

# HiFi-123: Towards High-fidelity One Image to 3D Content Generation

Wangbo Yu<sup>1,2</sup>, Li Yuan<sup>1,2</sup>, Yan-Pei Cao<sup>\*3</sup>, Xiangjun Gao<sup>5</sup>,  
Xiaoyu Li<sup>4</sup>, Wenbo Hu<sup>4</sup>, Long Quan<sup>5</sup>,  
Ying Shan<sup>4</sup>, and Yonghong Tian<sup>\*1,2</sup>

<sup>1</sup> School of Electronic and Computer Engineering, Peking University

<sup>2</sup> Peng Cheng Laboratory

<sup>3</sup> VAST

<sup>4</sup> Tencent AI Lab

<sup>5</sup> Hong Kong University of Science and Technology

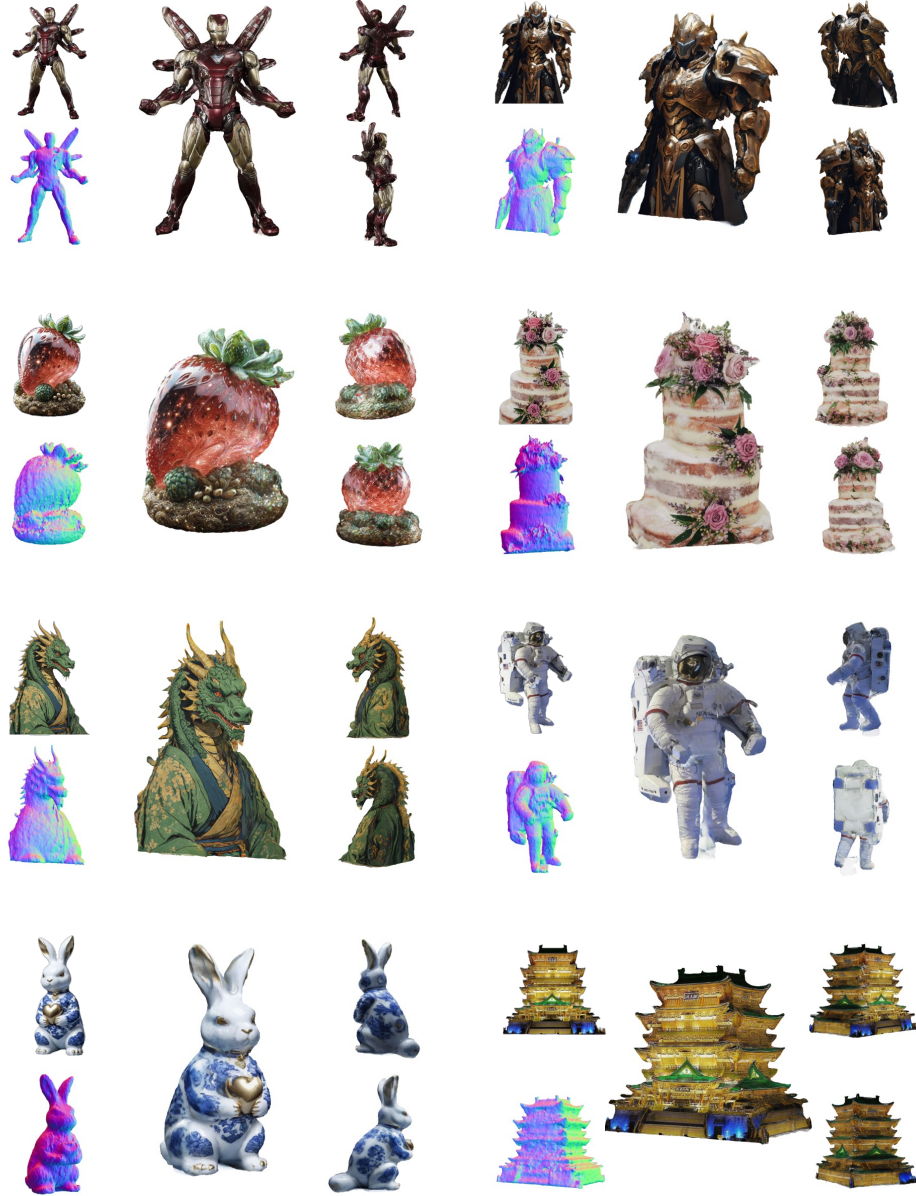
**Abstract.** Recent advances in diffusion models have enabled 3D generation from a single image. However, current methods often produce suboptimal results for novel views, with blurred textures and deviations from the reference image, limiting their practical applications. In this paper, we introduce **HiFi-123**, a method designed for high-fidelity and multi-view consistent 3D generation. Our contributions are twofold: First, we propose a Reference-Guided Novel View Enhancement (RGNV) technique that significantly improves the fidelity of diffusion-based zero-shot novel view synthesis methods. Second, capitalizing on the RGNV, we present a novel Reference-Guided State Distillation (RGSD) loss. When incorporated into the optimization-based image-to-3D pipeline, our method significantly improves 3D generation quality, achieving state-of-the-art performance. Comprehensive evaluations demonstrate the effectiveness of our approach over existing methods, both qualitatively and quantitatively. Video results are available on the project page.

## 1 Introduction

The generation of 3D digital content is a fundamental task in computer vision and computer graphics with applications in robotics, virtual reality, and augmented reality. Producing such 3D content often demands proficiency in specialized software tools, setting a high threshold in terms of skill and cost. An alternative approach is through 3D digitization, which often relies on a large set of multi-view images and their corresponding camera poses; however, acquiring such data is challenging. A more ambitious approach is to construct 3D content from only a single image, whether obtained from the web or generated. While humans can intuitively infer 3D shapes and textures from 2D images, creating 3D assets from a single image using computer vision techniques is difficult due to the limited 3D cues and ambiguities of a single viewpoint.

---

\* Corresponding Authors.



**Fig. 1:** HiFi-123 is capable of generating high-fidelity 3D content from a single reference image. In each block above, we display the reference image (top left corner) along with the rendered novel views and normal of the generated 3D content. The presented novel views demonstrate that our approach maintains consistency and high-fidelity with the reference image, even in views significantly deviating from the reference view.

Recent advances in diffusion models, trained on web-scale 2D image datasets, have led to significant improvements in text-to-image (T2I) generation [3, 26, 31, 33]. By leveraging the 3D priors inherent in T2I models, methods such as [27, 42] have utilized score distillation sampling (SDS) to achieve notable results in text-to-3D generation. This progress has also influenced the image-to-3D domain, with works such as [20, 39, 46] employing SDS loss combined with reference-view pixel losses to optimize neural representations from a single image. While these optimization-based image-to-3D methods can produce reasonable 3D structures from a single viewpoint, the visual quality in novel views often lacks fidelity, exhibiting inconsistencies with the reference image and oversmoothed textures, as shown in Fig. 6. The primary challenge arises when, for novel views outside the reference view, the optimization becomes overly reliant on, and thus closely tied to, the inferred text prompt of reference image. These text prompts, even those derived through textual inversion [4, 11], often fail to capture the full visual details of the reference image, leading to inconsistent optimization results in novel views. What’s more, the strong CFG guidance present in the SDS loss [27] further amplifies the issue. These problems not only compromise the realism of the generated content but also limit its potential for broader applications.

Apart from optimization-based image-to-3D generation, Zero-1-to-3 [17] introduced an approach that demonstrated the efficacy of fine-tuning T2I models for zero-shot novel view synthesis, highlighting their ability to produce novel views in an optimization-free manner. However, models like Zero-1-to-3 require fine-tuning on synthetic multi-view datasets [7], which can lead to a noticeable degradation in model performance, particularly in generating unnatural and low-quality novel views that deviates from the reference image, as shown in Fig. 5.

In this work, we aim to enhance the fidelity and consistency for both zero-shot novel-view synthesis and optimization-based image-to-3D generation, endowing generation of photo-realistic 3D assets. To this purpose, we devise a method that can simultaneously generates consistent novel views from a single image while maintaining high image quality. Our primary insight lies in the application of the diffusion-based image inversion technique [37] to retain the detailed structure and textures of a specific object, enabling the generation of novel views and the subsequent 3D representation of the object with consistent details. One of the key insight is we observe that by integrating depth information into the DDIM inversion [37] and the sampling process based on a depth-conditioned stable diffusion model [2], the reconstruction quality of the object is significantly improved and near perfect (illustrated in the Supplementary). Leveraging this observation, we introduce **HiFi-123**, a method that, while intuitive, effectively generates high-fidelity novel views and 3D content from a single reference image. Specifically, we design a Reference-Guided Novel View Enhancement (RGNV) pipeline in which both the reference image and a “coarse” estimation of the target novel view are inverted and reconstructed simultaneously, with the inversion process capturing fine details of the reference image, and the sampling process transferring texture details to the coarse novel view. This RGNV pipeline can be seamlessly integrated into the recent zero-shot novel view synthesis methods [17,

18]. Moreover, the inversion process’s unique properties enable us to re-formulate and re-derive the SDS loss [27], resulting in a Reference-Guided State Distillation (RGSD) loss that is easy to implement and efficient to optimize. As a result, we can also achieve high-fidelity in optimization-based image-to-3D generation that significantly exceeds prior techniques.

We comprehensively evaluated **HiFi-123** on both zero-shot novel view synthesis and image-to-3D generation tasks. Both qualitative and quantitative results indicate that our approach excels in generating high-fidelity and consistent novel views from a single reference image and further produces high-quality 3D content. Compared to state-of-the-art approaches, our method shows significant improvements in visual quality, marking an important step towards more accessible and democratized 3D content creation.

In summary, the main contributions of our work are two-fold:

- We introduce a Reference-Guided Novel View Enhancement (RGNV) pipeline grounded in a depth-based DDIM inversion. This approach can function as a plug-and-play module to improve the fidelity of results derived from diffusion-based zero-shot novel view synthesis methods.
- Leveraging the RGNV pipeline, we present a novel Reference-Guided State Distillation (RGSD) loss. When incorporated into the optimization-based image-to-3D framework, it significantly enhances the quality of 3D generation, achieving state-of-the-art performance.

## 2 Related Work

### 2.1 Optimization-based image-to-3D generation.

Based on the powerful text-to-image diffusion models [26, 30, 33, 34] in recent years, text-to-3D generation has also made great progress. DreamFields [14] uses aligned image and text models to optimize NeRF [23] without 3D shape or multi-view data. DreamFusion [27] proposes a Score Distillation Sampling (SDS) method that replaces CLIP loss from DreamField with a loss derived from the distillation of a 2D diffusion model to optimize a parametric NeRF model, which becomes a paradigm for 3D generation using 2D diffusion. To improve the text-to-3D generation results, Magic3D [15] builds upon DreamFusion that introduces several design choices like coarse-to-fine optimization, using Instant NGP representation in the coarse stage and 3D mesh representation in the fine stage. Fantasia3D [6] further disentangles the modeling of geometry and appearance, and ProlificDreamer [42] proposes to modify score distillation sampling to variational score distillation which models the 3D parameters as a random variable instead of a constant. Apart from text-to-3D generation, 3D generation based on a single image using diffusion models (image-to-3D) has also made rapid progress. NeuralLift-360 [46] learns to recover a 3D object from a single reference image with CLIP-guided diffusion prior. In addition to using the SDS loss for distillation, RealFusion [20] and NeRDv [8] also adopt textual inversion to condition the diffusion model on a prompt with a token inverted by



the reference image. Recently, Make-It-3D [39] employs textured point clouds as the representation in the fine stage to achieve high-quality results, Magic123 [28] and DreamCraft3D [38] suggests using an additional 3D diffusion prior trained on large-scale multi-view dataset for score distillation sampling. These methods often suffer from inconsistency between reference view and novel views.

## 2.2 Diffusion-based zero-shot novel view synthesis

Trained on large-scale 2D image datasets, the 2D text-to-image diffusion models could generalize to unseen scenes and different viewing angles that could be used for distilling 3D assets. However, due to the data bias of 2D images, e.g., most images are captured from front views, the 2D diffusion model may lack multi-view knowledge for 3D generation. Some efforts have been made to train the diffusion with 3D awareness. 3DiM [44] and Zero-1-to-3 [17] present viewpoint-conditioned diffusion model for novel view synthesis trained on multi-view images. Utilizing large-scale 3D data, Zero-1-to-3 achieves zero-shot generalization ability to unseen images. One-2-3-45 [16] uses the model from Zero-1-to-3 to generate multi-view images from the input view and leverage the generated results for 3D reconstruction. The recently works [4, 18, 40] try to generate multiview consistent images from a single view [18]. However, these methods usually produce lower-quality results compared with the input view, limiting their broader applications.

## 3 Methodology

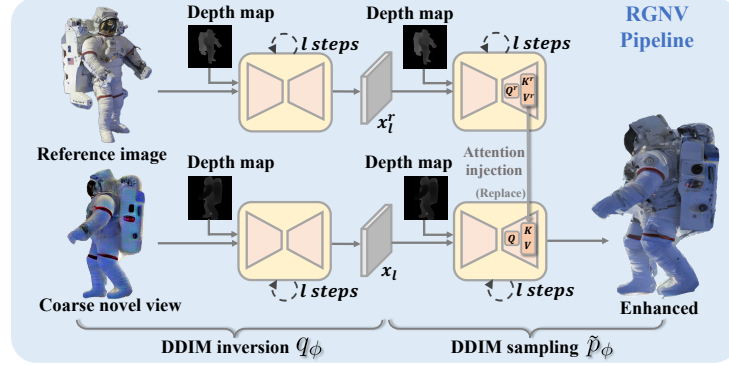
### 3.1 Preliminary

**Diffusion models.** A diffusion model consists of a forward process  $q$  and a reverse process  $p$ . In the forward process, starting from a clean data  $\mathbf{x}_0 \sim q_0(\mathbf{x}_0)$ , noise is gradually added to the data point  $\mathbf{x}_0$  to construct noisy state at different time steps, formulated as  $\mathbf{x}_t = \alpha_t \mathbf{x}_0 + \sigma_t \epsilon$ , where  $\alpha_t$  and  $\sigma_t$  are hyper-parameters satisfying  $\alpha_t^2 + \sigma_t^2 = 1$ ,  $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ . The reverse process  $p_\phi$  is defined by removing noise added on the clean data using a U-Net noise predictor  $\epsilon_\phi$ . In text-to-image diffusion models [30, 33, 34],  $\epsilon_\phi$  is trained by minimizing the score matching objective:

$$\mathcal{L}_{\text{Diff}}(\phi) = \mathbb{E}_{t \sim \mathcal{U}(0,1)} [w(t) \|\epsilon_\phi(\mathbf{x}_t; y, t) - \epsilon\|_2^2], \quad (1)$$

where  $w(t)$  is a time-dependent weighting function and  $y$  is conditional text embedding. To balance the quality and diversity of the generated images, classifier-free guidance (CFG [12]) is adopted to modify the estimated noise as a combination of conditional and unconditional output:  $\hat{\epsilon}_\phi(\mathbf{x}_t; y, t) = (1 + s)\epsilon_\phi(\mathbf{x}_t; y, t) - s\epsilon_\phi(\mathbf{x}_t; t)$ , where  $s > 0$  is the guidance scale. Increasing the guidance scale typically enhances the alignment between text and image, but at the cost of reduced diversity.

**DDIM inversion.** In the reverse process  $p_\phi$ , diffusion models often utilize deterministic DDIM sampling [37] to speed up inference. DDIM sampling converts



**Fig. 2:** Illustration of the RGNV pipeline. It performs depth-based DDIM inversion and sampling on both the reference image and coarse novel view, and utilizes attention injection to transfer detail textures from the reference image to the coarse novel view.

random noise  $\mathbf{x}_T$  into clean data  $\mathbf{x}_0$  over a sequence of discrete time steps, from  $t = T$  to  $t = 1$ , formulated as:  $\mathbf{x}_{t-1} = (\alpha_{t-1}/\alpha_t)(\mathbf{x}_t - \sigma_t\epsilon_\phi) + \sigma_{t-1}\epsilon_\phi$ . In contrast, DDIM inversion [9, 37] is a forward process that gradually converts a clean data  $\mathbf{x}_0$  back to a noisy state  $\mathbf{x}_T$  using denoising U-Net  $\epsilon_\phi$ . From  $t = 1$  to  $t = T$ , we have  $\mathbf{x}_t = (\alpha_t/\alpha_{t-1})(\mathbf{x}_{t-1} - \sigma_{t-1}\epsilon_\phi) + \sigma_t\epsilon_\phi$ . In the case of unconditional generation, the DDIM inversion process  $q_\phi$  is completely consistent with the sampling process  $p_\phi$ , so that the original data  $\mathbf{x}_0$  can be precisely reconstructed by applying DDIM sampling on the inverted  $\mathbf{x}_T$ . However, for the text-conditioned generation with classifier-free guidance, the two processes are not consistent and the reconstruction quality will significantly decrease [24].

**Score distillation sampling (SDS).** SDS [27] is an optimization method commonly used in recent text-to-3D generation [6, 15, 22, 27, 41, 41] and image-to-3D generation methods [20, 28, 39, 46]. The core idea of SDS is to distill prior knowledge from pre-trained T2I models by minimizing:

$$\mathcal{L}_{\text{SDS}}(\theta) = \mathbb{E}_t [(\sigma_t/\alpha_t)w(t)\text{KL}(q(\mathbf{x}_t|g(\theta, c); y, t) \| p_\phi(\mathbf{x}_t; y, t))], \quad (2)$$

where  $\theta$  denotes the parameters of a trainable 3D representation (e.g., NeRF [23] or DMTet [35]) and  $g(\theta, c)$  is a rendered image given a camera pose  $c$ . By minimizing the KL divergence between distributions of noisy renderings and denoised images at different time steps, the 3D representation will be optimized to match the distribution of the images synthesized by the text-to-image diffusion model. In practice, the gradients of Eq. 2 is approximated by [27]:

$$\nabla_\theta \mathcal{L}_{\text{SDS}}(\theta) \approx \mathbb{E}_{t, \epsilon} \left[ w(t) (\epsilon_\phi(\mathbf{x}_t; y, t) - \epsilon) \frac{\partial \mathbf{x}_0}{\partial \theta} \right]. \quad (3)$$

Although optimizing with SDS loss can result in overall reasonable geometry, the generated 3D model often exhibits over-saturated colors and over-smoothed tex-

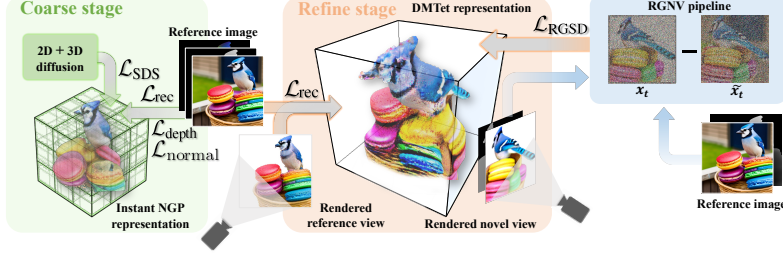
tures [27], which could lead to inconsistent results compared with the reference image when applied to image-to-3D generation tasks.

### 3.2 Reference-Guided Novel View Enhancement for zero-shot novel view synthesis

Given a reference image, previous diffusion-based zero-shot novel view synthesis methods [17, 27] prone to produce degraded and inconsistent results in novel views compared with the reference view. To tackle this problem, we propose a Reference-Guided Novel view Enhancement (RGNV) pipeline to transfer the detailed textures of the reference image to the coarse novel view. Our pipeline is built upon a discovery that incorporates depth map into the DDIM inversion and sampling process using a depth-conditioned diffusion model [2] will near perfectly reconstruct the reference image, achieving comparable performance with optimization-based inversion [24] (discussed in the Supplementary). With this discovery, we can obtain the initial noise and the reverse processes that can faithfully reconstruct the detailed textures of the reference image in a zero-shot manner. Then, inspired by the progressive generation property of the reverse process where the geometry structure emerges first at the early denoising steps while texture details appear at the late denoising steps, we design a dual-branch pipeline to transfer fine textures of the reference image to the coarse novel views.

As shown in Fig. 2, our pipeline performs DDIM inversion and sampling on both the reference image and coarse novel view simultaneously. In the forward process  $q_\phi$ , we separately map the reference image and coarse novel view back to the initial noisy state, denoted as  $\mathbf{x}_T^r$  and  $\mathbf{x}_T$ , with  $t = T$  steps’ DDIM inversion. Subsequently, in the reverse process  $\tilde{p}_\phi$  (differs from the regular reverse process  $p_\phi$ ) of the pipeline, we first perform DDIM sampling separately on the two states to denoise them for  $t = T - l$  steps, where coarse geometry structure has emerged. Then, in the following  $t = l$  denoising steps where fine textures will gradually appear in the reference image branch, inspired by recent works on consistent video generation [5, 45], we replace the  $K, V$  matrices of denoising U-Net’s self-attention in the coarse novel view branch with the corresponding matrices  $K^r, V^r$  in the reference image branch, which we term as attention injection. Through attention injection, fine textures of the reference image will be transferred to the coarse novel view. Thanks to the nearly perfect reconstruction quality of depth-based DDIM inversion, the inversion process and the sampling process are nearly consistent at every time step, we can thus simplify the pipeline to directly invert the two inputs for  $t = l$  steps, and then symmetrically adopt  $t = l$  denoising steps with attention injection to propagate textures of the reference view to the coarse novel view.

The RGNV pipeline can serve as a plug-and-play method for enhancing the quality of diffusion-based zero-shot novel view synthesis methods [17, 18], as shown in Fig. 5. We also demonstrate it can improve optimization-based image-to-3D generation in the following section.



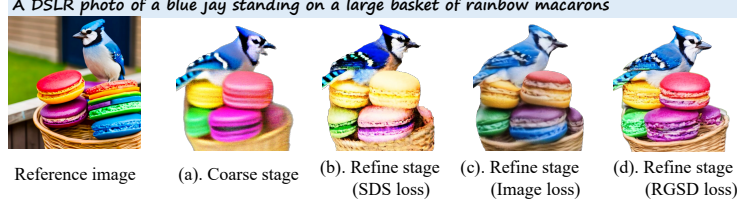
**Fig. 3:** Image-to-3D generation pipeline. We utilize two stages to generate high-fidelity 3D contents. In the coarse stage, we optimize an Instant-NGP representation using SDS loss, reference view reconstruction loss, depth loss, and normal loss. In the refine stage, we export DMTet representation and use our proposed RGSD loss to supervise training.

### 3.3 Reference-Guided State Distillation for image-to-3D generation.

As shown in Fig. 3, we adopt a coarse-to-fine optimization strategy to create 3D content from a single reference image. In the coarse stage, we use hybrid SDS loss provided by a 2D image diffusion model [1] and a 3D novel view synthesis diffusion model [17] to optimize a coarse Instant NGP [25]. The reference view reconstruction loss, depth loss, and normal loss are also involved to supervise training. As shown in Fig. 4. (a), after the coarse stage training, the resulting 3D representation already possesses reasonable geometry and colors. However, it suffers from over-smoothing and over-saturation of textures produced by the SDS loss.

In the refine stage, we convert the implicit NeRF into an explicit DMTet representation [36] with learnable parameter  $\theta$  for higher rendering resolution and efficient training. In particular, we fix the geometry of the DMTet and focus on refining texture details in this stage. Several works [28, 39] continue optimizing the texture details using SDS loss in the refine stage. Nonetheless, it can be observed in Eq. 2 that SDS loss leads to an optimization direction that forces the forward process  $q$  of rendered novel views to approach the distribution of the reverse process  $p_\phi$  of text-to-image generation. Due to the ambiguity of the inferred text descriptions and the large CFG guidance, the optimized novel views are often inconsistent with the reference image, as shown in Fig. 4. (b).

To address the inferior textures caused by SDS loss and ensure high fidelity in novel views, we integrate our proposed RGVN pipeline into the refine stage. One naive approach would be to randomly render coarse novel views, utilize the RGVN pipeline for enhancement, and subsequently apply reconstruction loss using the enhanced images. We refer to it as image loss. As shown in Fig. 4. (c), we found this approach produces oversmoothed textures, as even slight inconsistencies in the overlapping areas between enhanced images can accumulate and lead to blurry optimization results. Inspired by SDS loss (Eq. 2) that distills from the noisy states of text-to-image generation process for 3D generation, we propose a Reference-Guided State Distillation (RGSD) loss to distill from the generation process of our RGVN pipeline for high-fidelity and consistent texture synthe-



**Fig. 4:** Comparison of using different losses in the refine stage.

sis. Specifically, we construct a series of optimization targets using intermediate states from the RGNV pipeline, the resulting objective can be formulated as:

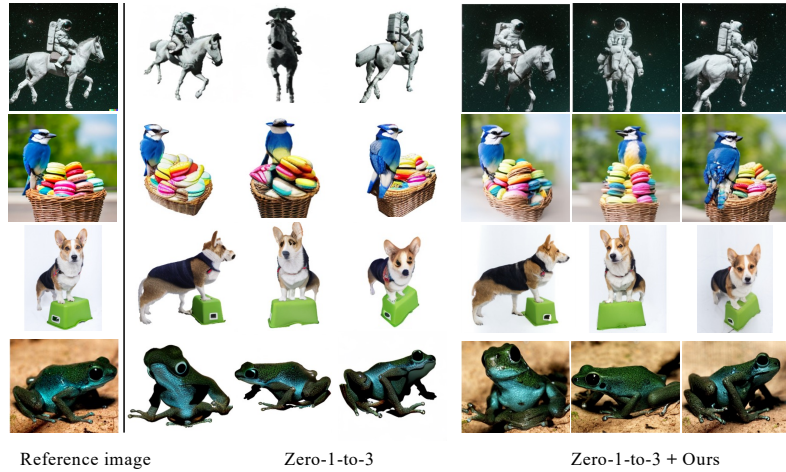
$$\min_{\theta} \mathbb{E}_{t \sim \mathcal{U}[0, l/T]} [\text{KL}(q_{\phi}(\mathbf{x}_t | g(\theta, c); y, m, t) \| \tilde{p}_{\phi}(\mathbf{x}_t; y, m, r, t))], \quad (4)$$

where  $m$  denotes the conditioned depth map,  $r$  denotes the reference image,  $q_{\phi}$  and  $\tilde{p}_{\phi}$  are the inversion and sampling process of the RGNV pipeline. Compared with the SDS loss (Eq. 2) that relies on the inferred high-level text prompts for optimization, this improved objective forces noisy states  $\mathbf{x}_t$  of the coarse novel views to approach their enhanced states  $\tilde{\mathbf{x}}_t$  produced by the reference-conditioned RGNV pipeline, ensuring that the supervision from the reference image can cover all the camera views, thereby endowing an accurate optimization direction towards the distribution of the 3D object that is consistent with the reference image.

Since  $q_{\phi}$  and  $\tilde{p}_{\phi}$  are deterministic processes given a specific reference image, we can solve Eq. 4 using a distance metric [13] such as L2 distance. In this way, Eq. 4 can be simplified as:

$$\mathcal{L}_{\text{RGSD}}(\theta) = \mathbb{E}_{t \sim \mathcal{U}[0, l/T]} [\|\mathbf{x}_t - \tilde{\mathbf{x}}_t\|_2^2]. \quad (5)$$

There are two inefficiencies in solving this objective. First, it requires multiple estimation of the U-Net  $\epsilon_{\phi}$  to get the optimization target  $\tilde{\mathbf{x}}_t$ . As depicted in Fig. 2, we need to invert the rendered novel view  $\mathbf{x}_0 = g(\theta, c)$  into a noisy state  $\mathbf{x}_l \sim q_{\phi}(\mathbf{x}_l)$  with  $l$ -step DDIM inversion, then perform with attention injection to denoise  $\mathbf{x}_l$  to an enhanced latent  $\tilde{\mathbf{x}}_t \sim \tilde{p}_{\phi}(\tilde{\mathbf{x}}_t)$  and detach it from the computation graph to make it the final optimization target. To accelerate training, we pre-select two fixed camera views and derive their  $\tilde{\mathbf{x}}_0$  states through the RGNV pipeline, using them as the optimization target at the  $t = 0$  time step. During training, we alternate between sampling  $t = 0$  to optimize the pre-defined  $\tilde{\mathbf{x}}_0$  states with fixed camera poses and sampling  $t \sim \mathcal{U}(0, l/T)$  to optimize the intermediate  $\tilde{\mathbf{x}}_t$  states with random camera poses. We found this leads to faster convergence compared with SDS loss, and results in superior results in novel views. Second, as shown in Eq. 4, unlike regular forward process  $q$  where the gradients of  $\mathbf{x}_t = \alpha_t \mathbf{x}_0 + \sigma_t \epsilon$  can be efficiently calculated, it requires multiple forward-pass of U-Net  $\epsilon_{\phi}$  to get  $\mathbf{x}_t$  in DDIM inversion  $q_{\phi}$ , in which the gradient is expensive to compute. We therefore turn to an approximate solution to compute



**Fig. 5:** Qualitative comparison with Zero-1-to-3 [17] on zero-shot novel view synthesis. “+Ours” denotes enhanced by our RGNV pipeline. Our method helps to generate novel views with higher fidelity and finer texture details.

gradients of  $\mathbf{x}_t$ . Instead of constructing  $\mathbf{x}_t$  by adding noise to  $\mathbf{x}_0$  step by step using DDIM inversion, we use the deterministic noise  $\tilde{\epsilon}_t$  predicted from  $\tilde{\mathbf{x}}_t$  in the DDIM sampling process to construct noisy states for  $\mathbf{x}_0$ , so that the resulted  $\mathbf{x}_t = \alpha_t \mathbf{x}_0 + \sigma_t \tilde{\epsilon}_t$  will have the same noisy level with  $\tilde{\mathbf{x}}_t$ . By this means, the gradients of  $\mathbf{x}_t$  can be efficiently computed.

As shown in Fig. 4. (d), optimizing with RGSD loss effectively resolves the issues of inconsistent color and oversmoothed textures, resulting in consistent appearance with the reference image. Please refer to Supplementary for a summarized algorithm of RGSD loss and more implementation details of the two training stages.

## 4 Experiments

### 4.1 Implementation details

**Zero-shot novel view synthesis.** For the RGNV pipeline, we use MiDaS [32] to estimate depth maps for both the reference image and coarse novel view, and normalize the depth map into  $[-1, 1]$  to align with the depth-conditioned SD model [2]. We adopt  $T = 50$  steps’ DDIM inversion, and set  $l = 30$  for attention injection.

**Image-to-3D generation.** In the coarse stage, we use an Instant NGP [25] representation optimized from 64 to 128 resolution. In the refine stage, we export DMTet [35] and use a rendering resolution of 1024. For the RGSD loss implementation, we use  $T = 20$  steps’ DDIM inversion with the attention injection start step set to  $l = 12$ . The coarse stage training takes about 30 minutes, and the refine stage training takes 10 minutes, both tested on a 40G A100 GPU.

**Table 1:** Comparison with Zero-1-to-3 [17] and SyncDreamer [18] on single view dataset and 3D dataset. “+Ours” denotes enhanced by our RGNV pipeline.

Methods	Single view dataset		3D dataset		
	Contextual↓	CLIP↑	PSNR↑	SSIM↑	LPIPS↓
Zero-1-to-3 [17]	1.742	0.825	18.95	0.782	0.163
Zero-1-to-3+Ours	<b>1.605</b>	<b>0.884</b>	<b>20.45</b>	<b>0.810</b>	<b>0.149</b>
SyncDreamer [18]	1.709	0.851	19.98	0.816	0.142
SyncDreamer+Ours	<b>1.598</b>	<b>0.896</b>	<b>21.08</b>	<b>0.849</b>	<b>0.123</b>

## 4.2 Zero-shot novel view synthesis comparison

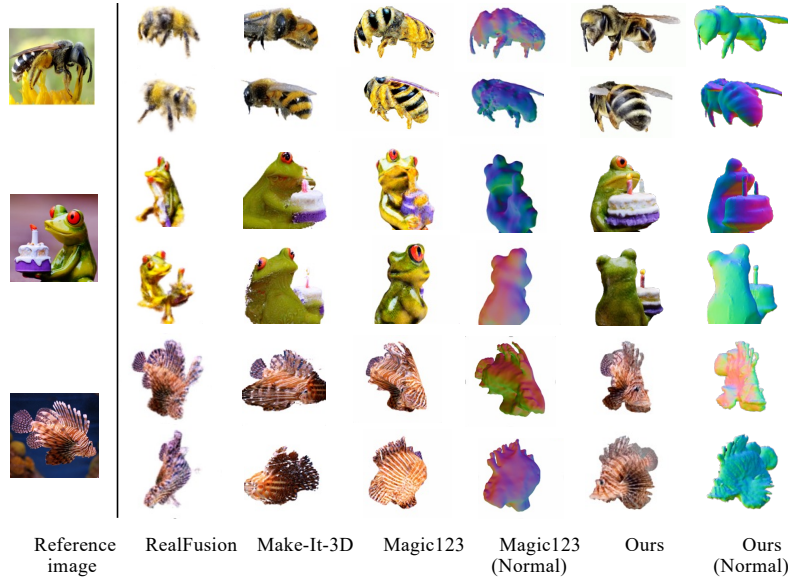
**Baselines.** We use Zero-1-to-3 [17] and SyncDreamer [18] as the baseline methods to assess our RGNV pipeline, both of which are diffusion-based zero-shot novel view synthesis methods. Specifically, Zero-1-to-3 allows for explicit control over the generation of novel views through relative camera poses. SyncDreamer is capable of simultaneously generating 16 novel views from a single image, with pre-defined camera poses.

**Comparison on single view dataset.** We compare our method with the baselines using 400 images, including challenging real-world images and realistic images generated by a T2I model [1]. Fig. 5 presents the qualitative comparison with Zero-1-to-3, please refer to the Supplementary for qualitative comparison with SyncDreamer. We found that although the novel views generated by the baselines exhibit reasonable geometry, their textures lack details and appear to be unreasonable, resulting in poor consistency with the reference image. In comparison, by applying the RGNV pipeline on the baselines, the fidelity and texture quality of the generated novel views are significantly improved. For quantitative evaluation, referring [28, 39, 46], we adopt contextual distance [19] and CLIP-similarity [29] to measure the consistency between reference image and novel views. Since the baselines cannot generate images with background, to ensure a fair comparison, we mask out the background generated by our method when computing the metrics. The results are listed in Tab. 1, which reflects the effectiveness of the RGNV pipeline.

**Comparison on 3D dataset.** For 3D evaluation, our evaluation dataset is the same with that of SyncDreamer [18], comprising of 30 objects from the Google Scanned Object dataset [10], each with 16 rendered novel views for evaluation. We adopt PSNR, SSIM [43] and LPIPS [47] to quantitatively evaluate the novel view synthesis quality, the results are shown in Tab. 1, validating that the RGNV pipeline helps to improve novel views synthesis quality. Qualitative results are displayed in the Supplementary.

## 4.3 Image-to-3D generation comparison

**Baselines.** We compare our image-to-3D generation framework against three baselines: RealFusion [20], Make-It-3D [39] and Magic123 [28]. RealFusion is a one-stage method that reconstructs NeRF representation from the reference



**Fig. 6:** Qualitative comparison with image-to-3D baselines. For each case, We show two novel views with a large angle from the reference image. It can be found that Our method outperforms baselines in maintaining texture details under significantly deviating viewpoints. Please refer to the video comparison in the Supplementary for more details.

image using L2 reconstruction loss and 2D SDS loss. Make-It-3D is a two-stage method that leverages point cloud representation in the second stage for training at higher resolution. 2D SDS loss is adopted in its two stages for geometry sculpturing and texture refining. Magic123 is also a two-stage method that uses both 2D SDS loss and 3D SDS loss provided by Zero-1-to-3 [17] to balance between geometry and texture quality.

**Comparison on single view dataset.** We firstly conduct comparisons against baseline methods on the aforementioned single view dataset. Fig. 6 displays the qualitative comparison between our method and the baselines, where we showcase two novel views for each generated object. We also present a comparison of the normal map optimized by Magic123 [28] and our method. It can be observed that, under the viewpoint that deviates significantly from the reference image, all the baseline methods fail to generate reasonable textures. The inconsistency is particularly evident at the boundaries between invisible and occluded areas, resulting in noticeable seams. In contrast, our method can maintain the same texture details as the reference image, which greatly improves the fidelity of the generated 3D assets. Please refer to the supplementary videos for a more comprehensive comparison. For quantitative evaluation, except for adopting CLIP-similarity and contextual distance for evaluating novel views, we also use LPIPS [47] to evaluate the reference view reconstruction quality. The re-



**Table 2:** Comparison with image-to-3D generation baselines on single view dataset and 3D dataset.

Methods	Single view dataset			3D dataset				
	LPIPS↓	Contextual↓	CLIP↑	PSNR↑	SSIM↑	LPIPS↓	CD↓	IoU↑
RealFusion [20]	0.195	2.180	0.767	15.37	0.715	0.288	0.082	0.274
Make-It-3D [39]	0.097	1.978	0.898	17.08	0.783	0.225	0.064	0.401
Magic123 [28]	0.085	1.882	0.883	19.33	0.801	0.156	0.052	0.453
Ours	<b>0.081</b>	<b>1.627</b>	<b>0.916</b>	<b>23.68</b>	<b>0.875</b>	<b>0.101</b>	<b>0.025</b>	<b>0.577</b>

**Fig. 7:** Ablation on design space of the RGNV pipeline.

sults are reported in Tab. 2, where the CLIP-similarity and contextual distance validates that our method can generate 3D objects with better 3D consistency.

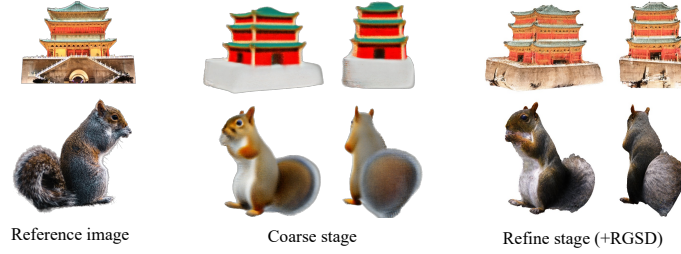
**Comparison on 3D dataset.** Following the 3D evaluation settings in Sec. 4.2, we adopt the Google Scanned Object dataset [10] and use 30 objects for evaluation, and use PSNR, SSIM [43] and LPIPS [47] to quantitatively evaluate the novel view synthesis quality. Referring [18], we also utilize the Chamfer Distance and Volume IoU to evaluate the generated geometry. Tab. 2 shows the quantitative results, which validates that our method is capable of generating 3D contents with better texture details as well as reasonable geometry. Qualitative results are in the Supplementary.

#### 4.4 Ablation study

**Design space of the RGNV pipeline.** There are two key designs in the RGNV pipeline: the depth-based DDIM inversion and the attention injection. We qualitatively validate the effectiveness of these designs. As shown in Fig. 7, given a reference image and a generated coarse novel view, a naive approach to improve the novel view quality is adopting SDEdit [21], which introduces random noise on the coarse novel view and denoise it to a clean image using pretrained diffusion model. However, we found enhancement results of SDEdit (Fig. 7. (a)) presents color and textures inconsistent with the reference image, because it didn’t make use of the reference information. In Fig. 7. (b), performing RGNV without the depth condition [2] also leads to inconsistent enhanced results. The reason lies in that the regular DDIM inversion (without depth condition) cannot precisely reconstruct the reference image (illustrated in the Supplementary), thus failing to transfer fine textures of reference image to the coarse novel view. Further, as shown in Fig. 7. (c), directly using the depth-conditioned SD model [2] without reference attention injection also leads to view inconsistent results. In contrast, as shown in Fig. 7. (d), with depth-based DDIM inversion that capture fine details

**Table 3:** Quantitative ablation on the effectiveness of RGSD loss.

Settings	Single view dataset		3D dataset		
	Contextual↓	CLIP↑	PSNR↑	SSIM↑	LPIPS↓
Coarse stage	1.901	0.836	17.82	0.794	0.215
Refine stage (SDS loss)	1.925	0.855	19.25	0.810	0.188
Refine stage (RGSD loss)	<b>1.627</b>	<b>0.916</b>	<b>23.68</b>	<b>0.875</b>	<b>0.101</b>

**Fig. 8:** Qualitative ablation on the effectiveness of RGSD loss.

of the reference image and attention injection that transfer fine textures to the coarse novel views, our RGNV pipeline can produce enhanced images consistent with the reference image. More ablations are presented in the supplementary.

**Effectiveness of the RGSD loss.** We adopt a coarse-to-fine strategy for image-to-3D generation. In the refine stage, we propose a RGSD loss to improve texture quality and consistency. Qualitative results of the coarse stage and refine stage are shown in Fig. 8. It can be found that although the coarse stage can provide a reasonable geometry, its texture details are different from the reference image. Through refine stage optimization using our proposed RGSD loss, the texture of the novel views are significantly improved. We further conduct quantitative ablation on datasets adopted in previous experiments, and evaluate the following settings: coarse stage, refine stage using SDS loss, and refine stage using RGSD loss, results are reported in Tab. 3. The results further demonstrate the effectiveness of refine stage training using RGSD loss, and validate that RGSD achieves better performance in enhancing texture quality and consistency than SDS loss.

## 5 Conclusion and Discussion

We introduce HiFi-123, a method that can be applied for generating high-fidelity novel views in a zero-shot manner as well as high-quality 3D contents. Our approach has two key contributions. Firstly, we propose an RGNV pipeline, which narrows the quality gap between synthesized and reference views in zero-shot novel view synthesis. Based on this pipeline, we further derive an RGSD loss to supervise and optimize 3D representations, resulting in highly realistic 3D assets.

## Acknowledgement

This work was supported in part by Natural Science Foundation of China (No.62332002, 62202014), and Shenzhen Basic Research Program (No.JCYJ20220813151736001).

## References

- [1] Deepfloyd. <https://www.deepfloyd.ai/deepfloyd-if>, 2023.9.29
- [2] Sd-2-depth. <https://huggingface.co/stabilityai/stable-diffusion-2-depth>, 2023.9.29
- [3] Balaji, Y., Nah, S., Huang, X., Vahdat, A., Song, J., Kreis, K., Aittala, M., Aila, T., Laine, S., Catanzaro, B., et al.: ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. arXiv preprint arXiv:2211.01324 (2022)
- [4] Burgess, J., Wang, K.C., Yeung, S.: Viewpoint textual inversion: Unleashing novel view synthesis with pretrained 2d diffusion models. arXiv preprint arXiv:2309.07986 (2023)
- [5] Cao, M., Wang, X., Qi, Z., Shan, Y., Qie, X., Zheng, Y.: Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2023)
- [6] Chen, R., Chen, Y., Jiao, N., Jia, K.: Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. arXiv preprint arXiv:2303.13873 (2023)
- [7] Deitke, M., Liu, R., Wallingford, M., Ngo, H., Michel, O., Kusupati, A., Fan, A., Laforte, C., Voleti, V., Gadre, S.Y., et al.: Objaverse-xl: A universe of 10m+ 3d objects. arXiv preprint arXiv:2307.05663 (2023)
- [8] Deng, C., Jiang, C., Qi, C.R., Yan, X., Zhou, Y., Guibas, L., Angelov, D., et al.: Nerdi: Single-view nerf synthesis with language-guided diffusion as general image priors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20637–20647 (2023)
- [9] Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. *Advances in neural information processing systems* **34**, 8780–8794 (2021)
- [10] Downs, L., Francis, A., Koenig, N., Kinman, B., Hickman, R., Reymann, K., McHugh, T.B., Vanhoucke, V.: Google scanned objects: A high-quality dataset of 3d scanned household items. In: 2022 International Conference on Robotics and Automation (ICRA) (2022)
- [11] Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-Or, D.: An image is worth one word: Personalizing text-to-image generation using textual inversion. arXiv preprint arXiv:2208.01618 (2022)
- [12] Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022)
- [13] Huang, C.W., Ahmed, F., Kumar, K., Lacoste, A., Courville, A.: Probability distillation: A caveat and alternatives. In: *Uncertainty in Artificial Intelligence*. pp. 1212–1221 (2020)
- [14] Jain, A., Mildenhall, B., Barron, J.T., Abbeel, P., Poole, B.: Zero-shot text-guided object generation with dream fields. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 867–876 (2022)

- [15] Lin, C.H., Gao, J., Tang, L., Takikawa, T., Zeng, X., Huang, X., Kreis, K., Fidler, S., Liu, M.Y., Lin, T.Y.: Magic3d: High-resolution text-to-3d content creation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 300–309 (2023)
- [16] Liu, M., Xu, C., Jin, H., Chen, L., Xu, Z., Su, H., 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)
- [17] Liu, R., Wu, R., Van Hoorick, B., Tokmakov, P., Zakharov, S., Vondrick, C.: Zero-1-to-3: Zero-shot one image to 3d object. ICCV (2023)
- [18] Liu, Y., Lin, C., Zeng, Z., Long, X., Liu, L., Komura, T., Wang, W.: Sync-dreamer: Generating multiview-consistent images from a single-view image. ICLR (2024)
- [19] Mechrez, R., Talmi, I., Zelnik-Manor, L.: The contextual loss for image transformation with non-aligned data. In: Proceedings of the European conference on computer vision (ECCV). pp. 768–783 (2018)
- [20] Melas-Kyriazi, L., Laina, I., Rupperecht, C., Vedaldi, A.: Realfusion: 360deg reconstruction of any object from a single image. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8446–8455 (2023)
- [21] Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.Y., Ermon, S.: SDEdit: Guided image synthesis and editing with stochastic differential equations. In: International Conference on Learning Representations (2022)
- [22] Metzer, G., Richardson, E., Patashnik, O., Giryes, R., Cohen-Or, D.: Latent-nerf for shape-guided generation of 3d shapes and textures. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12663–12673 (2023)
- [23] Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM **65**(1), 99–106 (2021)
- [24] Mokady, R., Hertz, A., Aberman, K., Pritch, Y., Cohen-Or, D.: Null-text inversion for editing real images using guided diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6038–6047 (2023)
- [25] Müller, T., Evans, A., Schied, C., Keller, A.: Instant neural graphics primitives with a multiresolution hash encoding. ACM Transactions on Graphics (ToG) (2022)
- [26] Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., Chen, M.: Glide: Towards photorealistic image generation and editing with text-guided diffusion models. arXiv preprint arXiv:2112.10741 (2021)
- [27] Poole, B., Jain, A., Barron, J.T., Mildenhall, B.: Dreamfusion: Text-to-3d using 2d diffusion. arXiv preprint arXiv:2209.14988 (2022)
- [28] Qian, G., Mai, J., Hamdi, A., Ren, J., Siarohin, A., Li, B., Lee, H.Y., Skorokhodov, I., Wonka, P., Tulyakov, S., et al.: Magic123: One image to high-quality 3d object generation using both 2d and 3d diffusion priors. In: ICLR (2024)

- [29] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sasttry, G., Aspell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763 (2021)
- [30] Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 1(2), 3 (2022)
- [31] Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation. In: International Conference on Machine Learning. pp. 8821–8831. PMLR (2021)
- [32] Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. ICCV (2021)
- [33] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
- [34] Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems* **35**, 36479–36494 (2022)
- [35] Shen, T., Gao, J., Yin, K., Liu, M.Y., Fidler, S.: Deep marching tetrahedra: a hybrid representation for high-resolution 3d shape synthesis. In: *Advances in Neural Information Processing Systems (NeurIPS)* (2021)
- [36] Shen, T., Gao, J., Yin, K., Liu, M.Y., Fidler, S.: Deep marching tetrahedra: a hybrid representation for high-resolution 3d shape synthesis. *Advances in Neural Information Processing Systems* (2021)
- [37] Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)
- [38] Sun, J., Zhang, B., Shao, R., Wang, L., Liu, W., Xie, Z., Liu, Y.: Dreamcraft3d: Hierarchical 3d generation with bootstrapped diffusion prior. arXiv preprint arXiv:2310.16818 (2023)
- [39] Tang, J., Wang, T., Zhang, B., Zhang, T., Yi, R., Ma, L., Chen, D.: Make-it-3d: High-fidelity 3d creation from a single image with diffusion prior. In: ICCV (2023)
- [40] Tang, S., Zhang, F., Chen, J., Wang, P., Furukawa, Y.: Mvdifffusion: Enabling holistic multi-view image generation with correspondence-aware diffusion. *NeurIPS* (2023)
- [41] Wang, H., Du, X., Li, J., Yeh, R.A., Shakhnarovich, G.: Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 12619–12629 (2023)
- [42] Wang, Z., Lu, C., Wang, Y., Bao, F., Li, C., Su, H., Zhu, J.: Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. arXiv preprint arXiv:2305.16213 (2023)

- [43] Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing* **13**(4), 600–612 (2004)
- [44] Watson, D., Chan, W., Martin-Brualla, R., Ho, J., Tagliasacchi, A., Norouzi, M.: Novel view synthesis with diffusion models. *arXiv preprint arXiv:2210.04628* (2022)
- [45] Wu, J.Z., Ge, Y., Wang, X., Lei, W., Gu, Y., Hsu, W., Shan, Y., Qie, X., Shou, M.Z.: Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. *ICCV* (2023)
- [46] Xu, D., Jiang, Y., Wang, P., Fan, Z., Wang, Y., Wang, Z.: Neurallift-360: Lifting an in-the-wild 2d photo to a 3d object with 360deg views. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 4479–4489 (2023)
- [47] Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 586–595 (2018)