

Voxify3D: Pixel Art Meets Volumetric Rendering

Yi-Chuan Huang Jiewen Chan Hao-Jen Chien Yu-Lun Liu

National Yang Ming Chiao Tung University



Figure 1. **Stylized voxel art with controllable abstraction.** Voxify3D converts 3D meshes into stylized voxel art using discrete color palettes, pixel art supervision, and voxel-based radiance fields. This teaser showcases the flexibility and quality of our method. (a) Diverse voxel art outputs across object types and use cases. (b) Comparison of different palette selection methods. (c) Control over the resolution of the voxel grid ($20\times$, $30\times$, $50\times$) allows a balance of detail and abstraction. (d) The variation in color count (2, 4, 8) shows the impact of palette size on expressiveness. (e) Input-output comparisons on multiple objects demonstrate faithful voxel structure, semantic clarity, and voxel art aesthetics.

Abstract

Voxel art is a distinctive stylization widely used in games and digital media, yet automated generation from 3D meshes remains challenging due to conflicting requirements of geometric abstraction, semantic preservation, and discrete color coherence. Existing methods either over-simplify geometry or fail to achieve the pixel-precise, palette-constrained aesthetics of voxel art. We introduce **Voxify3D**, a differentiable two-stage framework bridging 3D mesh optimization with 2D pixel art supervision. Our core innovation lies in the synergistic integration of three components: (1) **orthographic pixel art supervision** that eliminates perspective distortion for precise voxel-pixel alignment; (2) **patch-based CLIP alignment** that preserves semantics across discretization levels; (3)

palette-constrained Gumbel-Softmax quantization enabling differentiable optimization over discrete color spaces with controllable palette strategies. This integration addresses fundamental challenges: semantic preservation under extreme discretization, pixel-art aesthetics through volumetric rendering, and end-to-end discrete optimization. Experiments show superior performance (37.12 CLIP-IQA, 77.90% user preference) across diverse characters and controllable abstraction (2-8 colors, $20\times$ - $50\times$ resolutions). Project page: <https://yichuanh.github.io/Voxify-3D/>

1. Introduction

Voxel art is a distinctive form of 3D digital artwork, characterized by its minimalist aesthetic and discrete volumet-

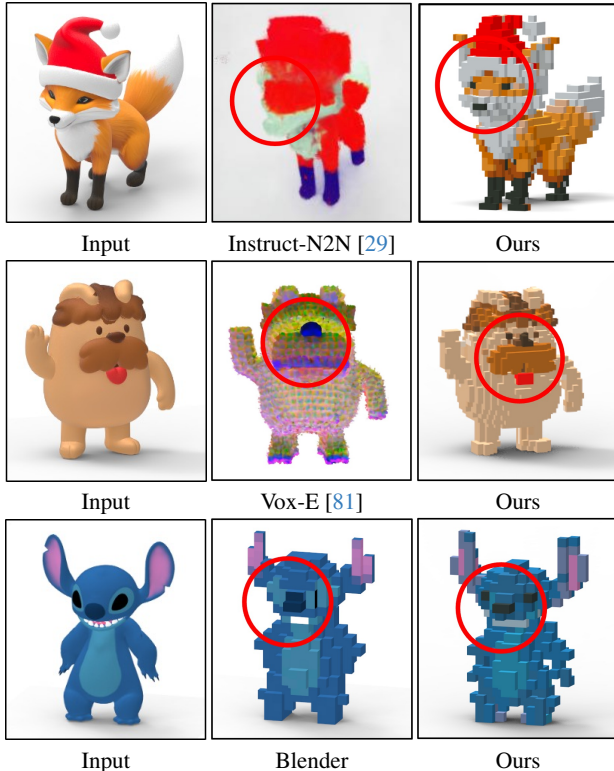


Figure 2. **Existing methods often miss key features in voxelization.** While IN2N [29], Vox-E [81], and Blender (Geometry Nodes) generate outputs that are coarse, blurry, or semantically inconsistent, they frequently lose critical elements such as facial features. In contrast, our method preserves structural details and produces visually appealing voxel art with sharp abstraction.

ric structure. Despite its growing popularity in games and digital media, creating high-quality voxel art remains challenging, requiring significant artistic expertise and manual effort. While recent works have achieved promising results in 2D pixel art stylization [4, 18, 28, 100], these techniques do not trivially extend to 3D voxel art. Directly using 2D pixel art for 3D reconstruction faces fundamental obstacles: projection-induced misalignment, multi-view inconsistencies, and ambiguous color representations.

Current voxel art generation from 3D meshes is limited. Simple downsampling loses semantic features, yielding overly coarse outputs. Voxel-based neural radiance fields [7, 22, 85] target photorealistic rendering, not stylistic abstraction. Neural editing methods [9, 29, 91] struggle with clean, discrete representations. Procedural tools like Blender’s Geometry Nodes require extensive manual tuning and lack unified optimization for discrete color control and semantic preservation, both critical for voxel art aesthetics. As Fig. 2 shows, existing methods miss key features.

Voxel art generation poses three interrelated challenges that cannot be addressed by naively combining existing tech-

niques: **(1) Alignment:** Perspective projection causes pixel-voxel misalignment, producing blurry gradients during optimization. Prior neural stylization [54, 93] uses perspective rendering, unsuited for discrete art styles. **(2) Semantic Preservation:** As resolution decreases, critical features (facial details, limb articulation) collapse. Standard perceptual losses on full images fail to capture local semantic importance. **(3) Discrete Optimization:** Voxel art requires small palettes (2-8 colors), but gradient-based methods produce continuous values. Existing quantization [20] lacks differentiability or user-controllable palette extraction.

We present **Voxify3D**, a principled framework addressing these challenges through synergistic technical design. We bridge 3D optimization with 2D pixel art supervision via: (1) six-view *orthographic* rendering that eliminates perspective distortion for precise alignment; (2) patch-based CLIP loss adapted to preserve semantics across discretization levels; (3) palette-constrained Gumbel-Softmax enabling differentiable discrete optimization with flexible extraction strategies. This integration requires careful synchronization of rendering strategy, loss formulation, and quantization timing, not a simple combination.

Our two-stage pipeline first initializes coarse voxel geometry and color via neural volume rendering, then refines using orthographic pixel art supervision with semantic and discrete color constraints. Our technical contributions include:

- **Orthographic pixel art supervision.** First framework to bridge 2D pixel art with 3D voxel optimization by eliminating perspective misalignment, enabling precise gradient flow for discrete stylization across six canonical views.
- **Resolution-adaptive semantic preservation.** Patch-based CLIP formulation maintaining object identity under extreme discretization (20x-50x), addressing semantic collapse that standard perceptual losses fail to prevent.
- **Palette-constrained differentiable quantization.** End-to-end optimization pipeline integrating Gumbel-Softmax with flexible palette extraction (4 strategies), temperature scheduling, and logit-based representation for controllable discrete color spaces (2-8 colors).

2. Related Work

3D Representations: From Pixels to Voxels. Pixel art generation evolved from interpolation [26] and content-aware downscaling [39, 42] to deep learning: paired [36] and unsupervised translation [28, 100], GANs [18, 82], diffusion [4], and vector methods [35, 37, 103]. For 3D, voxel-based methods accelerate neural fields [58, 64, 66] through explicit grids [7, 24, 69, 77, 80, 85, 106], differentiable voxelization [61], unified frameworks [99], hierarchical structures [78], sparse architectures [14], and compression [46, 110]. Multi-scale voxel representations [52], geometry-aware voxel features [88], tensorial decomposition [8], and MVS-based methods [84] further enhance

reconstruction quality. Voxels support geometry processing [17], storage [70], and simulation [60]. Recent feed-forward generation achieves scale through structured latents [101], hierarchical diffusion reaching 1024^3 [78], cascaded point clouds [109], transformers on voxelized shapes [67], and voxelized SDFs [47]. *Unlike* 2D stylization or 3D photorealism, we address *discrete, palette-constrained* voxel art by bridging pixel art supervision with volumetric optimization via orthographic alignment, extending voxel radiance fields [85] with palette quantization.

Stylization and Discrete Color Control. Neural 3D stylization progressed from score distillation [73] and CLIP guidance [65] to zero-shot transfer [54], painterly rendering [16, 23, 93, 112], high-resolution generation [12, 51], and *local* control [15, 19, 27, 56]. Gumbel-Softmax [38, 62, 83] enables discrete optimization in NAS [5, 53], VQ-VAE [86, 89], and neural fields [11, 57]. Score-based generative models [6] provide conditional generation through likelihood matching. Palette methods include 2D quantization [4], 3D color decomposition [43], material extraction [59], vector quantization [32], and interactive editing [44], with alternatives like VQGAN [20] and latent upsampling [63]. *In contrast to continuous* stylization and *fixed* codebooks, we integrate Gumbel-Softmax with *user-controllable* palette extraction (K-means, Max-Min, Median Cut, Simulated Annealing), synchronized scheduling, and logit-based representation for *pixel-precise* voxel art.

Multi-view Supervision and Semantic Preservation. Multi-view consistency uses RL refinement [102], view aggregation [84, 105], and latent diffusion [90]. Orthographic projection serves specialized domains: aerial orthophotos [13, 108], CAD reconstruction [113], and furniture assembly [31]. CLIP [76] enables semantic guidance [1, 10, 21, 41, 68, 71, 87, 92], with text supervision extending to semantic segmentation [97]. Semantic preservation under discretization uses masked autoencoders [49], context-aware transformers [107], semantic structures [50], geometry-aware downsampling [72], and hierarchical upsampling [79]. *Unlike* perspective stylization or orthographic reconstruction, we combine orthographic rendering with pixel art supervision, designing *resolution-adaptive patch-based* CLIP loss preventing semantic collapse at $20\times$ - $50\times$ discretization where image-level losses fail.

Applications and Datasets. Mesh generation exploits diffusion and sparse views [3, 30, 34, 48, 51, 55, 94, 96, 101, 104], with character datasets [95, 98]. Game assets require structural decomposition [33], PBR materials [111], and procedural libraries [40, 75]. Fabrication includes LEGO generation [25, 74], Earth voxelization [45], and 3D printing [2]. These inform our evaluation but don’t address *mesh-to-voxel-art* conversion with *semantic fidelity, palette constraints*, and *controllable* abstraction.

3. Method

We propose a two-stage framework for converting 3D meshes into stylized voxel art with high fidelity and semantic consistency (Fig. 3). Stage 1 (Sec. 3.1) builds a coarse voxel radiance field using DVGO [85] to establish geometric and color foundations. Stage 2 (Sec. 3.2) refines the grid under orthographic pixel-art supervision, with CLIP-based loss (Sec. 3.3) for semantic alignment and depth loss for geometric preservation. To achieve clean abstraction and a coherent palette, we replace the RGB grid with a learned color-logit grid and apply Gumbel-Softmax for differentiable palette quantization (Sec. 3.4). This pipeline retains abstract details, enforces a dominant palette, and conveys the distinctive style of voxel art across resolutions.

3.1. Coarse Voxel Grid Training

The first stage adapts DVGO [85] to build a coarse voxel representation. Unlike NeRFs using MLPs, DVGO directly optimizes two explicit voxel grids: a density grid d for spatial occupancy and a color grid $\mathbf{c} = (r, g, b)$ for appearance. This explicit structure enables faster training and efficient rendering.

We partition the object’s bounding box into a grid of resolution $(W/\text{cell_size})^3$, where W is the canonical orthographic image width (pixels) and cell_size is the number of pixels per voxel edge. Each voxel stores density d and RGB color \mathbf{c} . The rendered color $C(\mathbf{r})$ along a camera ray \mathbf{r} is computed as:

$$C(\mathbf{r}) = \sum_{k=1}^N T_k \alpha_k \mathbf{c}_k, \quad T_k = \exp\left(-\sum_{j=1}^{k-1} d_j \delta_j\right),$$

$$\alpha_k = 1 - \exp(-d_k \delta_k), \quad (1)$$

where N is the number of samples along the ray, d_k the density, δ_k the distance between consecutive samples, T_k the accumulated transmittance, and α_k the opacity at sample k .

The coarse voxel grid is optimized with:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{render}} + \lambda_d \mathcal{L}_{\text{density}} + \lambda_b \mathcal{L}_{\text{bg}}, \quad (2)$$

where $\mathcal{L}_{\text{render}}$ minimizes the MSE between rendered and target colors to ensure visual fidelity, $\mathcal{L}_{\text{density}}$ regularizes the density to suppress noise, prevent near-clip artifacts, and employs total variation (TV) regularization to enforce spatial smoothness, and \mathcal{L}_{bg} uses entropy loss to maintain clear geometry and reduce background artifacts. This stage provides a good initialization for color and density.

3.2. Orthographic Pixel Art Fine-tuning

To utilize the abstract features and clean edges of pixel art for 3D grid supervision, we fine-tune the voxel space by rendering orthographic projections from six axis-aligned views and

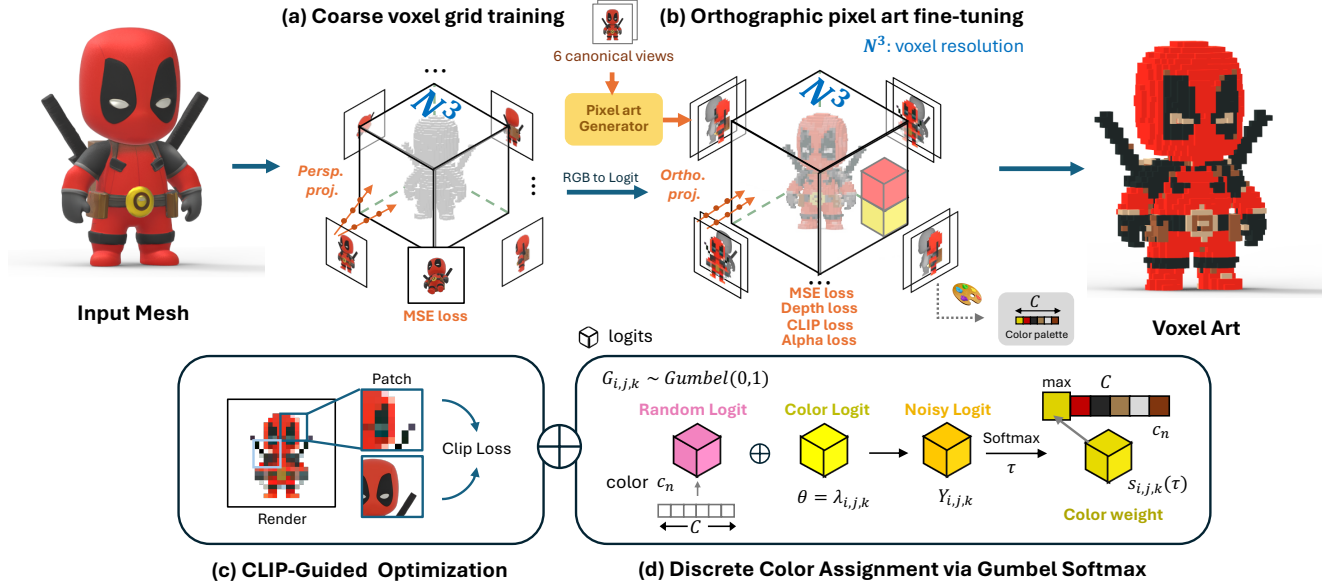


Figure 3. **Our two-stage voxel art generation pipeline.** (a) *Coarse voxel grid training*: Given a 3D mesh, we render multi-view images and optimize a voxel-based radiance field (DVGO [85]) using MSE loss to learn coarse RGB and density. (b) *Orthographic pixel art fine-tuning*: We refine the voxel grid using six orthographic pixel art views, which also serve to extract a discrete color palette (e.g., via k-means). Optimization includes appearance, depth, and alpha losses. (c) *CLIP-guided optimization*: A CLIP loss computed over rendered patches and mesh images encourages semantic alignment while being memory-efficient. (d) *Differentiable discrete color selection via Gumbel-Softmax*: Each voxel stores palette logits. Gumbel-Softmax enables differentiable sampling for end-to-end color optimization, yielding coherent, stylized voxel art.

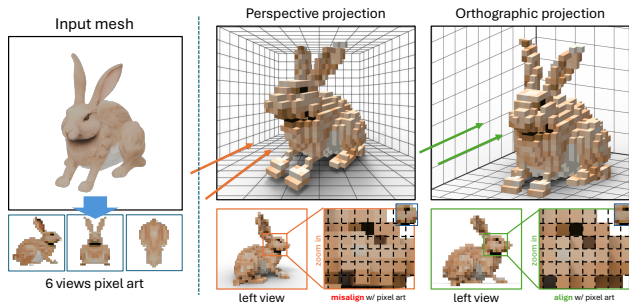


Figure 4. **Perspective vs. Orthographic.** (Left) Six-view pixel art pipeline. (Right) Perspective views (red) misalign pixels, while six orthographic views (green) enable precise pixel-voxel alignment.

comparing them against pixel art supervision generated by the pixel art generator [100]. This six-view setup compactly covers the major surfaces of the object, while orthographic rendering formulates parallel ray casting $\mathbf{r}_i(t) = \mathbf{o}_i + t\mathbf{d}$, where \mathbf{o}_i is the ray origin of pixel \mathbf{p}_i and \mathbf{d} is the fixed ray direction. All rays are parallel, ensuring pixel-to-voxel alignment without perspective distortions (Fig. 4).

We apply two foundational losses to supervise geometry and structure:

$$\mathcal{L}_{\text{pixel}} = \|\mathbf{C}(\mathbf{r}) - C_{\text{pixel}}\|_2^2 \quad (3)$$

$$\mathcal{L}_{\text{depth}} = \|\mathbf{D}(\mathbf{r}) - D_{\text{gt}}\|_1, \quad (4)$$

where $\mathbf{C}(\mathbf{r})$ and $\mathbf{D}(\mathbf{r})$ are the rendered color and depth along ray \mathbf{r} , C_{pixel} is the RGB color from the pixel-art supervision, and D_{gt} is the mesh-projected depth.

We also use an alpha loss to suppress density in background regions, enforcing background transparency to avoid floating density artifacts:

$$\mathcal{L}_{\alpha} = \|\mathcal{M}_{\alpha} \odot \bar{\alpha}\|^2, \quad (5)$$

where $\mathcal{M}_{\alpha} \in \{0, 1\}^{H \times W}$ is a binary mask from the pixel art alpha channel (1 for background), and $\bar{\alpha}$ denotes the accumulated ray opacity from volume rendering, which is encouraged to be 0 for background rays to allow full transparency. This encourages transparent regions in the pixel art to remain fully transmissive, preventing the formation of undesired voxels in areas without valid supervision.

By leveraging pixel art as the supervision signal, each voxel grid more effectively captures and expresses the most important structural and appearance information.

3.3. CLIP-based Semantic Loss

To incorporate semantic supervision, we sample half of the total rays to form patches for computing a CLIP-based perceptual loss. During training, we randomly sample patch rays

$(\mathbf{o}_{\text{patch}}, \mathbf{d}_{\text{patch}})$ from rendered images I_{mesh} of input mesh. Given the rendered patch \hat{I}_{patch} and the corresponding mesh-based patch $I_{\text{patch}}^{\text{mesh}}$, we extract CLIP features [21, 76] and compute a perceptual loss via cosine similarity:

$$\mathcal{L}_{\text{clip}} = 1 - \cos\left(\text{CLIP}(\hat{I}_{\text{patch}}), \text{CLIP}(I_{\text{patch}}^{\text{mesh}})\right), \quad (6)$$

where cosine similarity is defined as $\cos(a, b) = \frac{\langle a, b \rangle}{\|a\| \|b\|}$, and $\text{CLIP}(\cdot)$ denotes the CLIP image encoder output. This loss encourages voxel-rendered outputs to remain semantically aligned with the input mesh while supporting stylized abstraction, as illustrated in stage (c) of Fig. 3.

3.4. Discrete Color Selection via Gumbel-Softmax

To generate clean and stylized voxel appearances while allowing flexible color selection strategies, we adopt a palette-based quantization scheme where each voxel selects a color from a predefined palette. This palette is extracted from the six-view pixel art images using a chosen clustering method before Gumbel-Softmax quantization.

Instead of regressing RGB values, each voxel (i, j, k) stores a color logit vector $\lambda_{i,j,k} \in \mathbb{R}^C$, where C is the number of discrete colors in the predefined palette.

During training, Gumbel noise $\mathbf{G}_{i,j,k} \sim \text{Gumbel}(0, 1) \in \mathbb{R}^C$ is added to produce noisy logits:

$$\mathbf{Y}_{i,j,k} = \lambda_{i,j,k} + \mathbf{G}_{i,j,k}, \quad (7)$$

where $Y_{i,j,k,n}$ denotes the noisy logit for the n -th palette color at voxel (i, j, k) , with $n \in \{1, \dots, C\}$. A temperature-controlled softmax [38, 62] is then applied:

$$s_{i,j,k,n}(\tau) = \frac{\exp(Y_{i,j,k,n}/\tau)}{\sum_{n'=1}^C \exp(Y_{i,j,k,n'}/\tau)}, \quad (8)$$

where $s_{i,j,k,n}(\tau)$ is the probability of selecting the n -th color in the palette for voxel (i, j, k) , and τ is the temperature parameter controlling distribution sharpness.

In early training, we use the soft distribution $s_{i,j,k}$ directly. Later, we switch to the straight-through variant, where the forward pass uses a one-hot selection at $\arg \max_n s_{i,j,k}$, while gradients are backpropagated through the soft weights. We anneal the temperature τ during training to encourage smooth exploration in the early stages and sharper, more discrete selections later. The sampled RGB value is computed as:

$$\text{RGB}_{i,j,k} = \sum_{n=1}^C s_{i,j,k,n} \cdot \mathbf{c}_n, \quad (9)$$

where $\mathbf{c}_n \in \mathbb{R}^3$ is the n -th color in the palette.

After training, we directly select the color with the highest logit:

$$\text{RGB}_{i,j,k}^{\text{voxel}} = \mathbf{c}_{\arg \max_n \lambda_{i,j,k,n}}, \quad (10)$$

producing fully discrete voxel outputs. This process is illustrated in stage (d) of Fig. 3.

To enhance flexibility in stylization, this strategy allows users to choose the color selection method and number of colors, enabling explicit control over both color richness and overall style of the voxel art, making the design process more aligned with practical usage scenarios.

3.5. Loss Summary and Training Procedure

The overall loss optimized during fine-tuning is a weighted sum of multiple components that jointly supervise pixel-art faithfulness, geometry consistency, semantic alignment, and spatial regularity:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{pixel}} \cdot \mathcal{L}_{\text{pixel}} + \lambda_{\text{depth}} \cdot \mathcal{L}_{\text{depth}} + \lambda_{\text{alpha}} \cdot \mathcal{L}_{\text{alpha}} + \lambda_{\text{clip}} \cdot \mathcal{L}_{\text{clip}}, \quad (11)$$

where $\mathcal{L}_{\text{pixel}}$, $\mathcal{L}_{\text{depth}}$, and $\mathcal{L}_{\text{clip}}$ encourage pixel-level accuracy, depth consistency, and semantic alignment, respectively, while $\mathcal{L}_{\text{alpha}}$ suppresses background opacity to yield clean silhouettes. In Stage 2, rays are split into two groups: (1) $\mathcal{L}_{\text{pixel}}$, $\mathcal{L}_{\text{depth}}$, and $\mathcal{L}_{\text{alpha}}$, and (2) $\mathcal{L}_{\text{clip}}$ on rendered patches, all computed via volumetric rendering (Eq. (1)). Thus, geometric supervision of the density grid is provided by $\mathcal{L}_{\text{pixel}}$, $\mathcal{L}_{\text{depth}}$, and $\mathcal{L}_{\text{alpha}}$, while semantic supervision comes from $\mathcal{L}_{\text{clip}}$, which guides voxel appearance toward the intended pixel-art style.

4. Experiments

4.1. Experimental Setup

Dataset. We evaluate our method on three mesh datasets: **Rodin** [95], **Unique3D** [98], and **TRELLIS** [101]. Rodin and Unique3D primarily feature character 3D assets with rich semantic details, making them ideal for evaluating voxel abstraction and stylized representation. We also evaluate on diverse categories including architecture and vehicles; see supplementary material for details.

Implementation details. Training follows a two-stage schedule: **(a) Coarse Voxelization:** optimize the voxel grid for 8000 iterations to capture global structure; **(b) Pixel Art Supervision:** fine-tune for 6500 iterations with MSE, Depth, and CLIP losses on six orthographic views, rendered at a resolution of 1200×1200 , using fixed 80×80 patches randomly sampled each iteration for CLIP loss. In the final 2000 iterations, supervision is applied only to the front view to enhance key abstract features. Gumbel-Softmax sampling is performed over a fixed palette, with temperature τ annealed from 1.0 to 0.1.

Baseline methods. We compare against:

1. **Pixel art to 3D extension:** Render the input mesh into images, stylize them into pixel art, then train the original DVGO with these pixel-art images, using the coarse voxel grid as the final output.



Figure 5. **Qualitative comparisons on character models from the Rodin [95] dataset.** We compare our voxel art results with Pixel art to 3D extension, IN2N [29], Vox-E [81], and Blender’s voxelization. Our method produces stylized yet consistent voxel representations with pixel art aesthetics.

2. **IN2N [29]:** Language-guided mesh editing with view-consistent 3D stylization.
3. **Vox-E [81]:** Language-to-voxel generation prioritizing semantics over fine geometry.
4. **Blender Geometry Nodes:** Procedural mesh-to-voxel conversion, fast but without semantic or stylization control.

4.2. Qualitative Comparisons

We qualitatively compare our method with Pixel art to 3D extension, IN2N, Vox-E, and Blender on eight character meshes from the evaluation datasets (Fig. 5), with an additional eight groups of comparisons provided in the supplementary material.

IN2N preserves coarse structure but suffers from large variations across different guidance images, often failing to produce consistent voxelized results; Vox-E yields smoother volumes yet misses the discrete, blocky style of voxel art; Blender produces clean abstraction through procedural voxelization, which is akin to simple downsampling, but requires manual tuning and lacks semantic alignment.

Our method preserves key cues (e.g., ears, eyes) with sharp edges across $25\times$ – $50\times$ resolutions, achieving both expressive stylization and semantic fidelity. Additional results are provided in the supplementary material.

Table 1. **Average CLIP-IQA scores over all 35 examples.** Best scores are **highlighted**.

Method	Pixel	IN2N	Vox-E	Blender	Ours
CLIP-IQA	35.53	23.93	35.02	36.31	37.12

4.3. Quantitative Comparisons

To assess stylization fidelity and semantic preservation, we adopt the CLIP-IQA framework. For each character, we use GPT-4 to generate a detailed textual description based on the original mesh images, prepended with “A voxel art of...” (e.g., “A voxel art of a pink teddy bear with a red bow and heart-shaped feet”). We use OpenAI’s ViT-B/32 CLIP model and compute the average cosine similarity between each prompt and the rendered images from different methods.

As shown in Table 1, the reported CLIP-IQA scores are averaged over all 35 cases. Our method consistently achieves the highest score, demonstrating superior semantic alignment and stylized abstraction across a diverse set of character meshes.

4.4. Ablation Study

We analyze the contribution of each component by removing one module at a time. Qualitative results are shown in Fig. 6,

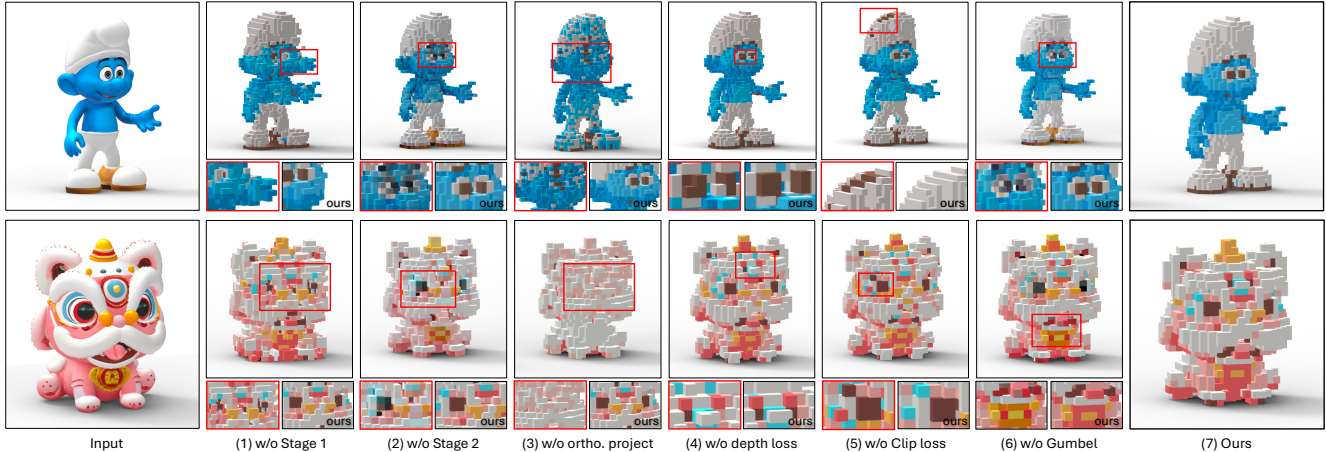


Figure 6. **Ablation study on key components.** We evaluate the impact of each module by removing it individually: (1) w/o Stage 1, (2) w/o Stage 2, (3) w/o orthographic projection (replaced with perspective projection), (4) w/o depth loss, (5) w/o CLIP loss, and (6) w/o Gumbel Softmax (resulting in continuous and unconstrained colors). Each row shows a different input, and zoom-in regions highlight local differences. Removing each component leads to noticeable degradation in geometry, color consistency, or semantic fidelity, while the full model produces coherent and stable voxel stylization.



Figure 7. **Effect of Palette Selection and Color Count.** Each row corresponds to a different palette extraction method: K-means, Max-Min, Median Cut, and Simulated Annealing. Each column shows increasing color counts (2, 3, 4, 8). Each method produces unique color clustering effects.

and quantitative results are reported in Table 2.

Stage-wise optimization. Without Stage 1, the model lacks proper geometric initialization, leading to distorted shapes. Without Stage 2, the result degenerates to a coarse DVGO grid without abstraction or semantic refinement. This shows

Table 2. **Ablation study of model components (CLIP-IQA).** We evaluate the impact of each component by removing it individually. Our full model achieves the best performance. Removing stage design or orthographic projection leads to significant degradation, while depth loss, CLIP loss, and Gumbel Softmax contribute to consistent improvements in semantic alignment.

Method	Stage 1	Stage 2	Ortho. Proj.	Depth loss	CLIP loss	Gumbel	CLIP-IQA
(1) w/o Stage 1							28.42
(2) w/o Stage 2	✓		✓	✓	✓	✓	34.32
(3) w/o ortho. proj.	✓	✓		✓	✓	✓	27.38
(4) w/o depth loss	✓	✓	✓		✓	✓	39.75
(5) w/o CLIP loss	✓	✓	✓	✓		✓	39.23
(6) w/o Gumbel	✓	✓	✓	✓	✓		39.31
(7) Ours	✓	✓	✓	✓	✓	✓	40.06

that Stage 1 provides geometric stability, while Stage 2 is the key component that enables semantic abstraction and stylized voxel representation.

Orthographic projection. Replacing orthographic projection with perspective projection causes severe color misalignment, as pixel colors no longer correspond to voxel locations, resulting in inconsistent appearance (Fig. 6).

Depth loss. Without depth loss, the model produces plausible views individually but fails to preserve global 3D structure, leading to geometric distortions.

CLIP loss. Removing CLIP loss reduces semantic clarity and leads to less coherent color regions, highlighting the importance of semantic alignment.

Gumbel-Softmax. Without Gumbel-Softmax, colors become mixed and lack clear boundaries, failing to produce clean and discrete voxel-style color patterns.

Summary. Fig. 6 and Table 2 show that all components contribute to geometry, semantic alignment, and discrete color



Figure 8. **Fabrication: LEGO render.** Rendered using KeyShot 2023. Our method extends to LEGO applications, where achieving rich visual results within the limited color palette is crucial for practical fabrication.

Table 3. **User studies.** (a) 35 examples (72 participants). (b) Color quantization (10 art-trained).

(a) Image quality (user votes, %)				(b) Color quantization preference (%)		
Metric	Abstract	Appeal	Geometry		w/o Gumbel	w/ Gumbel
Ours	77.90	80.36	96.55			
Others	22.10	19.64	3.45	Preferred	11.11	88.89

representation. Our full model achieves the best performance (40.06 CLIP-IQA). Removing any component consistently degrades quality. The quantitative results in Table 2 are averaged over five objects.

Palette controllability. We analyze palette design by varying the number of colors (2, 3, 4, 8) and extraction methods (K-means, K-means with rare color boosting, Median Cut, Max-Min, and Simulated Annealing).

The palette is extracted from all pixels across the six input pixel-art views to form a shared discrete color set for voxel optimization.

As shown in Fig. 7, increasing palette size improves color richness, while different strategies affect color distribution. K-means favors dominant tones, while the rare-color variant preserves infrequent details, and other methods (Median Cut, Max-Min, Simulated Annealing) promote balanced or diverse palettes.

Smaller palettes lead to stronger abstraction, whereas larger palettes improve detail. Our method maintains consistent voxel structure while enabling flexible control over color abstraction. Additional results and detailed palette selection strategies are provided in the supplementary material.

4.5. User Study

We conducted a user study with 72 participants to evaluate our method against four baselines: Pixel Art to 3D extension, IN2N [29], Vox-E [81], and Blender Geometry Nodes.

The study consists of two parts. In the first part (35 examples), participants evaluated *abstract detail* and *visual ap-*

peal. In the second part (geometry evaluation), participants compared grayscale renderings to assess shape preservation.

As shown in Table 3 (a), our method is consistently preferred, achieving 77.90%, 80.36%, and 96.55% in abstractness, appeal, and geometry, respectively.

Expert Study on Color Preference. We further conduct a focused evaluation on color quantization with 10 art-trained participants. As shown in Table 3 (b), 88.89% of participants prefer the results with Gumbel-Softmax, demonstrating its effectiveness in producing clearer structures and more distinct color regions.

More details on study design, questions, and additional results are provided in the supplementary material.

5. Conclusion

We introduce **Voxify3D**, a novel framework for transforming 3D meshes into stylized voxel art with strong semantic abstraction and structural consistency. By combining coarse voxel optimization, orthographic pixel art supervision, and palette-based color quantization, our method achieves expressive and visually appealing results across a variety of character assets. Extensive experiments and user studies confirm its advantages over existing baselines in both geometric faithfulness and artistic stylization.

In addition to digital results, we demonstrate the fabrication potential of our voxel outputs via LEGO-style assemblies (Fig. 8), enabled by their discrete structure and limited color palette.

Limitations and Future Work. Voxify3D struggles with highly intricate shapes, where thin structures or fine facial details may be lost at low voxel resolutions. Future work may explore integrating geometric priors or training strategies to enhance detail preservation and scalability, as well as adopting assembly-aware fabrication strategies inspired by LEGO brick design and connection principles to improve the physical realizability of voxel-based models.

Acknowledgements. This research was funded by the National Science and Technology Council, Taiwan, under Grants NSTC 112-2222-E-A49-004-MY2 and 113-2628-E-A49-023-. The authors are grateful to Google, NVIDIA, and MediaTek Inc. for their generous donations. Yu-Lun Liu acknowledges the Yushan Young Fellow Program by the MOE in Taiwan.

References

- [1] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3
- [2] Christoph Bader, Dominik Kolb, James C Weaver, Sunanda Sharma, Ahmed Hosny, João Costa, and Neri Oxman. Making data matter: Voxel printing for the digital fabrication of data across scales and domains. *Science advances*, 4(5): eaas8652, 2018. 3
- [3] Maciej Bala, Yin Cui, Yifan Ding, Yunhao Ge, Zekun Hao, Jon Hasselgren, Jacob Huffman, Jingyi Jin, J.P. Lewis, Zhaoshuo Li, Chen-Hsuan Lin, Yen-Chen Lin, Tsung-Yi Lin, Ming-Yu Liu, Alice Luo, Qianli Ma, Jacob Munkberg, Stella Shi, Fangyin Wei, Donglai Xiang, Jiashu Xu, Xiaohui Zeng, and Qinsheng Zhang. Edify 3d: Scalable high-quality 3d asset generation. *arXiv preprint arXiv:2411.07135*, 2024. 3
- [4] Alexandre Binniger and Olga Sorkine-Hornung. Sd- π xl: Generating low-resolution quantized imagery via score distillation. In *SIGGRAPH Asia Conference Papers*, pages 1–12, 2024. 2, 3
- [5] Han Cai, Ligeng Zhu, and Song Han. Proxylensnas: Direct neural architecture search on target task and hardware. In *International Conference on Learning Representations (ICLR)*, 2019. 3
- [6] Chen-Hao Chao, Wei-Fang Sun, Bo-Wun Cheng, Yi-Chen Lo, Chia-Che Chang, Yu-Lun Liu, Yu-Lin Chang, Chia-Ping Chen, and Chun-Yi Lee. Denoising likelihood score matching for conditional score-based data generation. *arXiv preprint arXiv:2203.14206*, 2022. 3
- [7] Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. Tensorf: Tensorial radiance fields. In *European conference on computer vision*, pages 333–350. Springer, 2022. 2
- [8] Bo-Yu Chen, Wei-Chen Chiu, and Yu-Lun Liu. Improving robustness for joint optimization of camera pose and decomposed low-rank tensorial radiance fields. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 990–1000, 2024. 2
- [9] Jun-Kun Chen, Jipeng Lyu, and Yu-Xiong Wang. Neuraleditor: Editing neural radiance fields via manipulating point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12300–12310, 2023. 2
- [10] Lianggangxu Chen, Xuejiao Wang, Jiale Lu, Shaohui Lin, Changbo Wang, and Gaoqi He. Clip-driven open-vocabulary 3d scene graph generation via cross-modality contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27863–27873, 2024. 3
- [11] Lei Chen, Yuan Meng, Chen Tang, Xinzhu Ma, Jingyan Jiang, Xin Wang, Zhi Wang, and Wenwu Zhu. Q-dit: Accurate post-training quantization for diffusion transformers. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 28306–28315, 2025. 3
- [12] Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 22246–22256, 2023. 3
- [13] Shihan Chen, Qingsong Yan, Yingjie Qu, Wang Gao, Junxing Yang, and Fei Deng. Ortho-nerf: generating a true digital orthophoto map using the neural radiance field from unmanned aerial vehicle images. *Geomatics, Natural Hazards and Risk*, 15(1):741–760, 2024. 3
- [14] Yukang Chen, Jianhui Liu, Xiangyu Zhang, Xiaojuan Qi, and Jiaya Jia. Voxelnex: Fully sparse voxelnet for 3d object detection and tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 21674–21683, 2023. 2
- [15] Yingshu Chen, Huajian Huang, Tuan-Anh Vu, Ka Chun Shum, and Sai-Kit Yeung. Stylecity: Large-scale 3d urban scenes stylization. In *European conference on computer vision*, pages 395–413. Springer, 2024. 3
- [16] SeungJeh Chung, JooHyun Park, and HyeonYeop Kang. 3dstyleglip: Part-tailored text-guided 3d neural stylization. *arXiv preprint arXiv:2404.02634*, 2024. 3
- [17] David Coeurjolly, Pierre Gueth, and Jacques-Olivier Lachaud. Regularization of voxel art. In *ACM SIGGRAPH 2018 Talks*, pages 1–2. 2018. 3
- [18] Flávio Coutinho and Luiz Chaimowicz. Generating pixel art character sprites using gans. *arXiv preprint arXiv:2208.06413*, 2022. 2
- [19] Dale Decatur, Itai Lang, Kfir Aberman, and Rana Hanocka. 3d paintbrush: Local stylization of 3d shapes with cascaded score distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4473–4483, 2024. 3
- [20] Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12873–12883, 2021. 2, 3
- [21] Kevin Frans, Lisa Soros, and Olaf Witkowski. Clipdraw: Exploring text-to-drawing synthesis through language-image encoders. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 3, 5
- [22] Sara Fridovich-Keil, Alex Yu, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance fields without neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5501–5510, 2022. 2
- [23] Haruo Fujiwara, Yusuke Mukuta, and Tatsuya Harada. Style-nerf2nerf: 3d style transfer from style-aligned multi-view

- images. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–10, 2024. 3
- [24] Stephan J. Garbin, Marek Kowalski, Matthew Johnson, Jamie Shotton, and Julien Valentin. Fastnerf: High-fidelity neural rendering at 200fps. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 2
- [25] Jiahao Ge, Mingjun Zhou, and Chi-wing Fu. Learn to create simple lego micro buildings. *ACM Transactions on Graphics (TOG)*, 43(6):1–13, 2024. 3
- [26] Timothy Gerstner, Doug DeCarlo, Marc Alexa, Adam Finkelstein, Yotam Gingold, and Andrew Nealen. Pixelated image abstraction with integrated user constraints. *Computers & Graphics*, 37(5):333–347, 2013. 2
- [27] Guilherme Gomes Haetinge, Jingwei Tang, Raphael Ortiz, Paul Kanyuk, and Vinicius Azevedo. Controllable neural style transfer for dynamic meshes. In *Acm siggraph 2024 conference papers*, pages 1–11, 2024. 3
- [28] Chu Han, Qiang Wen, Shengfeng He, Qianshu Zhu, Yinjie Tan, Guoqiang Han, and Tien-Tsin Wong. Deep unsupervised pixelization. *ACM Transactions on Graphics (TOG)*, 37(6):243:1–243:11, 2018. 2
- [29] Ayaan Haque, Matthew Tancik, Alexei A. Efros, Aleksander Holynski, and Angjoo Kanazawa. Instruct-nerf2nerf: Editing 3d scenes with instructions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8874–8884, 2023. 2, 6, 8
- [30] Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. Lrm: Large reconstruction model for single image to 3d. *arXiv preprint arXiv:2311.04400*, 2023. 3
- [31] Wentao Hu, Jia Zheng, Zixin Zhang, Xiaojun Yuan, Jian Yin, and Zihan Zhou. Plankassembly: Robust 3d reconstruction from three orthographic views with learnt shape programs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18495–18505, 2023. 3
- [32] Siyu Huang, Jie An, Donglai Wei, Jiebo Luo, and Hanspeter Pfister. Quantart: Quantizing image style transfer towards high visual fidelity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5947–5956, 2023. 3
- [33] Yuhang Huang, Shilong Zou, Xinwang Liu, and Kai Xu. Part-aware shape generation with latent 3d diffusion of neural voxel fields. *IEEE Transactions on Visualization and Computer Graphics*, 2025. 3
- [34] Zixuan Huang, Mark Boss, Aaryaman Vasishtha, James M Rehg, and Varun Jampani. Spar3d: Stable point-aware reconstruction of 3d objects from single images. *arXiv preprint arXiv:2501.04689*, 2025. 3
- [35] Yuki Igarashi and Takeo Igarashi. Pixel art adaptation for handicraft fabrication. *Computer Graphics Forum (Pacific Graphics 2022)*, 41(7):489–494, 2022. 2
- [36] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, pages 1125–1134, 2017. 2
- [37] Ajay Jain, Amber Xie, and Pieter Abbeel. Vectorfusion: Text-to-svg by abstracting pixel-based diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2
- [38] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. In *Proceedings of the 5th International Conference on Learning Representations (ICLR)*, 2017. 3, 5
- [39] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. *arXiv preprint arXiv:1603.08155*, 2016. 2
- [40] Bumsoo Kim, Sanghyun Byun, Yonghoon Jung, Wonseop Shin, Sareer UI Amin, and Sanghyun Seo. Minecraftify: Minecraft style image generation with text-guided image editing for in-game application. *arXiv preprint arXiv:2402.05448*, 2024. 3
- [41] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3
- [42] Johannes Kopf, Ariel Shamir, and Pieter Peers. Content-adaptive image downscaling. *ACM Transactions on Graphics (TOG)*, 32(6):1–8, 2013. 2
- [43] Zhengfei Kuang, Fujun Luan, Sai Bi, Zhixin Shu, Gordon Wetzstein, and Kalyan Sunkavalli. Palettenerf: Palette-based appearance editing of neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20691–20700, 2023. 3
- [44] Jae-Hyeok Lee and Dae-Shik Kim. Ice-nerf: Interactive color editing of nerfs via decomposition-aware weight optimization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3491–3501, 2023. 3
- [45] Ryan Hardesty Lewis. Voxelizing google earth: A pipeline for new virtual worlds. In *ACM SIGGRAPH 2024 Labs*, pages 1–2, 2024. 3
- [46] Lingzhi Li, Zhen Shen, Zhongshu Wang, Li Shen, and Liefeng Bo. Compressing volumetric radiance fields to 1 mb. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4222–4231, 2023. 2
- [47] Muheng Li, Yueqi Duan, Jie Zhou, and Jiwen Lu. Diffusion-sdf: Text-to-shape via voxelized diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12642–12651, 2023. 3
- [48] Ming-Feng Li, Yueh-Feng Ku, Hong-Xuan Yen, Chi Liu, Yu-Lun Liu, Albert YC Chen, Cheng-Hao Kuo, and Min Sun. Genrc: Generative 3d room completion from sparse image collections. In *European Conference on Computer Vision*, pages 146–163. Springer, 2024. 3
- [49] Yiming Li, Zhiding Yu, Christopher Choy, Chaowei Xiao, Jose M Alvarez, Sanja Fidler, Chen Feng, and Anima Anandkumar. Voxformer: Sparse voxel transformer for camera-based 3d semantic scene completion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9087–9098, 2023. 3
- [50] Yuan Li, Zhihao Liu, Bedrich Benes, Xiaopeng Zhang, and Jianwei Guo. Svdtree: Semantic voxel diffusion for single image tree reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4692–4702, 2024. 3

- [51] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 300–309, 2023. 3
- [52] Chin-Yang Lin, Chung-Ho Wu, Chang-Han Yeh, Shih-Han Yen, Cheng Sun, and Yu-Lun Liu. Frugalnerf: Fast convergence for extreme few-shot novel view synthesis without learned priors. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 11227–11238, 2025. 2
- [53] Hanxiao Liu, Karen Simonyan, and Yiming Yang. Darts: Differentiable architecture search. In *International Conference on Learning Representations (ICLR)*, 2019. 3
- [54] Kunhao Liu, Fangneng Zhan, Yiwen Chen, Jiahui Zhang, Yingchen Yu, Abdulmotaleb El Saddik, Shijian Lu, and Eric P Xing. Stylerf: Zero-shot 3d style transfer of neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8338–8348, 2023. 2, 3
- [55] Minghua Liu, Chao Xu, Haian Jin, Linghao Chen, Mukund Varma T, Zexiang Xu, and Hao Su. One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization. *Advances in Neural Information Processing Systems*, 36, 2024. 3
- [56] Richard Liu, Daniel Fu, Noah Tan, Itai Lang, and Rana Hanocka. Wir3d: Visually-informed and geometry-aware 3d shape abstraction. *arXiv preprint arXiv:2505.04813*, 2025. 3
- [57] Weihang Liu, Xue Xian Zheng, Jingyi Yu, and Xin Lou. Content-aware radiance fields: Aligning model complexity with scene intricacy through learned bitwidth quantization. In *European Conference on Computer Vision (ECCV)*, pages 239–256. Springer, 2024. 3
- [58] Yu-Lun Liu, Chen Gao, Andreas Meuleman, Hung-Yu Tseng, Ayush Saraf, Changil Kim, Yung-Yu Chuang, Johannes Kopf, and Jia-Bin Huang. Robust dynamic radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13–23, 2023. 2
- [59] Ivan Lopes, Fabio Pizzati, and Raoul de Charette. Material palette: Extraction of materials from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4379–4388, 2024. 3
- [60] Frank Losasso, Frédéric Gibou, and Ronald Fedkiw. Simulating water and smoke with an octree data structure. *ACM Transactions on Graphics (TOG)*, 23(3):457–462, 2004. 3
- [61] Yihao Luo, Yikai Wang, Zhengrui Xiang, Yuliang Xiu, Guang Yang, and ChoonHwai Yap. Differentiable voxelization and mesh morphing. *arXiv preprint arXiv:2407.11272*, 2024. 2
- [62] Chris J Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. In *Proceedings of the 5th International Conference on Learning Representations (ICLR)*, 2017. 3, 5
- [63] Sachit Menon, Alexandru Damian, Shijia Hu, Nikhil Ravi, and Cynthia Rudin. Pulse: Self-supervised photo upsampling via latent space exploration of generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3
- [64] Andreas Meuleman, Yu-Lun Liu, Chen Gao, Jia-Bin Huang, Changil Kim, Min H Kim, and Johannes Kopf. Progressively optimized local radiance fields for robust view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16539–16548, 2023. 2
- [65] Oscar Michel, Roi Bar-On, Richard Liu, Sagie Benaim, and Rana Hanocka. Text2mesh: Text-driven neural stylization for meshes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13492–13502, 2022. 3
- [66] 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
- [67] Shentong Mo, Enze Xie, Ruihang Chu, Lanqing Hong, Matthias Niessner, and Zhenguo Li. Dit-3d: Exploring plain diffusion transformers for 3d shape generation. *Advances in neural information processing systems*, 36:67960–67971, 2023. 3
- [68] Ron Mokady, Amir Hertz, and Amit H Bermano. Clipcap: Clip prefix for image captioning. In *European Conference on Computer Vision (ECCV)*, pages 531–547. Springer, 2022. 3
- [69] 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. 2
- [70] Ken Museth. Vdb: High-resolution sparse volumes with dynamic topology. *ACM Transactions on Graphics (TOG)*, 32(3):27:1–27:22, 2013. 3
- [71] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. *arXiv preprint arXiv:2103.17249*, 2021. 3
- [72] Sai Karthikey Pentapati, Anshul Rai, Arkady Ten, Chaitanya Atluru, and Alan Bovik. Geoscaler: Geometry and rendering-aware downsampling of 3d mesh textures. In *2025 IEEE International Conference on Image Processing (ICIP)*, pages 1007–1012. IEEE, 2025. 3
- [73] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 3
- [74] Ava Pun, Kangle Deng, Ruixuan Liu, Deva Ramanan, Changliu Liu, and Jun-Yan Zhu. Generating physically stable and buildable lego designs from text. *arXiv preprint arXiv:2505.05469*, 2025. 3
- [75] Adarsh Pyarelal, Aditya Banerjee, and Kobus Barnard. Modular procedural generation for voxel maps. In *AAAI Fall Symposium*, pages 85–101. Springer, 2021. 3
- [76] 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 supervi-

- sion. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3, 5
- [77] Christian Reiser, Songyou Peng, Yiyi Liao, and Andreas Geiger. Kilonerf: Speeding up neural radiance fields with thousands of tiny mlps. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 2
- [78] Xuanchi Ren, Jiahui Huang, Xiaohui Zeng, Ken Museth, Sanja Fidler, and Francis Williams. Xcube: Large-scale 3d generative modeling using sparse voxel hierarchies. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2, 3
- [79] Xuanchi Ren, Yifan Lu, Hanxue Liang, Zhangjie Wu, Huan Ling, Mike Chen, Sanja Fidler, Francis Williams, and Jiahui Huang. Scube: Instant large-scale scene reconstruction using voxplats. *Advances in Neural Information Processing Systems*, 37:97670–97698, 2024. 3
- [80] Katja Schwarz, Axel Sauer, Michael Niemeyer, Yiyi Liao, and Andreas Geiger. Voxgraf: Fast 3d-aware image synthesis with sparse voxel grids. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 2
- [81] Etai Sella, Gal Fiebelman, Peter Hedman, and Hadar Averbuch-Elor. Vox-e: Text-guided voxel editing of 3d objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 2, 6, 8
- [82] Ygor Rebouças Serpa and Maria Andréia Formico Rodrigues. Towards machine-learning assisted asset generation for games: A study on pixel art sprite sheets. In *2019 18th Brazilian Symposium on Computer Games and Digital Entertainment (SBGames)*, pages 219–228. IEEE, 2019. 2
- [83] Rushi Shah, Mingyuan Yan, Michael Curtis Mozer, and Dianbo Liu. Improving discrete optimisation via decoupled straight-through estimator. *arXiv preprint arXiv:2410.13331*, 2024. 3
- [84] Chih-Hai Su, Chih-Yao Hu, Shr-Ruei Tsai, Jie-Ying Lee, Chin-Yang Lin, and Yu-Lun Liu. Boostmvsnerfs: Boosting mvs-based nerfs to generalizable view synthesis in large-scale scenes. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–12, 2024. 2, 3
- [85] Cheng Sun, Min Sun, and Hwann-Tzong Chen. Direct voxel grid optimization: Super-fast convergence for radiance fields reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5459–5469, 2022. 2, 3, 4
- [86] Yuhta Takida, Yukara Ikemiya, Takashi Shibuya, Kazuki Shimada, Woosung Choi, Chieh-Hsin Lai, Naoki Murata, Toshimitsu Uesaka, Kengo Uchida, Wei-Hsiang Liao, et al. Hq-vae: Hierarchical discrete representation learning with variational bayes. *arXiv preprint arXiv:2401.00365*, 2023. 3
- [87] Jiaxiang Tang, Jiawei Ren, Hang Zhou, Ziwei Liu, and Gang Zeng. Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. In *International Conference on Learning Representations (ICLR)*, 2024. 3
- [88] Tao Tu, Shun-Po Chuang, Yu-Lun Liu, Cheng Sun, Ke Zhang, Donna Roy, Cheng-Hao Kuo, and Min Sun. Imgeonet: Image-induced geometry-aware voxel representation for multi-view 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6996–7007, 2023. 2
- [89] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 3
- [90] Vikram Voleti, Chun-Han Yao, Mark Boss, Adam Letts, David Pankratz, Dmitry Tochilkin, Christian Laforte, Robin Rombach, and Varun Jampani. Sv3d: Novel multi-view synthesis and 3d generation from a single image using latent video diffusion. In *European Conference on Computer Vision*, pages 439–457. Springer, 2024. 3
- [91] Can Wang, Menglei Chai, Mingming He, Dongdong Chen, and Jing Liao. Clip-nerf: Text-and-image driven manipulation of neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2
- [92] Can Wang, Menglei Chai, Mingming He, Dongdong Chen, and Jing Liao. Clip-nerf: Text-and-image driven manipulation of neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3
- [93] Can Wang, Ruixiang Jiang, Menglei Chai, Mingming He, Dongdong Chen, and Jing Liao. Nerf-art: Text-driven neural radiance fields stylization. *IEEE Transactions on Visualization and Computer Graphics*, 30(8):4983–4996, 2023. 2, 3
- [94] Nanyang Wang, Yinda Zhang, Zhuwen Li, Yanwei Fu, Wei Liu, and Yu-Gang Jiang. Pixel2mesh: Generating 3d mesh models from single rgb images. In *Proceedings of the European conference on computer vision (ECCV)*, pages 52–67, 2018. 3
- [95] Tengfei Wang, Bo Zhang, Ting Zhang, Shuyang Gu, Jianmin Bao, Tadas Baltrusaitis, Jingjing Shen, Dong Chen, Fang Wen, Qifeng Chen, and Baining Guo. Rodin: A generative model for sculpting 3d digital avatars using diffusion. *arXiv preprint arXiv:2212.06135*, 2022. 3, 5, 6
- [96] 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. *Advances in Neural Information Processing Systems*, 36, 2024. 3
- [97] Ji-Jia Wu, Andy Chia-Hao Chang, Chieh-Yu Chuang, Chun-Pei Chen, Yu-Lun Liu, Min-Hung Chen, Hou-Ning Hu, Yung-Yu Chuang, and Yen-Yu Lin. Image-text co-decomposition for text-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26794–26803, 2024. 3
- [98] Kailu Wu, Fangfu Liu, Zhihan Cai, Runjie Yan, Hanyang Wang, Yating Hu, Yueqi Duan, and Kaisheng Ma. Unique3d: High-quality and efficient 3d mesh generation from a single image. *arXiv preprint arXiv:2405.20343*, 2024. 3, 5
- [99] Shuang Wu, Songlin Tang, Guangming Lu, Jianzhuang Liu, and Wenjie Pei. Univoxel: Fast inverse rendering by unified voxelization of scene representation. In *European Conference on Computer Vision*, pages 360–376. Springer, 2024. 2

- [100] Zongwei Wu, Liangyu Chai, Nanxuan Zhao, Bailin Deng, Yongtuo Liu, Qiang Wen, Junle Wang, and Shengfeng He. Make your own sprites: Aliasing-aware and cell-controllable pixelization. *ACM Transactions on Graphics (TOG)*, 41(6): 1–16, 2022. 2, 4
- [101] Jianfeng Xiang, Zelong Lv, Sicheng Xu, Yu Deng, Ruicheng Wang, Bowen Zhang, Dong Chen, Xin Tong, and Jiaolong Yang. Structured 3d latents for scalable and versatile 3d generation. *arXiv preprint arXiv:2412.01506*, 2024. 3, 5
- [102] Desai Xie, Jiahao Li, Hao Tan, Xin Sun, Zhixin Shu, Yi Zhou, Sai Bi, Sören Pirk, and Arie E Kaufman. Carve3d: Improving multi-view reconstruction consistency for diffusion models with rl finetuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6369–6379, 2024. 3
- [103] Ximing Xing, Haitao Zhou, Chuang Wang, Jing Zhang, Dong Xu, and Qian Yu. Svgdreamer: Text guided svg generation with diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. Accepted by CVPR 2024; arXiv preprint arXiv:2312.16476. 2
- [104] Jiale Xu, Weihao Cheng, Yiming Gao, Xintao Wang, Shenghua Gao, and Ying Shan. Instantmesh: Efficient 3d mesh generation from a single image with sparse-view large reconstruction models. *arXiv preprint arXiv:2404.07191*, 2024. 3
- [105] Jiayu Yang, Ziang Cheng, Yunfei Duan, Pan Ji, and Hongdong Li. Consistnet: Enforcing 3d consistency for multi-view images diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7079–7088, 2024. 3
- [106] Alex Yu, Sara Fridovich-Keil, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance fields without neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2
- [107] Zhu Yu, Runmin Zhang, Jiacheng Ying, Junchen Yu, Xiaohai Hu, Lun Luo, Si-Yuan Cao, and Hui-Liang Shen. Context and geometry aware voxel transformer for semantic scene completion. *Advances in Neural Information Processing Systems*, 37:1531–1555, 2024. 3
- [108] Dongdong Yue, Xinyi Liu, Yi Wan, Yongjun Zhang, Maoteng Zheng, Weiwei Fan, and Jiachen Zhong. Nerfortho: Orthographic projection images generation based on neural radiance fields. *International Journal of Applied Earth Observation and Geoinformation*, 136:104378, 2025. 3
- [109] LAN Yushi, Shangchen Zhou, Zhaoyang Lyu, Fangzhou Hong, Shuai Yang, Bo Dai, Xingang Pan, and Chen Change Loy. Gaussiananything: Interactive point cloud flow matching for 3d generation. In *The Thirteenth International Conference on Learning Representations*, 2025. 3
- [110] Yu-Ting Zhan, Cheng-Yuan Ho, Hebi Yang, Yi-Hsin Chen, Jui Chiu Chiang, Yu-Lun Liu, and Wen-Hsiao Peng. Cat-3dgs: A context-adaptive triplane approach to rate-distortion-optimized 3dgs compression. *arXiv preprint arXiv:2503.00357*, 2025. 2
- [111] Longwen Zhang, Ziyu Wang, Qixuan Zhang, Qiwei Qiu, Anqi Pang, Haoran Jiang, Wei Yang, Lan Xu, and Jingyi Yu. Clay: A controllable large-scale generative model for creating high-quality 3d assets. *ACM Transactions on Graphics (TOG)*, 43(4):1–20, 2024. 3
- [112] Yuechen Zhang, Zexin He, Jinbo Xing, Xufeng Yao, and Jiaya Jia. Ref-npr: Reference-based non-photorealistic radiance fields for controllable scene stylization. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 4242–4251, 2023. 3
- [113] Zheng Zhou, Zhe Li, Bo Yu, Lina Hu, Liang Dong, Zijian Yang, Xiaoqi Liu, Ning Xu, Ziwei Wang, Yonghao Dang, et al. Gaussiancad: Robust self-supervised cad reconstruction from three orthographic views using 3d gaussian splatting. *arXiv preprint arXiv:2503.05161*, 2025. 3