Recon3D: High Quality 3D Reconstruction from a Single Image Using Generated Back-View Explicit Priors

Ruiyang Chen Mohan Yin Jiawei Shen
{chenruiyang, yinmohan, shenjiawei}@emails.bjut.edu.cn
Wei Ma†
mawei@bjut.edu.cn
Beijing University of Technology

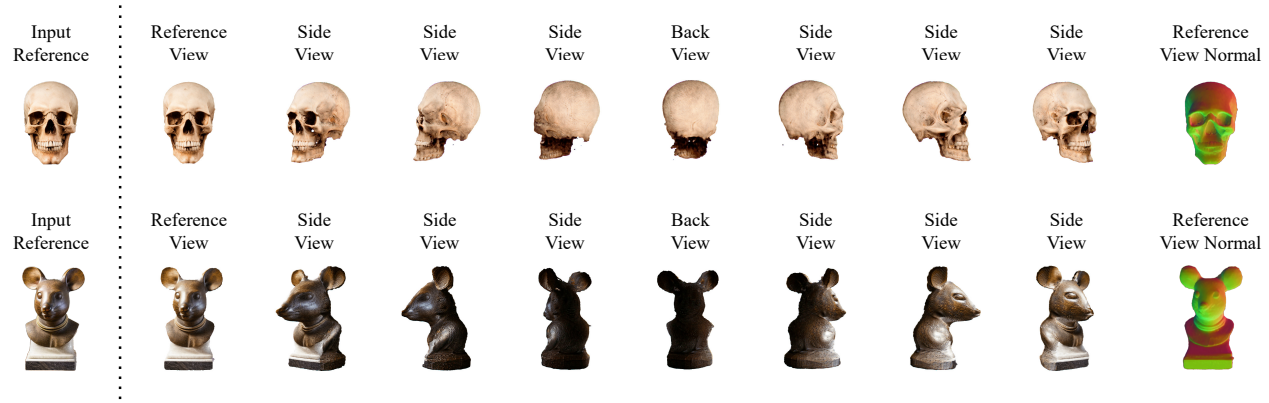

Figure 1. *Recon3D* has the ability to make the reference image into 3D content. We present a series of generated novel views and the normal map, showing high quality on geometry and textures of 3D models.

Abstract

Significant progress has been achieved in deep 3D reconstruction from a single frontal view with the aid of generative models; however, the unreliable nature of generated multi-views continues to present challenges in this domain. In this study, we propose Recon3D, a novel framework for 3D reconstruction. Recon3D exclusively utilizes a generated back view, which can be obtained more reliably through generative models based on the frontal reference image, as explicit priors. By incorporating these priors and guidance from a generative model, which is fine-tuned with Dreambooth and then enhanced with ControlNet, we effectively supervise NeRF rendering in the latent space. Subsequently, we convert the NeRF representation into an explicit point cloud and further optimize the explicit representation by referencing high-quality textured reference views. Extensive experiments demonstrate that our method achieves state-of-the-art performance in rendering novel views with superior geometry and texture quality.

1. Introduction

3D reconstruction from a single image has attracted extensive attention from researchers, due to its convenience to use. However, the task remains highly challenging because of the limited observations. In recent years, many researchers have employed generative models [21][38][17][27] or their generated views [18][34][19] as assistance for supervising NeRF rendering [22]. Nevertheless, these methodologies are still constrained by the unreliable generated multi-view clues.

Following our investigation, we have observed that generative models tend to produce more accurate back-view images for most objects by referencing their frontal views, possibly due to the similarity in silhouettes between these views. Based on this finding, we propose *Recon3D*, a framework that surpasses existing methods by achieving higher quality 3D reconstruction. Specifically, *Recon3D* leverages generated back-view explicit priors and incorpo-

† Corresponding author

rates guidance from fine-tuned generative models to train NeRF for single-reference image-based 3D modeling. Inspired by [38], our approach adopts a two-stage pipeline with an additional preprocessing stage.

In the preprocess stage, we use Zero-1-to-3 [18], a multi-view image generative model, to synthesize a back view by referencing the frontal view. This synthesized back view serves as a reliable prior to enhance the accuracy of the implicit expression of NeRF. Additionally, we use Depth Anything [40] and Omnidata [4] to produce depth and normal maps for both reference and back views respectively. To facilitate 3D reconstruction, we also introduce BLIP[14] to caption the reference image.

In the coarse stage, we fine-tune Stable Diffusion (SD) [30] using Dreambooth [31] and integrate it with the ControlNet [46]. We design a multi-view loss to supervise the training process, leveraging the explicit back view. Training NeRF with limited geometry information from the available views often leads to vague 3D representations. To eliminate this issue, we propose a new strategy for NeRF training that enhances both geometry and texture details in resulting 3D models.

In the refine stage, to eliminate this issue that the point cloud directly transformed from the NeRF representation is noisy due to the multi-source guidance in the coarse stage, we first transform the NeRF representation into a mesh and then sample mesh surfaces based on a Poisson distribution to obtain the point cloud. Subsequently, we train a UNet [44] with SD guidance and a mask loss to refine the texture and silhouette of the object. Finally, we obtain a high-quality 3D model of the object.

Experiments on 3D reconstruction of various objects show that our method surpasses State-of-the-art (SOTA) methods in generating high-quality 3D models from a single image. Our main contributions can be summarized as follows:

- We propose $Recon3D$, a new framework for 3D reconstruction from a single image that effectively exploits generative priors. In contrast to existing approaches, $Recon3D$ utilizes generative priors in a more robust manner, thereby enabling the generation of higher-quality 3D models.
- We propose an approach to enhance the quality of NeRF rendering by incorporating guidance from the generated back views.
- We propose a loss function and a training strategy to optimize the learning of both reference view and back view information in NeRF, thereby maximizing its potential for knowledge acquisition.
- We employ a distinct approach in utilizing SD, specifically by integrating ControlNet [46] with fine-tuned SD to generate images based on reference views. This enhances the applicability of SD priors in 3D reconstruction.

2. Related Work

3D reconstruction based on multiple images. In earlier studied works [2][5][32][9][13], image-based 3D reconstruction often requires the input of multi-view images. With the appearance of Neural Radiance Fields (NeRF) [22], 3D reconstruction has higher quality on geometry and textures. Subsequent related research works [11][3][25][29] minimizes the number of input images while ensuring the high-quality synthesis of novel views. For example, PixelNeRF [43] attempted to generate high-quality images with few input images, using input images as conditions for continuous neural representation to be inferred. In 2023, 3D Gaussian Splatting (3DGS) [10] and its related works [1][45] overcame the detrimental effect of noise on rendering quality and improved rendering speed.

3D reconstruction based on a single image. Single image often does not provide enough information to support computers achieve high-quality generating of novel viewpoints. At present, more and more research works are focusing on this problem, some of which [39][35][33][24] rely on depth maps to improve the quality of the results generated, while others [36][7][41][48] use multi-planar image representations. Among them [39] realizes novel views generating by transforming the depth information of the image and the image content. However, the method relies too much on the accuracy of the depth map and therefore has high limitations. Meanwhile, some research works [37][6] applied 3DGS technique to achieve 3D reconstruction from a single reference image. However, the current novel views generated by the above solutions are deficient in geometric consistency and textures. This is because neural networks cannot obtain prior knowledge of other perspectives of real or near real objects in advance.

3D reconstruction from a single image using generative prior guidance. Recently, using generative prior to supervise NeRF rendering for 3D reconstruction from a single reference image has become a new mainstream trend. Subsequent research works [17][27][15][26] attempt to use textual description and generative model to guide NeRF rendering, thereby enabling the generation of high-quality 3D models. There are also several research works [19][18][34] use the generative capabilities of the generative models and add input reference images constraints to generate a 3D model of the reference object. However, its drawback is the poor geometric consistency of the generated images because no uniform way of representing objects on a three-dimensional level is used.

The appearance of Make-It-3D [38] and Customize-It-3D [8] provides new solutions to the problem of 3D reconstruction using a single image. The above works uses generative prior as a 3D perceptual supervisor, combining score distillation sampling (SDS), CLIP loss, and texture point cloud enhancement to generate high-quality 3D mod-

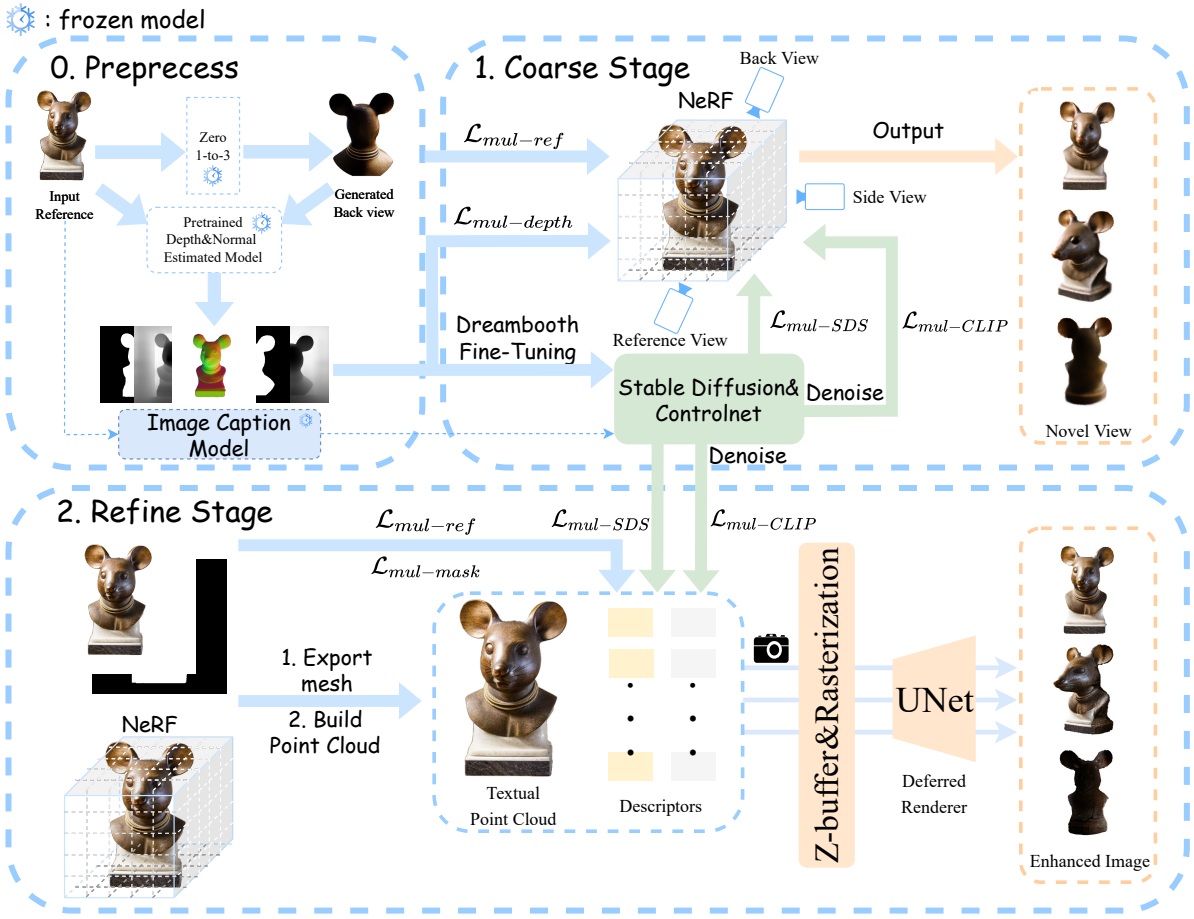

Figure 2. We propose a two-stage framework *Recon3D* for high-quality 3D generation from a single reference image with generated back-view explicit priors. At the preprocess stage, we leverage generative model to generate the back view from the references image and use depth and normal maps prediction models to generate multi-modal images (Section 3.1). At the coarse stage, we optimize a NeRF for reconstructing the geometry and texture of the reference image with generative model and multi-view images guided (Section 3.2). At the refine stage, we build textured explicit point clouds from NeRF, and jointly optimize the geometry and texture of invisible points and a learnable deferred renderer to generate high fidelity and view-consistent geometric and textures (Section 3.3).

els. Customize-It-3D also uses Dreambooth [31] for fine-tuning of SD[30]. However, both works still have some shortcomings in generating novel views of the sides and back of objects, and the quality of geometry and textures of generating images are not well. In the solution proposed by our team, neither the generative model alone is used to generate novel perspectives, nor the generative model alone is used as a guide for NeRF implicit representation. Instead, we combine the two, take advantage of the generative model's ability to generate explicit priorities, and use the prior together with the generative model to guide NeRF rendering, so that NeRF can obtain as many priors as possible, thus generating better quality 3D models on geometry and textures.

3. Method

We propose a two-stage coarse-to-fine framework *Recon3D*, aimed at improving the geometry quality of 3D models while minimizing texture deviation from any perspective. The pipeline of the framework is shown in Figure 2. In the coarse stage, we use the explicit prior of the back view to guide the generation of the 3D model, and propose a novel phased progressive method for training. In the refine stage, we convert the rough NeRF representation obtained previously into a point cloud and perform texture enhancement to obtain a 3D model with reliable semantics and higher quality.

3.1. Preprocess

It is worth noting that for most objects, their front view and corresponding back view often have similar geometric

shapes, so the generated back view is generally more reliable than other views, and can contain most of the geometry and texture information of the object. The recent Zero-1-to-3 [18] realized the generation of a single image to a consistent multi-views by adopting fixed absolute elevation and relative azimuth as the positions of the new perspective. Therefore, we adopt its strategy to generate the back view image from the reference image, enhancing the quality and quantity of explicit priors that NeRF can acquire to better help it learn more features of the image.

To better fine-tune SD [30] with DreamBooth [31] in the coarse stage, we need to obtain the multi-modal images first. Specifically, we use Depth Anything [40], a monocular depth prediction model and Omnidata [4], a single-view normal prediction model to obtain the depth and normal maps of the reference image and its back view image generated by Zero-1-to-3 [18]. It should be emphasized that Omnidata pre-trained model is trained with single RGB images, which may accumulate errors when used in combination with the depth map, so it performs normal mapping estimation independently and uses RGB images as input.

Before using diffusion priors as perceptual supervision for 3D reconstruction, it is also necessary to generate a text prompt that is faithful to the 3D representation of the image to achieve fidelity to the geometry and texture of the 3D model. Therefore, we use BLIP [14] image captioning model to generate a detailed text description for the reference image, which can facilitate its guidance on the training of the coarse stage.

3.2. Coarse Stage: Single View 3D Reconstruction

Stable Diffusion fine-tuning. Previous researches show that the diffusion model fine-tuned with DreamBooth [31] can generate images that are more in line with the theme of the reference image, which can better guide NeRF [22] in learning more features. Therefore, we choose to use DreamBooth to fine-tune SD to improve the supervising ability of generative model.

In addition, to further improve the stability and controllability of the generation process, we add Image-to-Image ControlNet [46] to the DreamBooth fine-tuned diffusion model for control. As an auxiliary neural network model, ControlNet controls the global view generation process, which makes its output results more consistent with the reference image, optimizing the guidance process.

Diffusion prior. To obtain a semantically reliable 3D model, we need to impose additional constraints on the rendering of new views. [38][8] both use the text-to-image diffusion model as the 3D perceptual prior. Through the detailed text description y generated in the preprocess stage, we perform score distillation sampling (SDS) on the fine-tuned SD and use the reference images close to their respective perspectives for control at different rendering angles, so as to further introduce the rendered images into a high-density area. Unlike the original SDS, we add the Image-to-Image ControlNet $\boldsymbol{I_c}$ to control the generation of the global view:

$$\nabla_\theta \mathcal{L}_{mul-SDS}(\phi, \mathcal{G}_\theta) = \mathbb{E}_{t,\boldsymbol{\epsilon}}\left[w(t)(\boldsymbol{\epsilon}_\phi(\boldsymbol{z_t}; \boldsymbol{y}, \boldsymbol{I_c}, t) - \boldsymbol{\epsilon})\frac{\partial \boldsymbol{z}}{\partial \boldsymbol{x}}\frac{\partial \boldsymbol{x}}{\partial \theta}\right], \tag{1}$$

where ϵ is the random amount of noise introduced by the diffusion model ϵ_ϕ with different time steps t in the rendered image. z_t is the potential noise space obtained from the new view rendering. And $w(t)$ is the weight function used to measure the noise level.

Although the 3D model generated by $\mathcal{L}_{mul-SDS}$ matches the text description well, it lacks fidelity to the reference image, because text alone cannot present all the geometry and texture information in the reference image, which makes it difficult for NeRF to learn the fine information. To address this issue, we use $\mathcal{L}_{mul-CLIP}$ to match the rendered images from each perspective with the reference images close to their respective perspectives during the training process, where the reference images include the actual input reference images and the generated back view reference images. As mentioned earlier, we believe that the generated back view reference image can provide real geometry and texture information, while the information in the generated image which is inconsistent with the reference image will not affect the implicit NeRF expression too much, so we take the multi-view Clip-score [28] for training guidance, *i.e.*:

$$\begin{aligned}\mathcal{L}_{mul-CLIP}(\mathcal{X}, \mathcal{G}_\theta(\beta)) = -\lambda_r \mathcal{E}_{CLIP}(\mathcal{X}_{ref}) \cdot \mathcal{E}_{CLIP}(\hat{\mathcal{X}}_0(\beta, t)) \\ -\lambda_b \mathcal{E}_{CLIP}(\mathcal{X}_{ger-back}) \cdot \mathcal{E}_{CLIP}(\hat{\mathcal{X}}_0(\beta, t))\end{aligned}, \tag{2}$$

where $\mathcal{E}_{CLIP}(\cdot)$ is the CLIP image encoder that encodes the rendered image $\mathcal{G}_\theta(\beta)$ into the potential noise space z_t and then denoise to obtain a clean image $\hat{\mathcal{X}}_0(\beta, t)$. λ_r and λ_b are its two weights, which take 1 when the rendering perspective is frontal and backside respectively, otherwise take 0.

Through CLIP's comparative learning on image encoders and text encoders, we can obtain rendering results that are further aligned with the reference image.

Groundtruth knowledge for the reference view and its back view. For the reference view β_{ref}, the rendered image $\mathcal{G}_\theta(\beta_{ref})$ by NeRF should theoretically be highly consistent with the input image x, so we consider the pixel-wise difference between the rendered image and the original input image under the reference view as one of the major losses.

In addition, to maximize the use of the prior information of the generated back view image to better guide the back view generation of the object, we construct the pixel-wise difference between the rendered image and the refer-

ence image under the back view. To reduce the impact of uncertain information in the generated back view image on training, we reduce the weight of this loss function. The resulting pixel-wise loss $\mathcal{L}_{mul-ref}$ is:

$$\begin{aligned}\mathcal{L}_{mul-ref} = \lambda_r \left\| x \odot m - \mathcal{G}_\theta(\beta_{ref}) \right\|_1 \\ +\lambda_b \left\| x \odot m - \mathcal{G}_\theta(\beta_{ger-back}) \right\|_1\end{aligned}, \tag{3}$$

where $\odot$ is Hadamard product. $\mathcal{G}$ is the differentiable rendering function for the 3D representation parameterized by θ. β_{ref} and $\beta_{ger-back}$ are the reference view and the generated back view. λ_r and λ_b take 1000 and 100 when the rendering perspective is frontal and backside respectively, otherwise take 0. Referring to [42], we apply the foreground mask m to get the foreground object to simplify geometric reconstruction.

However, using $\mathcal{L}_{mul-ref}$ alone cannot solve the problems of depth ambiguity and over-flat geometry in 3D reconstruction, so we regularize the negative Pearson correlation between the estimated depth d_{ref} of the reference image obtained in the preprocess stage and the actual depth d modeled by NeRF to ensure that the depth estimation is consistent with the depth prior.

Since we used Depth Anything in the preprocess stage to generate the depth map of the back view, which contains a lot of depth information from the back view, we want to use the information to construct a better image of the object's back. The resulting depth loss $\mathcal{L}_{mul-depth}$ is:

$$\begin{aligned}\mathcal{L}_{mul-depth} = -\lambda_r \frac{\mathrm{Cov}\left(d_{ref}, d\right)}{\mathrm{Var}\left(d_{ref}\right)\mathrm{Var}(d)} \\ -\lambda_b \frac{\mathrm{Cov}\left(d_{ger-back}, d\right)}{\mathrm{Var}\left(d_{ger-back}\right)\mathrm{Var}(d)}\end{aligned}, \tag{4}$$

where $\mathrm{Cov}(\cdot)$ represents the covariance and $\mathrm{Var}(\cdot)$ computes the standard deviation. d_{ref} and $d_{ger-back}$ are the depth maps of the reference view and the back view. Similarly,λ_r and λ_b take 1 when the rendering perspective is frontal and backside respectively, otherwise take 0. With $\mathcal{L}_{mul-ref}$ and $\mathcal{L}_{mul-depth}$(see results in the left of Figure 3), our novel views have higher quality on geometry and texture.

Overall training. Therefore, the losses in the coarse stage are mainly composed of $\mathcal{L}_{mul-SDS}$, $\mathcal{L}_{mul-CLIP}$, $\mathcal{L}_{mul-ref}$ and $\mathcal{L}_{mul-depth}$. To make the process of generating rough 3D models more stable while improving the geometric consistency of the overall model, we follow [38], adopting a progressive training strategy and creatively using different losses for guidance in different training stages. Each training session has a total of 100 epochs. In the 1st to 20th epochs, we first train a narrow perspective in the range of 90° near the reference view. In the 20th to 50th epochs, we increase the training range to 360°. At the same time, to allow NeRF to learn as much information as possible about the reference view and its back view, and to reduce the influence of the illusion information generated by the generative model on it, we only use SDS instead of SDS and CLIP which baseline used (see results in the right of Figure 3) as the 3D perception prior for this training stage. This stage allows NeRF to learn the fuzzy geometric and texture information of the object, and reduces the deviation of rendered images from new perspectives. In the last 50 epochs, since the previous training has obtained the fuzzy geometric and texture information, we re-add CLIP to further clarify the implicit expression of NeRF.

3.3. Refine Stage: Neural Texture Enhancement

After the reconstruction in the coarse stage, we initially obtain a 3D model with reasonable geometry, but due to the limitations of NeRF, the resolution and texture quality of the 3D model are not satisfactory. Therefore, the main objective of the refine stage is to enhance the rough texture quality while maintaining reasonable geometry. For visible textures in the reference view, we can project them directly, so the main goal of texture enhancement is to target textures that are not visible on the reference image. NeRF has good applicability in the coarse stage, and the ability to continuously process complex topological changes allows it to obtain 3D models with better consistency, but its performance on image projection is relatively poor. We choose to convert it into a point cloud for direct projection to meet its need for generating high-quality 3D models.

To avoid the rendering noise in RGBD images, we follow the approach of Customize-It-3D [8] to convert the NeRF in the coarse stage into a mesh. We use Poisson sampling to obtain the dense point cloud, and then use the front perspective point cloud P_{ref} constructed by reference view β_{ref} to gradually project the points on the new view to obtain a clean point cloud $P = P_{ref}, P_1, \ldots, P_n$ without 3D point color conflict. Finally, we render K times for each new perspective β_i using a 19 dimensional descriptor in multi-scale delayed rendering scheme and connect it using 2D-UNet to obtain the final image.

The overall losses in the refine stage are like Section 3.2, consisting of $\mathcal{L}_{mul-SDS}$, $\mathcal{L}_{mul-CLIP}$ and $\mathcal{L}_{mul-ref}$, with an additional $\mathcal{L}_{mul-mask}$ regularization term to prevent the texture-enhanced geometry from deviating too much from the initial geometry:

$$\begin{aligned}\mathcal{L}_{mul-mask} = \lambda_r MSE(\mathcal{M}_{ref}, \mathcal{M}_{pred}) \\ +\lambda_b MSE(\mathcal{M}_{ger-back}, \mathcal{M}_{pred})\end{aligned}, \tag{5}$$

where M_{ref}, M_{pred} and $M_{ger-back}$ are the reference image mask, the rendered image mask, and the generated back image mask. λ_r and λ_b take 1000 when the rendering perspective is frontal and backside respectively, otherwise take 0.

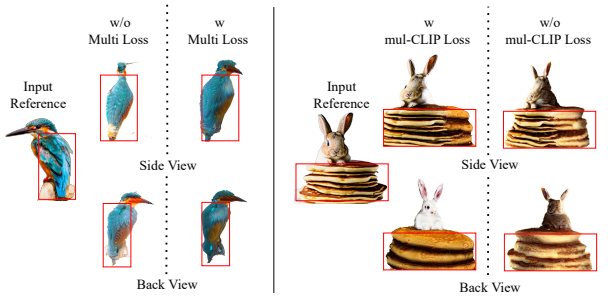

Figure 3. Ablation study results. Left: The effect of using $\mathcal{L}_{mul-ref}$ and $\mathcal{L}_{mul-depth}$. Right: The effect of not using $\mathcal{L}_{mul-CLIP}$ in the 20th-50th training epochs.

4. Experiments

4.1. Implementation Details

NeRF rendering. We use Instant-NGP [23], which uses multi-resolution hash coding and can significantly speed up the overall training and inference of neural network and save computational resources. Like Instant-NGP, we also maintain an occupancy grid to efficiently sample rays and skip invalid sampling of blank areas. In addition, we adopt some shading modes to enhance the quality of the rendered images, such as albedo, normal, and Lambertion shading mode, akin to Make-It-3D [38].

Overall training optimisation process. We randomly sampled *t* from 200 to 600 and applied classifier-free guidance to calculate SDS loss, and set the guidance scale to 4 to refer to the paper of ControlNet. We use Adam [12] as the learning rate scheduler for the training process. The overall preprocessing and training process takes about 3 hours and uses only one Tesla 32GB V100 GPU.

4.2. Comparisons with the State of the Arts

Baselines. We compare our technique with four recent representative techniques: (1) Realfusion [21]: This technique firstly exploits the generative power of generative models to guide the training of NeRF. (2) Make-It-3D [38]: This technique combines Realfusion technique and pioneered the conversion of implicit NeRF into explicit point clouds to better enhance the quality of 3D models. (3) Zero-1-to-3 [18]: This technique uses only generative models for the task of generating 3D models from a single image. (4) Customize-It-3D [8]: This technique is more effective in generating 360-degree objects, especially the back side. We used Realfusion15, Customize-It-3D self-created dataset and our self-created dataset (https://github.com/richardchen225/Recon3D.git) for the experimental comparison. The Realfusion15 dataset includes some naturalistic images, the Customize-It-3D self-created dataset includes some synthetic images, and our self-created dataset includes some pictures of real and complex objects. Our experiments use the official code of baseline for comparisons.

Qualitative comparison. Our qualitative comparison results are shown in Figure 4. We select some results of the baseline from their papers respectively. This qualitative comparison figure fully demonstrates the advantages of our technique over the previous techniques, i.e., our technique is better at generating 360-degree novel views of the object. Realfusion technique, due to the stronger guidance it receives from the generated model, results in a lack of geometrical quality in both the side and back views, and a larger difference between the color and the outline with the reference view. Make-It-3D technique generates a better frontal half-face 3D model, but due to the lack of direct prior of backside geometry information during the overall training process, there is a large gap between the color and structure of the generated results and the reference view such as the fifth results (house) in figure. Although Zero-1-to-3 technique can generate clearer images of the object from all viewpoints, there are often problems of poor geometric consistency in the side viewpoints such as the second and third results (disney castle and jay) in figure. These are the result of using only a generative model and not a stable neural network to represent the 3D object. Customize-It-3D technique is better at generating 3D models of objects than other techniques, but it produces less realistic back views and lacks some of the details of the objects such as the third and fourth results (jay and bunny cake) in figure. In contrast, our technique generates high-quality 3D models of objects with textured details, especially for the backside of objects such as the third and first results (jay and bird) of the image of results, which have the best backside detail and the best overall silhouette and colors. Our results are at an impressive and credible level, with guaranteed results in 360-degree.

Quantitative comparison. We quantitatively compare our techniques in Table 1. We use the metrics LPIPS [47], CLIP-score [28] and Contextual loss [20] to follow Make-It-3D [38]. LPIPS reflects how good the quality of the generated reference view is. The lower of the metric is better. CLIP-score are divided into CLIP-score text&image and CLIP-score image&image. These metrics can better reflect the quality of the generated novel view images. The higher of the metric is better. Contextual loss can reflect the semantic difference between the generated novel view and the real view. The lower of the metric is better. We randomly selected a class of objects from our own dataset, and a class of objects from the Realfusion15 dataset, and finally a class of objects from the Customize-It-3D dataset to calculate the above metrics, and then took the average of the corresponding values as the results. As shown in the Table 1, our technique achieves top-1 performance in all

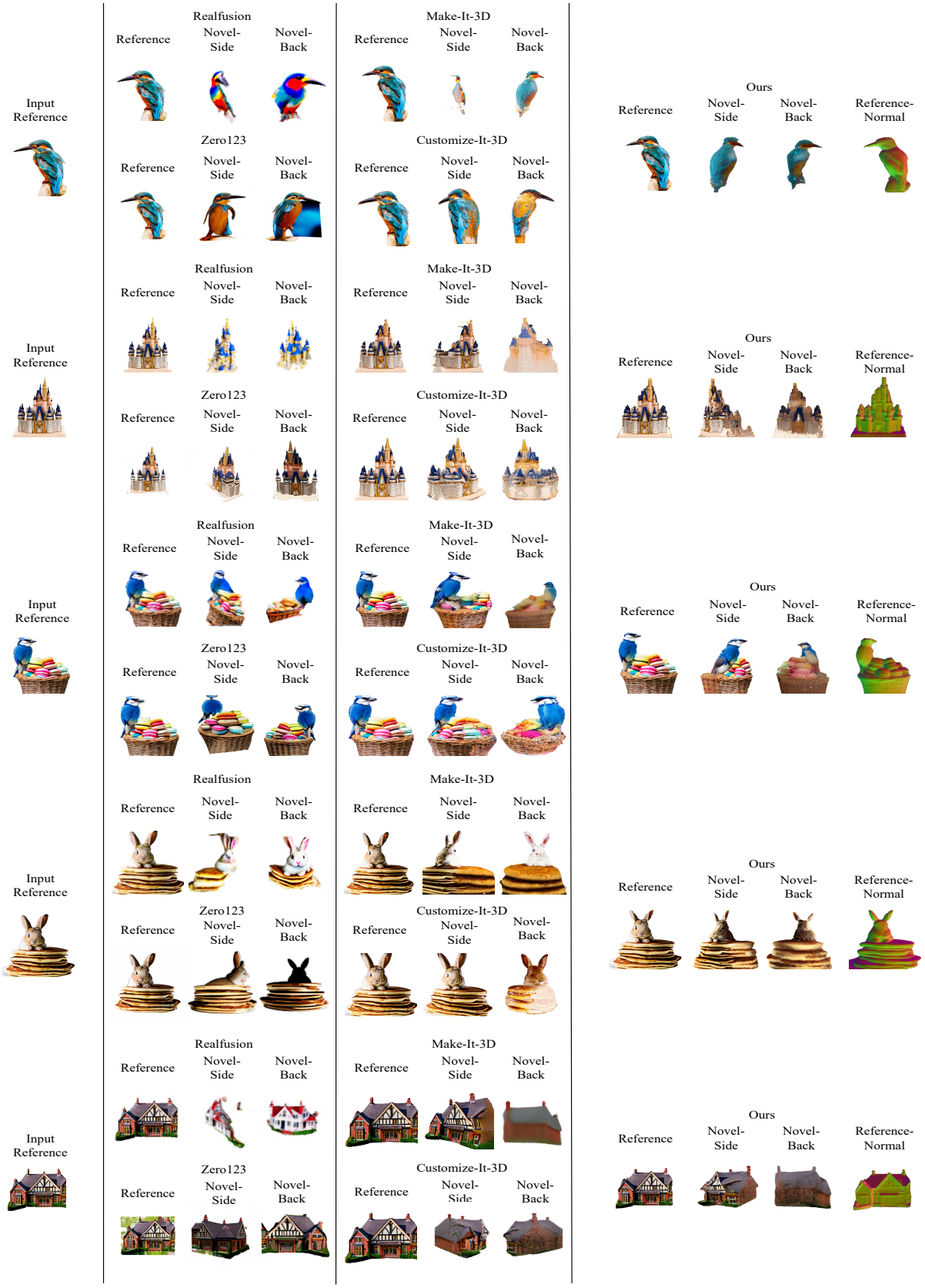

Figure 4. Qualitative comparison on image-to-3D generation. We compare our *Recon3D* to Realfusion [21], Make-It-3D [38], Zero-1-to-3 [18] and Customize-It-3D [8] for creating 3D objects from a single image (the leftmost column).

the evaluation metrics, which indicates that our technique can generate 360-degree 3D models of objects better, and the geometry and textures quality of the 3D models is very well.

	LPIPS ↓	CLIP-score Text & Image ↑	CLIP-score Image & Image ↑	Contextual ↓
Realfusion	0.3298	27.11	83.23%	3.60
Make-It-3D	0.2923	27.54	91.28%	3.26
Zero123	0.4233	27.00	81.84%	3.36
Customize-It-3D	0.3732	27.75	91.83%	3.22
Recon3D(Ours)	**0.2190**	**28.03**	**92.11%**	**3.06**

Table 1. Quantitative comparison on three datasets. We compute LPIPS under the reference view, CLIP-score and Contextual loss under novel views.

5. Ablations and Analysis

With or without generating back-view by generative model. Our first ablation study is to verify that using the Zero-1-to-3 [18] generative model in the preprocess stage to generate the back view of the reference viewpoint has an improved effect on subsequent training, as shown in the left of Figure 5. We found that when we using Zero-1-to-3 to generate back view, NeRF can learn more valid and fidelity geometry and textures information, so that NeRF [22] can better converge towards the real situation instead of completely converging towards the generated content of the generative model. As can be seen from the figure, after using Zero-1-to-3 to generate the back view as a prior, the generated novel views of the sides and back of the object by us are more in line with human cognition, with better geometry and textures quality, which solves the problem of poor generation of the sides and back of the object due to the lack of a prior of the object.

With or without the Image-to-Image ControlNet [46]. The second ablation study we did was to verify the fine-tuning effect of using Image-to-Image ControlNet for Stable Diffusion [30], as shown in the right of Figure 5. We found that the fine-tuning for SD with Image-to-Image ControlNet produces side and back view details that are more in line with human perception, as well as generating objects with more complete content, and the geometry and textures quality are better to the without fine-tuning. The overall quality and texture of the generated 3D model are better.

6. Applications

High-quality text-to-3D generation. To achieve high-quality text-to-3D generation, we use T2I diffusion model to convert the detailed text description generated by the BLIP [14] image captioning model into a reference image, and then transfer it to the image-based 3D creation method, as shown in Figure 6. It can be seen that $Recon3D$ demonstrates excellent quality on text-to-3D generation.

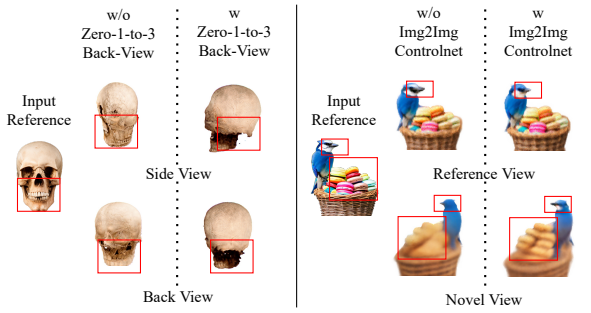

Figure 5. Ablation study results. Left: The effect of using generated back-view explicit prior. Right: The effect of Image-to-Image ControlNet.

7. Conclusion and Limitation

We propose a two-stage coarse-to-fine method $Recon3D$, which optimizes the learning process of NeRF through explicit priors of the generated back view, thus improving the geometry and texture quality of 3D models. Compared to methods that rely heavily on explicit priors like [17] [16], the number of explicit priors is greatly reduced to one. $Recon3D$ can be applied to general objects, providing reliable solutions for most application scenarios.

However, the use of explicit priors will inevitably lead to problems like over-flat geometry. Additionally, the geometry and texture of side views may not be satisfactory if useful side information cannot be extracted from the back view. Furthermore, if the quality of multi-modal images is not ideal, the training results may also deviate from expectations. In the future, as generative prior technique continues to develop, we expect $Recon3D$ to obtain 3D models with higher quality.

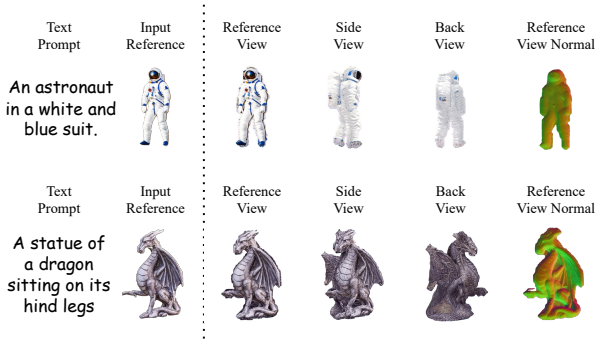

Figure 6. $Recon3D$ can generate high-quality 3D model from an caption.

References

[1] Jiazhong Cen, Jiemin Fang, Chen Yang, Lingxi Xie, Xiaopeng Zhang, Wei Shen, and Qi Tian. Segment any 3d gaussians. *ArXiv*, abs/2312.00860, 2023. 2

[2] Brian Curless and Marc Levoy. A volumetric method for building complex models from range images. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, 1996. 2

[3] Yilun Du, Cameron Smith, Ayush Kumar Tewari, and Vincent Sitzmann. Learning to render novel views from wide-baseline stereo pairs. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4970–4980, 2023. 2

[4] Ainaz Eftekhar, Alexander Sax, Roman Bachmann, Jitendra Malik, and Amir Zamir. Omnidata: A scalable pipeline for making multi-task mid-level vision datasets from 3d scans. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10766–10776, 2021. 2, 4

[5] Steven J. Gortler, Radek Grzeszczuk, Richard Szeliski, and Michael F. Cohen. The lumigraph. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, 1996. 2

[6] Antoine Gu'edon and Vincent Lepetit. Sugar: Surface-aligned gaussian splatting for efficient 3d mesh reconstruction and high-quality mesh rendering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[7] Yuxuan Han, Ruicheng Wang, and Jiaolong Yang. Single-view view synthesis in the wild with learned adaptive multi-plane images. *ACM SIGGRAPH 2022 Conference Proceedings*, 2022. 2

[8] Nan Huang, Ting Zhang, Yuhui Yuan, Dong Chen, and Shanghang Zhang. Customize-it-3d: High-quality 3d creation from a single image using subject-specific knowledge prior. *ArXiv*, abs/2312.11535, 2023. 2, 4, 5, 6, 7

[9] Po-Han Huang, Kevin Matzen, Johannes Kopf, Narendra Ahuja, and Jia-Bin Huang. Deepmvs: Learning multi-view stereopsis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2821–2830, 2018. 2

[10] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkuehler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics (TOG)*, 42:1 – 14, 2023. 2

[11] Mijeong Kim, Seonguk Seo, and Bohyung Han. Infonerf: Ray entropy minimization for few-shot neural volume rendering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12902–12911, 2021. 2

[12] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, 2014. 6

[13] John J. Leonard and Hugh F. Durrant-Whyte. Simultaneous map building and localization for an autonomous mobile robot. In *Proceedings IROS '91:IEEE/RSJ International Workshop on Intelligent Robots and Systems '91*, pages 1442–1447 vol.3, 1991. 2

[14] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning (ICML)*, 2022. 2, 4, 8

[15] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 300–309, 2022. 2

[16] 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*, abs/2311.07885, 2023. 8

[17] Minghua Liu, Chao Xu, Haian Jin, Ling Chen, T Mukund-Varma, Zexiang Xu, and Hao Su. One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization. In *The Thirty-seventh Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2023. 1, 2, 8

[18] 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 *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9264–9275, 2023. 1, 2, 4, 6, 7, 8

[19] Yuan Liu, Chu-Hsing 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 (ICLR)*, 2024. 1, 2

[20] Roey Mechrez, Itamar Talmi, and Lihi Zelnik-Manor. The contextual loss for image transformation with non-aligned data. In *European conference on computer vision (ECCV)*, pages 768–783, 2018. 6

[21] Luke Melas-Kyriazi, Iro Laina, C. Rupprecht, and Andrea Vedaldi. Realfusion 360° reconstruction of any object from a single image. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8446–8455, 2023. 1, 6, 7

[22] 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. 1, 2, 4, 8

[23] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics (TOG)*, 41:1 – 15, 2022. 6

[24] Phong Nguyen, Animesh Karnewar, Lam Huynh, Esa Rahtu, Jiri Matas, and J. Heikkila. Rgbd-net: Predicting color and depth images for novel views synthesis. In *International Conference on 3D Vision (3DV)*, pages 1095–1105, 2020. 2

[25] Michael Niemeyer, Jonathan T. Barron, Ben Mildenhall, Mehdi S. M. Sajjadi, Andreas Geiger, and Noha Radwan. Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5470–5480, 2021. 2

[26] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *ArXiv*, abs/2209.14988, 2022. 2

[27] Guocheng Qian, Jinjie Mai, Abdullah Hamdi, Jian Ren, Aliaksandr Siarohin, Bing Li, Hsin-Ying Lee, Ivan Skorokhodov, Peter Wonka, Sergey Tulyakov, and Bernard Ghanem. Magic123: One image to high-quality 3d object generation using both 2d and 3d diffusion priors. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024. 1, 2

[28] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 2021. 4, 6

[29] Barbara Roessle, Jonathan T. Barron, Ben Mildenhall, Pratul P. Srinivasan, and Matthias Nießner. Dense depth priors for neural radiance fields from sparse input views. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12882–12891, 2021. 2

[30] Robin Rombach, A. Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10674–10685, 2021. 2, 3, 4, 8

[31] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22500–22510, 2022. 2, 3, 4

[32] Johannes L. Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4104–4113, 2016. 2

[33] Jonathan Shade, Steven J. Gortler, Li wei He, and Richard Szeliski. Layered depth images. In *Proceedings of the 25th annual conference on Computer graphics and interactive techniques*, 1998. 2

[34] Ruoxi Shi, Hansheng Chen, Zhuoyang Zhang, Minghua Liu, Chao Xu, Xinyue Wei, Linghao Chen, Chong Zeng, and Hao Su. Zero123++: a single image to consistent multi-view diffusion base model. *ArXiv*, abs/2310.15110, 2023. 1, 2

[35] Meng-Li Shih, Shih-Yang Su, Johannes Kopf, and Jia-Bin Huang. 3d photography using context-aware layered depth inpainting. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8025–8035, 2020. 2

[36] Pratul P. Srinivasan, Richard Tucker, Jonathan T. Barron, Ravi Ramamoorthi, Ren Ng, and Noah Snavely. Pushing the boundaries of view extrapolation with multiplane images. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 175–184, 2019. 2

[37] Stanislaw Szymanowicz, C. Rupprecht, and Andrea Vedaldi. Splatter image: Ultra-fast single-view 3d reconstruction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[38] 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. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22762–22772, 2023. 1, 2, 4, 5, 6, 7

[39] Richard Tucker and Noah Snavely. Single-view view synthesis with multiplane images. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 548–557, 2020. 2

[40] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2, 4

[41] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo. In *European conference on computer vision (ECCV)*, 2018. 2

[42] Lior Yariv, Yoni Kasten, Dror Moran, Meirav Galun, Matan Atzmon, Ronen Basri, and Yaron Lipman. Multiview neural surface reconstruction by disentangling geometry and appearance. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 5

[43] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4576–4585, 2020. 2

[44] Jiahui Yu, Zhe L. Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S. Huang. Free-form image inpainting with gated convolution. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4470–4479, 2018. 2

[45] Zehao Yu, Anpei Chen, Binbin Huang, Torsten Sattler, and Andreas Geiger. Mip-splatting: Alias-free 3d gaussian splatting. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[46] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3813–3824, 2023. 2, 4, 8

[47] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018. 6

[48] Xiaoyu Zhou, Zhiwei Lin, Xiaojun (Gene) Shan, Yongtao Wang, Deqing Sun, and Ming Yang. Sampling: Scene-adaptive hierarchical multiplane images representation for novel view synthesis from a single image. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22773–22783, 2023. 2