

Make-It-3D: High-Fidelity 3D Creation from A Single Image with Diffusion Prior

Junshu Tang^{1†} Tengfei Wang^{2†} Bo Zhang^{3‡} Ting Zhang³

Ran Yi¹ Lizhuang Ma^{1‡} Dong Chen³

¹Shanghai Jiao Tong University ²HKUST ³Microsoft Research

<https://make-it-3d.github.io/>



Figure 1: *Make-It-3D* can create high-fidelity 3D content from only a single image. We show the normal map and novel-view renderings of created 3D content, showcasing fine geometry and faithful textures with stunning quality at novel views.

Abstract

In this work, we investigate the problem of creating high-fidelity 3D content from only a single image. This is inherently challenging: it essentially involves estimating the underlying 3D geometry while simultaneously hallucinating unseen textures. To address this challenge, we leverage prior knowledge from a well-trained 2D diffusion model to act as 3D-aware supervision for 3D creation. Our approach, **Make-It-3D**, employs a two-stage optimization pipeline: the first stage optimizes a neural radiance field by incorporating constraints from the reference image at the frontal view and diffusion prior at novel views; the second stage transforms the coarse model into textured point clouds and further elevates the realism with diffusion prior while leveraging the high-quality textures from the reference image. Extensive experiments demonstrate that our method outperforms prior works by a large margin, resulting in faithful reconstructions and impressive visual quality. Our method presents the first attempt to achieve high-quality 3D creation from a single image for general objects and enables various applications such as text-to-3D creation and texture editing.

1. Introduction

Given a single image as in Figure 1, how would the object portrayed in the image look like from a different perspective? Humans possess an innate ability to effortlessly imagine 3D geometry and hallucinate the appearance of novel views with a glance at the picture based on their prior knowledge about the world. In this work, we aim to achieve a similar goal: creating high-fidelity 3D content from a real or artificially generated single image. This will open up new avenues for artistic expression and creativity, such as bringing 3D effects to the fantasy images created by the cutting-edge 2D generative models like Stable Diffusion [39]. By offering a more accessible and automated way to create visually stunning 3D content, we hope to engage a broader audience with the world of 3D modeling with ease.

The creation of 3D objects from a single image presents a significant challenge due to the limited information that can be inferred from a single viewpoint. One categories of works aim to produce 3D photo effect [28, 42, 11, 48] in the manner of image-based rendering or single-view 3D reconstruction with neural rendering [55, 57, 38]. However, these methods often struggle with reconstructing fine geometry and fall short of rendering in large views. Another line of research [26, 52, 60, 56] projects the input image into

[†]Work is done during the internship at Microsoft Research.

[‡]Corresponding authors.

the latent space of the pretrained 3D-aware generative networks. Despite their impressive performance, existing 3D generative networks mainly model objects from a specific class and are therefore incapable of handling general 3D objects. In our case, we aim for general 3D creation from an arbitrary image, yet constructing a sufficiently large and diverse dataset for estimating the novel views or building a powerful 3D foundation model for general objects remains insurmountable.

Unlike the scarcity of 3D models, images are much more readily available, and recent advancements in diffusion models have sparked a revolution in 2D image generation [34, 40, 39, 53, 4]. Interestingly, we observed that well-trained image diffusion models can generate images under various views, which implies that they have already incorporated 3D knowledge. This has motivated us to explore the possibility of cultivating prior knowledge in a 2D diffusion model to reconstruct 3D objects. With diffusion prior, we propose *Make-It-3D*, a two-stage 3D content creation method that can generate a high-fidelity 3D object with superior quality from only one image.

In the first stage, we leverage diffusion prior to optimize a neural radiance field (NeRF) [23] by applying score distillation sampling (SDS) [32], and constrain this optimization with reference-view supervision. Different from prior text-to-3D works [32, 18, 22], we focus on image-based 3D creation so that we need to prioritize the faithfulness to the reference image. However, we observed that while 3D models generated with SDS match text prompts well, they often fail to align faithfully with reference images since textual descriptions do not capture all object details. To address this issue, we go beyond SDS by simultaneously maximizing the image-level similarity between the reference and the novel view rendering denoised by a diffusion model. Also, as images inherently capture more geometry-related information than textual descriptions, we can thus incorporate the depth of the reference image as an extra geometry prior to alleviate the shape ambiguity of NeRF optimization.

While the first stage generates a coarse model with plausible geometry, its appearance often deviates from the quality of the reference, exhibiting over-smooth textures and saturated colors [32]. This has limited its overall realism, and it is imperative to further bridge the gap between coarse model and reference image. As texture is more critical than geometry for human perception in the context of high-quality rendering, we choose to prioritize texture enhancement in the second stage, while inheriting the geometry from the first stage. We refine the model by leveraging the availability of ground-truth textures for regions that are observable in the reference image. To achieve this, we export the coarse NeRF model to textured point clouds and project reference textures onto their corresponding areas in the point clouds. We then utilize diffusion prior to enhance

the texture of the remaining points by jointly optimizing the point feature and a point cloud renderer, resulting in a clearly improved texture of the generated 3D model.

With diffusion prior as multi-view supervision, our approach can be applied to general objects without being limited to specific categories. To evaluate the method, we create a benchmark consisting of 400 images including both real images and generated images from 2D diffusion. We evaluate the proposed method on public DTU dataset [1] and our benchmark, and extensive experiments show a clear improvement over previous works. Furthermore, our method enables a range of applications beyond image-to-3D creation such as texture editing and high-quality text-to-3D creation. Our main contributions are summarized as:

- We propose *Make-It-3D*, a framework to create a high-fidelity 3D object from a single image, using a 2D diffusion model as 3D-aware prior. It does not require multi-view images for training and can be applied to any input image, whether it is real or generated.
- With a two-stage creation scheme, *Make-It-3D* represents the first work to achieve high-fidelity 3D creation for general objects. The resulting 3D models exhibit detailed geometry and realistic textures that accurately conform to the reference images.
- Beyond image-to-3D creation, our method enables multiple applications such as high-quality text-to-3D creation and texture editing.

2. Related Work

Novel view synthesis from a few images. Early attempts [14, 6, 46, 31] usually require dense observations of a scene from uniformly sampled poses. Recent emergence of implicit representations [43, 23] significantly advances the synthesis quality of novel views, whereas they tend to find a degenerate solution when given only very few input views. To enable novel view from sparse input views, a growing body of works [20, 47, 27] hence turn to extra prior knowledge as additional regularizations. PixelNeRF [60] predicts a continuous neural representation conditioned on the input images rather than only leveraging input views for supervision. DietNeRF [10] penalizes a semantic consistency loss by minimizing distance between CLIP [33] features of different views. With recent rapid progress of diffusion models, 3DiM [54] introduces a pose-conditional diffusion model that generates a novel view conditioned on a source view and a target pose. RenderDiffusion [3] presents a diffusion model for 3D generation that incorporates a tri-plane rendering mode into the denoiser.

Single-image 3D photography. Synthesizing novel views from a single image is quite challenging as it is a highly ill-posed problem, requiring precise geometry estimation and disocclusion of both geometry and texture. Increasing effort has been dedicated to this problem [63, 7, 19, 29, 31, 37, 28,

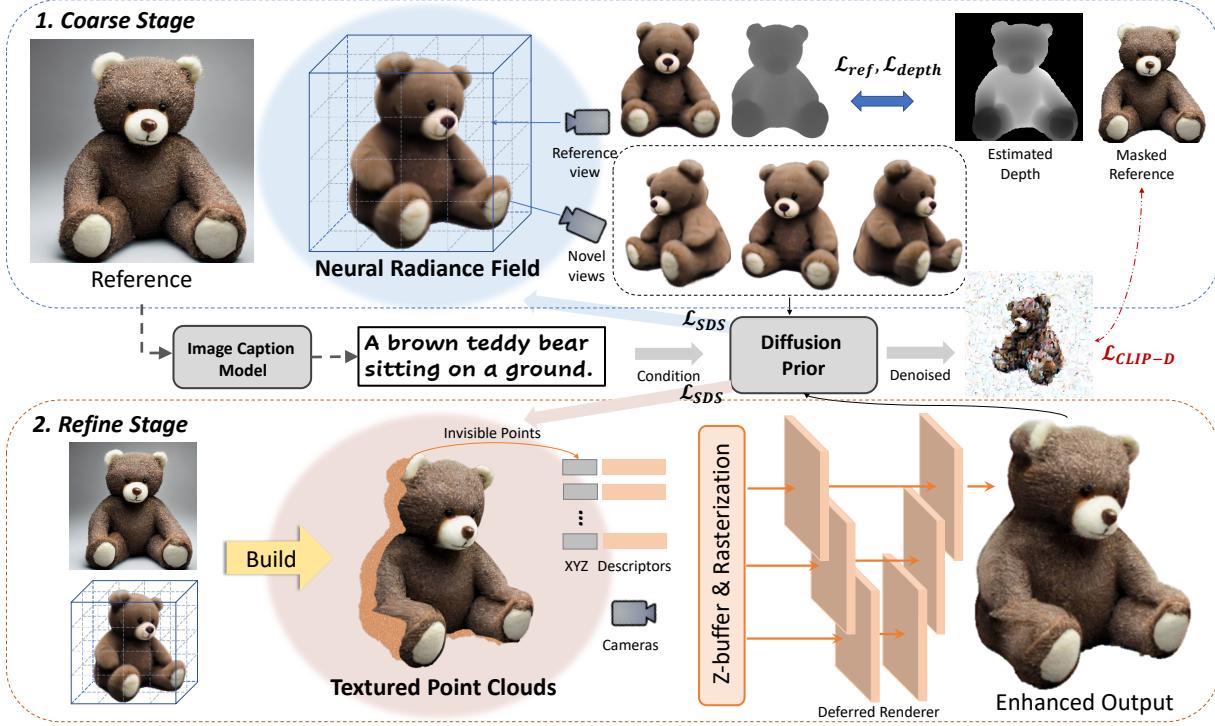


Figure 2: Overview architecture. We propose a two-stage framework for creating a high-quality 3D model from a reference image with diffusion prior (Sec. 3.1). At the coarse stage, we optimize a NeRF for reconstructing the geometry of the reference image (Sec. 3.2). We further build textured point clouds from NeRF and the reference image, and jointly optimize the texture of invisible points and a learnable deferred renderer to generate realistic and view-consistent textures (Sec. 3.3).

45], many of which can only handle specific types. Among them, a number of methods rely on layered representations such as layered depth images [41, 49, 42] and multi-plane images (MPIs) [44, 48, 17, 8]. For example, [48] predicts MPIs for view synthesis from single image without requiring ground truth 3D. [42] generates a 3D photo from a given RGB-D input through layered depth image with inpainted color and depth. Yet such a solution is limited by the number of planes and sensitive to discontinuities. Subsequent efforts generalize MPIs to continuous 3D representations such as NeRF [15] and latent 3D point cloud [55].

Lift 2D pretrained model to 3D. With the emergence of recent advances in modeling natural image manifold, how to exploit such powerful 2D pretrained model to recover 3D object structure has received considerable research interest. [30] attempts to reconstruct the 3D shape using pretrained 2D GANs. Subsequently, some works [9, 13, 24] explore zero-shot text-guided 3D content creation utilizing the guidance from CLIP [33]. Recent efforts such as DreamFusion [32], Magic3D [18] and Score Jacobian Chaining [51] explore text-to-3D generation by exploiting a score distillation sampling (SDS) loss derived from a 2D text-to-image diffusion model instead, showing impressive results. LatentNeRF [22] proposes to use a shape prior to guide and assist the 3D generation directly in the latent space

of the diffusion model. Prior works NeuralLift-360 [58] and NeRDi [5] also leverage the generative prior for 3D reconstruction from a single view. Yet the reconstructed 3D model has limited quality and is poorly aligned with the input image. In contrast, we propose a two-stage 3D synthesis framework with a relaxed SDS loss, yielding high-quality 3D representation faithful to the given input image.

3. Method

Generating novel views for general scenes or objects from only a single image is inherently challenging due to the difficulty of inferring both geometry and missing texture. We therefore tackle this challenge by cultivating the dark knowledge of pretrained 2D diffusion models. Specifically, given an input image \mathbf{x} , we first hallucinate its underlying 3D representation, neural radiance field (NeRF), whose rendering appears as a plausible sample to a pretrained denoising diffusion model, and we constrain this optimization process with the texture and depth supervision at the reference view. To further improve the rendering realism, we keep the learned geometry and enhance the textures with the reference image. As such, in the second stage, we lift the input image to textured point clouds and focus on refining the color of the points occluded in the reference view.

We leverage prior knowledge of the text-to-image generative model and the text-image contrastive model for both stages. In this way, we achieve a faithful 3D representation of the input image with restored high-fidelity texture and geometry. The proposed two-stage 3D learning framework is illustrated in Figure 2. We will subsequently brief the preliminaries and then detail our method.

3.1. Preliminaries

Recent findings show that pretrained 2D generative models offer rich 3D geometry knowledge for their 2D generation samples. Notably, DreamFusion [32] uses a text-to-image diffusion model to guide the optimization of the 3D representation. Let $\mathcal{G}_\theta(\beta)$ be the rendered image at the given viewpoint β , where \mathcal{G} is the differentiable rendering function for the 3D representation parameterized by θ and is amenable to choice. DreamFusion optimizes the neural radiance field such that its multi-view renderings look like high-quality samples from a frozen diffusion model.

Specifically, a diffusion model ϵ_ϕ introduces a random amount of noise ϵ to the rendered image $\mathbf{x}_0 := \mathcal{G}_\theta(\beta)$ at different timestep t , i.e., $\mathbf{x}_t = \alpha_t \mathbf{x}_0 + \sigma_t \epsilon$, where $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$; α_t and σ_t define a noise schedule whose log signal-to-noise ratio $\lambda_t = \log[\alpha_t^2/\sigma_t^2]$ linearly decreases with the timestep t . A pretrained text-conditioned diffusion model is trained to reverse this noising process given the text embedding \mathbf{y} . To optimize the 3D representation parameters to render images as close as good generation samples, a *score distillation sampling* (SDS) loss \mathcal{L}_{SDS} is introduced to push rendered images toward higher density region conditioned on the text embedding. Specifically, \mathcal{L}_{SDS} computes the difference of predicted noise and the added noise as per-pixel gradient which is used to update the scene parameters, i.e.,

$$\nabla_\theta \mathcal{L}_{\text{SDS}}(\phi, \mathcal{G}_\theta) = \mathbb{E}_{t, \epsilon} \left[w(t)(\epsilon_\phi(\mathbf{x}_t; \mathbf{y}, t) - \epsilon) \frac{\partial \mathbf{x}}{\partial \theta} \right], \quad (1)$$

where $w(t)$ is a weight function of different noise levels. It can be proved that this loss essentially measures the similarity between the image and the text prompt. The diffusion model acts as a critic and the gradient of \mathcal{L}_{SDS} will not be back-propagated through the diffusion network, resulting in efficient computation. As training proceeds, the NeRF parameters are updated during which the 3D object gradually reveals its texture and geometry. In practice, it is found that using a diffusion model with a strong classifier-free guidance strength leads to higher-quality 3D samples.

While DreamFusion uses the Imagen [40] to reverse the noising process at the pixel level, we use the publicly available Stable Diffusion [39] that models the latent space of the VQ-VAE [50] with an encoder \mathcal{E} and a decoder \mathcal{D} . Hence, the used diffusion model digests the latent $\mathbf{z}_0 := \mathcal{E}(\mathcal{G}_\theta(\beta))$ and the reconstructed latent $\hat{\mathbf{z}}_0$ can be mapped to image space through $\hat{\mathbf{x}} = \mathcal{D}(\hat{\mathbf{z}}_0)$.



Figure 3: Analysis on the coarse stage. Baseline is a naive solution using only \mathcal{L}_{SDS} and \mathcal{L}_{ref} , and it does not match the reference well. With $\mathcal{L}_{\text{CLIP-D}}$, the result aligns better with the reference. The depth prior further improves faithfulness.

3.2. Coarse Stage: Single-view 3D Reconstruction

As the first stage, we reconstruct a coarse NeRF from the single reference image \mathbf{x} with the diffusion prior constraining the novel views. Our optimization is expected to meet the following requirements simultaneously: 1) the optimized 3D representation should closely resemble the rendering appearance of the input observation \mathbf{x} at the reference view; 2) the novel view renderings should demonstrate consistent semantics with the input and appear as plausible as possible; 3) the generated 3D model should exhibit compelling geometry. In view of these, we randomly sample the camera poses around the reference view and enforce constraints upon the rendered images \mathcal{G}_θ for both the reference view and unseen views.

Reference view per-pixel loss. To encourage consistent appearance with the input image, we penalize the pixel-wise difference between the rendering and the input image at the reference view β_{ref} :

$$\mathcal{L}_{\text{ref}} = \|\mathbf{x} \odot \mathbf{m} - \mathcal{G}_\theta(\beta_{\text{ref}})\|_1. \quad (2)$$

Here we apply the foreground matting mask \mathbf{m} to segment out the foreground as we empirically find that this eases the geometry reconstruction, which conforms to [59].

Diffusion prior. Optimizing with the aforementioned losses can be unstable and may lead to implausible results, due to the ill-posed nature of the problem. In order to encourage semantically plausible results, additional constraints are needed on the novel view rendering. To tackle this challenge, we resort to the diffusion prior. Prior works on text-to-3D [32, 18] applied \mathcal{L}_{SDS} to leverage text-conditioned diffusion models as 3D-aware prior. To utilize \mathcal{L}_{SDS} in our case, we use an image captioning model [16], to generate a detailed text description \mathbf{y} for the reference image. With the text prompt \mathbf{y} , we can perform the SDS on the latent space of Stable Diffusion,

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

where the noisy latent \mathbf{z}_t is obtained from a novel view rendering \mathbf{x} by Stable Diffusion encoder.

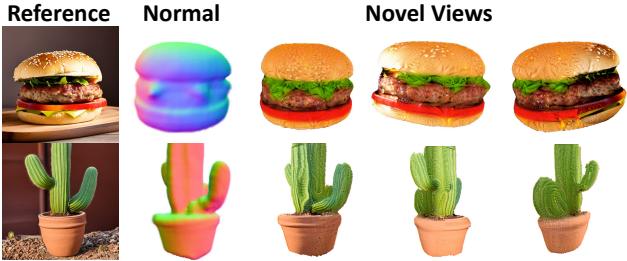


Figure 4: 360° object reconstruction from real images.

However, as discussed before, \mathcal{L}_{SDS} essentially measures the similarity between the image and the given text prompt. While \mathcal{L}_{SDS} can generate 3D models that are faithful to the text prompt, they do not align perfectly with the reference image (see baseline in Figure 3), since text prompts cannot capture all object details. We go beyond this by a diffusion CLIP loss, denoted as $\mathcal{L}_{\text{CLIP-D}}$, that additionally enforces the generated model to match the reference image:

$$\mathcal{L}_{\text{CLIP-D}}(\mathcal{X}, \mathcal{G}_\theta(\beta)) = -\mathcal{E}_{\text{CLIP}}(\mathcal{X}) \cdot \mathcal{E}_{\text{CLIP}}(\hat{\mathcal{X}}_0(\beta, t)), \quad (4)$$

where $\mathcal{E}_{\text{CLIP}}(\cdot)$ is a CLIP image encoder [33]. Rather than directly measuring CLIP loss on the rendered images $\mathcal{G}_\theta(\beta)$, we encode the novel view rendering $\mathcal{G}_\theta(\beta)$ to noisy latent z_t and then denoise it to a clean image $\hat{\mathcal{X}}_0(\beta, t)$ with 2D diffusion. By imposing the similarity loss on denoised images sampled from diffusion models, we encourage the rendering to align with the reference image, while resembling high-quality samples from a frozen diffusion.

In detail, we do not optimize $\mathcal{L}_{\text{CLIP-D}}$ and \mathcal{L}_{SDS} at the same time. We use $\mathcal{L}_{\text{CLIP-D}}$ at small timesteps and switch to \mathcal{L}_{SDS} at large timesteps. More details and analysis are in the *Supplement*. Combining \mathcal{L}_{SDS} and $\mathcal{L}_{\text{CLIP-D}}$, our diffusion prior ensures that the resulting 3D model appears visually appealing and plausible while also conforming to the given image (see Figure 3).

Depth prior. Nonetheless, even if the rendered image appears meaningful to the diffusion model, there still exists shape ambiguity that brings about issues such as sunken faces, over-flat geometry [32] or depth ambiguity (see Figure 3). We mitigate these by leveraging depth prior learned from abundant external images and directly enforcing the supervision in 3D. To be specific, we utilize an off-the-shelf single-view depth estimator [35] to estimate the depth d for the input image. While the estimated depth may not accurately characterize the geometric detail, it suffices to ensure plausible geometry and resolve most of the ambiguity. To account for the inaccuracy and the scale mismatch in d , akin to [5], we regularize the negative Pearson correlation between the estimated depth and the depth $d(\beta_{\text{ref}})$ modeled by NeRF at the reference viewpoint, *i.e.*,

$$\mathcal{L}_{\text{depth}} = -\frac{\text{Cov}(d(\beta_{\text{ref}}), d)}{\text{Var}(d(\beta_{\text{ref}}))\text{Var}(d)}, \quad (5)$$

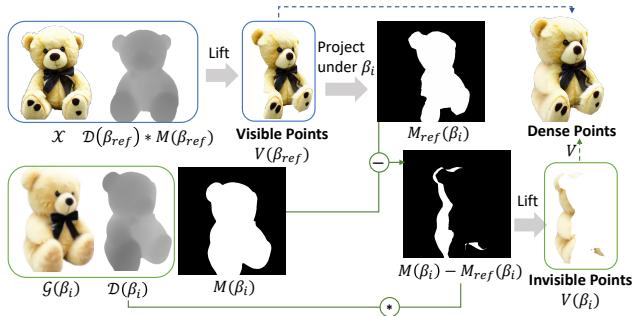


Figure 5: Illustration of textured point cloud building. We aim at building dense points and texturing visible points using reference image, and invisible points from NeRF.

where $\text{Cov}(\cdot)$ denotes the covariance, $\text{Var}(\cdot)$ computes the standard deviation. With this regularization, the NeRF depth estimation is encouraged to be linearly correlated with the depth prior.

Overall training. The overall loss can be formulated as a combination of \mathcal{L}_{ref} , \mathcal{L}_{SDS} , $\mathcal{L}_{\text{CLIP-D}}$ and $\mathcal{L}_{\text{depth}}$. To stabilize the optimization process, we adopt a progressive training strategy, where we start with a narrow range of views near the reference view and gradually expand the range during training. With progressive training, we can achieve a 360° reconstruction of an object, as shown in Figure 4.

3.3. Refine Stage: Neural Texture Enhancement

After the coarse stage, we obtained a 3D model with plausible geometry, but it often displays coarse textures that can bottleneck the overall quality in Figure 6. Further refinement is thus desired for high-fidelity 3D models. Given that humans are more discerning when it comes to texture quality than geometry, we prioritize texture enhancement while preserving the geometry of the coarse model.

Our key insight for texture enhancement is that for a novel view, certain pixels can be observable in both the novel and reference views. Consequently, we can exploit this overlap to project the high-quality texture of the reference image onto the corresponding areas of the 3D representation. We then focus on enhancing the textures of regions that are occluded in the reference view.

While NeRF is a suitable representation in the coarse stage as it can handle topological changes continuously, projecting the reference image onto it is challenging. We thus opt to export the neural radiance field to an explicit representation, specifically point clouds. Compared to the noisy mesh exported by marching cube, point clouds offer a cleaner and more straightforward projection.

Textured point cloud building. A naive attempt to build point clouds is to render multi-view RGBD images from NeRF and lift them to textured points in 3D space. However, we found this simple method leads to noisy point

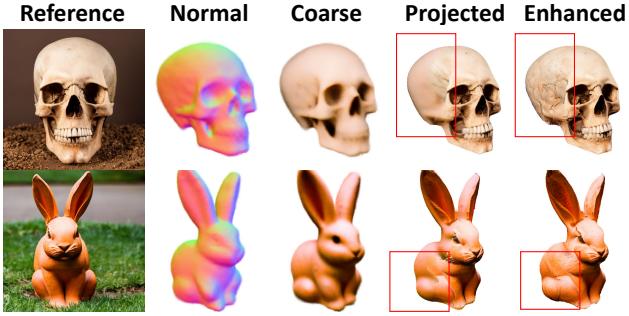


Figure 6: Visualization of neural texture enhancement. We project the reference textures to the coarse model and enhance the remaining regions to high-frequency details consistent with the reference. Best viewed with zoom-in.

clouds due to the conflict among different views: a 3D point may possess different RGB colors in NeRF rendering under different views [56]. We thus propose an iterative strategy to build clean point clouds from multi-view observations.

As in Figure 5, we first build point clouds from the reference view β_{ref} according to the rendered depth $\mathcal{D}(\beta_{\text{ref}})$ and alpha mask $M(\beta_{\text{ref}})$ of NeRF,

$$V(\beta_{\text{ref}}) = R_{\text{ref}} K^{-1} \mathcal{P}(\mathcal{D}(\beta_{\text{ref}}) * M(\beta_{\text{ref}})), \quad (6)$$

where R_{ref} and K are the extrinsic and intrinsic matrices of the camera, and \mathcal{P} denotes depth-to-point projection. These points are visible under the reference view and thus colorized with ground-truth textures. For the projection of the remaining views β_i , it is important to avoid introducing points that overlap with existing points but have conflicting colors. To this end, we project the existing points $V(\beta_{\text{ref}})$ to the novel view β_i to yield a mask indicating the presence of existing points. With this mask as guidance, we only lift those points $V(\beta_i)$ that have not been observed yet, as shown in Figure 5. These invisible points are then initialized with coarse textures from NeRF rendering $\mathcal{G}(\beta_i)$ and integrated into the dense point clouds.

Deferred point cloud rendering. So far, we have built a set of textured point clouds $V = \{V(\beta_{\text{ref}}), V(\beta_1), \dots, V(\beta_N)\}$. Though $V(\beta_{\text{ref}})$ already have high-fidelity textures projected from the reference image, the other points that are occluded in the reference view still suffer smooth textures from the coarse NeRF, as shown in Figure 6. To enhance the texture, we optimize the texture of the other points and constrain novel-view rendering with diffusion prior. Specifically, we optimize a 19-dimensional descriptor F for each point, whose first three dimensions are initialized with the initial RGB colors. To avoid noisy colors and bleeding artifacts [2], we adopt a multi-scale deferred rendering scheme. In particular, given a novel view β , we rasterize the point cloud V for K times to obtain K feature maps I_i with varying sizes of $[W/2^i, H/2^i]$, where $i \in [0, K]$. These feature maps are then concatenated and rendered into an image \mathbf{I}

	Views	LPIPS \downarrow	Contextual \downarrow	CLIP \uparrow
DietNeRF [10]	3	0.1831	5.34	64.90%
SinNeRF [57]	1	0.2059	4.28	73.24%
DreamFusion+ [32]	1	0.4075	2.15	82.81%
Point-E [26]	1	-	2.23	71.31%
3D-Photo [42]	1	0	3.43	87.65%
Ours-coarse	1	0.1427	1.74	87.50%
Ours-enhanced	1	0.0908	1.59	95.65%

Table 1: Quantitative comparison on DTU. We compute LPIPS under the reference view, and other two metrics under novel views. LPIPS of Point-E is not reported due to the lack of a defined reference view.

	LPIPS \downarrow	Contextual \downarrow	CLIP \uparrow
DreamFusion+ [32]	0.5649	3.07	84.08%
Point-E [26]	-	5.37	64.36%
Ours-coarse	0.2354	1.98	89.06%
Ours-enhanced	0.0780	1.33	95.12%

Table 2: Quantitative comparison on the test benchmark. using a U-Net renderer \mathcal{R}_θ [2] that is jointly optimized:

$$\begin{aligned} I_i(\beta) &= \mathcal{S}(i, V, F, \beta), \quad i \in [0, K], \\ \mathbf{I}(\beta) &= \mathcal{R}_\theta(I_0(\beta), I_1(\beta), \dots, I_{K-1}(\beta)), \end{aligned} \quad (7)$$

where \mathcal{S} is a differentiable point rasterizer. The objective of the texture enhancement process is similar to that of the geometry creation discussed in Sec. 3.2, but we additionally include a regularization term that penalizes large differences between the optimized texture and the initial texture.

4. Experiments

4.1. Implementation Details

NeRF rendering. We use the multi-scale hash encoding from Instant-NGP [25] to implement the NeRF representation in the coarse optimization stage, which enables neural rendering at a computational cost. Similar to Instant-NGP, we maintain an occupancy grid to enable efficient ray sampling by skipping empty space. Additionally, we adopt several shading augmentations on the rendered images, such as Lambertian and normal shading, akin to [32].

Point cloud rendering. For deferred rendering, we use a 2D U-Net architecture with gated convolutions [61]. The dimension of the point descriptor is 19, where the first 3 dimensions are initialized RGB colors and the remaining dimensions are randomly initialized. We also set a learnable descriptor for the background.

Camera setting. Following the camera sampling method used in [32], we randomly sample novel views with a 75% probability and sample the pre-defined reference view with a 25% probability. We also randomly enlarge the FOV when rendering with NeRF, following [18].

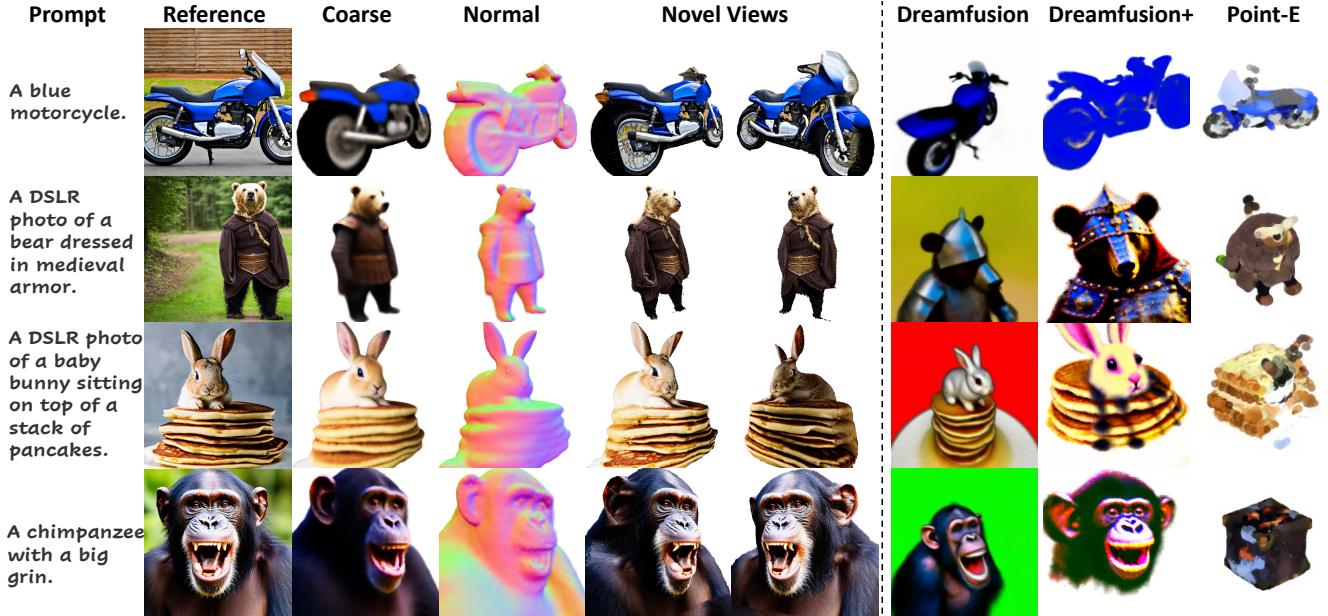


Figure 7: Qualitative comparison on the test benchmark with two diffusion-based 3D content creation models, Dreamfusion and Point-E. We show our results with high-fidelity geometry and texture. The results of Dreamfusion are from its website.

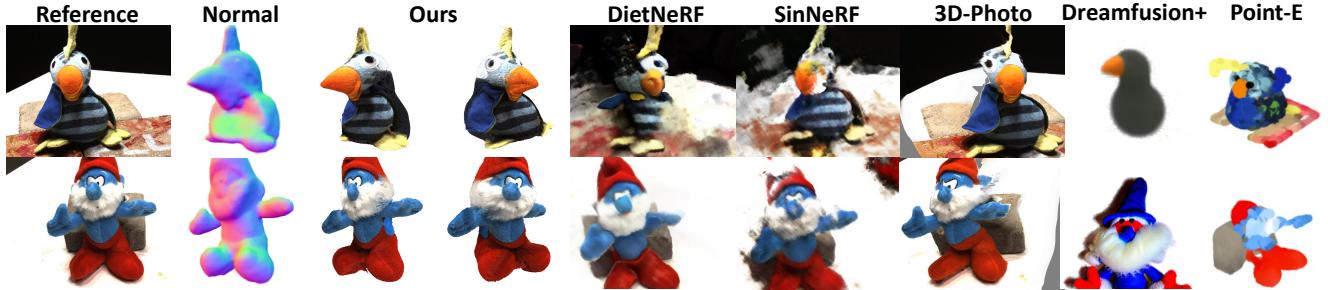


Figure 8: Qualitative comparison of novel view synthesis on DTU with state of the arts. Our method generates sharper and more plausible details in both geometry and texture.

Score distillation sampling. We randomly sample t from 200 to 600, and set $w(t)$ as a uniform weighting depending on the timestep. We also use classifier-free guidance with a guidance weight $\omega : \hat{\epsilon}_\phi(z_t; y, t) = (1 + \omega)\epsilon_\phi(z_t; y, t) - \omega\epsilon_\phi(z_t; t)$. Our method aims to align the created 3D model with the input image, and we use a guidance weight $\omega = 10$.

Training speed. We use Adam [12] with a learning rate of 0.001 for both stages. The coarse stage is trained for 5,000 iterations at a rendering resolution of 100×100 . The refine stage then takes another 5,000 iterations at a rendering resolution of 800×800 . The entire training process takes approximately 2 hours on a single Tesla 32GB V100 GPU.

Test Benchmark. To the best of our knowledge, we are the first method focusing on high-fidelity 3D creation from an arbitrary image. So we build a test benchmark consisting of 400 images, comprising both real images and images generated by Stable Diffusion [39]. Each image in the benchmark is accompanied by a foreground alpha mask, an estimated

depth map, and a text prompt. The text prompts for real images are obtained from an image caption model [16]. We will make this test benchmark publicly available.

4.2. Comparisons with the State of the Arts

Baselines. We compare our method with five representative baselines. 1) DietNeRF [10], a few-shot NeRF. We train it with three input views. 2) SinNeRF [57], a single-view NeRF method. 3) DreamFusion [32]. As it is originally conditioned on text prompts, we also modify it with image reconstruction loss at the reference view, referred as *DreamFusion+* for fair comparison. 4) Point-E [26], point cloud generation conditioned on image. 5) 3D-Photo [42], depth-based image warping and inpainting method.

Qualitative comparison. We first compare our method with 3D generation baselines, where DreamFusion and DreamFusion+ leverage 2D diffusion as 3D prior and PointE is a 3D diffusion model. As shown in Figure 7,

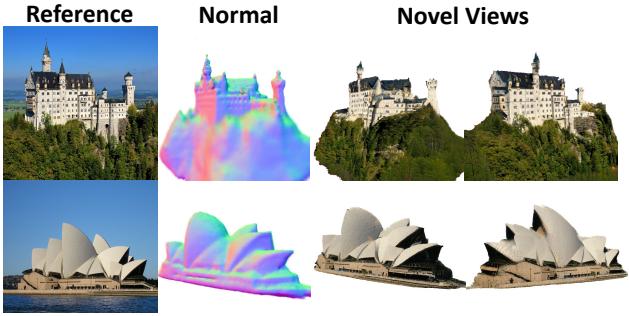


Figure 9: *Make-It-3D* enables high-fidelity 3D creation on real complex scenes.

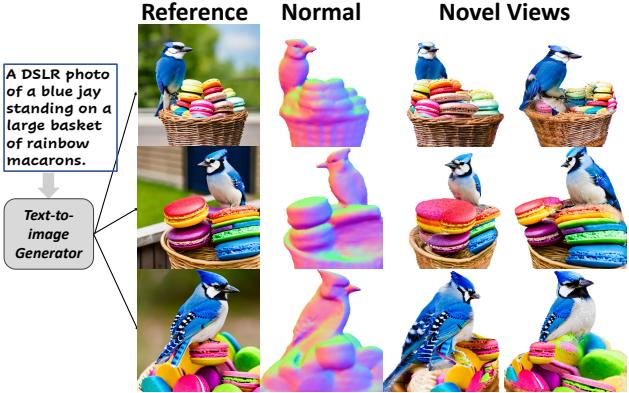


Figure 10: *Make-It-3D* generates diverse and visually stunning 3D models given a text description.

their generated models fail to align faithfully with the reference image and suffer smooth textures. In contrast, our method produces high-fidelity 3D models with fine geometry and realistic textures. Figure 8 shows additional comparison on novel view synthesis. SinNeRF and DietNeRF encounter difficulties in reconstructing complex objects due to the lack of multi-view supervision. 3D-Photo fails to reconstruct underlying geometry and produces visible artifacts in large views. In comparison, our method achieves remarkably faithful geometry and visually pleasing textures under novel views.

Quantitative comparison. A compelling generated 3D model should closely resemble the input image at the reference view, and demonstrate consistent semantics with the reference under novel views. We evaluate these two aspects using the following metrics: 1) LPIPS [62], which assesses the reconstruction quality at the reference view, 2) contextual distance [21], which measures pixel-level similarity between novel-view rendering and the reference, and 3) CLIP score [33], which evaluates the semantic similarity between the novel view and the reference. As shown in Table 1 and Table 2, our approach substantially outperforms baselines in terms of both reference-view and novel-view quality.



Figure 11: *Make-It-3D* achieves 3D-aware texture modification such as tattoo drawing and stylization.

5. Applications

Real scene modeling. As shown in Figure 9, *Make-It-3D* can successfully convert a single photo of a complex scene to a 3D model, such as buildings and landscapes. This empowers users to model a scene with ease, which could be difficult for some traditional 3D modeling techniques.

High-quality text-to-3D generation with diversity. Prior arts [32, 18] often produce models with limited diversity and excessively smooth textures. To perform high-quality text-to-3D creation, we first convert the text prompt to a reference image using 2D diffusion, and proceed with our image-based 3D creation method. As shown in Figure 10, *Make-It-3D* is capable of producing diverse examples from a text prompt that exhibit stunning quality.

3D-aware texture modification. *Make-It-3D* enables view-consistent texture editing by manipulating the reference image in the refine stage while freezing the geometry. Figure 11 shows that we can add a tattoo and apply stylization to the generated 3D model.

6. Conclusions

We introduce *Make-It-3D*, a novel two-stage method for creating high-fidelity 3D content from one single image. Leveraging diffusion prior as 3D-aware supervision, the generated 3D models exhibit faithful geometry and realistic textures with the diffusion CLIP loss and textured point cloud enhancement. *Make-It-3D* is applicable to general objects, empowering versatile fascinating applications. We believe our method takes a big step in extending the success of 2D content creation to 3D, providing users with a fresh 3D creation experience.

References

- [1] Henrik Aanæs, Rasmus Ramsbøl Jensen, George Vogiatzis, Engin Tola, and Anders Bjorholm Dahl. Large-scale data for multiple-view stereopsis. *International Journal of Computer Vision*, pages 1–16, 2016. [2](#)
- [2] Kara-Ali Aliev, Artem Sevastopolsky, Maria Kolos, Dmitry Ulyanov, and Victor Lempitsky. Neural point-based graphics. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16*, pages 696–712. Springer, 2020. [6, 12](#)
- [3] Titas Anciukevičius, Zexiang Xu, Matthew Fisher, Paul Henderson, Hakan Bilen, Niloy J Mitra, and Paul Guerrero. Renderdiffusion: Image diffusion for 3d reconstruction, inpainting and generation. *arXiv preprint arXiv:2211.09869*, 2022. [2](#)
- [4] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, Tero Karras, and Ming-Yu Liu. ediff-i: Text-to-image diffusion models with an ensemble of expert denoisers. *ArXiv*, abs/2211.01324, 2022. [2](#)
- [5] Congye Deng, Chiyu Jiang, Charles R Qi, Xinchen Yan, Yin Zhou, Leonidas Guibas, Dragomir Anguelov, et al. Nerdi: Single-view nerf synthesis with language-guided diffusion as general image priors. *arXiv preprint arXiv:2212.03267*, 2022. [3, 5](#)
- [6] 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*, pages 43–54, 1996. [2](#)
- [7] Tewodros Habtegebrial, Kiran Varanasi, Christian Bailer, and Didier Stricker. Fast view synthesis with deep stereo vision. *arXiv preprint arXiv:1804.09690*, 2018. [3](#)
- [8] Yuxuan Han, Ruicheng Wang, and Jiaolong Yang. Single-view view synthesis in the wild with learned adaptive multiplane images. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–8, 2022. [3](#)
- [9] Ajay Jain, Ben Mildenhall, Jonathan T Barron, Pieter Abbeel, and Ben Poole. Zero-shot text-guided object generation with dream fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 867–876, 2022. [3](#)
- [10] Ajay Jain, Matthew Tancik, and Pieter Abbeel. Putting nerf on a diet: Semantically consistent few-shot view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5885–5894, 2021. [2, 6, 7](#)
- [11] Varun Jampani, Huiwen Chang, Kyle Sargent, Abhishek Kar, Richard Tucker, Michael Krainin, Dominik Kaeser, William T Freeman, David Salesin, Brian Curless, et al. Slide: Single image 3d photography with soft layering and depth-aware inpainting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12518–12527, 2021. [1](#)
- [12] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. [7](#)
- [13] Han-Hung Lee and Angel X Chang. Understanding pure clip guidance for voxel grid nerf models. *arXiv preprint arXiv:2209.15172*, 2022. [3](#)
- [14] Marc Levoy and Pat Hanrahan. Light field rendering. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, pages 31–42, 1996. [2](#)
- [15] Jiaxin Li, Zijian Feng, Qi She, Henghui Ding, Changhu Wang, and Gim Hee Lee. Mine: Towards continuous depth mpi with nerf for novel view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12578–12588, 2021. [3](#)
- [16] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. [4, 7](#)
- [17] Qinbo Li and Nima Khademi Kalantari. Synthesizing light field from a single image with variable mpi and two network fusion. *ACM Trans. Graph.*, 39(6):229–1, 2020. [3](#)
- [18] 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. *arXiv preprint arXiv:2211.10440*, 2022. [2, 3, 4, 6, 8](#)
- [19] Miaomiao Liu, Xuming He, and Mathieu Salzmann. Geometry-aware deep network for single-image novel view synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4616–4624, 2018. [3](#)
- [20] Stephen Lombardi, Tomas Simon, Jason Saragih, Gabriel Schwartz, Andreas Lehrmann, and Yaser Sheikh. Neural volumes: Learning dynamic renderable volumes from images. *arXiv preprint arXiv:1906.07751*, 2019. [2](#)
- [21] Roey Mechrez, Itamar Talmi, and Lihi Zelnik-Manor. The contextual loss for image transformation with non-aligned data. In *Proceedings of the European conference on computer vision (ECCV)*, pages 768–783, 2018. [8](#)
- [22] Gal Metzer, Elad Richardson, Or Patashnik, Raja Giryes, and Daniel Cohen-Or. Latent-nerf for shape-guided generation of 3d shapes and textures. *arXiv preprint arXiv:2211.07600*, 2022. [2, 3](#)
- [23] 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. [2](#)
- [24] Nasir Mohammad Khalid, Tianhao Xie, Eugene Belilovsky, and Tiberiu Popa. Clip-mesh: Generating textured meshes from text using pretrained image-text models. In *SIGGRAPH Asia 2022 Conference Papers*, pages 1–8, 2022. [3](#)
- [25] 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. [6, 12](#)
- [26] 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. [1, 6, 7](#)

- [27] Michael Niemeyer, Jonathan T Barron, Ben Mildenhall, Mehdi SM Sajjadi, Andreas Geiger, and Noha Radwan. Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5480–5490, 2022. 2
- [28] Simon Niklaus, Long Mai, Jimei Yang, and Feng Liu. 3d ken burns effect from a single image. *ACM Transactions on Graphics (ToG)*, 38(6):1–15, 2019. 1, 3
- [29] Kyle Olszewski, Sergey Tulyakov, Oliver Woodford, Hao Li, and Linjie Luo. Transformable bottleneck networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7648–7657, 2019. 3
- [30] Xingang Pan, Bo Dai, Ziwei Liu, Chen Change Loy, and Ping Luo. Do 2d gans know 3d shape? unsupervised 3d shape reconstruction from 2d image gans. *arXiv preprint arXiv:2011.00844*, 2020. 3
- [31] Eunbyung Park, Jimei Yang, Ersin Yumer, Duygu Ceylan, and Alexander C Berg. Transformation-grounded image generation network for novel 3d view synthesis. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 3500–3509, 2017. 2, 3
- [32] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 2, 3, 4, 5, 6, 7, 8, 12, 14
- [33] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2, 3, 5, 8
- [34] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 2
- [35] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. *ArXiv preprint*, 2021. 5
- [36] Nikhila Ravi, Jeremy Reizenstein, David Novotny, Taylor Gordon, Wan-Yen Lo, Justin Johnson, and Georgia Gkioxari. Accelerating 3d deep learning with pytorch3d. *arXiv preprint arXiv:2007.08501*, 2020. 12
- [37] Konstantinos Rematas, Chuong H Nguyen, Tobias Ritschel, Mario Fritz, and Tinne Tuytelaars. Novel views of objects from a single image. *IEEE transactions on pattern analysis and machine intelligence*, 39(8):1576–1590, 2016. 3
- [38] Chris Rockwell, David F. Fouhey, and Justin Johnson. Pixelsynth: Generating a 3d-consistent experience from a single image. In *ICCV*, 2021. 1
- [39] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 1, 2, 4, 7
- [40] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022. 2, 4
- [41] Jonathan Shade, Steven Gortler, Li-wei He, and Richard Szeliski. Layered depth images. In *Proceedings of the 25th annual conference on Computer graphics and interactive techniques*, pages 231–242, 1998. 3
- [42] Meng-Li Shih, Shih-Yang Su, Johannes Kopf, and Jia-Bin Huang. 3d photography using context-aware layered depth inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8028–8038, 2020. 1, 3, 6, 7
- [43] Vincent Sitzmann, Michael Zollhöfer, and Gordon Wetzstein. Scene representation networks: Continuous 3d-structure-aware neural scene representations. *Advances in Neural Information Processing Systems*, 32, 2019. 2
- [44] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 175–184, 2019. 3
- [45] Pratul P Srinivasan, Tongzhou Wang, Ashwin Sreelal, Ravi Ramamoorthi, and Ren Ng. Learning to synthesize a 4d rgbd light field from a single image. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2243–2251, 2017. 3
- [46] Shao-Hua Sun, Minyoung Huh, Yuan-Hong Liao, Ning Zhang, and Joseph J Lim. Multi-view to novel view: Synthesizing novel views with self-learned confidence. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 155–171, 2018. 2
- [47] Alex Trevithick and Bo Yang. Grf: Learning a general radiance field for 3d representation and rendering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15182–15192, 2021. 2
- [48] Richard Tucker and Noah Snavely. Single-view view synthesis with multiplane images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 551–560, 2020. 1, 3
- [49] Shubham Tulsiani, Richard Tucker, and Noah Snavely. Layer-structured 3d scene inference via view synthesis. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 302–317, 2018. 3
- [50] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 4
- [51] Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A Yeh, and Greg Shakhnarovich. Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. *arXiv preprint arXiv:2212.00774*, 2022. 3
- [52] Tengfei Wang, Bo Zhang, Ting Zhang, Shuyang Gu, Jianmin Bao, Tadas Baltrusaitis, Jingjing Shen, Dong Chen, Fang Wen, Qifeng Chen, et al. Rodin: A generative model for sculpting 3d digital avatars using diffusion. *CVPR*, 2023. 1
- [53] Tengfei Wang, Ting Zhang, Bo Zhang, Hao Ouyang, Dong Chen, Qifeng Chen, and Fang Wen. Pretraining is all you need for image-to-image translation. In *arXiv*, 2022. 2

- [54] Daniel Watson, William Chan, Ricardo Martin-Brualla, Jonathan Ho, Andrea Tagliasacchi, and Mohammad Norouzi. Novel view synthesis with diffusion models. *arXiv preprint arXiv:2210.04628*, 2022. 2
- [55] Olivia Wiles, Georgia Gkioxari, Richard Szeliski, and Justin Johnson. Synsin: End-to-end view synthesis from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7467–7477, 2020. 1, 3
- [56] Jiaxin Xie, Hao Ouyang, Jingtian Piao, Chenyang Lei, and Qifeng Chen. High-fidelity 3d gan inversion by pseudo-multi-view optimization. *CVPR*, 2023. 1, 6
- [57] Dejia Xu, Yifan Jiang, Peihao Wang, Zhiwen Fan, Humphrey Shi, and Zhangyang Wang. Sinnerf: Training neural radiance fields on complex scenes from a single image. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXII*, pages 736–753. Springer, 2022. 1, 6, 7
- [58] Dejia Xu, Yifan Jiang, Peihao Wang, Zhiwen Fan, Yi Wang, and Zhangyang Wang. Neurallift-360: Lifting an in-the-wild 2d photo to a 3d object with 360 $\{\backslash\deg\}$ views. *arXiv preprint arXiv:2211.16431*, 2022. 3
- [59] Lior Yariv, Yoni Kasten, Dror Moran, Meirav Galun, Matan Atzmon, Basri Ronen, and Yaron Lipman. Multiview neural surface reconstruction by disentangling geometry and appearance. *Advances in Neural Information Processing Systems*, 33:2492–2502, 2020. 4
- [60] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4578–4587, 2021. 1, 2
- [61] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Free-form image inpainting with gated convolution. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4471–4480, 2019. 6, 12
- [62] 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. 8
- [63] Tinghui Zhou, Shubham Tulsiani, Weilun Sun, Jitendra Malik, and Alexei A Efros. View synthesis by appearance flow. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 286–301. Springer, 2016. 3

Appendix

A. Broad Impact

We have presented *Make-It-3D*, a novel approach to create novel views from a single image of general genre. *Make-It-3D* first hallucinates the 3D geometry by the usage of depth prior at the frontal view and the geometry prior of a pretrained diffusion model to ensure plausibility at novel views. Motivated by the fact that human eyes are more sensitive to texture over geometry, we thus reuse the coarse 3D geometry estimated from the implicit representation as well as the texture from the reference image, and specifically “in-paints” the texture of explicit 3D representation at occluded regions, ultimately producing compelling novel view renderings with highly-detailed texture.

Our primary aim is to advance the research of generative modeling from 2D to 3D. Without relying on 3D training data that is hardly accessible in scale, this work tackles the 3D synthesis problem by lifting 2D generated images to 3D. This way essentially builds on the assumption that a diffusion model not only generates 2D observations but also implicitly contains rich 3D understanding of the scene. Thus, using our technique, one can generate a 3D scene that can be immersively viewed by merely using a 2D diffusion model. Compared to DreamFusion and Magic3D, our work produces more diverse 3D synthesis results with significantly improved realism. On top of creatively generated images, this work also performs well on real images with complicated structures.

We hope this work opens the door towards high-quality 3D synthesis and inspires more following works along this way. While we have demonstrated the ability to synthesize novel views in 360 degree, it is still non-trivial to produce holistically plausible 3D objects when viewed from large viewpoints. Moreover, while this work aims for 3D synthesis from a single image, the same pipeline is applicable to the few-shot scenario where a few multi-view images can be obtained. In addition, it would be fruitful to generalize the proposed technique to augment the quality of 4D synthesis. We will release the code to facilitate the research in this emerging area.

B. Additional Implementation Details

B.1. Coarse stage

Scene representation and rendering. We use the explicit-implicit representation from Instant-NGP [25] to implement the NeRF representation in the coarse optimization stage, where we choose 16-level hash encoding of size 2^{19} and dimension 32, with a 3-layer MLP with 64 hidden units to decode the density and color for each spatial location. During volumetric rendering, we sample 96 points for each ray, including 64 points for uniform sampling and 32 for impor-

tance sampling. We initialize the density field as a Gaussian sphere, which leads to faster convergence and more stable training. Specifically, we initialize the density as $\sigma_{\text{init}} = d * \exp(-||x||^2/(2\mu^2))$, where we set density bias $d = 5$ and $\mu = 0.2$; x denotes the distance between the ray point and the scene center.

Camera setting. Following the camera sampling method used in [32], we randomly sample camera distance from 0.8 to 1.2, and the field-of-view (FOV) from 40 to 80 degrees. We find that randomly sampling FOV is instrumental to mitigate the artifacts that arise in large rendering view angles.

Augmentation and Regularization. To encourage the network to focus more on the foreground and avoid adversarial samples that hack the pretrained diffusion model, we train NeRF with a random background augmentation. Specifically, during training, we randomly jitters the background color of both the reference alpha image and NeRF rendering. During inference, we render the scene with a white background. Furthermore, following [32], we use three types of geometric regularization including sparsity, opacity and smoothness.

B.2. Refine stage

Point cloud rasterization. Following [2], we rasterize neural points V to multi-scale feature maps $\mathcal{S}(i, V)$, $i \in [0, K]$, $K = 3$. We use a differentiable point rasterizer implemented by PyTorch3D [36] to assign every pixel a neural descriptor and a binary scalar that indicates a non-empty pixel. We consider the binary mask as a point-based occupancy mask.

Background regularization. To handle pixels without corresponding point cloud projection, we assign a learnable descriptor as the background. During texture enhancement optimization, we additionally add a regularization to encourage the scene to be rendered with a white background according to the binary occupancy mask mentioned above.

Deferred neural rendering. For deferred rendering of the point clouds, we use a 2D U-Net architecture with gated convolutions [61]. It contains 3 down- and up-sampling layers to integrate multi-scale feature maps and output the final RGB image.

C. Additional Ablation Study and Analysis

C.1. Analysis of SDS and CLIP-D loss

As mentioned in Sec 3.1, in the coarse stage, we use the diffusion prior by applying score distillation sampling (SDS) scheme on novel view renderings. It can successfully encourage the generated scene to match the conditioned text prompt. However, as an image-based 3D content creation model, we need to prioritize the faithfulness between created 3D and the reference image. Although we add pixel-wise constrain under the reference view for optimization,

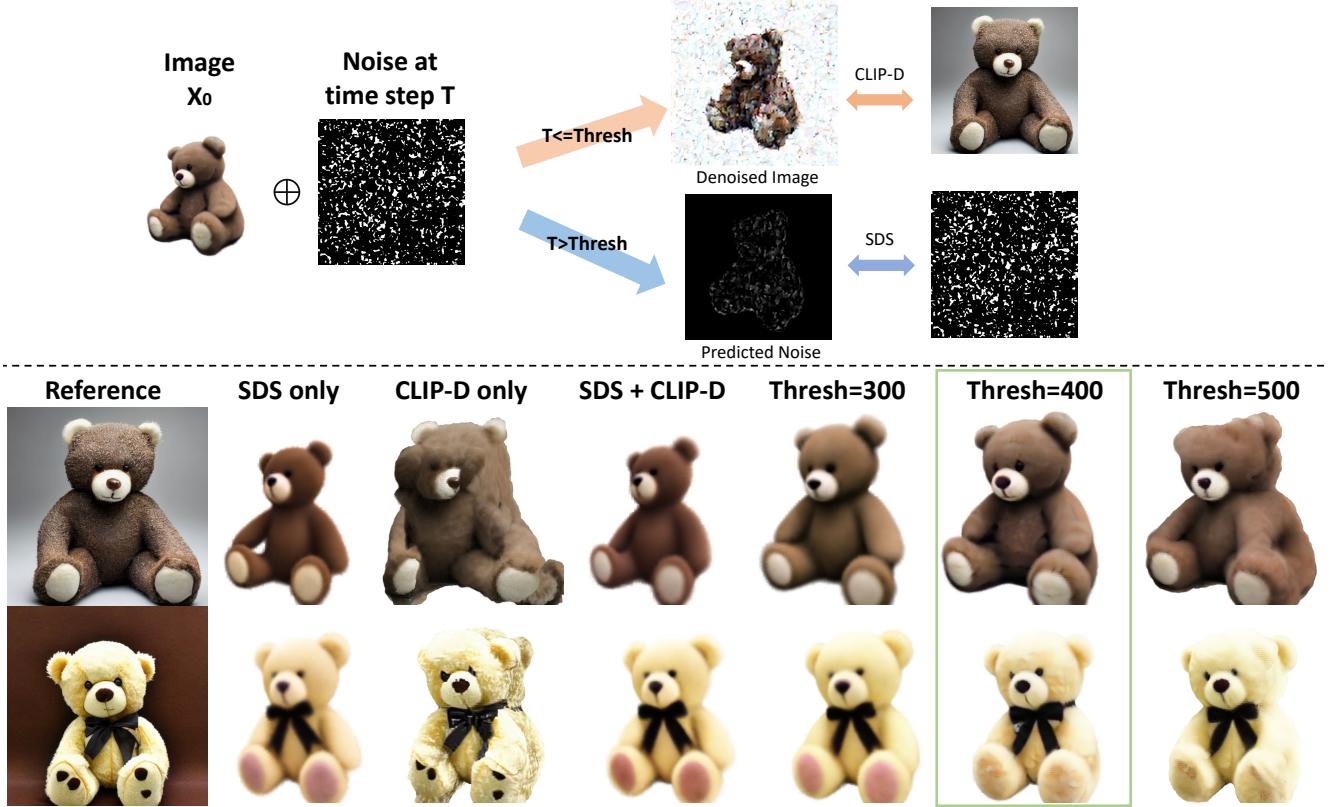


Figure 12: Analysis of SDS and CLIP-D loss.

	LPIPS↓	Contextual↓	CLIP↑
SDS	0.3045	2.29	86.04%
CLIP-D	0.1260	2.43	80.27%
SDS+CLIP-D	0.2772	2.32	84.01%
Thresh=300	0.1757	2.19	87.40%
Thresh=400	0.1427	1.74	87.50%
Thresh=500	0.1696	2.23	86.09%

Table 3: Ablation study on SDS and CLIP-D loss on the test benchmark. We compute LPIPS under the reference view, and the other two metrics under novel views. “Thresh” denotes the boundary of time steps using SDS or CLIP-D in the denoising process.

SDS provides a strong geometric prior and enforces the optimized scene to be a plausible result according to the text condition. Constraints under a single view can be limited. Thus the created results may not be rigorously aligned with the reference image (See Figure 12).

Therefore, we need to relax the strong geometric guidance provided by SDS and add more image-level constraints under multi-views. We achieve this goal by simultaneously maximizing the image-level similarity between the reference image and the novel view renderings denoised by the

diffusion model, named as a diffusion CLIP loss $\mathcal{L}_{\text{CLIP-D}}$. Compared with introducing this constraint directly on novel view renderings, the CLIP-D encourages the pretrained diffusion model to provide better guidance to generate more faithful 3D content with the reference image.

In view of this, we conduct several experiments to study the effect of SDS and CLIP-D loss during optimization, which is shown in Figure 12. Results show that using only SDS generates high-quality and plausible geometry, but the optimized 3D does not align with the image. On the contrary, using only CLIP-D preserves the appearance of the reference image, but fails to generate good geometry. A simple solution is to combine the two losses, but this does not fully address the non-alignment issue. To achieve a balance between geometric quality and appearance alignment, we introduce an optimization strategy by setting a threshold of sampling steps. Specifically, we optimize CLIP-D loss at small timesteps and optimize SDS at large steps. We conduct several qualitative and quantitative studies on different threshold settings, which are shown in Figure 12 and Table 3. During training, we randomly sample noise step T from 200 to 600, and we find that $T = 400$ could balance the geometric quality and the appearance alignment.

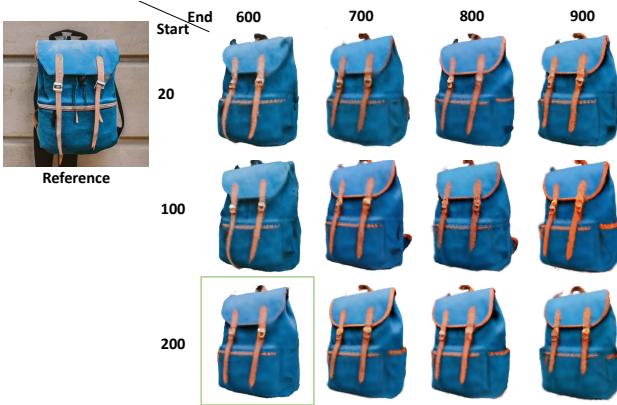


Figure 13: Analysis of the time step range in SDS process. We visualize novel view results in the coarse stage that are trained with different time step ranges (from start to end).

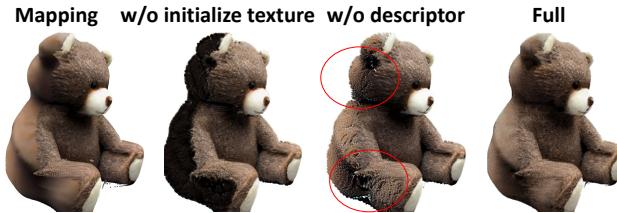


Figure 14: Analysis of texture initialization and point descriptors.

C.2. Analysis of various sampling time step ranges

We also investigate the effect of various sampling time steps in SDS process. The experimental results are shown in Figure 13. We conduct several experiments using different sampling ranges. We observe that adding noise at large time steps can improve the geometry quality but reduce the alignment and potentially saturate textures. And the diffusion prior does not provide adequate supervision at small time steps. In our method, we exclude small and large time steps and instead randomly sample time step T from 200 to 600.

C.3. Analysis of texture initialization and point descriptors

We conduct ablation studies on texture enhancement process. We explore the importance of the initialized unseen texture from NeRF and point descriptor. The qualitative results are shown in Figure 14. We can see that texture initialization is crucial for global texture enhancement. And only optimizing point color without descriptor outputs artifacts and cannot produce reasonable results.

D. Additional Results

In this section, we provide additional results of creating 3D models from different reference images using our method. The results are shown on Figure 16, Figure 17, and

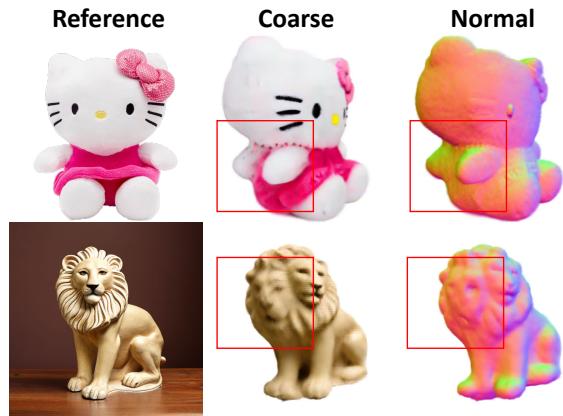


Figure 15: Failure cases due to the geometry ambiguity.

Figure 18. Results show that our method has a strong ability on creating high-fidelity 3D content including high-quality geometries and textures using a single reference image.

E. Limitations

Our method suffers from some geometry ambiguity, such as Janus problem or over-flat geometry [32]. A depth prior can reduce this issue. However, since we only add depth constrain at a single view, the geometry ambiguity may still exist under other views. We show some failure cases in Figure 15.

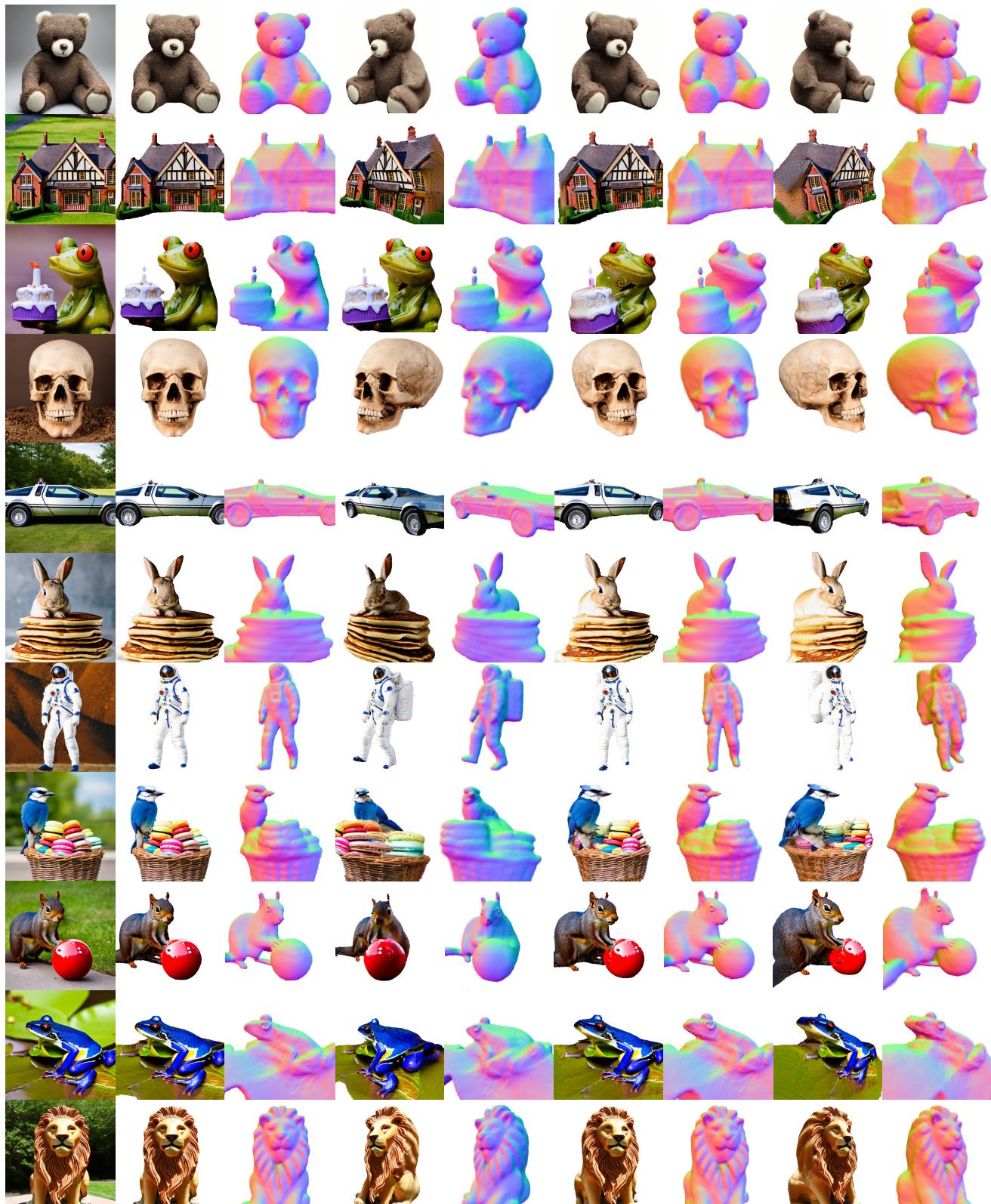


Figure 16: Additional results by *Make-It-3D*. The first column is the reference image. We show high-fidelity results including normal maps under the reference view and novel views.

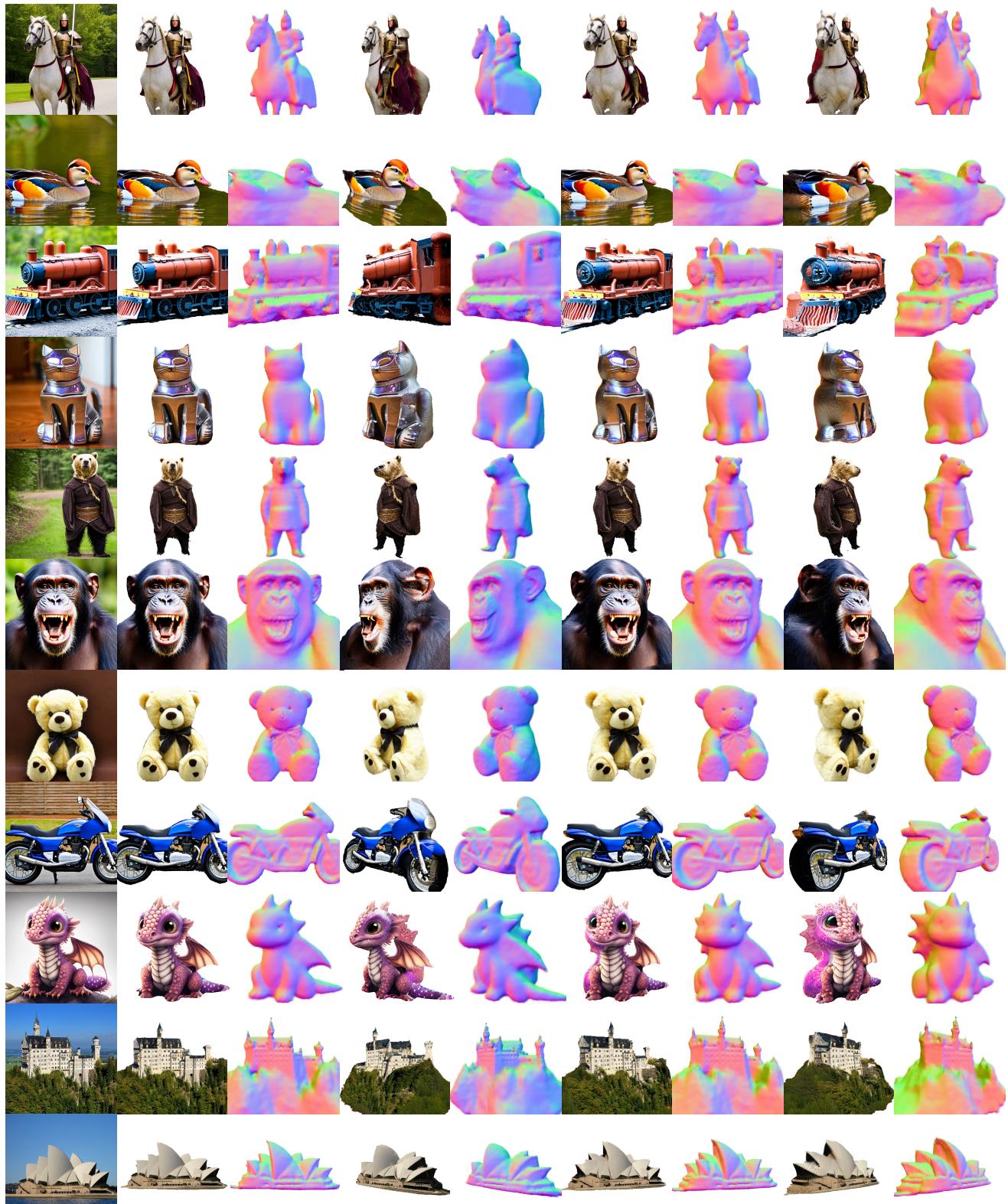


Figure 17: Additional results by *Make-It-3D*. The first column is the reference image.



Figure 18: Additional results by *Make-It-3D*. The first column is the reference image.