CustomTex: High-fidelity Indoor Scene Texturing via Multi-Reference Customization

Weilin Chen, Jiahao Rao, Wenhao Wang, Xinyang Li, Xuan Cheng*, Liujuan Cao
Key Laboratory of Multimedia Trusted Perception and Efficient Computing,
Ministry of Education of China, Xiamen University
https://chenweilinx.github.io/CustomTex/

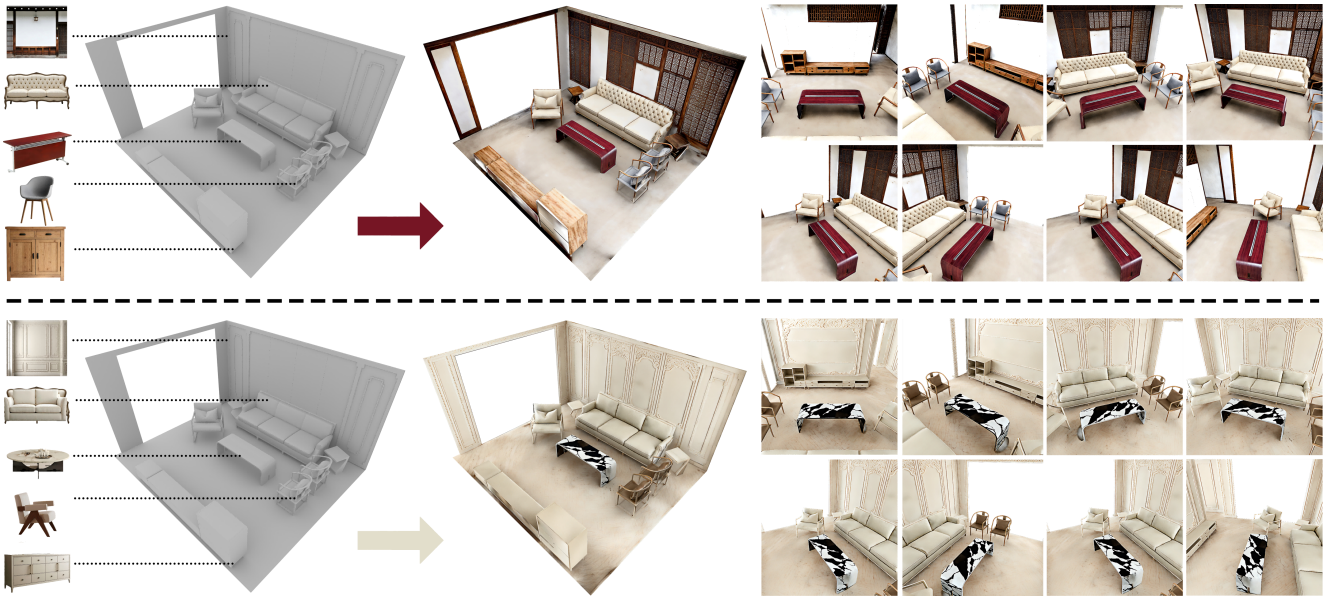


Figure 1. CustomTex is capable of generating high-fidelity texture for a 3D scene mesh, driven by instance-specific reference images.

Abstract

The creation of high-fidelity, customizable 3D indoor scene textures remains a significant challenge. While text-driven methods offer flexibility, they lack the precision for fine-grained, instance-level control, and often produce textures with insufficient quality, artifacts, and baked-in shading. To overcome these limitations, we introduce CustomTex, a novel framework for instance-level, high-fidelity scene texturing driven by reference images. CustomTex takes an untextured 3D scene and a set of reference images specifying the desired appearance for each object instance, and generates a unified, high-resolution texture map. The core of our method is a dual-distillation approach that separates semantic control from pixel-level enhancement. We employ semantic-level distillation, equipped with an instance cross-attention, to ensure semantic plausibility and "reference-instance" alignment, and pixel-level distillation to enforce high visual fidelity. Both are unified within a Variational Score Distillation (VSD) optimization framework. Experiments demonstrate that CustomTex achieves precise instance-level consistency with reference images and produces textures with superior sharpness, reduced artifacts, and minimal baked-in shading compared to state-of-the-art methods. Our work establishes a more direct and user-friendly path to high-quality, customizable 3D scene appearance editing.

1. Introduction

The creation of photorealistic and stylistically consistent 3D indoor scenes is a cornerstone of applications in virtual and augmented reality, architectural visualization, and film production. A critical factor in achieving this realism is texturing, which entails assigning surface materials and colors to 3D geometry. While recent advancements in 3D reconstruction, such as Neural Radiance Fields [49] and 3D Gaussian

*Corresponding author: Xuan Cheng (chengxuan@xmu.edu.cn)

Splatting [38], have made remarkable progress in capturing geometry and view-dependent appearance, they often produce "baked-in" textures that are entangled with lighting and lack material properties. Consequently, re-texturing these scenes with new, consistent and high-quality materials remains a significant challenge.

Most recent 3D indoor scene texturing methods [14, 35, 72] address this problem by leveraging the power of pre-trained text-to-image diffusion models [33, 59]. By using text prompts as guidance, these methods can, in principle, generate a vast array of materials and styles. However, this approach is often insufficient for indoor scene texturing, as text prompts alone are inherently ambiguous and struggle to convey precise visual characteristics. Using a reference image as prompt [36] can provide more direct visual guidance, such as the photographs of different home decoration styles. Yet, this approach typically offers only global, coarse-level control, still falling short of allowing users to specify fine-grained attributes like the specific weave of a fabric, the exact grain of wood, or the subtle pattern of a wallpaper.

Moreover, state-of-the-art 3D indoor scene texturing methods [14, 35, 72] often produce textures of insufficient quality for high-fidelity rendering. Their outputs lack the sharpness and richness found in artist-created or physically scanned textures, appearing instead soft, blurry or unnaturally uniform upon close inspection. This limitation arises from the entanglement of pixel-level fidelity and semantic-level perception in the diffusion process [63]. Additionally, these methods tend to produce textures with "baked-in" shading, as diffusion models learn and replicate the lighting and depth cues prevalent in the training datasets. The resulting textures that contain obvious highlights and shadows are not suitable for differently lighted renderings.

To overcome these limitations, we present **CustomTex**, a novel framework designed for instance-level controllable and high-fidelity texturing of 3D indoor scenes. As shown in Fig. 1, CustomTex generates a customized texture for a 3D scene mesh, based on a set of reference images that specify the desired appearance for each instance (e.g., sofa, cabinet, chair and walls) in the scene. The texture generated by CustomTex maintains instance-level consistency with the reference images, and yields a visually compelling appearance with significantly reduced blurriness, artifacts, and "baked-in" shading compared to existing SOTA methods. To this end, CustomTex separates semantic generation and pixel enhancement into two distinct distillation processes with two pre-trained stable diffusion models [63, 80]. The semantic-level distillation, equipped with an instance cross-attention mechanism, ensures sufficient semantic plausibility and enables precise instance-specific control over the generated textures. The pixel-level distillation focuses on boosting the visual fidelity and quality of the generated texture, while preserving the underlying structural and semantic information. Both processes are unified within a single optimization framework based on Variational Score Distillation (VSD) [66].

The technical contributions of this paper are summarized as follows:

- A novel texture generation framework that enables flexible, instance-level customization using multiple reference images as prompts.
- A dual-distillation training approach that effectively preserves semantic content while enhancing pixel-level fidelity.
- An instance-guided Variational Score Distillation approach that captures the multiple modes present in the reference imagery.

2. Related Work

2.1. Neural 3D Generation

Early research on 3D generation neural methods primarily relied on curated 3D datasets and explicit geometric representations, including voxels [16, 41, 60, 62, 68], point clouds [2, 19, 70, 81], meshes [21, 29, 43, 55, 71], and signed distance fields [17, 20, 22, 32, 53], which established foundational pipelines for learning 3D structures. However, the scarcity and high cost of high-quality 3D datasets limit current generative models, hindering their scalability, shape diversity, and visual fidelity due to limited scale and category coverage.

Driven by the progress in large-scale vision-language models [9–12, 34], recent methods treat 3D as a rendering target rather than a domain requiring explicit 3D supervision. By leveraging differentiable rendering and volumetric neural representations, these methods can propagate gradients directly from 2D observations [1, 7, 8, 51, 82], expanding the effective data regime to infer 3D structure from vast collections of unlabeled images [28]. However, these methods still faces challenges in reconstructing fine-grained geometry and maintaining high-frequency appearance details.

2.2. Feed-Forward Texturing

Texture synthesis aims to produce coherent and high-fidelity appearances on 3D surfaces, ensuring consistency in color [39, 46] and geometry cues [4, 30, 31] that reflect shading, material response, and view-dependent effects.

Unlike 2D texture transfer [23, 24], 3D texturing must maintain cross-view and geometric consistency over complex surfaces. With the advancement of learning-based 3D representations, UV-based networks [5, 18], convolution-based operations [3, 61], transformer-based methods [65, 69] and StyleGAN-inspired architectures [37, 42] have jointly modeled geometry and appearance, achieving more faithful texturing for 3D objects and small-scale scenes. Nevertheless, these methods often rely on carefully curated

textured datasets and generalize poorly to complex real-world settings, which exhibit diverse geometries, cluttered compositions, and intricate light–material interactions.

2.3. Diffusion-based Texturing

Diffusion models [33, 44, 59, 78, 79] have revolutionized generative modeling, providing powerful 2D priors that enable high-quality 3D texture generation for both objects [25, 66, 75] and scenes [35, 77] without large-scale 3D supervision, building upon their success in realistic image synthesis.

Existing diffusion-based 3D texturing frameworks can be broadly categorized into three major streams. (1) Inpainting-based methods, such as TEXTure [58], Text2Tex [13] and TexFusion [6], progressively generate or refine textures from each rendered view with depth-aware diffusion models [50, 80]. Then, they project and blend the results onto the 3D surface to form a global texture map progressively. While effective, this sequential process often introduces artifacts such as cross-view inconsistencies, seams, and texture drift. (2) Texture-space sampling methods like GenesisTex [27] and GenesisTex2 [47], address the coherence issue by performing diffusion directly on UV maps or latent textures, utilizing cross-view latent buffers and multiview score feedback. This strategy yields superior global consistency but introduces a dependence on high-quality UV parameterization and faces scalability issues with large or irregular assets. (3) Score Distillation Sampling (SDS) based methods enable direct 3D optimization through gradients from pre-trained diffusion models. Originally introduced by DreamFusion [56] for text-driven NeRF optimization, SDS was later extended to latent-space training in Latent-NeRF [48] and texture optimization in Latent-Paint. Subsequent works [15, 40, 74] incorporated geometry-aware constraints, enhanced sampling strategies, and accelerated convergence, evolving SDS from slow per-scene optimization toward more scalable and generalizable 3D generation frameworks.

2.4. 3D Indoor Scene Texturing

SceneTex [14] is a representative work in this field, which marks an important step toward realistic indoor environment texturing using VSD. RoomPainter [35] features a zero-shot technique that effectively adapts diffusion model for 3D-consistent texture synthesis. However, both methods rely on text prompts, which can't convey precise visual information, and offer only global, coarse control over the texture generation process. The only existing method for instance-level control is InstanceTex [72], which uses multiple text prompts to specify textures for different objects. Unfortunately, it shares the limitations of other text-driven methods and struggles to produce textures of sufficient quality.

Our method explores image-driven, instance-level texturing for indoor scenes. This paradigm provides fine-grained supervision, material-level alignment and richer style controllability, creating a more direct and user-friendly path to high-fidelity scene appearance editing.

3. Method

3.1. Overview

Our work aims to texture a complete 3D indoor scene from a collection of reference images. The framework takes two inputs: an untextured 3D scene composed of multiple object instances $\{\mathcal{O}_i\}_{i=1}^N$, and a set of reference images $\{\mathcal{I}_i^{ref}\}_{i=1}^N$ that define the desired appearance of each instance. We assume that the input 3D scene is unwrapped, where UV coordinates map each vertex of the mesh to a texel in a texture map. The main requirements for the output texture $\mathcal{T}$ are twofold: 1) each instance's texture must faithfully harmonize with the appearance of its assigned reference image; 2) the entire texture must be of high quality and form a stylistically coherent whole across all instances.

We formulate this texture synthesis task as an optimization problem in the UV space, and propose a dual-distillation training approach comprising semantic-level and pixel-level distillation. As shown in Fig. 2, the framework distills a pre-trained depth-to-image diffusion model [80] to provide a prior on semantic plausibility based on the reference images prompts. Concurrently, the framework also distills a pre-trained super-resolution diffusion model [63] to provide a prior on visual fidelity and quality of the rendered texture. The optimized texture is represented by an implicit multi-resolution texture field that encodes the texture features at different scales in the UV space.

3.2. Semantic-level Distillation

To ensure the generated texture for each instance semantically aligns with its reference image, we employ a semantic-level distillation process using instance-guided Variational Score Distillation (InsVSD). InsVSD conditions the distillation on instance-level images prompts, thus allowing the generated texture to learn a comprehensive distribution that encapsulates the multiple modes present in the reference imagery.

The 3D target mesh, with its optimized texture $\mathcal{T}$, is projected to a randomly sampled viewpoint via a differentiable rasterizer, rendering an RGB image I, a depth image d and a set of instance masks $\{m_i\}_{i=1}^N$ for different object instances. The depth image d is fed into the depth-to-image diffusion model [80], which is conditioned on the features $\{f_i^{ref}\}_{i=1}^N$ of reference images $\{\mathcal{I}_i^{ref}\}_{i=1}^N$ to provide semantic guidance. These features are extracted from the reference images by the image encoder in IP-Adapter [73]. To align each feature f_i^{ref} with its corresponding in-

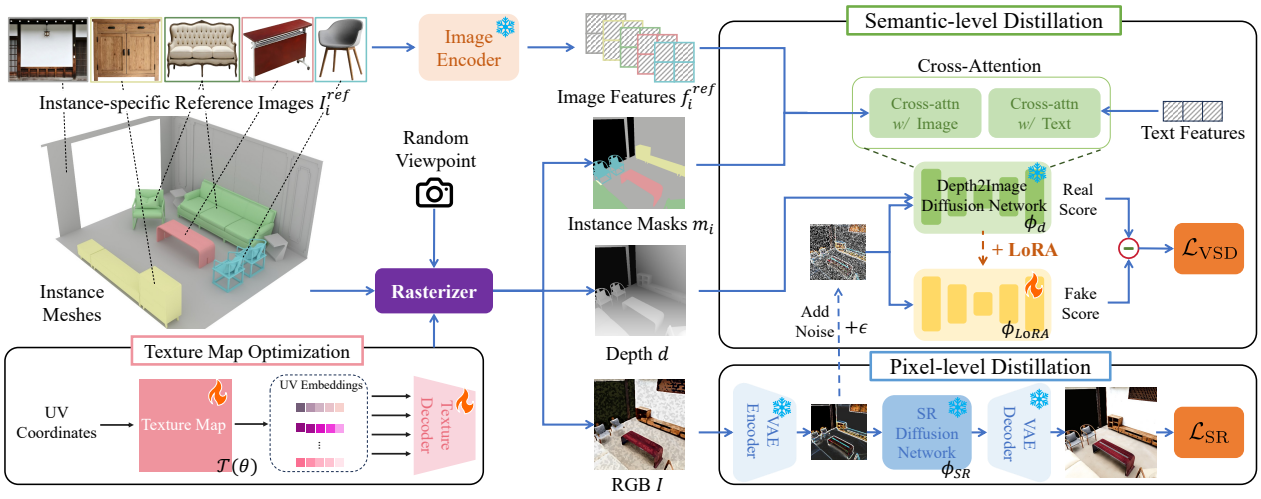


Figure 2. **Pipeline of CustomTex.** CustomTex textures a complete 3D indoor scene by optimizing a texture map in UV space through a dual-distillation training approach. In each iteration, the 3D scene with optimized texture is rendered from a random viewpoint, producing an RGB image, a depth map and instance masks. Instance masks are used to align each reference image's features with the correct object instance in the rendered RGB image via a specialized cross-attention. The Variational Score Distillation gradient and the Super-Resolution gradient are computed based on the well-aligned reference images condition to update the texture field.

stance's position on the rendered RGB image I, we use the instance mask m_i to adjust the computation of cross attention corresponding to f_i^{ref}. This process is formulated as:

$$Z' = \frac{1}{N}\sum_{i=1}^{N} m_i \cdot \text{Softmax}(\frac{\mathbf{Q}\mathbf{K}_i^\top}{\sqrt{d_k}})\mathbf{V}_i, \tag{1}$$

where $\mathbf{Q} = Z\mathbf{W}_q, \mathbf{K} = f_i^{ref}\mathbf{W}_k, \mathbf{V} = f_i^{ref}\mathbf{W}_v$ represent the queries, keys and values within the cross-attention module, Z, Z' denote the input and output features of the module, and $\mathbf{W}_q, \mathbf{W}_k, \mathbf{W}_v$ are the projection matrices used for linear transformations. As the mask m_i can accurately align each f_i^{ref} with specific positions of the rendered RGB image I, the appearance of the original reference image I_i^{ref} is precisely positioned on the relevant instance's parts in the texture UV map $\mathcal{T}$.

In the distillation, we transfer the semantic and stylistic content of each I_i^{ref} into the corresponding parts in $\mathcal{T}$. To achieve this, we train a lightweight and learnable LoRA model ϕ_{LoRA} to learn the prior from the frozen depth-to-image model ϕ_d. This prior defines the characteristics of a plausible rendered RGB image based on the provided image prompts I_i^{ref}. As both the parameters ϕ in the LoRA model ϕ_{LoRA} and the parameters θ in texture $\mathcal{T}(\theta)$ need to be optimized, we adopt an alternative two-step optimization strategy. Firstly, the LoRA model is frozen and θ is optimized via the VSD gradient:

$$\nabla_\theta \mathcal{L}_{\text{VSD}}(\theta, d, c^{ref}) = \mathbb{E}_{t,\epsilon}[w(t)(\epsilon_{\phi_d}(\mathcal{T}(\theta); d, c^{ref}, t) - \epsilon_{\phi_{LoRA}}(\mathcal{T}(\theta); d, c^{ref}, t))\frac{\partial \mathcal{T}(\theta)}{\partial \theta}], \tag{2}$$

where c^{ref} denotes the reference image prompts condition, t denotes the time step and $w(t)$ denotes the weighting function. After $\mathcal{T}(\theta)$ is updated via the VSD gradient, we unfreeze LoRA model and update ϕ with θ fixed. The training objective for the LoRA model is defined as:

$$\mathcal{L}_{\text{LoRA}}(\phi, d, c^{ref}) = \min_{\phi} \mathbb{E}_{t,\epsilon}[||\epsilon_{\phi_{LoRA}}(\mathcal{T}(\theta); d, c^{ref}, t) - \epsilon||_2^2]. \tag{3}$$

where the added noise is $\epsilon \sim N(0, 1)$.

3.3. Pixel-level Distillation

While the semantic guidance ensures the instance-level consistency with the reference images, achieving high-fidelity texture details is paramount for visual realism. To this end, our pixel-level distillation leverages a pre-trained image super-resolution model [63] to enhance the clarity and resolution of the synthesized texture. This model [63] achieves state-of-the-art super-resolution performance and is itself trained with VSD, allowing for the seamless integration of its components into our framework.

As shown in Fig. 2, the super-resolution (SR) model contains VAE encoder, SR diffusion network and VAE decoder, which are all frozen in the optimization. We denote the SR diffusion network as ϕ_{SR}. The rendered RGB image I is fed into the VAE encoder to generate latent feature I_{emb}. After being added with noise, I_{emb} is denoised by the depth-to-image diffusion network ϕ_d. Additionally, I_{emb} is also denoised by the SR diffusion network. The SR gradient

is defined as:

$$\nabla_\theta \mathcal{L}_{\text{SR}}(\theta, d, c^{ref}) = \mathbb{E}_{t,\epsilon}[w(t)(\epsilon_{\phi_{SR}}(\mathcal{T}(\theta); t) - \epsilon_{\phi_{LoRA}}(\mathcal{T}(\theta); d, c^{ref}, t))\frac{\partial \mathcal{T}(\theta)}{\partial \theta}]. \quad (4)$$

When optimizing the parameters θ in texture $\mathcal{T}(\theta)$, the SR gradient is combined with the VSD gradient defined in Eq. 2, and back-propagated together through the rasterizer to update θ. The final gradient is defined as:

$$\nabla_\theta \mathcal{L} = \nabla_\theta \mathcal{L}_{\text{VSD}}(\theta, d, c^{ref}) + \lambda_{SR} \nabla_\theta \mathcal{L}_{\text{SR}}(\theta, d, c^{ref}), \quad (5)$$

where λ_{SR} is the weighting parameter used to balance the two gradients in optimization. This gradient combination ensures the final texture possesses both correct large-scale structures and rich high-frequency details, yielding a sharp and visually compelling appearance.

3.4. Texture Representation

We use multi-resolution hash grid derived from Instant-NGP [52] to represent the implicit texture $\mathcal{T}(\theta)$. In this representation, $\mathcal{T}(\theta)$ first takes UV coordinates from the rasterized renderer, quantizing these coordinates into multi-scale grid levels through a hash mapping function. The feature values from all levels are then concatenated along the feature dimension to form high-dimensional UV embeddings. The UV embeddings are then decoded into the final RGB image I via a cross-attention texture decoder from [14]. The parameters θ encompass both the parameters in texture map and the parameters in texture decoder, all of which need to be optimized.

3.5. Implementation Details

Our training uses 5,000 spherically distributed viewpoints sampled within the scene space and runs for 30,000 iterations. The learning rate is set to 0.001 in the texture field updating and 0.0001 in the LoRA module fine-tuning. We employ a time annealing strategy: uniformly sample timesteps $t \sim U(0.02, 0.98)$ for the first 5,000 iterations and then gradually shift the sampling distribution to $t \sim U(0.02, 0.5)$. To ensure training stability, $\mathcal{L}_{\text{SR}}$ is set to 0 for the initial 5,000 iterations and then increased to 1.2 for the remainder of the training. The whole training takes approximately 48 hours on a single NVIDIA RTX A800 GPU. All generated textures have a resolution of $4,096 \times 4,096$. Our implementation is built upon the PyTorch, utilizing the Parameter-Efficient Fine-Tuning (PEFT) library for LoRA injection, and PyTorch3D for all differentiable rendering and texture projection.

4. Experiment

4.1. Experimental Setup

Baselines. We compare CustomTex against a comprehensive set of texture synthesis baselines to demonstrate its effectiveness. These baselines include image-to-texture methods like Paint3D [76], HY3D-2.1 [64] and SceneTex-IPA [14], and text-to-texture methods like TEXture [58], Paint3D [76], SyncMVD [45], HY3D-2.1 [64] and SceneTex [14]. Paint3D and HY3D-2.1 support both image and text prompts. As the original SceneTex is limited to text prompt, we adapt it by integrating an IP-Adapter [73] to accommodate image prompt and name this variant as SceneTex-IPA. Since all image-to-texture baselines don't support multiple image prompts, we stitch the reference images into a single large image to serve as the image prompt for these baselines. As for the text-to-texture comparison, we utilize GPT-4v to generate the set of reference images for our CustomTex from a fine-grained textual prompt.

Evaluation Metrics. We employ a multi-faceted evaluation protocol tailored to the two comparison scenarios. For comparison with the image-to-texture baselines, we adopt CLIP-Score (CLIP-I) [57] and CLIP-FID (a CLIP version of FID [54]) to quantify the distributional similarity between renderings of the textured meshes and the reference images. Furthermore, to incorporate a human-centric perspective, we leverage the latest LMM-based model Q-Align [67] for comprehensive Image Quality Assessment (Q-Align IQA) and Image Aesthetic Assessment (Q-Align IAA) of the final textured mesh renderings. For comparison with text-to-texture baselines, we follow SceneTex [14] that utilizes CLIP-Score (CLIP-T) and Inception Score (IS) [62] to measure semantic fidelity to the input prompts and generated texture quality respectively. We conduct quantitative experiments on the same evaluation dataset as SceneTex [14], which comprises 10 scenes sampled from 3D-FRONT [26].

4.2. Method Comparison

Quantitative Results. Our CustomTex demonstrates quantitative superiority in both image-to-texture and text-to-texture generation. As shown in Tab. 1, CustomTex achieves top performance across all four metrics in the image-to-texture task. Its leading CLIP-I and CLIP-FID scores indicate that the textures produced by CustomTex exhibit greater distributional similarity to the reference images, and its leading Q-Align IQA and IAA scores reflect its higher-quality texture generation. This superiority extends to the text-to-texture task. As evidenced by Tab. 2, CustomTex achieves the highest scores across all metrics, demonstrating strong semantic fidelity to text prompts (CLIP-T) while also attaining superior image quality (IS, Q-Align IQA and Q-Align IAA).

Figure 3. Qualitative comparison on image-to-texture generation. All generated textures are rendered by 3ds Max software at a resolution of 2000 $\times$ 2000 for visualization. CustomTex demonstrates instance-level consistency with the reference images, while also exhibiting greater sharpness with fewer shading effects and artifacts compared with the baselines.

Figure 4. Qualitative comparison on close-up texture renderings.

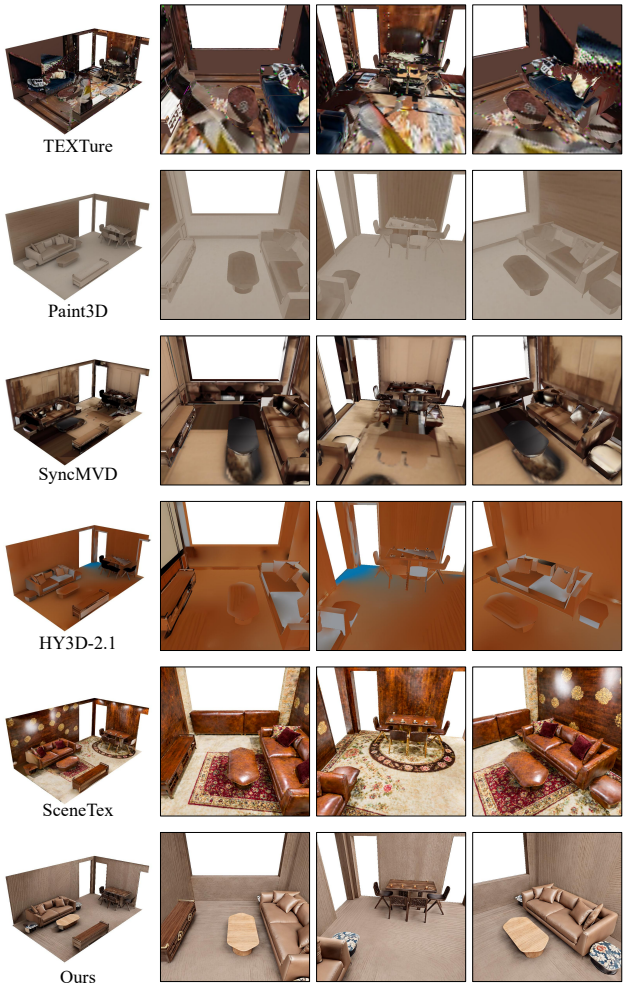


Figure 5. Qualitative comparison on text-to-texture generation. The text prompt is: "*The Nanyang vintage-style living room equipped with walls featuring dark wood panel textures, a brown leather sofa, a round fabric stool with floral patterns, a TV stand made of dark wood with golden handles, dark brown wooden chairs and a light-color wood coffee table.*" GPT-4v is used to convert this text prompt into reference image prompts for our CustomTex. All generated textures are rendered to 768 × 768 resolution images for visualization. Only CustomTex demonstrates instance-level consistency with the text prompt.

Method	CLIP-I↑	CLIP-FID↓	Q-Align IQA↑	Q-Align IAA↑
Paint3D [76]	0.694	130.138	2.896	2.401
HY3D-2.1 [64]	0.682	134.680	2.187	1.838
SceneTex-IPA [14]	0.741	121.118	4.009	3.594
CustomTex (Ours)	**0.797**	**106.229**	**4.469**	**3.629**

Table 1. Quantitative comparison on image-to-texture generation.

Method	CLIP-T↑	IS↑	Q-Align IQA↑	Q-Align IAA↑
TEXture [58]	0.557	1.372	1.645	1.574
Paint3D [76]	0.734	2.330	2.442	1.792
SyncMVD [45]	0.712	2.467	2.409	2.328
HY3D-2.1 [64]	0.734	2.381	2.774	2.033
SceneTex [14]	0.639	3.009	3.824	2.681
CustomTex (Ours)	**0.766**	**3.311**	**4.252**	**3.343**

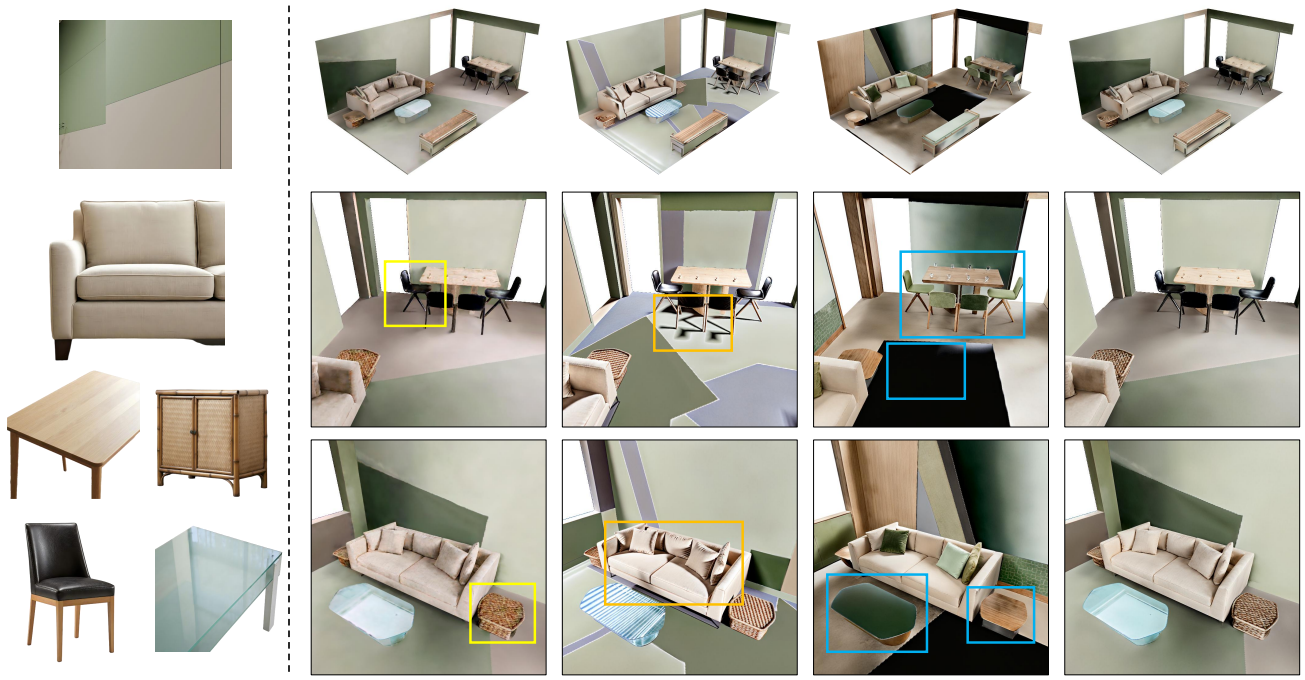
Table 2. Quantitative comparison on text-to-texture generation.

Qualitative Results. Fig. 3 presents the qualitative comparison on image-to-texture generation. Paint3D [76] and HY3D-2.1 [64] fail to correctly interpret stitched reference image prompts, resulting in repetitive patterns and an unnaturally uniform style. SceneTex-IPA [14] achieves greater similarity to the reference yet still deviates from it. In contrast, our CustomTex demonstrates precise instance-level consistency with the reference across objects like sofa, chair, cabinet, tea table and walls. Additionally, CustomTex produces textures with enhanced sharpness and a notable reduction in shading effects and artifacts compared to SceneTex-IPA [14]. Fig. 4 presents close-up visual comparisons of the rendered textures, highlighting differences in quality, sharpness and detail.

Fig. 5 presents the qualitative comparison on text-to-texture generation. The text prompt specifies fine-grained attributes for each object in the scene. However, TEXture [58], Paint3D [76], SyncMVD [45] and HY3D-2.1 [64] fail to correctly interpret this complex text prompt, resulting in visually irrational and low-quality textures. SceneTex [14] produces more globally reasonable and natural textures, but the textures of the walls and coffee table are not consistent with the text prompt, which specifies "walls featuring dark wood panel textures" and "light-color wood coffee table." CustomTex uses image prompt to produce textures with precise instance-level consistency to the complex text prompt. This comparison indicates that text prompt struggles to convey precise visual characteristics compared with image prompt.

4.3. Ablation Study

To comprehensively analyze the contribution of each module, we evaluate three variants of CustomTex: w/o $\mathcal{L}_{\text{SR}}$, w/o *feature-level mask* and w/o *multi-ref*. The qualitative and



Reference Images (a) *w/o* $\mathcal{L}_{SR}$ (b) *w/o feature-level mask* (c) *w/o multi-ref* (d) *full model*

Figure 6. Qualitative ablation study results.

Method	CLIP-I↑	CLIP-FID↓	Q-Align IQA↑	Q-Align IAA↑
w/o $\mathcal{L}_{\text{SR}}$	0.736	118.247	3.330	2.664
w/o f-mask	0.743	111.205	3.689	3.231
w/o multi-ref	0.757	109.243	4.053	3.519
full model	**0.797**	**106.229**	**4.469**	**3.629**

Table 3. Quantitative ablation study results. We validate the effectiveness of our method by testing three model variants: w/o $\mathcal{L}_{\text{SR}}$, w/o *feature-level mask* and w/o *multi-ref*, showcasing their distinct contributions.

quantitative ablation results are shown in Fig. 6 and Tab. 3.

Does Pixel-level Distillation truly improve texture quality? To answer this, we let λ_{SR} equal 0 in the Eq. 5 and conduct only semantic-level distillation, under the *w/o* $\mathcal{L}_{\text{SR}}$ setting. As shown in Tab. 3, this leads to a significant drop in two image quality assessment metrics (Q-Align IQA and IAA). Fig. 6 (a) further reveals that the results exhibit increased blurriness and artifacts, particularly along object boundaries. This confirms that pixel-level distillation provides crucial high-fidelity supervision, which helps preserve finer textural details and sharpness.

Is feature-level masking an effective way to fuse reference images and instance masks? In the *w/o feature-level mask* (f-mask) configuration, we apply instance masks at the noise-level instead of the feature-level of cross-attention layers, which can be represented by the formula: $\epsilon_{\phi_d} = \frac{1}{N}\sum_{i=1}^{N} m_i \epsilon_{\phi_d}(\mathcal{T}(\theta); d, c_i^{ref}, t)$, and the same operation is also applied to ϵ_{ϕ}. As shown in Fig. 6 (b), this causes unstable lighting around objects, indicating that feature-level masking better aligns the instance cues with the generated features while more effectively preserving lighting stability.

What are the effects of merging all reference images into a single large image instead of using multi-reference inputs? In the *w/o multi-ref* setting, we concatenate all reference images (e.g., sofa, bed, table, etc.) into one large composite input. As shown in Fig. 6 (c), this design makes it difficult for the model to distinguish between object instances, resulting in notable inconsistency between the reference images and the generated targets. These results highlight the necessity of maintaining multi-reference input for clear instance separation and fine-grained controllability.

5. Conclusion

In this paper, we present CustomTex, a novel framework for high-fidelity and instance-level controllable texturing of 3D indoor scenes. By providing a direct path from a set of reference images to a high-quality, unified texture map, CustomTex offers a more practical and user-friendly paradigm for 3D scene appearance editing. CustomTex has certain limitations we plan to address in future work. The training process to generate 4K-resolution textures is computationally intensive, typically taking several hours to complete. Furthermore, our method currently focuses on diffuse albedo texturing and does not generate other material maps, such as normal or roughness.

Acknowledgment

This work was partially supported by Natural Science Foundation of Fujian Province of China (No. 2023J05001) and National Science Fund for Distinguished Young Scholars (No. 62525605).

References

[1] Rameen Abdal, Hsin-Ying Lee, Peihao Zhu, Menglei Chai, Aliaksandr Siarohin, Peter Wonka, and Sergey Tulyakov. 3davatargan: Bridging domains for personalized editable avatars. In *Proc. of CVPR*, pages 4552–4562, 2023. 2

[2] Panos Achlioptas, Olga Diamanti, Ioannis Mitliagkas, and Leonidas J. Guibas. Learning representations and generative models for 3d point clouds. In *Proc. of ICLR*, 2018. 2

[3] Alexey Bokhovkin, Shubham Tulsiani, and Angela Dai. Mesh2tex: Generating mesh textures from image queries. In *Proc. of ICCV*, pages 8884–8894, 2023. 2

[4] Toby P Breckon and Robert B Fisher. A hierarchical extension to 3d non-parametric surface relief completion. *Pattern Recognit.*, 45:172–185, 2012. 2

[5] Haitao Cao, Baoping Cheng, Qiran Pu, Haocheng Zhang, Bin Luo, Yixiang Zhuang, Juncong Lin, Liyan Chen, and Xuan Cheng. DNPM: A neural parametric model for the synthesis of facial geometric details. In *Proc. of ICME*, pages 1–6, 2024. 2

[6] Tianshi Cao, Karsten Kreis, Sanja Fidler, Nicholas Sharp, and Kangxue Yin. Texfusion: Synthesizing 3d textures with text-guided image diffusion models. In *Proc. of ICCV*, pages 4146–4158, 2023. 3

[7] Eric R. Chan, Marco Monteiro, Petr Kellnhofer, Jiajun Wu, and Gordon Wetzstein. Pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In *Proc. of CVPR*, pages 5799–5809, 2021. 2

[8] Eric R. Chan, Connor Z. Lin, Matthew A. Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J. Guibas, Jonathan Tremblay, Sameh Khamis, Tero Karras, and Gordon Wetzstein. Efficient geometry-aware 3d generative adversarial networks. In *Proc. of CVPR*, pages 16102–16112, 2022. 2

[9] Dave Zhenyu Chen, Angel X. Chang, and Matthias Nießner. Scanrefer: 3d object localization in RGB-D scans using natural language. In *Proc. of ECCV*, pages 202–221, 2020. 2

[10] Dave Zhenyu Chen, Ali Gholami, Matthias Nießner, and Angel X. Chang. Scan2cap: Context-aware dense captioning in RGB-D scans. In *Proc. of CVPR*, pages 3193–3203, 2021.

[11] Dave Zhenyu Chen, Qirui Wu, Matthias Nießner, and Angel X. Chang. D^3net: A unified speaker-listener architecture for 3d dense captioning and visual grounding. In *Proc. of ECCV*, pages 487–505, 2022.

[12] Dave Zhenyu Chen, Ronghang Hu, Xinlei Chen, Matthias Nießner, and Angel X. Chang. Unit3d: A unified transformer for 3d dense captioning and visual grounding. In *Proc. of ICCV*, pages 18063–18073, 2023. 2

[13] Dave Zhenyu Chen, Yawar Siddiqui, Hsin-Ying Lee, Sergey Tulyakov, and Matthias Nießner. Text2tex: Text-driven texture synthesis via diffusion models. In *Proc. of ICCV*, pages 18512–18522, 2023. 3

[14] Dave Zhenyu Chen, Haoxuan Li, Hsin-Ying Lee, Sergey Tulyakov, and Matthias Nießner. Scenetex: High-quality texture synthesis for indoor scenes via diffusion priors. In *Proc. of CVPR*, pages 21081–21091, 2024. 2, 3, 5, 7

[15] Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. In *Proc. of ICCV*, pages 22189–22199, 2023. 3

[16] Wenzheng Chen, Huan Ling, Jun Gao, Edward J. Smith, Jaakko Lehtinen, Alec Jacobson, and Sanja Fidler. Learning to predict 3d objects with an interpolation-based differentiable renderer. In *Proc. of NeurIPS*, pages 9605–9616, 2019. 2

[17] Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. In *Proc. of CVPR*, pages 5939–5948, 2019. 2

[18] Zhiqin Chen, Kangxue Yin, and Sanja Fidler. Auv-net: Learning aligned uv maps for texture transfer and synthesis. In *Proc. of CVPR*, pages 1455–1464, 2022. 2

[19] Xuan Cheng, Ming Zeng, Jinpeng Lin, Zizhao Wu, and Xinguo Liu. Efficient l0 resampling of point sets. *Comput. Aided Geom. Des.*, 75, 2019. 2

[20] Yen-Chi Cheng, Hsin-Ying Lee, Sergey Tulyakov, Alexander G. Schwing, and Liangyan Gui. Sdfusion: Multimodal 3d shape completion, reconstruction, and generation. In *Proc. of CVPR*, pages 4456–4465, 2023. 2

[21] Yao Cheng, Weilin Chen, Yizhe Gu, Yue Sun, You Zhai, Xuan Cheng, and Juncong Lin. Size-aware indoor scene retargeting with generalized summarization. *Comput. Graph.*, 132:104315, 2025. 2

[22] Zezhou Cheng, Menglei Chai, Jian Ren, Hsin-Ying Lee, Kyle Olszewski, Zeng Huang, Subhransu Maji, and Sergey Tulyakov. Cross-modal 3d shape generation and manipulation. In *Proc. of ECCV*, pages 303–321, 2022. 2

[23] Jeremy S De Bonet. Multiresolution sampling procedure for analysis and synthesis of texture images. In *Proc. of SIGGRAPH*, pages 361–368, 1997. 2

[24] Alexei A Efros and Thomas K Leung. Texture synthesis by non-parametric sampling. In *Proc. of ICCV*, pages 1033–1038, 1999. 2

[25] Chaoran Feng, Wangbo Yu, Xinhua Cheng, Zhenyu Tang, Junwu Zhang, Li Yuan, and Yonghong Tian. Ae-nerf: Augmenting event-based neural radiance fields for non-ideal conditions and larger scenes. In *Proc. of AAAI*, pages 2924–2932, 2025. 3

[26] Huan Fu, Bowen Cai, Lin Gao, Lingxiao Zhang, Jiaming Wang, Cao Li, Qixun Zeng, Chengyue Sun, Rongfei Jia, Binqiang Zhao, and Hao Zhang. 3d-front: 3d furnished rooms with layouts and semantics. In *Proc. of ICCV*, pages 10913–10922, 2021. 5

[27] Chenjian Gao, Boyan Jiang, Xinghui Li, Yingpeng Zhang, and Qian Yu. Genesistex: adapting image denoising diffusion to texture space. In *Proc. of CVPR*, pages 4620–4629, 2024. 3

[28] Jun Gao, Tianchang Shen, Zian Wang, Wenzheng Chen, Kangxue Yin, Daiqing Li, Or Litany, Zan Gojcic, and Sanja Fidler. GET3D: A generative model of high quality 3d textured shapes learned from images. In *Proc. of NeurIPS*, 2022. 2

[29] Lin Gao, Jie Yang, Tong Wu, Yu-Jie Yuan, Hongbo Fu, Yu-Kun Lai, and Hao Zhang. SDM-NET: deep generative network for structured deformable mesh. *ACM Trans. Graph.*, 38:243:1–243:15, 2019. 2

[30] Aleksey Golovinskiy, Wojciech Matusik, Hanspeter Pfister, Szymon Rusinkiewicz, and Thomas Funkhouser. A statistical model for synthesis of detailed facial geometry. *ACM Trans. Graph.*, 25:1025–1034, 2006. 2

[31] Amir Hertz, Rana Hanocka, Raja Giryes, and Daniel Cohen-Or. Deep geometric texture synthesis. *ACM Trans. Graph.*, 39:108, 2020. 2

[32] Amir Hertz, Or Perel, Raja Giryes, Olga Sorkine-Hornung, and Daniel Cohen-Or. SPAGHETTI: editing implicit shapes through part aware generation. *ACM Trans. Graph.*, 41: 106:1–106:20, 2022. 2

[33] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Proc. of NeurIPS*, 2020. 2, 3

[34] Ronghang Hu and Amanpreet Singh. Unit: Multimodal multitask learning with a unified transformer. In *Proc. of ICCV*, pages 1419–1429. IEEE, 2021. 2

[35] Zhipeng Huang, Wangbo Yu, Xinhua Cheng, ChengShu Zhao, Yunyang Ge, Mingyi Guo, Li Yuan, and Yonghong Tian. Roompainter: View-integrated diffusion for consistent indoor scene texturing. In *Proc. of CVPR*, pages 574–584, 2025. 2, 3

[36] Dadong Jiang, Xianghui Yang, Zibo Zhao, Sheng Zhang, Jiaao Yu, Zeqiang Lai, Shaoxiong Yang, Chunchao Guo, Xiaobo Zhou, and Zhihui Ke. Flexitex: Enhancing texture generation via visual guidance. In *Proc. of AAAI*, pages 3967–3975, 2025. 2

[37] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proc. of CVPR*, pages 8107–8116, 2020. 2

[38] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139:1–139:14, 2023. 2

[39] Chengyang Li, Baoping Cheng, Yao Cheng, Haocheng Zhang, Renshuai Liu, Yinglin Zheng, Jing Liao, and Xuan Cheng. Facerefiner: High-fidelity facial texture refinement with differentiable rendering-based style transfer. *IEEE Trans. Multim.*, 26:7225–7236, 2024. 2

[40] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. In *Proc. of CVPR*, pages 300–309, 2023. 3

[41] Chieh Hubert Lin, Hsin-Ying Lee, Willi Menapace, Menglei Chai, Aliaksandr Siarohin, Ming-Hsuan Yang, and Sergey Tulyakov. Infinicity: Infinite-scale city synthesis. In *Proc. of ICCV*, pages 22751–22761, 2023. 2

[42] Renshuai Liu, Chengyang Li, Haitao Cao, Yinglin Zheng, Ming Zeng, and Xuan Cheng. EMEF: ensemble multi-exposure image fusion. In *Proc. of AAAI*, pages 1710–1718, 2023. 2

[43] Renshuai Liu, Yao Cheng, Sifei Huang, Chengyang Li, and Xuan Cheng. Transformer-based high-fidelity facial displacement completion for detailed 3d face reconstruction. *IEEE Trans. Multim.*, 26:799–810, 2024. 2

[44] Renshuai Liu, Bowen Ma, Wei Zhang, Zhipeng Hu, Changjie Fan, Tangjie Lv, Yu Ding, and Xuan Cheng. Towards a simultaneous and granular identity-expression control in personalized face generation. In *Proc. of CVPR*, pages 2114–2123, 2024. 3

[45] Yuxin Liu, Minshan Xie, Hanyuan Liu, and Tien-Tsin Wong. Text-guided texturing by synchronized multi-view diffusion. In *Proc. of SIGGRAPH Asia*, pages 60:1–60:11, 2024. 5, 7

[46] Jianye Lu, Athinodoros S. Georghiades, Andreas Glaser, Hongzhi Wu, Li-Yi Wei, Baining Guo, Julie Dorsey, and Holly E. Rushmeier. Context-aware textures. *ACM Trans. Graph.*, 26:3, 2007. 2

[47] Jiawei Lu, Yingpeng Zhang, Zengjun Zhao, He Wang, Kun Zhou, and Tianjia Shao. Genesistex2: Stable, consistent and high-quality text-to-texture generation. In *Proc. of AAAI*, pages 5820–5828, 2025. 3

[48] Gal Metzer, Elad Richardson, Or Patashnik, Raja Giryes, and Daniel Cohen-Or. Latent-nerf for shape-guided generation of 3d shapes and textures. In *Proc. of CVPR*, pages 12663–12673, 2023. 3

[49] 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. In *Proc. of ECCV*, pages 405–421, 2020. 1

[50] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proc. of AAAI*, number 5, pages 4296–4304, 2024. 3

[51] Norman Müller, Yawar Siddiqui, Lorenzo Porzi, Samuel Rota Bulò, Peter Kontschieder, and Matthias Nießner. Diffrf: Rendering-guided 3d radiance field diffusion. In *Proc. of CVPR*, pages 4328–4338, 2023. 2

[52] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Trans. Graph.*, 41(4):102:1–102:15, 2022. 5

[53] Jeong Joon Park, Peter R. Florence, Julian Straub, Richard A. Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proc. of CVPR*, pages 165–174, 2019. 2

[54] Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. On aliased resizing and surprising subtleties in gan evaluation. In *Proc. of CVPR*, pages 11400–11410, 2022. 5

[55] Dario Pavllo, Graham Spinks, Thomas Hofmann, Marie-Francine Moens, and Aurélien Lucchi. Convolutional generation of textured 3d meshes. In *Proc. of NeurIPS*, 2020. 2

[56] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *Proc. of ICLR*, 2023. 3

[57] 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 *Proc. of ICML*, pages 8748–8763, 2021. 5

[58] Elad Richardson, Gal Metzer, Yuval Alaluf, Raja Giryes, and Daniel Cohen-Or. Texture: Text-guided texturing of 3d shapes. In *Proc. of ACM SIGGRAPH*, pages 54:1–54:11, 2023. 3, 5, 7

[59] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proc. of CVPR*, pages 10674–10685, 2022. 2, 3

[60] Aliaksandr Siarohin, Willi Menapace, Ivan Skorokhodov, Kyle Olszewski, Jian Ren, Hsin-Ying Lee, Menglei Chai, and Sergey Tulyakov. Unsupervised volumetric animation. In *Proc. of CVPR*, pages 458–469, 2023. 2

[61] Yawar Siddiqui, Justus Thies, Fangchang Ma, Qi Shan, Matthias Nießner, and Angela Dai. Texturify: Generating textures on 3d shape surfaces. In *Proc. of ECCV*, pages 72–88, 2022. 2

[62] Edward J. Smith and David Meger. Improved adversarial systems for 3d object generation and reconstruction. In *Proc. of CORL*, pages 87–96, 2017. 2, 5

[63] Lingchen Sun, Rongyuan Wu, Zhiyuan Ma, Shuaizheng Liu, Qiaosi Yi, and Lei Zhang. Pixel-level and semantic-level adjustable super-resolution: A dual-lora approach. In *Proc. of CVPR*, pages 2333–2343, 2025. 2, 3, 4

[64] Tencent Hunyuan3D Team. Hunyuan3d 2.1: From images to high-fidelity 3d assets with production-ready pbr material. *arXiv preprint arXiv:2506.15442*, 2025. 5, 7

[65] Dmitry Tochilkin, David Pankratz, Zexiang Liu, Zixuan Huang, Adam Letts, Yangguang Li, Ding Liang, Christian Laforte, Varun Jampani, and Yan-Pei Cao. Triposr: Fast 3d object reconstruction from a single image. *arXiv preprint arXiv:2403.02151*, 2024. 2

[66] Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. In *Proc. of NeurIPS*, 2023. 2, 3

[67] Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, et al. Q-align: Teaching lmms for visual scoring via discrete text-defined levels. In *Proc. of ICML*, pages 54015–54029, 2024. 5

[68] Jianwen Xie, Zilong Zheng, Ruiqi Gao, Wenguan Wang, Song-Chun Zhu, and Ying Nian Wu. Learning descriptor networks for 3d shape synthesis and analysis. In *Proc. of CVPR*, pages 8629–8638, 2018. 2

[69] Yinghao Xu, Zifan Shi, Wang Yifan, Hansheng Chen, Ceyuan Yang, Sida Peng, Yujun Shen, and Gordon Wetzstein. Grm: Large gaussian reconstruction model for efficient 3d reconstruction and generation. In *Proc. of ECCV*, pages 1–20, 2024. 2

[70] Guandao Yang, Xun Huang, Zekun Hao, Ming-Yu Liu, Serge J. Belongie, and Bharath Hariharan. Pointflow: 3d point cloud generation with continuous normalizing flows. In *Proc. of ICCV*, pages 4540–4549, 2019. 2

[71] Jie Yang, Kaichun Mo, Yu-Kun Lai, Leonidas J. Guibas, and Lin Gao. Dsg-net: Learning disentangled structure and geometry for 3d shape generation. *ACM Trans. Graph.*, 42: 1:1–1:17, 2023. 2

[72] Mingxin Yang, Jianwei Guo, Yuzhi Chen, Lan Chen, Pu Li, Zhanglin Cheng, Xiaopeng Zhang, and Hui Huang. Instancetex: Instance-level controllable texture synthesis for 3d scenes via diffusion priors. In *Proc. of SIGGRAPH Asia*, pages 59:1–59:11, 2024. 2, 3

[73] Hu Ye, Jun Zhang, Siyi Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023. 3, 5

[74] Kim Youwang, Tae-Hyun Oh, and Gerard Pons-Moll. Paint-it: Text-to-texture synthesis via deep convolutional texture map optimization and physically-based rendering. In *Proc. of CVPR*, pages 4347–4356, 2024. 3

[75] Wangbo Yu, Chaoran Feng, Jianing Li, Jiye Tang, Jiashu Yang, Zhenyu Tang, Meng Cao, Xu Jia, Yuchao Yang, Li Yuan, et al. Evagaussians: Event stream assisted gaussian splatting from blurry images. In *Proc. of ICCV*, pages 24780–24790, 2025. 3

[76] Xianfang Zeng, Xin Chen, Zhongqi Qi, Wen Liu, Zibo Zhao, Zhibin Wang, Bin Fu, Yong Liu, and Gang Yu. Paint3d: Paint anything 3d with lighting-less texture diffusion models. In *Proc. of CVPR*, pages 4252–4262, 2024. 5, 7

[77] Jingbo Zhang, Xiaoyu Li, Ziyu Wan, Can Wang, and Jing Liao. Text2nerf: Text-driven 3d scene generation with neural radiance fields. *IEEE Trans. Vis. Comput. Graph.*, 30(12): 7749–7762, 2024. 3

[78] Jinlu Zhang, Yiyi Zhou, Qiancheng Zheng, Xiaoxiong Du, Gen Luo, Jun Peng, Xiaoshuai Sun, and Rongrong Ji. Fast text-to-3D-aware face generation and manipulation via direct cross-modal mapping and geometric regularization. In *Proc. of ICML*, pages 60605–60625, 2024. 3

[79] Jinlu Zhang, Jiji Tang, Rongsheng Zhang, Tangjie Lv, and Xiaoshuai Sun. Storyweaver: A unified world model for knowledge-enhanced story character customization. In *Proc. of AAAI*, pages 9951–9959, 2025. 3

[80] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proc. of ICCV*, pages 3813–3824, 2023. 2, 3

[81] Linqi Zhou, Yilun Du, and Jiajun Wu. 3d shape generation and completion through point-voxel diffusion. In *Proc. of ICCV*, pages 5806–5815, 2021. 2

[82] Yixiang Zhuang, Baoping Cheng, Yao Cheng, Yuntao Jin, Renshuai Liu, Chengyang Li, Xuan Cheng, Jing Liao, and Juncong Lin. Learn2talk: 3d talking face learns from 2d talking face. *IEEE Trans. Vis. Comput. Graph.*, 31(9):5829–5841, 2025. 2