HyperDreamer: Hyper-Realistic 3D Content Generation and Editing from a Single Image

Tong Wu*
wt020@ie.cuhk.edu.hk
The Chinese University of Hong Kong
China
Shanghai AI Laboratory
China

Zhibing Li*
lz022@ie.cuhk.edu.hk
The Chinese University of Hong Kong
China
Shanghai AI Laboratory
China

Shuai Yang*
yssss.mikey@gmail.com
Shanghai AI Laboratory
China
Shanghai Jiao Tong University
China

Pan Zhang
zhangpan@pjlab.org.cn
Shanghai AI Laboratory
China

Xingang Pan
xingang.pan@ntu.edu.sg
S-Lab, NTU
Singapore

Jiaqi Wang
wangjiaqi@pjlab.org.cn
Shanghai AI Laboratory
China

Dahua Lin[†]
dhlin@ie.cuhk.edu.hk
The Chinese University of Hong Kong
China
Shanghai AI Laboratory
China

Ziwei Liu[†]
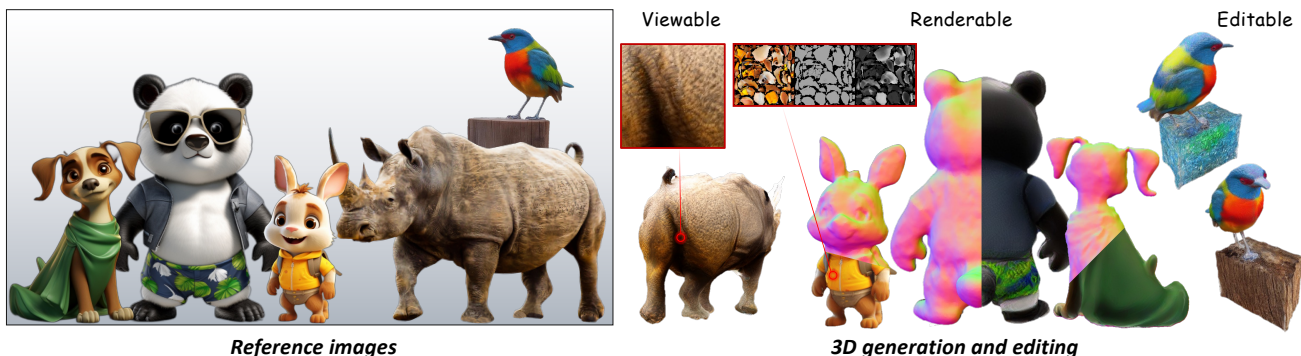ziwei.liu@ntu.edu.sg
S-Lab, NTU
Singapore

Figure 1: Overview. Given a single RGB image, we generate a realistic 3D model with rich details, which is full-range viewable, renderable, and editable.

ABSTRACT

3D content creation from a single image is a long-standing yet highly desirable task. Recent advances introduce 2D diffusion priors, yielding reasonable results. However, existing methods are not hyper-realistic enough for post-generation usage, as users cannot view, render and edit the resulting 3D content from a full range.

*Equal contribution.
[†]Corresponding Authors.

To address these challenges, we introduce **HyperDreamer** with several key designs and appealing properties: **1) Full-range viewable**: 360° mesh modeling with high-resolution textures enables the creation of visually compelling 3D models from a full range of observation points. **2) Full-range renderable:** Fine-grained semantic segmentation and data-driven priors are incorporated as guidance to learn reasonable albedo, roughness, and specular properties of the materials, enabling semantic-aware arbitrary material estimation. **3) Full-range editable:** For a generated model or their own data, users can interactively select any region via a few clicks and efficiently edit the texture with text-based guidance. Extensive experiments demonstrate the effectiveness of HyperDreamer in modeling region-aware materials with high-resolution textures and enabling user-friendly editing. We believe that HyperDreamer holds promise for advancing 3D content creation and finding applications in various domains.

CCS CONCEPTS

• **Computing methodologies** → **Computer vision**.

KEYWORDS

Single-image reconstruction, 3D generation, text-guided texturing.

1 INTRODUCTION

In light of the high costs associated with expert-assisted 3D content creation and the increasing demand across diverse applications, such as gaming, online conferencing, and virtual social presence, there has been growing attention on 3D content generation, particularly in the domain of controllable generation. Traditional approaches [Chan et al. 2021; Deng et al. 2021] in this field have predominantly relied on training category-specific models using large-scale 3D or 2D datasets, resulting in limited applications to specific categories. However, recent years have witnessed remarkable progress [Lin et al. 2023; Poole et al. 2022], notably through the incorporation of diffusion priors derived from state-of-the-art 2D generative models. These advancements have facilitated the generation of reasonably accurate 3D content, marking a significant breakthrough in the field.

In recent 2D diffusion-based 3D content generation methods [Poole et al. 2022; Tang et al. 2023], it becomes common practice to incorporate text or single image conditions to achieve controllable generation. Due to its inherent ill-posed nature, researchers rely on a 2D diffusion model [Rombach et al. 2021] as a guide prior to directing the rendering process, ensuring that all generated images are concentrated within the high-realism regions of the latent space. By confining the generated content to these regions, the overall realism of the produced 3D content is significantly enhanced.

Despite notable advancements, Current methods for 3D content generation suffer from two major drawbacks: limited post-generation usability and 2D diffusion bias. The former stems from the use of implicit 3D representations that trade off usability for fidelity. Users are unable to freely zoom, re-render, or edit the resulting 3D content to get the desired 3D content, which hampers its practical applicability and restricts creative possibilities. The latter arises from the training of the diffusion model on a 2D dataset that contains rich lighting and shading variations. These variations enhance the realism of the 2D images, but also introduce unwanted effects in the textures of the 3D models, as shown in Figure 4-d.

To address the above issues, we propose HyperDreamer, a 3D content generation and editing framework that is full-range viewable, renderable, and editable. 1) Full-range viewable: A novel custom super-resolution module is introduced, which incorporates pseudo multi-view images to facilitate high-resolution supervision. This module enables the generation of high-resolution textures for 360° content, allowing the creation of visually captivating 3D models from a full range of observation points. 2) Full-range renderable:

The Segment-Anything-Model [Kirillov et al. 2023] is integrated into our generation approach, enabling online 3D semantic segmentation. Leveraging the segmentation mask, we introduce a semantic-aware albedo regularization loss to mitigate the diffusion bias. To enable a more realistic rendering in downstream applications, we model the appearance using a spatially varying Bidirectional Reflectance Distribution Function (BRDF) [Chen et al. 2022] and learn reasonable albedo, roughness, and specular properties of the materials, enabling semantic-aware arbitrary material estimation. 3) Full-range editable: An interactive editing method is introduced, enabling users to perform interactive segmentation on 3D meshes effortlessly. By leveraging a normal-to-image model diffusion model, HyperDreamer allows users to edit textures of specific regions in 3D meshes using text-based guidance. With just a few clicks, users can efficiently modify the targeted region, enhancing the editability and flexibility of the HyperDreamer.

Extensive experiments demonstrate the effectiveness of Hyper-Dreamer in modeling region-aware materials with high-resolution textures, and facilitating user-friendly editing, and show that Hy-perDreamer surpasses state-of-the-art methods by a significant margin in terms of both 3D generation and editing quality. We believe that HyperDreamer, with its markedly superior quality and flexibility, effectively broadens the accessibility of AI-generated 3D content for practical applications.

2 RELATED WORKS

Text-guided 3D Generation. The text-guided 3D generation has gained significant attention following the remarkable success of text-to-image generation methods. Dream Fields [Jain et al. 2022] employed the text-image model CLIP [Radford et al. 2021] to optimize NeRFs [Mildenhall et al. 2020] by aligning the text and image embeddings. Building on the same principle, DreamFusion [Poole et al. 2022] replaced CLIP with diffusion models and devised an SDS loss to distill knowledge from the denoising procedures. Magic3D [Lin et al. 2023] further enhanced generation performance by employing a coarse-to-fine framework and using meshes as the 3D representation in the second stage. Fantasia3D [Chen et al. 2023] disentangled the geometry and appearance modeling and introduced the spatially varying bidirectional reflectance distribution function (BRDF) for photo-realistic texture. Our approach utilizes a single image as the guided condition instead of text, which provides more detailed and specific information and introduces additional challenges.

Single-image Reconstruction. Reconstructing 3D models from a single image has been a long-existing topic. Inference-based methods [Choy et al. 2016; Gu et al. 2023; Jun and Nichol 2023; Melas-Kyriazi et al. 2023a; Nichol et al. 2022; Pavllo et al. 2023; Tulsiani et al. 2017; Vasudev et al. 2022; Wu et al. 2023] heavily depend on the datasets used for training, many of which can not handle diverse and general objects. Optimization-based methods utilize priors from 2D text-to-image diffusion model to guide the reconstruction process. RealFusion [Melas-Kyriazi et al. 2023b] employs textual inversion technique to bridge the gap between the reference image and text-conditioned guidance. Make-it-3D [Tang et al. 2023] employs a two-stage framework and leverages high-quality textures extracted from the reference image. Zero-1-to-3 [Liu et al. 2023] synthesizes novel views by fine-tuning diffusion models [Rombach
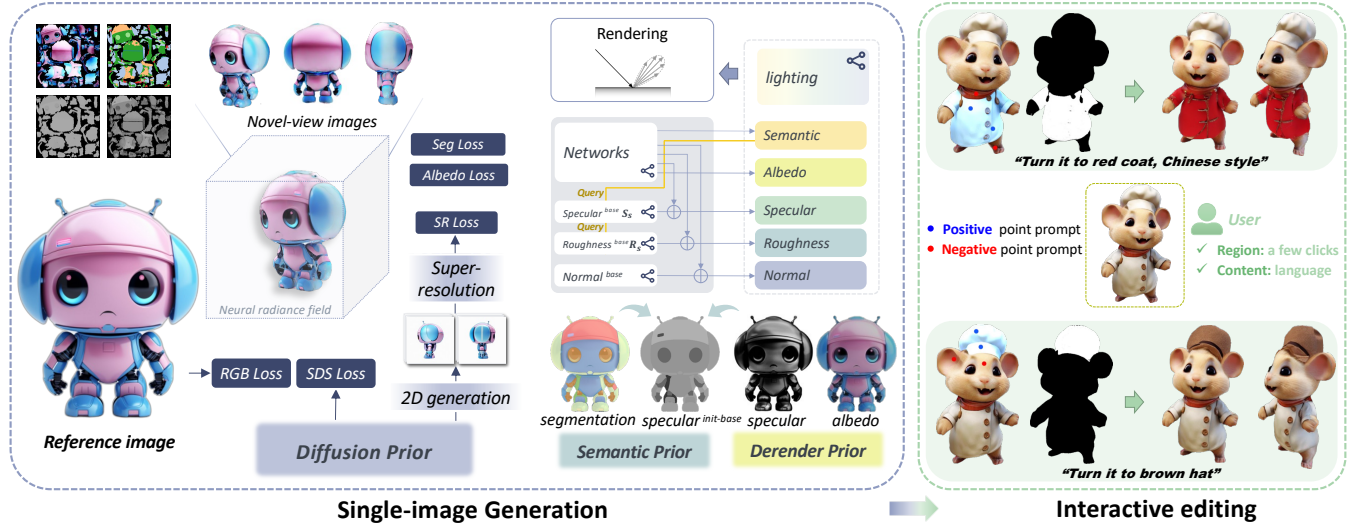
Figure 2: Overview of our 3D generation and editing pipeline. We introduce diffusion priors, semantic priors, and derendering priors into this highly under-constraint problem to enable high-resolution textures with material modeling and interactive editing after the generation.

et al. 2021] with multi-view data. It has also been applied to single-image 3D reconstruction by applying SJC [Wang et al. 2022]. Our work utilizes Zero-1-to-3 as the guidance model and incorporate several key designs to enable broader applications.

Material and Illumination Estimation. Multi-view reconstruction methods [Munkberg et al. 2022] benefited from separately modeling geometry, material, and illumination conditions, while it's a highly ill-posed problem for generation. Previous works like Fantasia3D [Chen et al. 2023] propose to learn globally varying roughness and metallic distributions, which may not always align with realistic material properties. Based on the material estimation approaches from a single image [Sang and Chandraker 2020; Wimbauer et al. 2022], we further propose a more plausible assumption that materials within the same semantic class share similar material properties, enabling spatially varying materials modelling while preventing degenerate solutions.

Text-guided 3D Editing. Recently, text-guided image processing has experienced rapid development in both quality and diversity. Text2Mesh [Michel et al. 2022] proposes a neural style field, which uses CLIP to guide the initial mesh based on text. TANGO [Chen et al. 2022] follows a similar scheme and uses a BRDF to optimize the appearance. However, there is a gap from the actual use due to insufficient accuracy. More recently, TEXTure [Richardson et al. 2023] leverages an improved depth-to-image diffusion process and applies an iterative scheme that paints a 3D model from different viewpoints. However, none of them enable text-guided editing of a local area on a 3D object. We propose an interactive editing method that users can edit textures based on text guidance in selected 3D regions with a few simple clicks or in a global manner.

3 PRELIMINARIES

3.1 3D Representation

Inspired by Magic3D [Lin et al. 2023], we adopt NeRF [Mildenhall et al. 2020] and DMTet [Shen et al. 2021] for the first and second stage training, respectively. In the first stage, NeRF represents the scene as an implicit function that maps a 3D location x and a 2D viewing direction d to a volume density τ and color c. To render a pixel, NeRF alpha-composites the densities and colors along the ray that is cast from the camera to the pixel:

$$C = \sum_k \alpha_k \prod_{k'<k} (1 - \alpha_{k'})c_k, \quad \alpha_k = 1 - \exp(-\tau_k \|x_{k+1} - x_k\|). \quad (1)$$

To accelerate the training, we employ the efficient hash grid encoding from Instant NGP [Müller et al. 2022] instead of pure MLPs.

In the second stage, we adopt DMTet to produce high-resolution outputs without high computational and memory requirements. DMTet is a hybrid representation that integrates implicit and explicit surface representations and can efficiently render high-resolution textured meshes with differentiable rasterization. Formally, DMTet models the 3D shape as a deformable tetrahedral grid (V_T, T), where V_T are the vertices in the tetrahedral grid T. Each tetrahedron $T_k \in T$ has four vertices $\{v_{i_k}|i \in \{a, b, c, d\}\}$, each associated with a SDF value $s(v_i)$ and a deformation Δv_{i_k}. The surface mesh is extracted by differentiable marching tetrahedra algorithm.

3.2 Score Distillation Sampling (SDS)

Previous works [Lin et al. 2023; Poole et al. 2022] have leveraged the 2D diffusion model [Rombach et al. 2021] as prior knowledge for text-to-3D generation. The diffusion model ϕ learns a denoising function $\epsilon_\phi(x_t; y, t)$ that estimates the noise ϵ based on the noisy image x_t, text embedding y and noise step t. It progressively reduces the noise and introduces image structure. To optimize the 3D scene θ, Score Distillation Sampling (SDS) guides all rendered images to

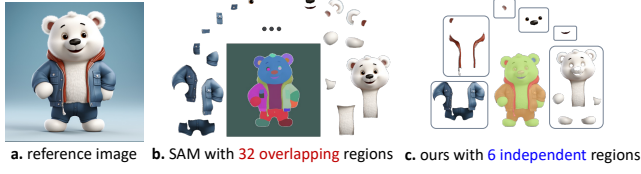**a.** reference image   **b.** SAM with 32 overlapping regions   **c.** ours with 6 independent regions

Figure 3: SAM at the generation stage. We effectively cluster concise semantic groups compared to the raw SAM results.

match the given text embedding y under diffusion priors:

$$\nabla_\theta \mathcal{L}_{SDS}(\phi, x = g(\theta)) = \mathbb{E}_{t,\epsilon}\left[\omega(t)(\epsilon_\phi(x_t; y, t) - \epsilon)\frac{\partial x}{\partial \theta}\right], \quad (2)$$

where g denotes the image renderer and $\omega(t)$ represents a weighting function. In addition to text conditional SDS, zero-1-to-3 [Liu et al. 2023] introduces a 3D-aware SDS that conditions on input view and relative camera extrinsic to exploit the 3D consistent priors:

$$\nabla_\theta \mathcal{L}_{SDS}(\phi, x = g(\theta)) = \mathbb{E}_{t,\epsilon}\left[\omega(t)(\epsilon_\phi(x_t; x_i, R, T, t) - \epsilon)\frac{\partial x}{\partial \theta}\right], \quad (3)$$

where x_i represents the input view, R and T are the relative camera rotation and translation from the input view to the desired viewpoint.

3.3 Segment Anything Model (SAM)

Segment Anything Model (SAM) [Kirillov et al. 2023] is the foundation model for general image segmentation, which supports various segmentation modes such as automatic everything and manual prompt. Taking point prompts as an example, SAM takes an image I and a set of user-specific prompts $\mathcal{P} = (p, l)$ as inputs, and the output is a corresponding segmentation mask M. Among them, $\mathcal{P}$ includes p and l, where p is the set of each point coordinate and l is the set corresponding to each point label. We use S to represent the SAM model, so we have $M_{I,\mathcal{P}} = S(I, \mathcal{P})$.

4 METHODOLOGY

This section elaborates on the proposed framework in detail. Despite the inherent challenges posed by the ill-posed nature of the problem, HyperDreamer capitalizes on the deep priors from the 2D diffusion model, semantic segmentation model, and material estimation model, which collectively empower the capability for full-range viewing, rendering, and interactive editing. Specifically, (1) to achieve high-fidelity texture generation, we utilize high-resolution pseudo multi-view images for auxiliary supervision, as detailed in Sec 4.1. (2) For material modeling, we introduce online 3D semantic segmentation and semantic-aware regularizations, which is initialized via material estimation results, as described in Sec. 4.2. (3) Furthermore, a novel interactive editing approach is proposed in Sec. 4.3 for effortless targeted modification of regions on 3D meshes via interactive segmentation.

a. reference image   **b.** baseline   **c.** w/ albedo regularization
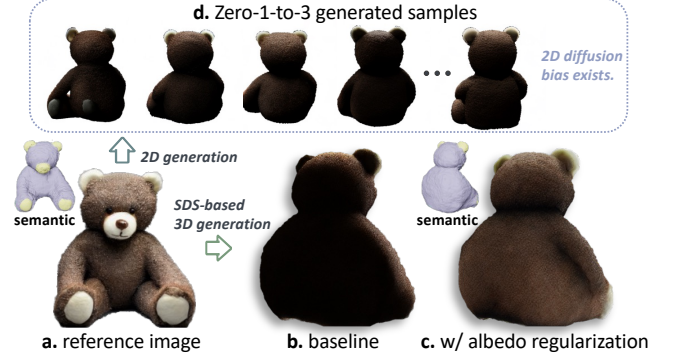
Figure 4: Diffusion bias. The 2D diffusion bias in d leads to 3D generation failures in b, which can be alleviated by the albedo regularization in c.

4.1 360° High-Resolution Texture Generation

In the second training stage, the mesh representation allows for rapid images rendering, unlocking the potential of achieving high-resolution texture maps. However, our guidance model, Zero-1-to-3 [Liu et al. 2023], was originally trained on low-resolution images (256×256). The resulting SDS loss fails to handle higher-resolution images, thereby limiting the benefits offered by mesh representation. The disparity between the resolutions used for training and inferencing leads to a relatively blurry texture map.

To overcome this challenge, we propose a high-resolution texture generation module. We first select a set of novel views and directly generate m images per view using Zero-1-to-3. Subsequently, we employ a super-resolution network [Rombach et al. 2021] to upscale the sampled images, enabling high-resolution supervision. Since the multi-view images generated by Zero-1-to-3 are not perfectly 3D consistent, directly applying per-pixel loss can lead to network instability. Instead, we employ perceptual loss [Johnson et al. 2016] in the feature space. By leveraging perceptual loss, we can minimize the content and style differences between two images without relying on pixel-level alignment, effectively alleviating inconsistencies during the training process.

4.2 Semantic-Aware Material Estimation

4.2.1 Online global semantic segmentation. During the second training stage, we also propose to integrate a new MLP-based branch upon the hash encoding and equip the framework with a globally-consistent mesh segmentation for further semantic regularization. We first use SAM to produce over-segmented results of the reference image (Figure 3 (b)), and then we cluster different semantic parts by thresholding the feature similarity among them before assigning the semantic labels, as shown in Figure 3 (c). We assume that the reference image already contains all of the semantic components of the generated 3D model. We also assign pseudo labels to novel view images by thresholding the feature similarities, and all these 2D labels are used to supervise the semantic branch training. We present detailed implementations in the supplementary materials.

4.2.2   Semantic-Aware Albedo Regularization. Recent approaches in the single-image 3D generation [Liu et al. 2023; Melas-Kyriazi et al. 2023b; Tang et al. 2023] optimize the model with diffusion priors and RGB reconstruction loss. They adopt different types of shading augmentations at novel views, including albedo, diffuse, and textureless, while only albedo shading is applied at the reference view. However, this pipeline introduces two inherent problems.

Firstly, the diffusion priors suffer from intrinsic shading and reflectance effects. For instance, Stable Diffusion and Zero-1-to-3 are trained on abundant images with lighting and shading variations, inevitably baking these effects into the textures of the generated 3D models. As shown in Figure 4, given the front view of the teddy bear, Zero-1-to-3 tends to generate a dark back view, as if the light source only exists in the front, leading to a black back of the 3D model as in Figure 4-b.

Secondly, the shading and reflectance characteristics in the reference image are integrated into the albedo color learning of the model via RGB reconstruction loss, making it challenging for re-rendering.

We aim to introduce several albedo losses to alleviate the aforementioned problems. For the diffusion bias, we assume that albedo colors in regions under the same semantic label are similar. For the N_s semantic labels, we maintain a albedo library called A_s, which is updated regularly according to the semantic-region-averaged albedo colors of the reference image along the training. For each novel view, we predict segmentation masks with the semantic branch, and then we use a Gaussian filter to gain a weighted average of predicted albedo colors inside each semantic group. We propose a *semantic-aware albedo regularization* as below:

$$L_a = \sum_{i=1}^{N_s} ||F_{gaussian}(A_{pred}^i) - A_s^i||_2^2. \tag{4}$$

Furthermore, we incorporate a state-of-the-art single-image derendering framework [Wimbauer et al. 2022] to generate the albedo map of the reference image as an additional albedo supervision.

4.2.3   Appearance Modeling. To enable a more realistic rendering, we introduce the Physically-Based Rendering (PBR) material model. Following TANGO [Chen et al. 2022] and PhySG [Zhang et al. 2021], we leverage spatially varying BRDF (SVBRDF) to parameterize the material, including roughness, specular, and normal. Based on the rendering equation [Kajiya 1986], given a location x and the surface normal n, the incident light intensity at this point is denoted as $L_i(\omega_i; x)$ along the direction ω_i; BRDF $f_r(\omega_o, \omega_i; x)$ denotes the reflectance coefficient of the material viewing from direction ω_o. The observed light intensity $L_o(\omega_0; x)$ is calculated over the hemisphere $\Omega = \{\omega_i : \omega_i \cdot n > 0\}$:

$$L_o(\omega_0; x) = \int_{\Omega} L_i(\omega_i; x) f_r(\omega_o, \omega_i; x)(\omega_i \cdot n) d\omega_i. \tag{5}$$

We utilize spherical Gaussians (SGs) [Yan et al. 2012] to approximate the rendering equation in closed form. For a spherical Gaussian with n dimensions, given the lobe axis $\xi \in \mathbb{S}^2$, lobe sharpness $\lambda \in \mathbb{R}_+$, and lobe amplitude $\mu \in \mathbb{R}_+^n$, the spherical function is formulated as:

$$G(v; \xi, \lambda, \mu) = \mu e^{\lambda(v \cdot \xi - 1)}, \tag{6}$$
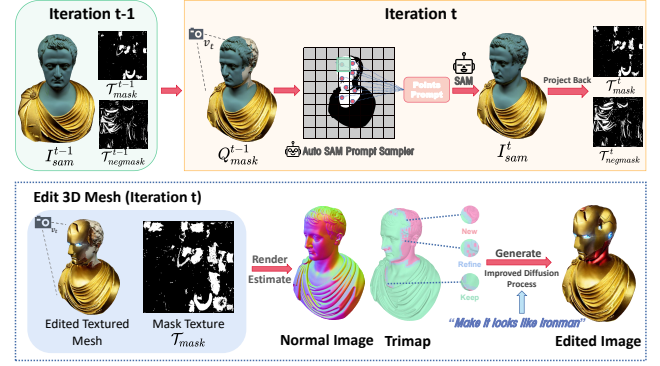
where $v \in \mathbb{S}^2$ denotes the input.

Figure 5: Interactive editing process. Users can select the interest regions and then our method output the texture mask of the target area to our texture synthesis pipeline for text-guided editing.

The environment map $L(\omega_i)$ is represented as a mixture of SGs:

$$L_i(\omega_i) = \sum_{k=1}^{M} G(\omega_i; \xi_k, \lambda_k, \mu_k). \tag{7}$$

The SVBRDF is divided into diffuse BRDF and specular BRDF: $f_r(\omega_o, \omega_i; x) = f_d(x)/\pi + f_s(\omega_o, \omega_i; x)$. The diffuse term is modeled as an MLP based on the multi-resolution hash input encoding. And the specular term at location x is formulated as:

$$f_s(\omega_o, \omega_i; x) = G(h; n_x, \frac{\lambda_x}{4h \cdot \omega_o}, \mathcal{M}_x \mu_x), \tag{8}$$

where h is half vector and $\mathcal{M}$ is Fresnel and shadowing effects.

The last term in Eqn. 5 is approximated as [Meder and Brüderlin 2018]: $(\omega_i \cdot n) = G(\omega_i; 0.0315, n, 32.7080) - 31.7003$.

Therefore, the rendering equation is represented as the multiplication of SGs and can be calculated in closed form. Learnable parameters above include $\{\xi_k, \lambda_k, \mu_k\}_{k=1}^{M}$ for the environmental map, diffuse albedo f_d, and the spatially varying $\{\lambda, \mu\}$. We assume that regions with the same semantic label usually share alike materials and enforce channel consistency in roughness and specular. Please refer to the supplementary materials for more details.

4.3   Interactive Editing

Editing a 3D model requires complex interaction with 3D shapes while maintaining global consistency to achieve the desired design. We propose an intelligent and user-friendly interactive 3D editing tool that allows users to quickly settle the target area in 3D space by one-shot selection and edit its texture based on text guidance.

4.3.1   Interactive Segmentation In Mesh. Interactive segmentation in the 3D mesh enables users to segment any region of the 3D object. In our method, as shown in Figure 5, we use two UV maps to represent the masks of the 3D mesh, where $\mathcal{T}_{mask}$ for selected regions and $\mathcal{T}_{negmask}$ for remaining regions, respectively. Given a target view v_t, we can render the masks Q_{mask}^{t-1} and $Q_{negmask}^{t-1}$ which are actually the point prompts cache from the previous t - 1 views but not complete segmentation results in the current view. Then,
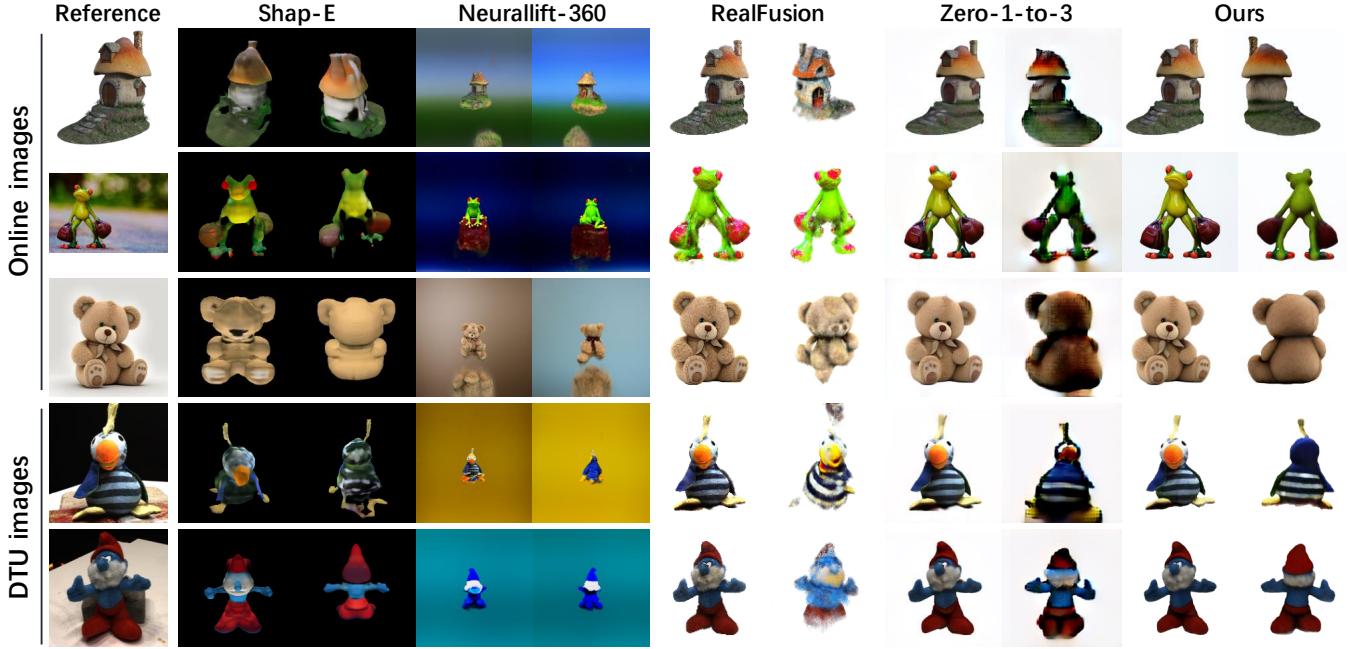
Figure 6: Qualitative comparisons. HyperDreamer generates a high-fidelity reference view and more realistic and reasonable results at the novel view.

we sample points with a patch sampling mechanism on Q_{mask}^{t-1} as positive prompts and $Q_{negmask}^{t-1}$ as negative prompts in each patch to generate refined segmentation results I_{sam} and I_{negsam} via SAM. Inverse rendering is then applied to project I_{sam} and I_{negsam} onto $\mathcal{T}_{mask}$ and $\mathcal{T}_{negmask}$, where we use a gradient-based optimization to $\mathcal{T}_{mask}$ for $\mathcal{L}_t$ over the values of $\mathcal{T}_{mask}$ when rendered through the differential renderer $\mathcal{R}$ [Jatavallabhula et al. 2019].That is,

$$\nabla_{\mathcal{T}_{mask}^t} \mathcal{L}_t = \left[\left(\mathcal{R}\left(\text{Mesh}, \mathcal{T}_{mask}^t, v_t\right) - I_{sam}^t\right) \odot m_t\right] \frac{\partial \mathcal{R} \odot m_t}{\partial \mathcal{T}_{mask}^t}, \quad (9)$$

where, m_t is the mask of mesh at the view v_t. Similarly, the method of projecting back to texture in each view in subsequent is the same.

4.3.2 Text-Guided Texture Synthesis. We apply Normal-to-Image model $\mathcal{M}_{normal}$ based on ControlNet [Zhang and Agrawala 2023] to paint textures that closely match the surface details on the 3D Mesh directly.

To address the inconsistency problem, we divide each rendered view into three partitions: M_{new}, M_{keep}, and M_{refine}. The M_{new} partition is the target region that needs to be painted for the first time. The M_{keep} partition is either a previously well-painted target region or a region that is out of the target region. The M_{refine} partition is the region painted from the previous views, but they are mainly the junction of adjacent views and need further refinement.

To attain M_{refine}, we first perform an opening operation, *Open*, on the mask M_{new} to eliminate out-lie small regions. We then performs erode $\mathcal{E}$ and dilate $\mathcal{D}$ as follows,

$$M_{refine} = \mathcal{D}(Open(M_{new})) - \mathcal{E}(Open(M_{new})). \quad (10)$$

Table 1: Quantitative results on our data.

Method	Contextual ↓	CLIP ↑	Perceptual ↓
Shap-E	4.95	0.68	-
NeuralLift-360	4.71	0.78	0.67
RealFusion	2.25	0.79	0.17
Zero-1-to-3	3.36	0.74	0.13
Ours	**2.11**	**0.86**	**0.10**

In the $\mathcal{M}_{normal}$, we modify the sampling process by blender diffusion to inject the information of region partition into the denoising process. The mask M_{paint} explicitly blends the noised latent z_{Q_t} and the denoised latent estimation z_t as follows:

$$M_{paint} = \begin{cases} 0 & M_{keep} \\ 1 & M_{new} \cup M_{refine}, \end{cases} \quad (11)$$

$$\hat{z}_t = \hat{z}_t \odot M_{paint} + z_t \odot (1 - M_{paint}). \quad (12)$$

Based on the above texture synthesis method, we can achieve local editing in the 3D mesh. In more detail, we can limit the editing area to the target region by doing the dot product with the original texture map $\mathcal{T}$ and the $\mathcal{T}_{mask}$ obtained by interactive segmentation in mesh. Finally, users can select any region in the 3D object to edit based on text guidance, as illustrated in Figure 5.

4.4 Implementation Details

4.4.1 Model and training. For the generation process, we follow Instant-NGP [Müller et al. 2022] to adopt a two-stage training pipeline: a coarse NeRF is trained for 50 epochs in the first stage, guided by Zero-1-to-3 based SDS loss and other regularisation terms

Table 2: Quantitative results on the DTU dataset.

Methods	Contextual ↓	CLIP ↑	Perceptual ↓
Shap-E	2.82	0.80	-
NeuralLift-360	4.74	0.78	0.72
RealFusion	4.84	0.82	0.47
Zero-1-to-3	4.50	0.80	0.43
Ours	**2.08**	**0.89**	**0.13**
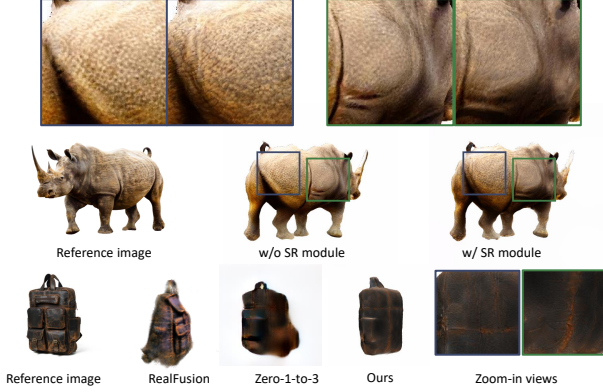
Figure 7: Ablation on the super-resolution (SR) module. High-frequency details on textures are generated under SR supervision.

like depth and normal loss. We use MiDaS [Ranftl et al. 2021] and OmniData [Eftekhar et al. 2021] to extract the depth and normal estimations, respectively. We train a DMTet [Shen et al. 2021] for 100 epochs in the second stage based on the first stage model. The SR, semantics, and material modules are integrated with the second stage only. More details are presented in the supplementary materials.

4.4.2 Methods for comparison. We compare four recent approaches for single-image based 3D generation for arbitrary images: **Shap-E** [Jun and Nichol 2023] is a conditional generative model for implicit representations trained on millions of 3D assets; **NeuralLift-360** [Xu et al. 2022] generates the neural radiance field based on the CLIP-guided diffusion priors. **RealFusion** [Melas-Kyriazi et al. 2023b] applies the SDS loss based on Stable Diffusion [Rombach et al. 2021] and RGB reconstruction loss with the reference view, in comparison, **Zero-1-to-3** [Liu et al. 2023] leverages the SJC loss based on its viewpoint-conditioned model as the guidance.

5 EXPERIMENTS

5.1 Qualitative Comparisons

We show some qualitative comparisons with several state-of-the-art works in Figure 6, where we present both the reference view and the back view of the object for each method. The results by Shap-E are relatively worse than the other optimization-based methods. The instances generated by NeuralLift-360 are small in size and low in quality, while the basic semantics is reserved. RealFusion and Zero-1-to-3, both leverage the reference view RGB reconstruction loss for constraint and thus keep a high fidelity with the reference image. While RealFusion suffers severely from the multi-face problem,
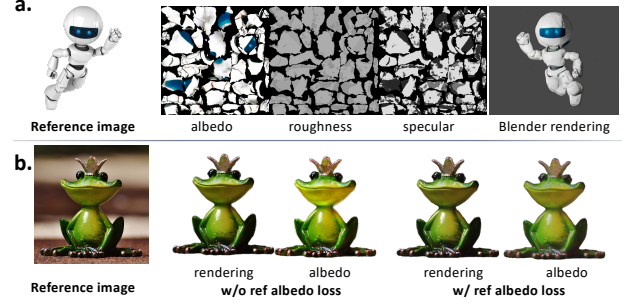
Figure 8: Analysis of the material modeling. We show an example of the output roughness and specular maps in a, together with its rendering results in Blender. We show how the albedo loss at the reference view help alleviate shading and reflectance learning in the albedo texture in b.

and the results from Zero-1-to-3 are blurry. Our method achieves the highest quality in both the reference view and the back view, presenting realistic and reasonable generations.

5.2 Quantitative Comparisons

We adopt three metrics for quantitative comparisons: 1) LPIPS [Johnson et al. 2016] evaluates the reconstruction quality of the reference view image; 2) Contextual distance [Mechrez et al. 2018] evaluates the pixel-level distance between the rendered novel view images and the reference image; 3) CLIP-Score [Radford et al. 2021] measures the semantic-level distance between the novel view images and the reference image. We select 20 images online with a wide range of diversity, and we select 10 instances from the DTU dataset [Aanæs et al. 2016] that are basically complete. The results for the two datasets are shown in Table 1 and Table 2, respectively. Our model outperforms the comparison methods in all three metrics by a large margin, quantitatively revealing the effectiveness of the pipeline.

5.3 Analysis and Ablations

5.3.1 Super Resolution. We show how our SR module works in Figure 7. It largely enhances the texture details and realism in our model, which enables it to support high-resolution zoom-in views in comparison with other methods.

5.3.2 Materials. Examples of the generated roughness and specular map are shown in Figure 8-a, where we observe that the material proprieties are highly correlated with the semantic label of the region. We also show how the albedo loss helps decompose the albedo texture out of the reference view.

5.3.3 Editing. We show in Figure 9 that the Naive method for segmentation in the mesh, which only inputs positive prompt and randomly samples the point prompts cache, has a high probability of failure especially in the condition of dealing with discrete and complex regions. However, adopting our scheme, it is more robust with patch sampling in the positive and negative prompts and input both into SAM.
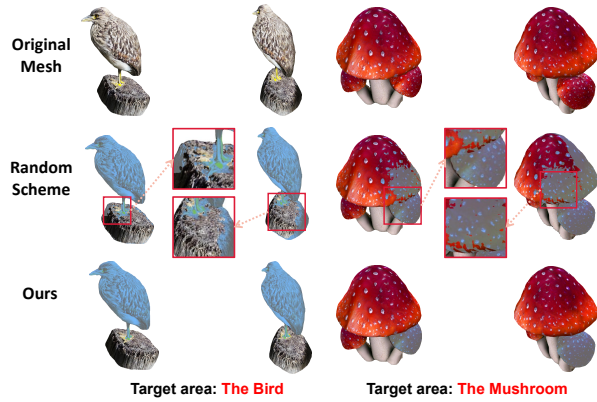
Figure 9: Analysis of our scheme of segmentation in mesh. Our method has better ability to handle complex circumstances.

6 CONCLUSION

This paper introduces a framework, **HyperDreamer**, which enables hyper-realistic 3d content generation and editing for a single image. In contrast to previous works, the 3D content generated by our method is full-range viewable, renderable, and editable. Extensive experiments demonstrate the effectiveness of HyperDreamer in modeling region-aware materials with high-resolution textures and enabling user-friendly editing. We believe that HyperDreamer holds promise for advancing 3D content creation and editing, which would be practical for both academic and industrial usage.

ACKNOWLEDGMENTS

REFERENCES

Henrik Aanæs, Rasmus Ramsbøl Jensen, George Vogiatzis, Engin Tola, and Anders Bjorholm Dahl. 2016. Large-scale data for multiple-view stereopsis. *International Journal of Computer Vision* 120 (2016), 153–168.

Eric R Chan, Marco Monteiro, Petr Kellnhofer, Jiajun Wu, and Gordon Wetzstein. 2021. pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition.* 5799–5809.

Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. 2023. Fantasia3D: Disentangling Geometry and Appearance for High-quality Text-to-3D Content Creation. *arXiv preprint arXiv:2303.13873* (2023).

Yongwei Chen, Rui Chen, Jiabao Lei, Yabin Zhang, and Kui Jia. 2022. TANGO: Text-driven Photorealistic and Robust 3D Stylization via Lighting Decomposition. In *Advances in Neural Information Processing Systems (NeurIPS).*

Christopher B Choy, Danfei Xu, JunYoung Gwak, Kevin Chen, and Silvio Savarese. 2016. 3D-R2N2: A Unified Approach for Single and Multi-view 3D Object Reconstruction. In *Proceedings of the European Conference on Computer Vision (ECCV).*

Yu Deng, Jiaolong Yang, and Xin Tong. 2021. Deformed implicit field: Modeling 3d shapes with learned dense correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 10286–10296.

Ainaz Eftekhar, Alexander Sax, Jitendra Malik, and Amir Zamir. 2021. Omnidata: A Scalable Pipeline for Making Multi-Task Mid-Level Vision Datasets From 3D Scans. In *Proceedings of the IEEE/CVF International Conference on Computer Vision.* 10786–10796.

Jiatao Gu, Alex Trevithick, Kai-En Lin, Josh Susskind, Christian Theobalt, Lingjie Liu, and Ravi Ramamoorthi. 2023. NerfDiff: Single-image View Synthesis with NeRF-guided Distillation from 3D-aware Diffusion. In *International Conference on Machine Learning.*

Ajay Jain, Ben Mildenhall, Jonathan T. Barron, Pieter Abbeel, and Ben Poole. 2022. Zero-Shot Text-Guided Object Generation with Dream Fields. (2022).

Krishna Murthy Jatavallabhula, Edward Smith, Jean-Francois Lafleche, Clement Fuji Tsang, Artem Rozantsev, Wenzheng Chen, Tommy Xiang, Rev Lebaredian, and Sanja Fidler. 2019. Kaolin: A pytorch library for accelerating 3d deep learning research. *arXiv preprint arXiv:1911.05063* (2019).

Justin Johnson, Alexandre Alahi, and Li Fei-Fei. 2016. Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14.* Springer, 694–711.

Heewoo Jun and Alex Nichol. 2023. Shap-E: Generating Conditional 3D Implicit Functions. *arXiv preprint arXiv:2305.02463* (2023).

James T Kajiya. 1986. The rendering equation. In *Proceedings of the 13th annual conference on Computer graphics and interactive techniques.* 143–150.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. 2023. Segment Anything. *arXiv:2304.02643* (2023).

Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. 2023. Magic3D: High-Resolution Text-to-3D Content Creation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. 2023. Zero-1-to-3: Zero-shot One Image to 3D Object. arXiv:2303.11328 [cs.CV]

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

Julian Meder and Beat Brüderlin. 2018. Hemispherical gaussians for accurate light integration. In *Computer Vision and Graphics: International Conference, ICCVG 2018, Warsaw, Poland, September 17-19, 2018, Proceedings.* Springer, 3–15.

Luke Melas-Kyriazi, Christian Rupprecht, Iro Laina, and Andrea Vedaldi. 2023b. RealFusion: 360 Reconstruction of Any Object from a Single Image. In *CVPR.*

Luke Melas-Kyriazi, Christian Rupprecht, and Andrea Vedaldi. 2023a. PC2: Projection-Conditioned Point Cloud Diffusion for Single-Image 3D Reconstruction. In *Arxiv.*

Oscar Michel, Roi Bar-On, Richard Liu, Sagie Benaim, and Rana Hanocka. 2022. Text2mesh: Text-driven neural stylization for meshes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 13492–13502.

Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. 2020. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. In *ECCV.*

Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. 2022. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics (ToG)* 41, 4 (2022), 1–15.

Jacob Munkberg, Jon Hasselgren, Tianchang Shen, Jun Gao, Wenzheng Chen, Alex Evans, Thomas Müller, and Sanja Fidler. 2022. Extracting triangular 3d models, materials, and lighting from images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 8280–8290.

Alex Nichol, Heewoo Jun, Prafulla Dhariwal, Pamela Mishkin, and Mark Chen. 2022. Point-E: A System for Generating 3D Point Clouds from Complex Prompts. *arXiv preprint arXiv:2212.08751* (2022).

Dario Pavllo, David Joseph Tan, Marie-Julie Rakotosaona, and Federico Tombari. 2023. Shape, Pose, and Appearance from a Single Image via Bootstrapped Radiance Field Inversion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).*

Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. 2022. DreamFusion: Text-to-3D using 2D Diffusion. *arXiv* (2022).

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning.* PMLR, 8748–8763.

René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. 2021. Vision Transformers for Dense Prediction. *ICCV* (2021).

Elad Richardson, Gal Metzer, Yuval Alaluf, Raja Giryes, and Daniel Cohen-Or. 2023. Texture: Text-guided texturing of 3d shapes. *arXiv preprint arXiv:2302.01721* (2023).

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2021. High-Resolution Image Synthesis with Latent Diffusion Models. arXiv:2112.10752 [cs.CV]

Shen Sang and M. Chandraker. 2020. Single-Shot Neural Relighting and SVBRDF Estimation. In *ECCV.*

Tianchang Shen, Jun Gao, Kangxue Yin, Ming-Yu Liu, and Sanja Fidler. 2021. Deep Marching Tetrahedra: a Hybrid Representation for High-Resolution 3D Shape Synthesis. In *Advances in Neural Information Processing Systems (NeurIPS).*

Junshu Tang, Tengfei Wang, Bo Zhang, Ting Zhang, Ran Yi, Lizhuang Ma, and Dong Chen. 2023. Make-It-3D: High-Fidelity 3D Creation from A Single Image with Diffusion Prior. *arXiv preprint arXiv:2303.14184* (2023).

Shubham Tulsiani, Tinghui Zhou, Alexei A Efros, and Jitendra Malik. 2017. Multi-view supervision for single-view reconstruction via differentiable ray consistency. In *Proceedings of the IEEE conference on computer vision and pattern recognition.* 2626–2634.

Kalyan Alwala Vasudev, Abhinav Gupta, and Shubham Tulsiani. 2022. Pre-train, Self-train, Distill: A simple recipe for Supersizing 3D Reconstruction. In *Computer Vision and Pattern Recognition (CVPR).*

Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A. Yeh, and Greg Shakhnarovich. 2022. Score Jacobian Chaining: Lifting Pretrained 2D Diffusion Models for 3D Generation. *arXiv preprint arXiv:2212.00774* (2022).

Felix Wimbauer, Shangzhe Wu, and Christian Rupprecht. 2022. De-rendering 3D Objects in the Wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 18490–18499.

Chao-Yuan Wu, Justin Johnson, Jitendra Malik, Christoph Feichtenhofer, and Georgia Gkioxari. 2023. Multiview Compressive Coding for 3D Reconstruction. *arXiv preprint arXiv:2301.08247* (2023).

Dejia Xu, Yifan Jiang, Peihao Wang, Zhiwen Fan, Yi Wang, and Zhangyang Wang. 2022. NeuralLift-360: Lifting An In-the-wild 2D Photo to A 3D Object with 360° Views. *arXiv preprint arXiv:2211.16431.*

Ling-Qi Yan, Yahan Zhou, Kun Xu, and Rui Wang. 2012. Accurate translucent material rendering under spherical Gaussian lights. In *Computer Graphics Forum,* Vol. 31. Wiley Online Library, 2267–2276.

Kai Zhang, Fujun Luan, Qianqian Wang, Kavita Bala, and Noah Snavely. 2021. PhySG: Inverse Rendering with Spherical Gaussians for Physics-based Material Editing and Relighting. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).*

Lvmin Zhang and Maneesh Agrawala. 2023. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543* (2023).

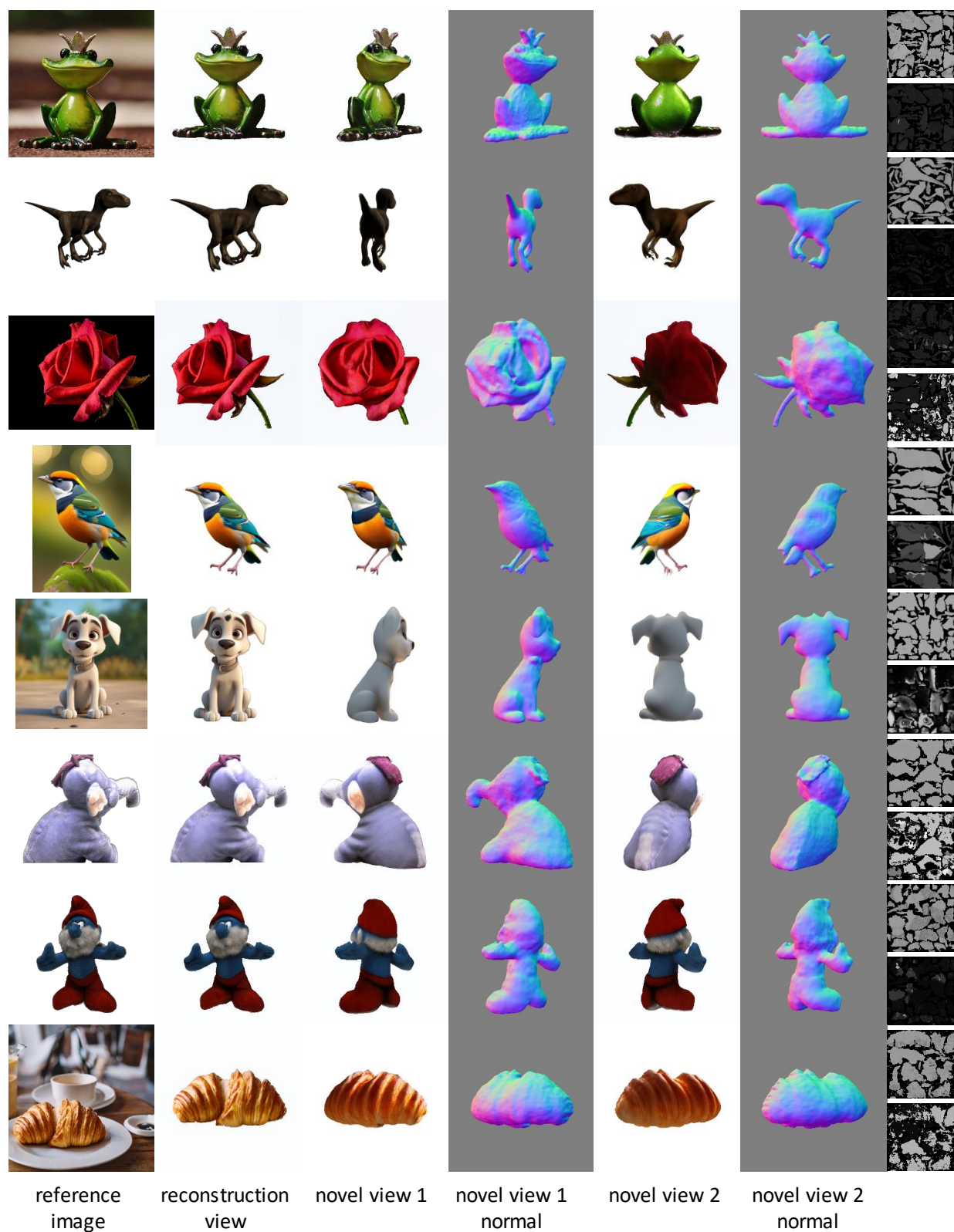| reference image | reconstruction view | novel view 1 | novel view 1 normal | novel view 2 | novel view 2 normal |

Figure 10: Additional results by HyperDreamer with more views. Images in the last column are specular and roughness map respectively (from top to bottom).