fed into three Multilayer Perceptron (MLP) layers \mathcal{R}_1 and \mathcal{R}_2 to regress the elevation $\tilde{\alpha}$ and focal lens \tilde{f}

$$\tilde{\alpha} = \mathcal{R}_1(\text{AvgPool}(\mathbf{H})), \tilde{f} = \mathcal{R}_2(\text{AvgPool}(\mathbf{H})). \quad (1)$$

The regressed elevation $\tilde{\alpha}$ and focal length \tilde{f} are supervised by the ground-truth elevation α and focal length f by

$$\ell_{\text{regress}} = \text{MSE}(\tilde{\alpha}, \alpha) + \text{MSE}(\tilde{f}, f). \quad (2)$$

Then, the regressed $\tilde{\alpha}$ and \tilde{f} are used as conditions in the diffusion process. We apply positional encoding on $\tilde{\alpha}$ and \tilde{f} and concatenate them with the time embeddings of the Stable Diffusion model. The concatenated feature vectors are used in all the upsampling layers of the UNet, providing information about the estimated focal lengths and elevations for better denoising.

3.3 Row-wise Multiview Attention (RMA)

To generate multiview-consistent images, multiview diffusion models typically rely on multiview attention layers to exchange information among generated images. Such layers are often implemented by extending the existing self-attention layers of Stable Diffusion to conduct the attention on all the generated multiview images [55, 84, 103, 94]. However, this dense multiview attention can be computationally expensive and memory-intensive, as it processes all pixels of all multiview images. This limitation hinders the scalability of multiview diffusion models to high resolutions, such as 512×512 .

Since the pixels of multiview images are related by epipolar geometry, considering epipolar lines in multiview attention could possibly reduce computational and memory complexity. However, strictly considering epipolar attention [98] in general two-view camera setting still consumes massive computation and memory. As shown in Fig. 2(b), for a pixel on camera O_1 , we find its corresponding epipolar line in camera O_2 by the relative camera pose. Then, we need to sample K points on the epipolar lines to conduct cross attention between these sample points of O_2 and the input pixel of O_1 . In the following, we propose an efficient and compact row-wise multiview attention, which is a special epipolar attention tailored to our canonical camera setting.

In our canonical camera setting, cameras are distributed at an elevation of 0° around an object. Therefore, we can easily demonstrate the following proposition.

Proposition 1 *If two orthogonal cameras look at the origin with their y coordinate aligned with gravity direction and their elevations of 0° as shown in Fig. 2(d), then for a pixel with coordinate $(x, y) = (u, v)$ on one camera, its corresponding epipolar line on other views is $y = v$.*

We leave the proof in the supplementary material. Proposition 1 is a simplification of epipolar constraint, revealing that all epipolar lines correspond to rows in the generated multiview images. Building on this insight, we leverage the epipolar constraint by applying new self-attention layers on the same row across generated images to learn multiview consistency. By exploiting this constraint, we avoid the computational expense of dense multiview attention and instead accurately focus attention on epipolar lines. Moreover, our row-wise attention layer only involves elements from the same row, rather than sampling multiple points on the epipolar line, thereby significantly reducing computational complexity and facilitating training even on high-resolution inputs such as 512×512 .

4 Experiments

Datasets. We trained Era3D on a subset of Objaverse [18]. To construct training images, we render 16 ground-truth images using orthogonal cameras with evenly distributed azimuth from 0° to 360° and a fixed elevation of 0° . Subsequently, for each azimuth, we render 3 more images using perspective cameras and one image using an orthogonal camera, both of which have random elevations sampled from the range $[-20, 40]$ degrees. The perspective camera has a focal length randomly selected from the set $\{35, 50, 85, 105, 135\}$ mm, which are commonly used camera parameters. All the renderings have the resolution of 512×512 . Following the previous methodologies [52, 53], we evaluate the performance of Era3D on the Google Scanned Object [21] dataset, widely regarded as a standard benchmark for 3D generation tasks. Moreover, we also evaluate our methods on in-the-wild images collected from the Internet or generated by image diffusion models [78] to show the generalization ability. The same as previous methods [84, 103], we remove backgrounds and center objects on these in-the-wild images.

Metrics. Our methodology is evaluated in two tasks, novel view synthesis (NVS) and 3D reconstruction. The NVS quality is evaluated by the Learned Perceptual Image Patch Similarity (LPIPS) [117]

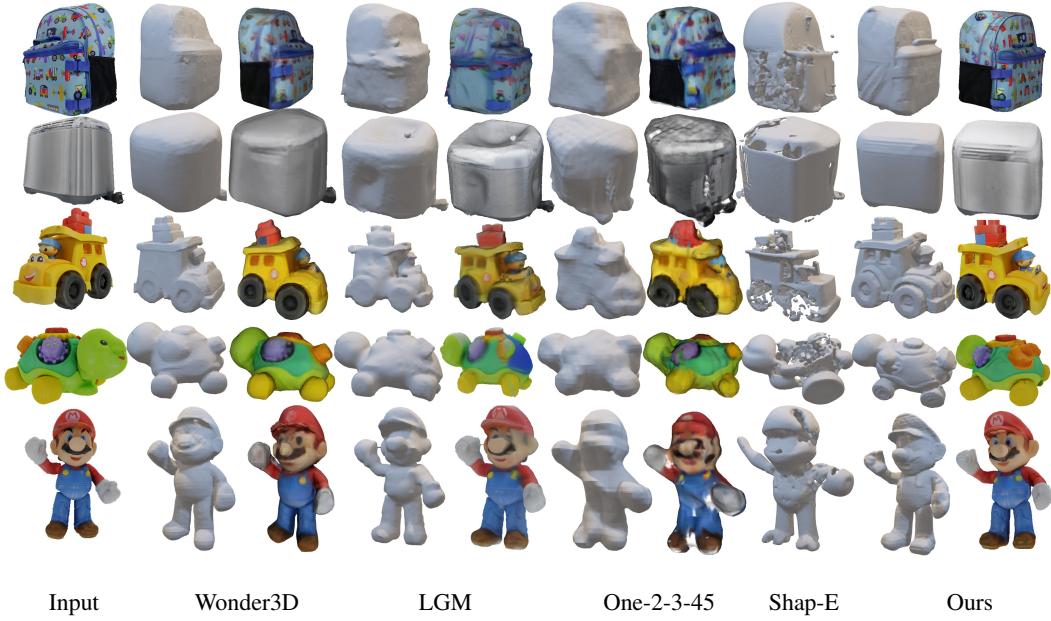


Figure 5: Qualitative comparison of 3D reconstruction results on the GSO dataset [21]. Era3D produces the most high-quality 3D meshes with more details than baseline methods.

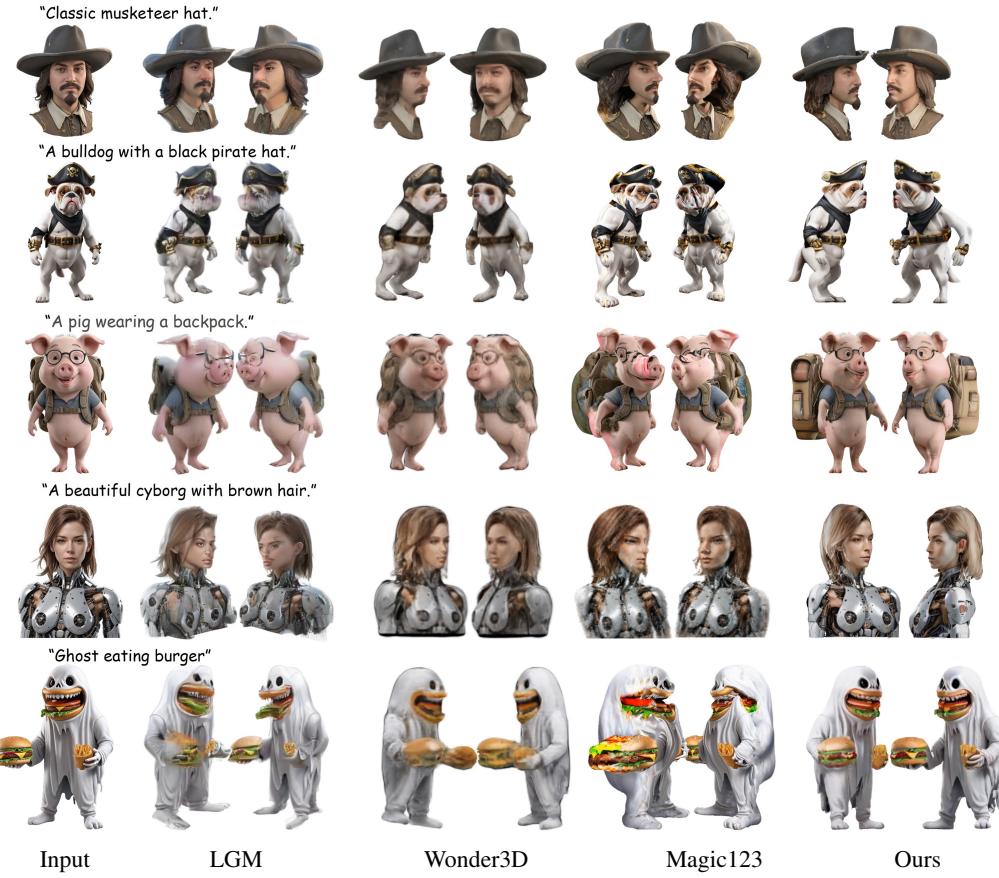


Figure 6: Qualitative comparisons of novel view synthesis quality of reconstructed 3D meshes with single-view images generated by SDXL [71].

between the generated and ground-truth images. LPIPS evaluates the perceptual consistency because there may be slight misalignment between generated and ground-truth images. The 3D reconstruction quality is evaluated by the Chamfer Distance (CD) and the Volume IOU between the reconstructed meshes and the ground truth ones.

Implementation details. Our implementation is built upon the open-source text-to-image model, SD2.1-unclip [77]. We train Era3D on 16 H800 GPUs (each with 80 GB) using a batch size of 128 for 40,000 steps. We set the initial learning rate as 1e-4 and decreased it to 5e-5 after 5,000 steps. The training process takes approximately 30 hours. To conduct classifier-free guidance (CFG) [29], we randomly omit the clip condition at a rate of 0.05. During inference, we employ the DDIM sampler [87] with 40 steps and a CFG scale of 3.0 for the generation. Following Wonder3D [55], we first perform reconstruction with NeuS and then apply a texture refinement step. For further details, please refer to the supplementary material.

4.1 Experimental Results

Novel view synthesis. First, several examples of the multiview images and normal maps generated by Era3D are shown in Fig. 1. The results demonstrate that given input images with varying focal lengths and viewpoints, Era3D can generate high-quality and consistent multiview images and normal maps. When the input image is captured by a perspective camera and its viewpoint is not on an elevation of 0°, Era3D can correctly perceive the elevation of the viewpoint and the perspective distortion. Then, our method learns to generate images of the same object with high fidelity using orthogonal cameras on canonical viewpoints, effectively reducing the artifacts caused by the perspective distortion and improving the reconstruction quality. Moreover, Era3D can produce images in the 512 × 512 resolution, which enables generating much more details like the fine-grained texture on the “Armor” and the complex structures on the “Mecha” in Fig. 1.

Furthermore, we provide a quantitative comparison with other single-view reconstruction methods including RealFusion [58], Zero-1-to-3 [52], SyncDreamer [53] and Wonder3D [55] in Tab. 1. The results show that our method outperforms previous approaches in terms of novel-view-synthesis quality by a significant margin, showcasing the effectiveness of our designs.

Reconstruction. We further conduct experiments to evaluate the quality of reconstructed 3D meshes. We compare our method with RealFusion [58], Zero-1-to-3 [52], One-2-3-45 [51], Shap-E [34], Magic123[73], Wonder3D [55], SyncDreamer [53], and LGM [91]. Reconstructed meshes and their textures on the GSO dataset are shown in Fig. 5 while the renderings of the reconstructed meshes on text-generated images are shown in Fig. 6. As shown in the results, Shap-E fails to generate completed structures. The meshes reconstructed by One-2-3-45 and LGM tend to be over-smoothed and lack details due to the multiview inconsistency in generated images by Zero-1-to-3 [52] or ImageDream [103]. The results of Wonder3D tend to be distorted on these input images rendered with a focal length of 35mm because it assumes the input images are captured by orthogonal cameras. In contrast, our results show significant improvements in terms of completeness and details than these baseline methods.

Quantitative comparisons of Chamfer Distance (CD) and Intersection over Union (IoU) are shown in Tab. 1. Era3D outperforms all other methods, exhibiting lower Chamfer Distance and higher Volume IoU, suggesting that the meshes it generates align more closely with the actual 3D models.

Table 1: Quantitative evaluation of Chamfer distance, IoU (for reconstruction), and LPIPS (for NVS).

Method	CD ↓	IoU↑	LPIPS ↓
RealFusion	0.0819	0.2714	0.283
Zero-1-to-3	0.0339	0.5035	0.166
One-2-3-45	0.0629	0.4086	-
Shap-E	0.0436	0.3584	-
Magic123	0.0516	0.4528	-
SyncDreamer	0.0261	0.5421	0.146
Wonder3D	0.0248	0.5678	0.141
LGM	0.0259	0.5628	-
Ours	0.0217	0.5973	0.126

Table 2: Comparison of pose estimation accuracy. $\tilde{\alpha}$: elevation, \tilde{f} : normalized focal length.

	Method	$\tilde{\alpha}$ / °	\tilde{f} / mm
Error	Dino	10.24	0.28
	One-2-3-45	10.14	-
	Ours	2.69	0.13
Variance	Dino	377.07	0.058
	One-2-3-45	267.44	-
	Ours	112.30	0.036

Table 3: Memory usage and running time of multiview attention with resolution of 256 and 512.

Multiview attention	w/o xFormers				w/ xFormers			
	Memory usage (G)	Running time (ms)	Memory usage (G)	Running time (ms)	256	512	256	512
Dense	3.02	35.32	8.88	220.41	0.99	1.42	1.77	22.96
Epipolar	2.43	24.20	3.57	60.89	1.02	1.71	1.78	20.03
Row-wise	0.95	1.66	0.91	2.23	0.99	1.08	0.28	1.86

4.2 Accuracy of Estimated Elevations and Focal Lengths

Beyond the tasks already discussed, we further evaluate the pose prediction of Era3D on the GSO dataset. We render the images with an elevation of $[-10, 40]$ degrees and focal lengths of $\{35, 50, 85, 105, 135, \infty\}$, respectively. As a baseline method, we employ dinov2_vitb14 feature [66] to predict the pose and train it with the same dataset. We compare our predictions with this baseline method and One-2-3-45. As shown in Tab. 2, Era3D achieves superior performance in error and variance. A more detailed analysis is provided in the supplementary materials.

4.3 Ablations and Discussions

Regression and condition scheme. In Fig. 7, we remove the EFReg and compare the results with our full model to demonstrate the effectiveness of our design. Without regressing a focal length and elevation as conditions during the denoising process, the resulting shape is distorted and fails to generate reasonable novel views in the canonical camera setting. In comparison, adding our EFReg module and conditioning on the predicted elevations and focal lengths provides effective guidance to generate undistorted cross-domain images, thereby resulting in more accurate 3D reconstructions.

Row-wise Multiview Attention. As illustrated in Fig. 1, our proposed RMA effectively facilitates information exchange among multiview images, yielding consistent results comparable to those achieved by dense multiview attention layers in [103, 55]. In a N -view camera system, assuming a latent feature with size of $S \times S$, our RMA design significantly improves training efficiency by reducing the computational complexity of attention layers from $O(N^2S^4)$ to $O(N^2S^3)$, as shown in Fig. 2. While epipolar attention also achieves a complexity reduction to $O(N^2S^2K)$, where K is the sample number, it does so at the cost of increased memory and time consumption due to the sampling process. To further highlight the efficiency of RMA over dense multiview attention, we present the memory usage and running time of both 256 and 512 resolutions. We use the epipolar attention implementation in [36]. As listed in Tab. 3, the advantage of RMA becomes increasingly obvious as the resolution grows. At a resolution of 512, RMA achieves a thirty-fold reduction in memory usage and a nearly hundred-fold reduction in running time. Even with xFormers [44], our method substantially improves training efficiency by a large margin (22.9 ms vs. 1.86 ms). This efficiency enables training models on higher resolutions or with denser views without significantly increasing computational efficiency and demand, thereby maintaining a lightweight framework.

5 Limitation and Conclusion

Limitations. Though Era3D achieves improvements on the multiview generation task, our method struggles to generate intricate geometries like thin structures because we only generate 6 multiview images and such sparse generated images have difficulty in modeling complex geometries. Since the reconstruction algorithm is based on Neural SDF, Era3D cannot reconstruct meshes with open surfaces. In future works, we could integrate our framework with other 3D representations, such as Gaussian splatting, to improve both rendering and geometry quality.

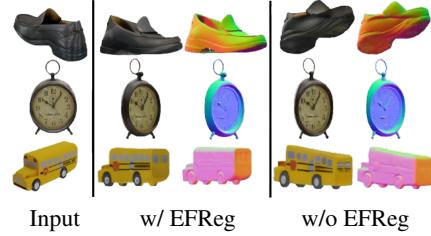


Figure 7: Ablation study of EFReg.

Conclusion. In this paper, we present Era3D, a high-quality multiview generation method for single-view 3D reconstruction. In Era3D, we propose to generate images in a canonical camera setting while allowing input images to have arbitrary camera intrinsics and viewpoints. To improve the generation quality, we design a regression and condition scheme to predict the focal length and elevation of input images, which are further conditioned to the diffusion process. Additionally, we employ row-wise multiview attention to replace dense attention, significantly reducing computational workloads and facilitating high-resolution cross-view generation. Compared with baseline methods, Era3D achieves superior geometry quality in single-view 3D reconstruction.

References

- [1] Rameen Abdal, Wang Yifan, Zifan Shi, Yinghao Xu, Ryan Po, Zhengfei Kuang, Qifeng Chen, Dit-Yan Yeung, and Gordon Wetzstein. Gaussian shell maps for efficient 3d human generation. [arXiv preprint arXiv:2311.17857](#), 2023.
- [2] Titas Auciukevičius, Zexiang Xu, Matthew Fisher, Paul Henderson, Hakan Bilen, Niloy J Mitra, and Paul Guerrero. Renderdiffusion: Image diffusion for 3d reconstruction, inpainting and generation. In [CVPR](#), 2023.
- [3] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In [Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition](#), pages 5470–5479, 2022.
- [4] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. [arXiv preprint arXiv:2311.15127](#), 2023.
- [5] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. [arXiv preprint arXiv:1809.11096](#), 2018.
- [6] Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. Efficient geometry-aware 3d generative adversarial networks. In [Proceedings of the IEEE/CVF conference on computer vision and pattern recognition](#), pages 16123–16133, 2022.
- [7] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. [arXiv preprint arXiv:1512.03012](#), 2015.
- [8] David Charatan, Sizhe Li, Andrea Tagliasacchi, and Vincent Sitzmann. pixelsplat: 3d gaussian splats from image pairs for scalable generalizable 3d reconstruction. [arXiv preprint arXiv:2312.12337](#), 2023.
- [9] Anpei Chen, Zexiang Xu, Fuqiang Zhao, Xiaoshuai Zhang, Fanbo Xiang, Jingyi Yu, and Hao Su. Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo. In [ICCV](#), 2021.
- [10] Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. In [Proceedings of the IEEE/CVF International Conference on Computer Vision](#), pages 22246–22256, 2023.
- [11] Ting Chen. On the importance of noise scheduling for diffusion models. [arXiv preprint arXiv:2301.10972](#), 2023.
- [12] Yabo Chen, Jiemin Fang, Yuyang Huang, Taoran Yi, Xiaopeng Zhang, Lingxi Xie, Xinggang Wang, Wenrui Dai, Hongkai Xiong, and Qi Tian. Cascade-zero123: One image to highly consistent 3d with self-prompted nearby views. [arXiv preprint arXiv:2312.04424](#), 2023.
- [13] Yiwen Chen, Chi Zhang, Xiaofeng Yang, Zhongang Cai, Gang Yu, Lei Yang, and Guosheng Lin. It3d: Improved text-to-3d generation with explicit view synthesis. In [Proceedings of the AAAI Conference on Artificial Intelligence](#), volume 38, pages 1237–1244, 2024.
- [14] Zilong Chen, Feng Wang, and Huaping Liu. Text-to-3d using gaussian splatting. [arXiv preprint arXiv:2309.16585](#), 2023.
- [15] Zilong Chen, Yikai Wang, Feng Wang, Zhengyi Wang, and Huaping Liu. V3d: Video diffusion models are effective 3d generators. [arXiv preprint arXiv:2403.06738](#), 2024.

- [16] Yen-Chi Cheng, Hsin-Ying Lee, Sergey Tulyakov, Alexander G Schwing, and Liang-Yan Gui. Sdfusion: Multimodal 3d shape completion, reconstruction, and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4456–4465, 2023.
- [17] Jaeyoung Chung, Suyoung Lee, Hyeongjin Nam, Jaerin Lee, and Kyoung Mu Lee. Luciddreamer: Domain-free generation of 3d gaussian splatting scenes. *arXiv preprint arXiv:2311.13384*, 2023.
- [18] 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.
- [19] Yu Deng, Jiaolong Yang, Jianfeng Xiang, and Xin Tong. Gram: Generative radiance manifolds for 3d-aware image generation. In *CVPR*, 2022.
- [20] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.
- [21] 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.
- [22] Jun Gao, Tianchang Shen, Zian Wang, Wenzheng Chen, Kangxue Yin, Daiqing Li, Or Litany, Zan Gojcic, and Sanja Fidler. Get3d: A generative model of high quality 3d textured shapes learned from images. *NeurIPS*, 2022.
- [23] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, 2014.
- [24] Jiatao Gu, Lingjie Liu, Peng Wang, and Christian Theobalt. Stylenet: A style-based 3d-aware generator for high-resolution image synthesis. *arXiv preprint arXiv:2110.08985*, 2021.
- [25] Jiatao Gu, Alex Trevithick, Kai-En Lin, Joshua M Susskind, Christian Theobalt, Lingjie Liu, and Ravi Ramamoorthi. Nerfdiff: Single-image view synthesis with nerf-guided distillation from 3d-aware diffusion. 2023.
- [26] Anchit Gupta, Wenhan Xiong, Yixin Nie, Ian Jones, and Barlas Oğuz. 3dgen: Triplane latent diffusion for textured mesh generation. *arXiv preprint arXiv:2303.05371*, 2023.
- [27] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.
- [28] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 2020.
- [29] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.
- [30] 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.
- [31] Yukun Huang, Jianan Wang, Yukai Shi, Xianbiao Qi, Zheng-Jun Zha, and Lei Zhang. Dreamtime: An improved optimization strategy for text-to-3d content creation. *arXiv preprint arXiv:2306.12422*, 2023.
- [32] Ajay Jain, Ben Mildenhall, Jonathan T. Barron, Pieter Abbeel, and Ben Poole. Zero-shot text-guided object generation with dream fields. In *CVPR*, 2022.
- [33] Hanwen Jiang, Zhenyu Jiang, Yue Zhao, and Qixing Huang. Leap: Liberate sparse-view 3d modeling from camera poses. *arXiv preprint arXiv:2310.01410*, 2023.
- [34] Heewoo Jun and Alex Nichol. Shap-e: Generating conditional 3d implicit functions. *arXiv preprint arXiv:2305.02463*, 2023.
- [35] Minguk Kang, Jun-Yan Zhu, Richard Zhang, Jaesik Park, Eli Shechtman, Sylvain Paris, and Taesung Park. Scaling up gans for text-to-image synthesis. In *CVPR*, 2023.
- [36] Yash Kant, Ziyi Wu, Michael Vasilkovsky, Guocheng Qian, Jian Ren, Riza Alp Guler, Bernard Ghanem, Sergey Tulyakov, Igor Gilitschenski, and Aliaksandr Siarohin. Spad: Spatially aware multiview diffusers. *arXiv preprint arXiv:2402.05235*, 2024.

- [37] Animesh Karnewar, Andrea Vedaldi, David Novotny, and Niloy J Mitra. Holodiffusion: Training a 3d diffusion model using 2d images. In *CVPR*, 2023.
- [38] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *ICLR*, 2018.
- [39] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. In *NeurIPS*, 2021.
- [40] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019.
- [41] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *CVPR*, 2020.
- [42] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4):1–14, 2023.
- [43] Samuli Laine, Janne Hellsten, Tero Karras, Yeongho Seol, Jaakko Lehtinen, and Timo Aila. Modular primitives for high-performance differentiable rendering. *ACM Transactions on Graphics*, 39(6), 2020.
- [44] Benjamin Lefauveux, Francisco Massa, Diana Liskovich, Wenhan Xiong, Vittorio Caggiano, Sean Naren, Min Xu, Jieru Hu, Marta Tintore, Susan Zhang, Patrick Labatut, Daniel Haziza, Luca Wehrstedt, Jeremy Reizenstein, and Grigory Sizov. xformers: A modular and hackable transformer modelling library. <https://github.com/facebookresearch/xformers>, 2022.
- [45] 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. *arXiv preprint arXiv:2311.06214*, 2023.
- [46] Xinhai Li, Huaibin Wang, and Kuo-Kun Tseng. Gaussiandiffusion: 3d gaussian splatting for denoising diffusion probabilistic models with structured noise. *arXiv preprint arXiv:2311.11221*, 2023.
- [47] Yixun Liang, Xin Yang, Jiantao Lin, Haodong Li, Xiaogang Xu, and Yingcong Chen. Luciddreamer: Towards high-fidelity text-to-3d generation via interval score matching. *arXiv preprint arXiv:2311.11284*, 2023.
- [48] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 300–309, 2023.
- [49] Huan Ling, Seung Wook Kim, Antonio Torralba, Sanja Fidler, and Karsten Kreis. Align your gaussians: Text-to-4d with dynamic 3d gaussians and composed diffusion models. *arXiv preprint arXiv:2312.13763*, 2023.
- [50] 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.
- [51] Minghua Liu, Chao Xu, Haian Jin, Linghao Chen, Zexiang Xu, Hao Su, et al. One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization. *arXiv preprint arXiv:2306.16928*, 2023.
- [52] 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.
- [53] 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.
- [54] Zexiang Liu, Yangguang Li, Youtian Lin, Xin Yu, Sida Peng, Yan-Pei Cao, Xiaojuan Qi, Xiaoshui Huang, Ding Liang, and Wanli Ouyang. Unidream: Unifying diffusion priors for relightable text-to-3d generation. *arXiv preprint arXiv:2312.08754*, 2023.
- [55] 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. *arXiv preprint arXiv:2310.15008*, 2023.

- [56] Xiaoxiao Long, Cheng Lin, Peng Wang, Taku Komura, and Wenping Wang. Sparseneus: Fast generalizable neural surface reconstruction from sparse views. In *ECCV*, 2022.
- [57] Luke Melas-Kyriazi, Iro Laina, Christian Rupprecht, Natalia Neverova, Andrea Vedaldi, Oran Gafni, and Filippos Kokkinos. Im-3d: Iterative multiview diffusion and reconstruction for high-quality 3d generation. *arXiv preprint arXiv:2402.08682*, 2024.
- [58] Luke Melas-Kyriazi, Christian Rupprecht, Iro Laina, and Andrea Vedaldi. Realfusion: 360° reconstruction of any object from a single image. In *Arxiv*, 2023.
- [59] 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.
- [60] Thomas Müller, 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.
- [61] Thu Nguyen-Phuoc, Chuan Li, Lucas Theis, Christian Richardt, and Yong-Liang Yang. Hologan: Unsupervised learning of 3d representations from natural images. In *ICCV*, 2019.
- [62] Alex Nichol, Heewoo Jun, Prafulla Dhariwal, Pamela Mishkin, and Mark Chen. Point-e: A system for generating 3d point clouds from complex prompts. *arXiv preprint arXiv:2212.08751*, 2022.
- [63] Michael Niemeyer and Andreas Geiger. Giraffe: Representing scenes as compositional generative neural feature fields. In *CVPR*, 2021.
- [64] Matthias Nießner, Michael Zollhöfer, Shahram Izadi, and Marc Stamminger. Real-time 3d reconstruction at scale using voxel hashing. *ACM Transactions on Graphics (ToG)*, 32(6):1–11, 2013.
- [65] Evangelos Ntavelis, Aliaksandr Siarohin, Kyle Olszewski, Chaoyang Wang, Luc Van Gool, and Sergey Tulyakov. Autodecoding latent 3d diffusion models. *arXiv preprint arXiv:2307.05445*, 2023.
- [66] Maxime Oquab, Timothée Darcret, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023.
- [67] Roy Or-El, Xuan Luo, Mengyi Shan, Eli Shechtman, Jeong Joon Park, and Ira Kemelmacher-Shlizerman. Stylesdf: High-resolution 3d-consistent image and geometry generation. In *CVPR*, 2022.
- [68] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deep sdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 165–174, 2019.
- [69] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deep sdf: Learning continuous signed distance functions for shape representation. In *CVPR*, 2019.
- [70] Songyou Peng, Michael Niemeyer, Lars Mescheder, Marc Pollefeys, and Andreas Geiger. Convolutional occupancy networks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pages 523–540. Springer, 2020.
- [71] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.
- [72] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022.
- [73] Guocheng Qian, Jinjie Mai, Abdullah Hamdi, Jian Ren, Aliaksandr Siarohin, Bing Li, Hsin-Ying Lee, Ivan Skorokhodov, Peter Wonka, Sergey Tulyakov, et al. Magic123: One image to high-quality 3d object generation using both 2d and 3d diffusion priors. *arXiv preprint arXiv:2306.17843*, 2023.
- [74] Jiawei Ren, Liang Pan, Jiaxiang Tang, Chi Zhang, Ang Cao, Gang Zeng, and Ziwei Liu. Dreamgaussian4d: Generative 4d gaussian splatting. *arXiv preprint arXiv:2312.17142*, 2023.
- [75] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.

- [76] 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.
- [77] 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 (CVPR)*, pages 10684–10695, June 2022.
- [78] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021.
- [79] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2304–2314, 2019.
- [80] Kyle Sargent, Zizhang Li, Tanmay Shah, Charles Herrmann, Hong-Xing Yu, Yunzhi Zhang, Eric Ryan Chan, Dmitry Lagun, Li Fei-Fei, Deqing Sun, et al. Zeronvs: Zero-shot 360-degree view synthesis from a single real image. [arXiv preprint arXiv:2310.17994](#), 2023.
- [81] Hoigi Seo, Hayeon Kim, Gwanghyun Kim, and Se Young Chun. Ditto-nerf: Diffusion-based iterative text to omni-directional 3d model. [arXiv preprint arXiv:2304.02827](#), 2023.
- [82] Junyoung Seo, Wooseok Jang, Min-Seop Kwak, Hyeonsu Kim, Jaehoon Ko, Junho Kim, Jin-Hwa Kim, Jiyoung Lee, and Seungryong Kim. Let 2d diffusion model know 3d-consistency for robust text-to-3d generation. [arXiv preprint arXiv:2303.07937](#), 2023.
- [83] Ruoxi Shi, Hansheng Chen, Zhuoyang Zhang, Minghua Liu, Chao Xu, Xinyue Wei, Linghao Chen, Chong Zeng, and Hao Su. Zerol123++: a single image to consistent multi-view diffusion base model. [arXiv preprint arXiv:2310.15110](#), 2023.
- [84] Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. [arXiv preprint arXiv:2308.16512](#), 2023.
- [85] J. Ryan Shue, Eric Ryan Chan, Ryan Po, Zachary Ankner, Jiajun Wu, and Gordon Wetzstein. 3d neural field generation using triplane diffusion. In *CVPR*, 2023.
- [86] Ivan Skorokhodov, Sergey Tulyakov, Yiqun Wang, and Peter Wonka. Epigraf: Rethinking training of 3d gans. In *NeurIPS*, 2022.
- [87] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. [arXiv preprint arXiv:2010.02502](#), 2020.
- [88] Cheng Sun, Min Sun, and Hwann-Tzong Chen. Direct voxel grid optimization: Super-fast convergence for radiance fields reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5459–5469, 2022.
- [89] Stanislaw Szymanowicz, Christian Rupprecht, and Andrea Vedaldi. Splatter image: Ultra-fast single-view 3d reconstruction. [arXiv preprint arXiv:2312.13150](#), 2023.
- [90] Stanislaw Szymanowicz, Christian Rupprecht, and Andrea Vedaldi. Viewset diffusion;(0-) image-conditioned 3d generative models from 2d data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8863–8873, 2023.
- [91] 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. [arXiv preprint arXiv:2402.05054](#), 2024.
- [92] 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.
- [93] Junshu Tang, Tengfei Wang, Bo Zhang, Ting Zhang, Ran Yi, Lizhuang Ma, and Dong Chen. Make-it-3d: High-fidelity 3d creation from a single image with diffusion prior. [arXiv preprint arXiv:2303.14184](#), 2023.
- [94] Shitao Tang, Jiacheng Chen, Dilin Wang, Chengzhou Tang, Fuyang Zhang, Yuchen Fan, Vikas Chandra, Yasutaka Furukawa, and Rakesh Ranjan. Mvdiffusion++: A dense high-resolution multi-view diffusion model for single or sparse-view 3d object reconstruction. [arXiv preprint arXiv:2402.12712](#), 2024.

- [95] Shitao Tang, Fuayng Zhang, Jiacheng Chen, Peng Wang, and Furukawa Yasutaka. Mvdiffusion: Enabling holistic multi-view image generation with correspondence-aware diffusion. [arXiv preprint 2307.01097](#), 2023.
- [96] Ayush Tewari, Tianwei Yin, George Cazenavette, Semon Rezhikov, Josh Tenenbaum, Frédo Durand, Bill Freeman, and Vincent Sitzmann. Diffusion with forward models: Solving stochastic inverse problems without direct supervision. [Advances in Neural Information Processing Systems](#), 36, 2024.
- [97] Christina Tsalicoglou, Fabian Manhardt, Alessio Tonioni, Michael Niemeyer, and Federico Tombari. Textmesh: Generation of realistic 3d meshes from text prompts. [arXiv preprint arXiv:2304.12439](#), 2023.
- [98] Hung-Yu Tseng, Qinbo Li, Changil Kim, Suhib Alsisan, Jia-Bin Huang, and Johannes Kopf. Consistent view synthesis with pose-guided diffusion models. In [Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition](#), pages 16773–16783, 2023.
- [99] 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. [arXiv preprint arXiv:2403.12008](#), 2024.
- [100] Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A Yeh, and Greg Shakhnarovich. Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. In [Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition](#), pages 12619–12629, 2023.
- [101] Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A Yeh, and Greg Shakhnarovich. Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. In [Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition](#), pages 12619–12629, 2023.
- [102] 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.
- [103] Peng Wang and Yichun Shi. Imagedream: Image-prompt multi-view diffusion for 3d generation. [arXiv preprint arXiv:2312.02201](#), 2023.
- [104] Peng Wang, Hao Tan, Sai Bi, Yinghao Xu, Fujun Luan, Kalyan Sunkavalli, Wenping Wang, Zexiang Xu, and Kai Zhang. Pf-lrm: Pose-free large reconstruction model for joint pose and shape prediction. [arXiv preprint arXiv:2311.12024](#), 2023.
- [105] Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul P Srinivasan, Howard Zhou, Jonathan T Barron, Ricardo Martin-Brualla, Noah Snavely, and Thomas Funkhouser. Ibrnet: Learning multi-view image-based rendering. In [CVPR](#), 2021.
- [106] 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.
- [107] 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.
- [108] Xinyue Wei, Kai Zhang, Sai Bi, Hao Tan, Fujun Luan, Valentin Deschaintre, Kalyan Sunkavalli, Hao Su, and Zexiang Xu. Meshlrm: Large reconstruction model for high-quality mesh. [arXiv preprint arXiv:2404.12385](#), 2024.
- [109] Dejia Xu, Ye Yuan, Morteza Mardani, Sifei Liu, Jiaming Song, Zhangyang Wang, and Arash Vahdat. Agg: Amortized generative 3d gaussians for single image to 3d. [arXiv preprint arXiv:2401.04099](#), 2024.
- [110] 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.
- [111] Yinghao Xu, Zifan Shi, Wang Yifan, Hansheng Chen, Ceyuan Yang, Sida Peng, Yujun Shen, and Gordon Wetzstein. Grm: Large gaussian reconstruction model for efficient 3d reconstruction and generation. [arXiv preprint arXiv:2403.14621](#), 2024.
- [112] Yinghao Xu, Hao Tan, Fujun Luan, Sai Bi, Peng Wang, Jiahao Li, Zifan Shi, Kalyan Sunkavalli, Gordon Wetzstein, Zexiang Xu, et al. Dmv3d: Denoising multi-view diffusion using 3d large reconstruction model. [arXiv preprint arXiv:2311.09217](#), 2023.

- [113] Taoran Yi, Jiemin Fang, Guanjun Wu, Lingxi Xie, Xiaopeng Zhang, Wenyu Liu, Qi Tian, and Xinggang Wang. Gaussiandreamer: Fast generation from text to 3d gaussian splatting with point cloud priors. [arXiv preprint arXiv:2310.08529](#), 2023.
- [114] Yuyang Yin, Dejia Xu, Zhangyang Wang, Yao Zhao, and Yunchao Wei. 4dgen: Grounded 4d content generation with spatial-temporal consistency. [arXiv preprint arXiv:2312.17225](#), 2023.
- [115] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In [CVPR](#), 2021.
- [116] Kai Zhang, Gernot Riegler, Noah Snavely, and Vladlen Koltun. Nerf++: Analyzing and improving neural radiance fields. [arXiv preprint arXiv:2010.07492](#), 2020.
- [117] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In [Proceedings of the IEEE conference on computer vision and pattern recognition](#), pages 586–595, 2018.
- [118] Ruichen Zheng, Peng Li, Haoqian Wang, and Tao Yu. Learning visibility field for detailed 3d human reconstruction and relighting. In [Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition](#), pages 216–226, 2023.
- [119] 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.

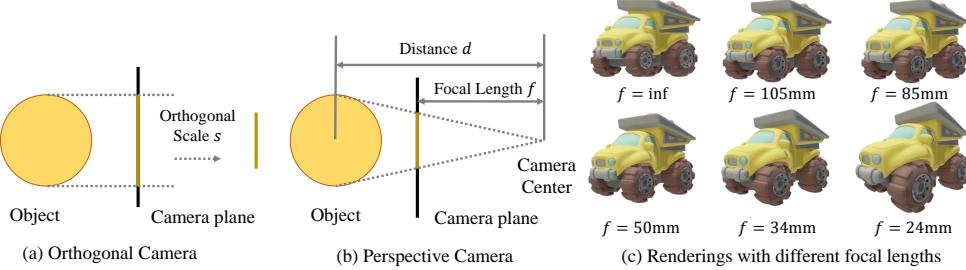


Figure 8: Equivalence between orthogonal and perspective camera models and our rendering samples in GSO dataset [21].

A Supplementary Material

A.1 Implementation Details

Data preparation. To render training data, we randomly choose a focal length for the input image from a predefined set of focal lengths and we always use the same orthogonal camera for all generated images. To train Era3D, we need to construct an orthogonal camera to render generation images and the equivalent perspective cameras to render input images. The equivalence here means that the renderings from these two kinds of cameras have almost the same size, for reducing training bias. As shown in Fig. 8 (a) and (b), given a predefined orthogonal scale s and a given focal length f , we compute the distance d by $d = f/s$, where s is the orthogonal scale to scale the size of the rendered image. We will adjust the distance from the camera center to the object to the value d to make the rendered image as similar as possible.

Normalized focal length. In our experiment, we choose several discrete focal lengths, {24, 35, 50, 85, 105, 135} mm, and an orthogonal camera to render multiview images for training. Due to the broad range of focal lengths, we regress the normalized focal lengths instead of the actual ones. Specifically, we normalize them with minimum focal,

$$\tilde{f} = 24/f. \quad (3)$$

For the orthogonal views, we set $\tilde{f} = 0$.

Cross-domain attention. In contrast to Wonder3D [55], we do not incorporate an additional attention layer to fuse information between cross-domain images. Instead, in earlier experiments, we combined both the cross-domain fusion module and the multiview fusion module within the same row-wise attention block to enhance training efficiency. This initial setup, however, leads to suboptimal outcomes, as shown in Fig. 4, where it tends to produce overly smoothed normal maps. We hypothesize that the significant differences between domain images are the primary cause of this issue, as the row-wise information alone is insufficient for learning fine-grained features across domains. In contrast, integrating the cross-domain module within the self-attention block enables the utilization of complete image features, allowing the model to better handle these disparities and achieve improved generalization. Our attention block consists of self-cross-domain attention, row-wise multiview attention, and cross-domain attention, which allows us to merge the two-stage training into a single-stage joint training, effectively aligning images of both domains.



Table 4: Ablation study on the position of the cross domain-block.

Reconstruction and texture refinement. We perform reconstruction from the generated multiview normal and color images using NeuS. However, since we only generate sparse views, the rendered images of the reconstructed meshes may not attain the quality of the generated ones, particularly for objects with complex textures. To address this issue, we employ differential rendering to refine the textures on the reconstructed meshes. Specifically, we preserve the geometry of NeuS and initialize the texture with vertex colors. We utilize [43] to render multiview images in predefined views and optimize the vertex colors with the corresponding generated images. This process takes less than 1 second and significantly improves the texture quality.

Noise scheduler. The original SD2.1-Unclip base model employs a scaled linear noise scheduler during training, which is effective for smaller sample sizes, such as 256. However, this scheduler tends to restrict the generative capabilities of models at larger sizes, such as 512. Drawing inspiration from recent research by Chen [11], we

implemented a linear noise scheduler throughout our experiments to enhance generation quality and accelerate the convergence of the background. This adjustment proves critical in supporting the model’s performance and efficiency.

A.2 Proof of Proposition 1

Given two cameras O_1 and O_2 with relative rotation R and translation t , let x_1 and x_2 denote the homogeneous coordinates of corresponding points in the image planes, respectively, epipolar constraint can be expressed as

$$x_2^T E x_1 = 0, \quad (4)$$

where E is the fundamental matrix. Then, the epipolar line in O_2 can be represented as

$$l = E x_1, E = [t]_{\times} R, \quad (5)$$

where $[t]_{\times}$ denotes the skew-symmetric matrix generated by t . However, epipolar attention needs to query dense points on epipolar lines, which leads to significant computational inefficiency, due to arbitrary directions and varying lengths of epipolar lines.

Row-wise multiview attention is a special epipolar attention in our defined canonical camera setting, as depicted in Fig. 2(d). Assuming the y-axis represents the gravity direction, the x-axis denotes the right-hand direction, and the z-axis points away from the camera to the origin of the object. The relative R and t between two cameras can be represented as

$$R = \begin{bmatrix} \cos(\theta) & 0 & \sin(\theta) \\ 0 & 1 & 0 \\ -\sin(\theta) & 0 & \cos(\theta) \end{bmatrix}, t = [t_x, 0, t_z]^T, \quad (6)$$

where θ is the azimuth angle of the cameras. Considering the point $P_1(x_1, y_1, z_1)$ in camera O'_1 , the coordinates of P_1 in camera O'_2 are given by

$$P_2(x_2, y_2, z_2) = RP_1 + t = \begin{bmatrix} \cos(\theta)x_1 + \sin(\theta)z_1 + t_x \\ y_1 \\ \cos(\theta)z_1 - \sin(\theta)x_1 + t_z \end{bmatrix}. \quad (7)$$

It is observed that the scene points from different views share the same y-coordinate. Assuming the identical orthographic scale, the y-coordinates of projections on the image plane are also equivalent. Extending a single sample to a line of points with the same y-coordinate, all projections of those points on two image planes are on the same row.

A.3 Pose Estimation

During inference, we obtain the final pose by averaging the class-free guidance results from all denoising steps,

$$\begin{aligned} \tilde{\alpha} &= \frac{1}{T} \sum_{t=1}^T [(1+w)\alpha_{\theta}^t(z, c) - w\alpha_{\theta}^t(z)], \\ \tilde{f} &= \frac{1}{T} \sum_{t=1}^T [(1+w)f_{\theta}^t(z, c) - wf_{\theta}^t(z)], \end{aligned} \quad (8)$$

where z is the latent, c is the condition, w is the CFG scale, θ is the parameters of UNet and MLP and T is the denoising step. We observe that the averaged class-free guidance predictions achieve the highest accuracy. We attribute this to the random dropping of condition images during training. We do not rely solely on the prediction at the final step because we constrain the estimated elevation and focal length with the ground pose rather than the noisy ones at each denoising step. We conduct extensive experiments to evaluate the accuracy of the estimated pose on the GSO dataset.

Elevation. As listed in Tab. 5, Dino and One-2-3-45 only learn the relative trend as the elevation increases, resulting in large errors for cases with high elevations. This is because our baseline only employs the feature of a single image, which is insufficient for predicting the absolute elevation. The prediction of One-2-3-45 mainly depends on the multiview images generated by Zero-1-to-3. The inconsistency of generated images leads to ambiguity in prediction. In comparison, our results are closer to the ground truth elevation. In most cases, the variance of our method is significantly lower than that of Dino and One-2-3-45, demonstrating remarkable robustness.

Focal. We report the error and variance of normalized focal predictions for the baseline and our method in Tab. 6. It is observed that the baseline cannot distinguish the differences between various focal lengths. In contrast, our model can predict the large distortion (e.g., focal=35, 50) of input images, which is advantageous for correcting

Table 5: Elevation accuracy on GSO dataset.

Method	Elevation / $^{\circ}$												
	-10		0		10		20		30		40		
	Pred	Err	Pred	Err	Pred	Err	Pred	Err	Pred	Err	Pred	Err	
Mean	Dino	-13.58	3.58	-6.83	6.83	4.63	5.37	10.21	9.79	16.72	14.28	18.41	
	One-2-3-45	-11.93	1.93	-5.23	5.23	-0.06	10.06	11.17	8.83	15.5	14.5	19.73	
	Ours	-9.71	0.29	-2.20	2.20	5.59	4.41	23.85	3.85	34.03	4.03	38.67	1.33
Var	Dino	202.37	-	252.51	-	153.67	-	365.32	-	463.92	-	824.63	-
	One-2-3-45	175.02	-	103.64	-	168.61	-	200.97	-	326.05	-	630.32	-
	Ours	32.1	-	83.63	-	163.38	-	113.55	-	134.69	-	146.46	-

Table 6: Focal length accuracy on GSO dataset.

Method	0.68 (35)		0.48 (50)		0.28 (85)		0.22 (105)		0.17 (135)		0.0 (ortho)	
	Pred	Err	Pred	Err	Pred	Err	Pred	Err	Pred	Err	Pred	Err
Mean	Dino	0.75	0.07	0.46	0.02	0.53	0.27	0.56	0.34	0.57	0.4	0.59
	Ours	0.69	0.01	0.54	0.06	0.37	0.09	0.33	0.11	0.34	0.17	0.32
var	Dino	0.069	-	0.073	-	0.054	-	0.049	-	0.052	-	0.049
	Ours	0.041	-	0.035	-	0.024	-	0.02	-	0.069	-	0.032

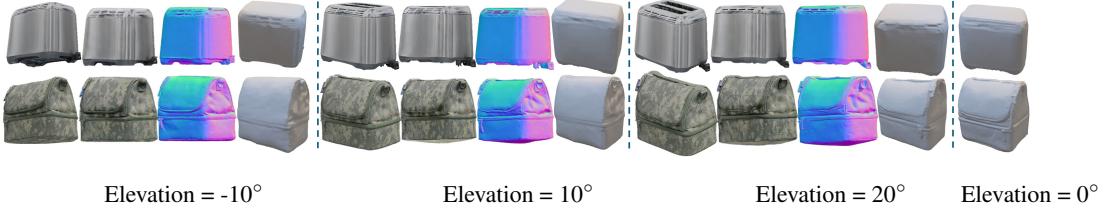


Figure 9: Generation results of various elevation. We use reconstructions from the view of Elevation = 0° as the reference. Different inputs generate consistent results.



Figure 10: Generation results of various distortions. We employ reconstructions from orthogonal inputs as the reference. For each case, we illustrate the input along with the generated color, normal, and mesh.

Table 7: Quantitative evaluation on images with various focal lengths. CD: Chamfer Distance.

Pose	$\alpha=0$					$f=\infty$					$\alpha=0$
	$f=35$	$f=50$	$f=85$	$f=105$	$f=135$	$\alpha=-10$	$\alpha=10$	$\alpha=20$	$f=30$	$f=40$	
CD	0.0223	0.0219	0.0216	0.0214	0.0214	0.0217	0.0216	0.216	0.0219	0.0217	0.0213

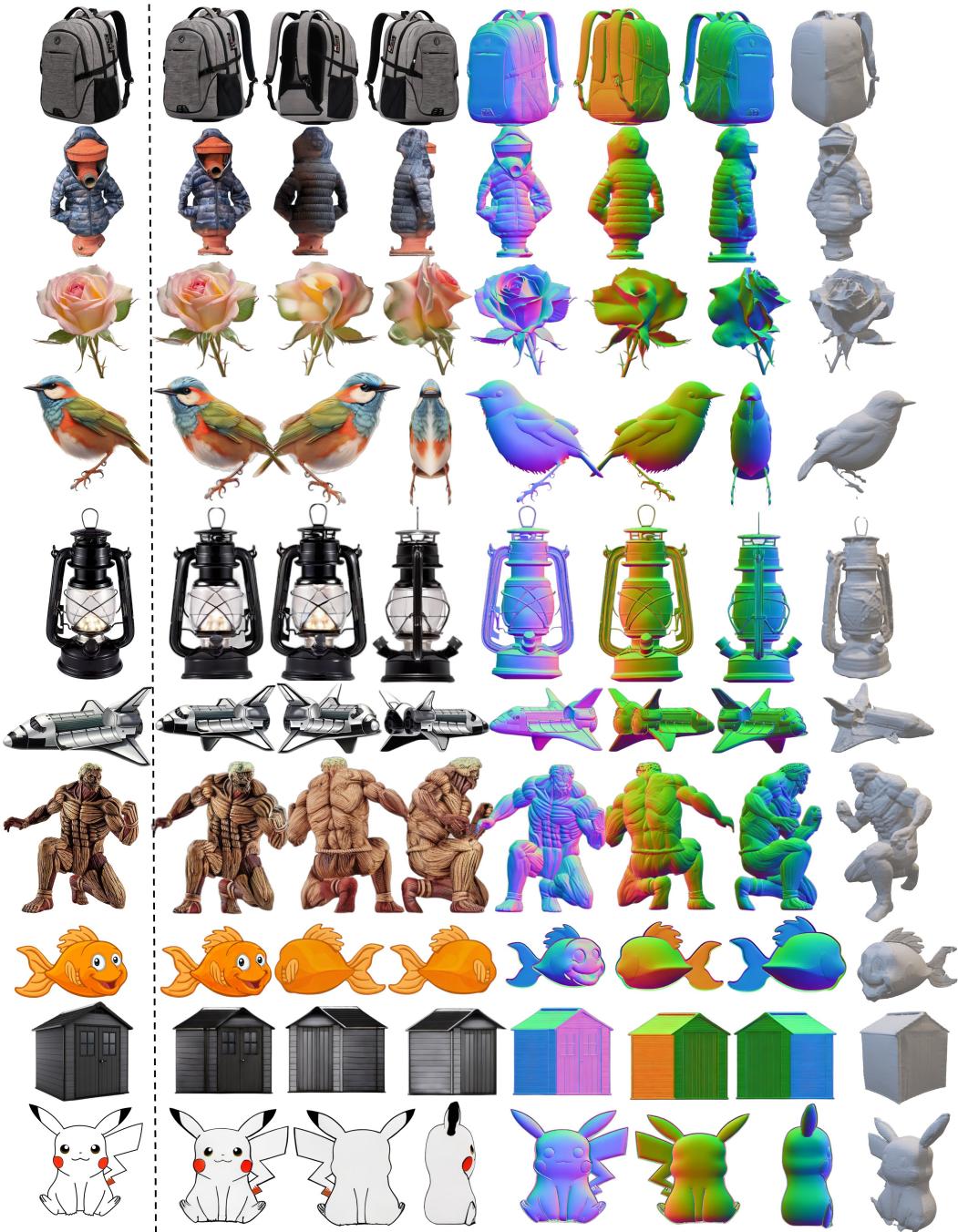


Figure 11: More results of images from the Internet.

them to some extent. Simultaneously, our method exhibits smaller errors for settings with large focal lengths than the baseline. We believe our method could be further explored and improved in the future.

Furthermore, we use orthogonal renderings at an elevation of 0 degree as the reference. We vary the elevation from -10 to 40 degrees and select focal lengths from $\{35, 50, 85, 105, 135, \infty\}$ to assess the system's robustness to elevations and focal distortions by reconstructed meshes. As indicated in Tab. 7, the orthogonal setting at elevation 0 achieves the best performance of CD, and other settings are on the same bar with the reference setting, suggesting that our method effectively handles these distortions, producing meshes that align well with the reference ones, even under significant focal distortion. We visualize some samples in Fig. 9 and Fig. 10.