CTR3D: Cross-View Token Reduction for Dense Multi-View Generation

Kunming Luo[1]*, Hongyu Yan[1]*, Yuan Liu[1]†,
Zihao Zhang[2], Manyuan Zhang[3], Wenping Wang[4], Ping Tan[1]†
[1]Hong Kong University of Science and Technology, [2]Institute of AI for Industries, CAS,
[3]The Chinese University of Hong Kong, [4]Texas A&M University.

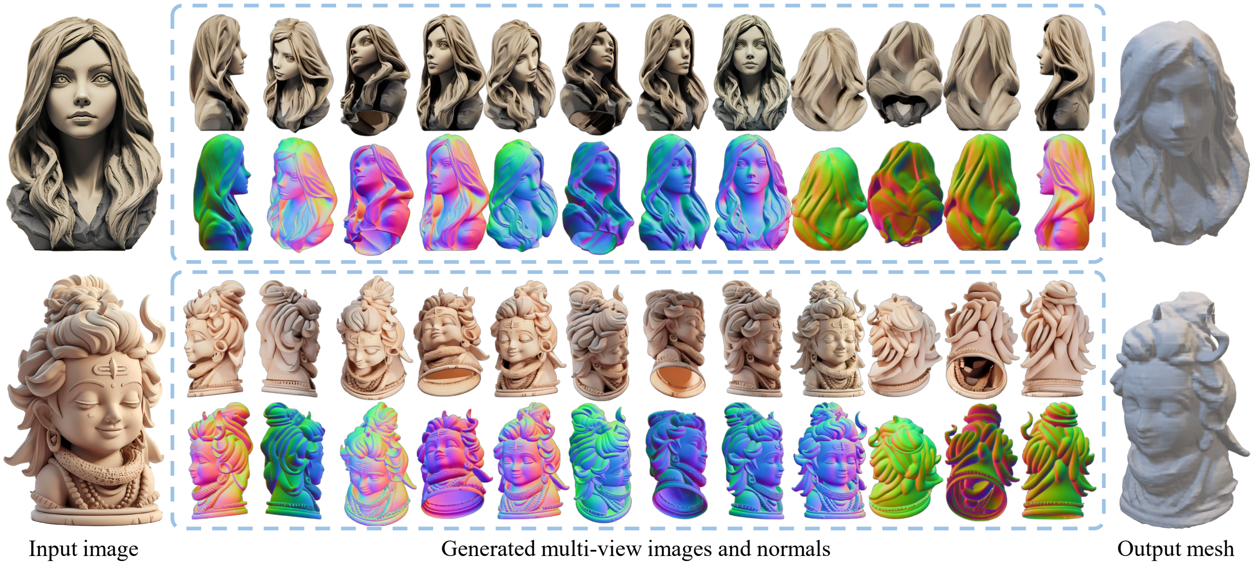

Figure 1. We propose CTR3D for dense multi-view generation from single-view images. Given a single view input image, our method can generate 12 views with normal maps in a resolution of 512 × 512. The generated multi-view images and normals can be directly reconstructed into 3D models.

Abstract

Recent multi-view diffusion (MVD) methods have utilized the generative capabilities of 2D image diffusion models to produce multi-view images from a single-view input. However, existing approaches often depend on dense cross-view attention layers, which hinder scalability and fidelity due to their high computational costs. In this paper, we propose CTR3D, a novel method that incorporates token reduction in multi-view attention layers to efficiently generate dense, high-resolution multi-view images without restricting the camera viewpoints of the generated views. Our approach is designed into three key steps: redundancy removal, attention interaction, and token recovery. These steps leverage lightweight, projection-based techniques for multi-view token reduction and recovery, significantly improving the computational efficiency of MVD. By reducing the number of tokens in attention layers while preserving multi-view consistency, our model achieves state-of-the-art performance in novel view synthesis and 3D reconstruction while keeping efficiency for generation of dense high-resolution images and normals. Experimental results demonstrate that our method surpasses existing approaches, providing a more efficient and effective solution for multi-view generation. https://github.com/HKUST-SAIL/CTR3D

1. Introduction

3D Reconstruction from single-view images is an important task in computer vision and graphics, as it has potential applications in game design, virtual reality, and robotics. This task is ill-posed and presents significant challenges be-

*These authors contributed equally to this work.
†Corresponding authors.

cause it requires extensive knowledge of the 3D world to imagine or generate the 3D structure and texture of the invisible parts. Recent research [39, 42, 51, 52] has proposed fine-tuning 2D image diffusion models into multi-view diffusion models (MVD) to explicitly generate multi-view images, which are then used for 3D reconstruction through neural reconstruction methods or large reconstruction models (LRM) [15, 60]. This explicit generation of multi-view images is more controllable, efficient, and stable than the distillation-based method [44, 61], making it more popular for single-view 3D reconstruction tasks.

The effectiveness of MVD strongly relies on dense cross-view attention layers to maintain multi-view consistency by exchanging information across different views. However, the attention layers have quadratic computation and memory complexity with the number of tokens while cross-view dense attention layers apply every pixel from all views as tokens in the attention. As the number of views and the resolution increase, the computational cost and memory consumption of this dense cross-view attention become a bottleneck, limiting all existing methods to generate only limited images with low resolutions. This severely harms the detailed quality of single-view 3D reconstruction results.

To address this issue, the recent Era3D [27] introduced row-wise attention, which sets all target views as horizontal views, allowing each pixel only to correlate pixels in the same row from other views and thereby limiting the attention computation to pixels within the same row. In this context, Era3D can generate high-resolution multi-view images and achieve high-fidelity 3D generation results. However, adopting such row-wise attention layers severely restricts the generated camera views to be located solely on the horizontal plane, which significantly reduces the flexibility in capturing diverse perspectives and ultimately harms the completeness of the generated 3D models, as essential parts of the object that are not visible from horizontal viewpoints remain missing or inadequately represented.

In this paper, we present a novel method for efficient multi-view diffusion to generate high-resolution images from more views without additional constraints on the camera viewpoints as shown in Fig. 1. Our method is based on our observation that there is a lot of redundancy among different views, for example, the same regions often show very similar patterns from nearby viewpoints, resulting in similar features in the denoising neural network. However, in the multi-view attention layer, previous methods treat all features equally as tokens in the attention computation, which actually wastes a huge amount of computation in processing repetitive similar features. This redundancy phenomenon becomes even more severe with the increase of the view number and resolution. This observation motivates us to adopt token reduction to merge similar features in the cross-view attention layers, which could decrease the computational complexity of multi-view diffusion, thereby enabling the generation of a large number of high-resolution multi-view images in one diffusion process.

However, designing such a token reduction mechanism in Multi-view Diffusion is non-trivial with two challenges. First, unlike most existing token reduction methods [3, 11, 46, 70] that only reduce the tokens for subsequent tasks and do not recover them, MVD needs to retrieve all the tokens to generate all pixels of all target views. How to design such a token recovery scheme for multi-view generation is still unexplored. Second, the token reduction should be trainable without imposing too much computation overhead. Many existing methods [2, 31] simply merge tokens according to feature similarity in a post-processing manner, which requires exhaustively and inefficiently computing similarity between all tokens. In MVD, we need to train the token reduction and recovery modules along with the diffusion process instead of using a post-process for token reduction.

In this paper, we propose a simple and effective multi-view token reduction approach. Specifically, as shown in Fig. 2, our cross-view token reduced attention consists of three steps: 1) Multi-view redundancy removal to reduce redundant tokens; 2) Attention to exchange information among different views for consistency; 3) Token recovery to recover all tokens. In the redundancy reduction, given N tokens with a size of $N \times F$, we utilize the instance normalization layers and MLP layers to compute a weight matrix of size $K \times N$ ($K \ll N$). Then, we compute the product between the weight matrix and the input tokens to get reduced tokens of a size $K \times F$. The instance normalization layers normalize all the tokens into a standard normal distribution with similar tokens distributing around 0 and thus enable us to find the redundancy in the tokens when computing the weight matrix. Then, the cross-view attention is conducted on the reduced tokens, which are further recovered back to the size of input tokens by multiplying the transposed and re-normalized weight matrix. By using our proposed token reduction mechanism, CTR3D enables the generation of 12 RGB images and normal maps at a resolution of 512 simultaneously in a single diffusion step with 0.35 seconds and 14.6G memory while previous methods are typically limited to only 4 to 6 views.

To validate the effectiveness of our token reduction design in the multiview diffusion model, we conduct extensive experiments on the single-view 3D generation task. The results demonstrate that by generating dense multiview images with a high resolution, our method achieves superior 3D generation quality than baseline methods on the GSO benchmark. Though generating 12 views with a 512 resolution, our method still has similar computation efficiency as baseline methods.

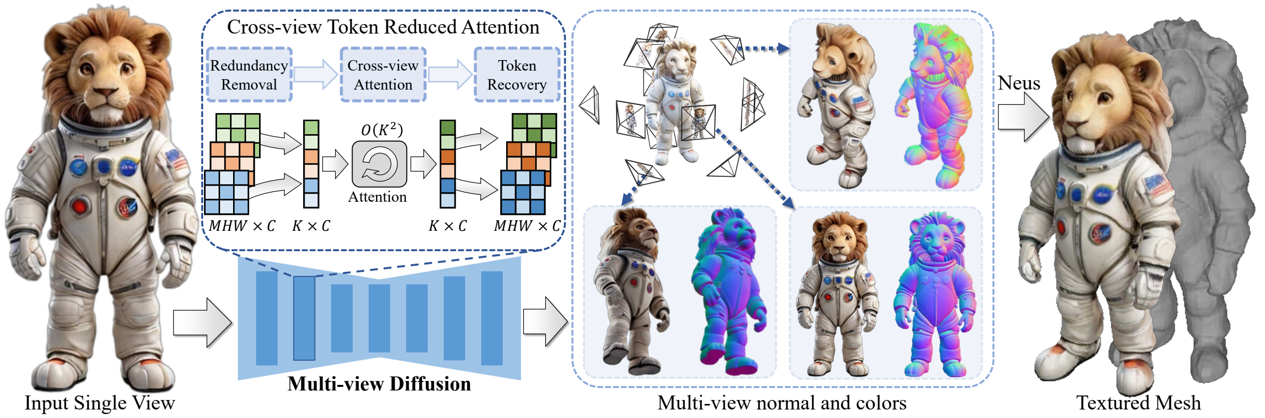

Figure 2. Illustration of the overall pipeline of CTR3D. We propose our Cross-view Token Reduced Attention (CTRA) to achieve efficient multi-view generation. CTRA consists of three main steps: multi-view token reduction for redundancy removal, consistency checking by attention, and multi-view token recovery. We apply CTRA to each layer of the U-Net in Multi-view Diffusion (MVD), enabling dense, high-resolution, and consistent multi-view generation. Finally, we use the generated images and normals to reconstruct the 3D model.

2. Related Work

Different from previous image-to-3D generation methods [6, 13, 28, 30, 33, 34, 37, 40, 44, 54, 61, 69] based on Score Distillation Sampling (SDS) or 3D generation methods [17, 18, 29, 62–64, 66, 68, 71] based on native 3D diffusion, our method focuses on multi-view generation by introducing the token reduction strategy into the 2D stable diffusion model [48]. Therefore, in this section, we start by reviewing the methods based on the multi-view diffusion model and illustrating the noticeable difference between these approaches and our method. Then, we introduce the advanced works in terms of token reduction to demonstrate our practical and creative application in consistent dense multi-view image synthesis.

2.1. Multi-view Diffusion Model

Based on Stable Diffusion [47], the pioneering work MVDream [52] proposed to distribute a text-to-image diffusion model to support novel view synthesis. There are two challenges in the multi-view diffusion model: one is multi-view consistency, and another is computational complexity. To solve the problem of multi-view consistency, recent methods [10, 16, 26, 32, 36, 38, 39, 42, 50, 55–57, 60, 65] introduced cross-view attention. While cross-view attention can align information from all views to achieve high consistency, it faces the problem of high computation complexity. Therefore, these methods struggle to generate dense and high-resolution multi-view images. Although recent works [7, 58, 67] proposed using stable video diffusion [1] for dense multi-view image generation, the temporal convolutions in video models are insufficient to ensure the consistency of the generated multi-view images. As a result, these methods [58] require complex post-processing to produce 3D models. To reduce the commutative complexity of dense cross-view attention and achieve high-resolution consistent multi-view image synthesis, the recent method Era3D [27] proposed a novel attention strategy. Specifically, Era3D fixes the generated target view to a horizontal canonical camera perspective, allowing each pixel to be related only to the pixels on its horizontal line. This reduces dense attention to row-wise attention, enabling high-resolution multi-view generation. However, they still struggle to form dense multi-view images and have a pivot limitation in that their generated images tend to be horizontal even if the input image is viewed from above or below. In addition, since each pixel only interacts with the pixels horizontally and vertically, it generates low-consistency multi-view images. Unlike these methods, our method aims to filter redundant tokens based on the fact that there is a large amount of redundant information in the dense-view images.

2.2. Token reduction

Token reduction [11, 41, 46, 70] is mainly studied as a task to reduce computational and memory requirements of large vision and language transformers. Existing token reduction methods can be divided into two categories: dynamic token pruning [4, 14, 24, 46] and token merging [3, 12, 49, 73]. The approach of token pruning primarily focuses on learning the importance of tokens and excluding unimportant tokens from the computation process, thereby improving efficiency while maintaining accuracy. However, completely discarding tokens can lead to a loss of detailed information [23]. Unlike token pruning, the main idea of token merging is to continuously merge similar tokens during the computation process of transformers to reduce redundant calculations. One of the most classic methods is ToMe [3], where tokens are randomly divided into two groups, and

Bipartite Soft Matching (BSM) is used to calculate the similarity between the two groups of tokens. Based on this similarity, the most similar tokens are merged. Although BSM-based methods can be easily influenced by the distribution of the matching groups in BSM, token merging can better preserve detailed information compared to token pruning schemes. As a result, the recent ToMeSD [2] has applied this approach to accelerate text-to-image generation in SD. Additionally, recent work VidTome [31] has proposed using token merging to enhance the consistency of zero-shot video editing. However, existing token reduction methods still focus on single image processing, where redundancy between pixels mainly exists within local windows. In this paper, we focus on reducing the redundancy among multiple views and propose a new approach for multi-view token reduction so as to facilitate dense multi-view generation.

3. Method

We propose CTR3D to generate dense multi-view images from a single-view input image. The overall pipeline of CTR3D is shown in Fig. 2. We first use Multi-view Diffusion (MVD) to generate multi-view images and normal maps conditioned on the input single-view image. Then, we extract a textured mesh from the generated images and normal maps using existing reconstruction methods like NeuS [59]. The key idea of our proposed method is to reduce the token redundancy in the multi-view attention layer of MVD to improve the computational efficiency so that we can generate high-resolution and dense multi-view images and normal maps.

To achieve this, we design the multi-view attention layer as three parts: redundancy removal (Sec. 3.2), cross-view attention, and multi-view token recovery (Sec. 3.3). Specifically, we first design a lightweight multi-view reduction layer to project multi-view tokens into a set of representative token. Then, cross-view information exchanging is performed only on those representative tokens by a self-attention layer. Finally, those enhanced tokens yield from the self-attention layer are consistently projected back to the original amount by our re-weighting recovery module.

3.1. Multi-view Diffusion

At the foundation of our proposed method is the Multi-view Diffusion (MVD), which is an extension of the stable diffusion model [48] for text-to-image generation. Similar to previous methods, we use pretrained VAE to extract latent z_c of the input image I_c as a condition and perform denoising simultaneously on multi-view by an UNet ϵ_θ to generate multi-view latents z_{mv} that contains color latents $z_c = \{z_{c1}, z_{c2}, ..., z_{cM}\}$ and normal latents $z_n = \{z_{n1}, z_{n2}, ..., z_{nM}\}$, where M is the number of generated views. Finally, multi-view images and normal maps can be obtained from those latents through VAE decoding. The UNet ϵ_θ is trained as follows:

$$\mathcal{L}_{MVD} = \|\epsilon^t_{mv} - \epsilon_\theta(x^t_{mv}, t, z_c)\|^2_2, \quad \text{where } \epsilon_t \sim \mathcal{N}(0, I) \tag{1}$$

where x^t_{mv} is the noisy sample of z_{mv} for time step t. The UNet ϵ_θ is designed as a sequence of blocks with four levels of feature pyramid. In each block, there are four modules: 1) a resnet block and a self-attention layer on each single view; 2) a cross-attention layer to incorporate information from the condition CLIP embedding to each single view; 3) a cross-view attention layer on all multi-view tokens to keep multi-view consistency.

3.2. Recap of Cross-view Attention

Dense cross-view attention. MVD achieves consistency by applying dense cross-view attention layers within each block of the UNet ϵ_θ to exchange information among all tokens from all views. Let $F \in \mathbb{R}^{(Mhw)\times d}$ be the flattened input color feature or normal feature of one UNet block, where d is the feature dimension, h, w are spatial dimensions and $F(i) \in \mathbb{R}^{1\times d}$ is the i-th token in F. Then the dense cross-view attention layer can be formulated as follows:

$$\hat{F}(i) = Softmax(\frac{\mathcal{Q}(F(i)) \cdot \mathcal{K}(F)^\top}{\sqrt{d}}) \cdot \mathcal{V}(F) \tag{2}$$

where $\mathcal{Q}$, $\mathcal{K}$ and $\mathcal{V}$ are linear projections in the attention layer and $\hat{F(i)}$ is the output token for $F(i)$. Although this dense multi-view attention layer can ensure the multi-view consistency by information propagation across all multi-view tokens, its computational complexity is $O(M^2h^2w^2)$, which is very expensive and memory-intensive and limits the scalability of MVD model.

Row-wise cross-view attention. To reduce complexity, Era3D simplifies the cross-view attention layer into a row-wise attention mechanism by setting the target views all at the horizontal plane, achieving a complexity of $(O(M^2hw^2))$. However, this design restricts its applicability, as generating images from only horizontal perspectives cannot totally ensure the completeness of the resulting 3D models. For example, a bucket's horizontal top and bottom plane can never be seen from these generated views. In this paper, we introduce our CTR3D to enhance computational efficiency and improve generated view number and resolution by eliminating redundancy among multiple views, without imposing any restrictions on camera settings.

3.3. Multi-view Token Reduction

Weighted token reduction. We adopt a fusion weight matrix to conduct the token reduction. Let K represent the number of tokens after reduction. The token reduction is

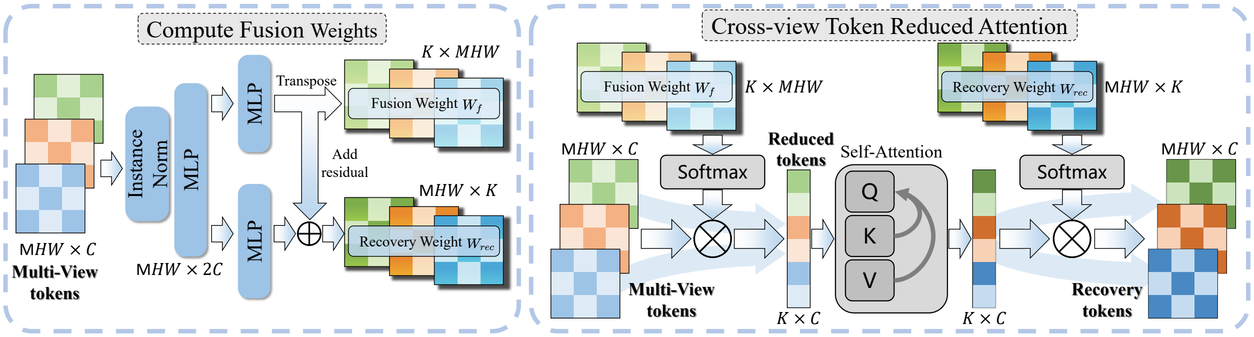

Figure 3. Detailed structure of our proposed Cross-view Token Reduced Attention. We first obtain fusion weight W_f and recovery weight W_{rec} by two branches of MLPs (left). Then we perform multi-view token reduction and use a self-attention layer to exchange multi-view information. Finally, we use the recovery weight to recover the number of tokens.

defined as a weighted sum

$$F_r(j) = \sum_{i=1}^{Mhw} W_f(j,i) \cdot F(i), \quad j \in [0, K), \tag{3}$$

where $F_r(j)$ is the j-th token of the reduced tokens $F_r \in \mathbb{R}^{K \times d}$ and $W_f \in \mathbb{R}^{K \times Mhw}$ is the fusion weight. The fusion weight here determines the redundancy of each token, which plays an essential role in eliminating redundancy and retaining key information. However, computing the fusion weight should be as lightweight as possible for efficiency consideration. We next introduce how we compute the fusion weight with an instance normalization layer.

Fusion weight computation. The fusion weights are required to reveal the redundancy of all tokens such that similar tokens should be fused into a single reduced token while distinct tokens should be retained. However, determining the similarity or distinctness of tokens requires us to compare each token to each other. Thus, a naive way to implement such relationship extraction would lead to a quadratic computation complexity again in fusion weight computation by comparing every token with each other. To avoid this, we propose an efficient but simple way to model the relative relationship between all tokens by applying an instance normalization layer. This instance normalization layer normalizes all tokens F to $\bar{F}$ conforming to a standard Gaussian distribution. In this case, the distances of these normalized tokens to the origin reveal their importance in the whole token set, which helps us to determine their fusion weight. Fig. 3 gives an detailed illustration. For each normalized token $\bar{F}(i), i \in [0, Mhw)$, we employ a module with two MLP layers to estimate its contribution score $W_c \in \mathbb{R}^{1 \times K}$. Then, its fusion weight can be obtained by applying softmax normalization to the contribution scores by

$$W_f(j) = Softmax(W_c^\top(j)), \quad W_c(i) = \mathcal{M}(\bar{F}(i)), \tag{4}$$

where $\mathcal{M}$ represents our MLP network, $W_c^\top$ is the transpose of matrix W_c.

Discussion about existing token reduction solution. A straightforward approach to implementing multi-view token reduction is to directly apply single-image token reduction methods, such as ToMeSD, to multi-view tokens. In ToMeSD, tokens are first divided into two groups, and the similarity between each pair of tokens from these groups is calculated. Following this, similar token pairs are merged based on their similarity scores. However, when applied to multi-view reduction, this token merging method has a computational complexity of $(O((Mhw/2)^2) = O(M^2h^2w^2)$. As we increase the number or resolution of the generated views, the number of tokens grows exponentially. Consequently, the computational cost of similarity calculations also increases quadratically, resulting in inefficiencies in this approach. In comparison, the overall computational complexity of our reduction process is $O(MhwK)$. When K is relatively small, our method is more lightweight and computationally efficient, making it better suited for multi-view reduction. In our work, we set K as $[4096, 1536, 384, 96]$ separately for four different pyramid levels in the UNet structure.

Attention with reduced tokens. Once we obtained K reduced tokens F_r, we employ a self-attention layer to exchange information among these reduced tokens so as to ensure consistency across multiple views. The results of this attention layer are K tokens, named F_{ra}. The computational complexity of this attention layer is only $O(K^2)$, which is significantly lower than that of the previous dense cross-view attention layer.

3.4. Token Recovery

After conducting cross-view attention layers with reduced tokens, we need to recover these tokens for the generation task. To implement this, we again apply another weight matrix $W_{rec} \in \mathbb{R}^{Mhw \times K}$ to multiply F_{ra} to get back the

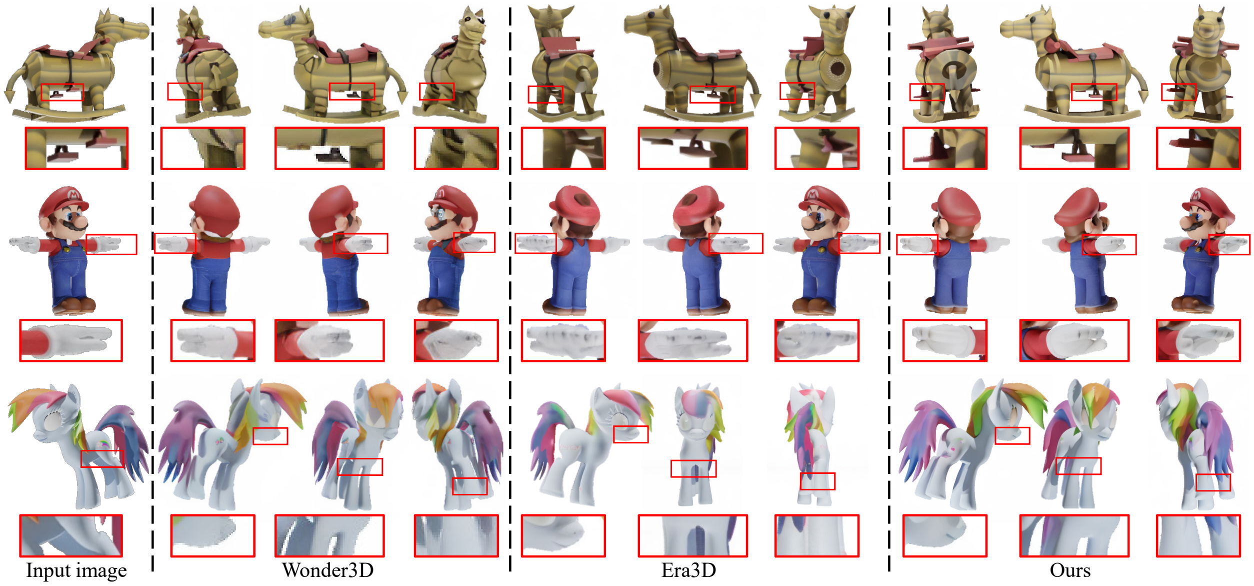

Figure 4. The qualitative comparison with baseline models on novel view synthesis.

processed full tokens $\hat{F}$. A naive way for this is to adopt the transposed fusion weights W_f in Eq. 3 as the recovery weights. However, during the token reduction process, some less important tokens may have low fusion weights close to zero because they come from some repeated regions such as backgrounds. If we directly use fusion weights W_f for recovery, those tokens may be assigned values close to 0 after the recovery process, thus losing multi-view consistency. To address this issue, we propose a re-weighting method to adjust the recovery weights. As shown in Fig. 3, we introduce an additional branch on $\mathcal{M}$ with a single MLP layer, which outputs a residual weight $W_{res} \in \mathbb{R}^{Mhw \times K}$ for recovery. Thus, the token recovery layer is defined by

$$\begin{aligned} \hat{F}(i) &= \sum_{j=1}^{K} W_{rec}(i,j) \cdot F_{ra}(j), \\ W_{rec}(i) &= Softmax(W_{res}(i) + W_f^T(i)) \end{aligned} \tag{5}$$

where $\hat{F}(i)$ is the i-th token in our recovered tokens $\hat{F} \in \mathbb{R}^{Mhw \times d}$.

4. Experiments

Datasets. We followed the previous methods [27, 42] and trained our CTR3D on the LVIS subset of Objaverse [8], which contains approximately 32,000 high-quality 3D models. To construct the multi-view training dataset, we normalized each model to a unit scale and positioned it at the center. Then we rendered images and normals at a resolution of 512×512 from twelve different viewpoints including six horizontal views: front, back, left, right, front-right, and front-left views; three 45° top-down views: back-top, front-right-top, and front-left-top; as well as three 45° up-looking views: back-down, front-right-down, and front-left-down. During the rendering process, random rotations were also applied to the 3D models to enhance diversity. Following the previous methods, we evaluate the performance of CTR3D on the Google Scanned Object (GSO) dataset [9], which is a standard benchmark widely used to evaluate 3D generation tasks. To demonstrate the generalization ability of our CTR3D, we also evaluate our method on images collected from the Internet.

Metrics. We evaluate our method on two tasks: novel view synthesis (NVS) and 3D reconstruction. The quality of NVS is assessed using the Learned Perceptual Image Patch Similarity (LPIPS) metric [72], which measures the perceptual consistency between the generated images and the ground truth. For the evaluation of 3D reconstruction quality, we use the Chamfer Distance (CD) and the Volume Intersection over Union (IOU) metrics to compare the reconstructed meshes with the ground truth models.

Implementation details. We implement our model based on the open-source text-to-image model SD2.1-unclip [48]. We use 16 A100 GPUs to train CTR3D for 40,000 training steps with a total batch size of 256. The initial learning rate is set to 1e-4, which is reduced to 5e-5 after 5,000 steps. To implement classifier-free guidance (CFG), we randomly omit the clip condition at a rate of 0.05. During the multi-view diffusion generation process, we use the DDIM[53] method with 40 steps and a CFG scale of 3.0. After multi-view generation, we follow Wonder3D to perform reconstruction using NeuS, followed by a texture refinement step. The entire pipeline takes approximately 4 minutes to com-

Methods	CD ↓	IoU ↑	LPIPS ↓	SSIM ↑	PSNR ↑
RealFusion	0.0819	0.2714	0.283	0.722	15.26
Zero-1-to-3	0.0339	0.5035	0.166	0.779	18.93
One-2-3-45	0.0629	0.4086	-	-	-
Shap-E	0.0436	0.3584	-	-	-
Magic123	0.0516	0.4528	-	-	-
SyncDreamer	0.0261	0.5421	0.146	0.798	20.05
Wonder3D	0.0248	0.5678	0.141	0.811	20.83
LGM	0.0259	0.5628	-	-	-
ERA3D	0.0217	0.5973	0.126	0.837	22.74
Ours	**0.0206**	**0.6447**	**0.081**	**0.914**	**23.47**

Table 1. Quantitative evaluation results of Chamfer distance (CD), IoU (for reconstruction), LPIPS, SSIM and PSNR,(for NVS) on GSO benchmark.

plete, consisting of 13.8s for multi-view diffusion generation (40 diffusion steps), 3 minutes for the NeuS reconstruction, and 10s for the texture refinement.

4.1. Experimental Results

In this section, we quantitatively and qualitatively compare our method with recent multi-view generation methods, including RealFusion [43], Zero-1-to-3 [38], One-2-3-45 [35], Shap-E [21], Magic123 [45], SyncDreamer [39], Wonder3D [42], Era3D [27].

Novel view synthesis. As shown in Table 1, we first quantitatively compare our method with recent methods in terms of LPIPS, SSIM, and PNSR metrics on the GSO dataset. The results indicates that our method achieves the best performance on all metrics. In addition to a quantitative comparison, we provide visualized results in Fig. 4. Compared to previous state-of-the-art method Era3D [27], our method can achieve comparable novel view generation on normal examples (as in row 3) with high resolution and fidelity. Note that Era3D needs to regress the elevation and focal length so as to fulfill its camera constraint (Canonical cameras on the horizontal plane). However, this perspective correction process in Era3D may introduce errors and degrade the results of Era3D (as in row 2). More importantly, the fixed horizontal viewpoint in Era3D may prevent the observation of certain areas of the object, such as the surface of the chair in the first row. This limitation can significantly impact the completeness of the reconstructed 3D model. In contrast to existing methods, our approach imposes no constraints on the camera system and enables the generation of more high-resolution views, resulting in higher quality and more complete 3D generation.

Reconstruction. To verify our superiority in the 3D reconstruction task, we then present the quantitative reconstruction comparison on the GSO dataset in Table 1. Similar to Wonder3D [42] and Era3D [27], we use the NeuS [59] to form a 3D mesh based on our generated multi-view images. Since we can generate more views with high resolution, the geometric quality of our reconstructed meshes can be better than that of existing methods. Thus our method produces the best performance with the lowest CD and highest IoU. We also provide a visualized comparison of our generated 3D mesh with previous methods in Fig. 5. As can be seen, our method produces 3D models of higher quality and greater completeness compared to previous methods. We also test our method on in-the-wild images, and the results are shown in Fig. 6. The results indicate that a limited number (e.g. 6 views) from horizontal viewpoints is insufficient to generate objects with complete structures, leading to lower quality reconstructions. In contrast, our method generates more high-resolution views, allowing us to produce complete, high-quality 3D models.

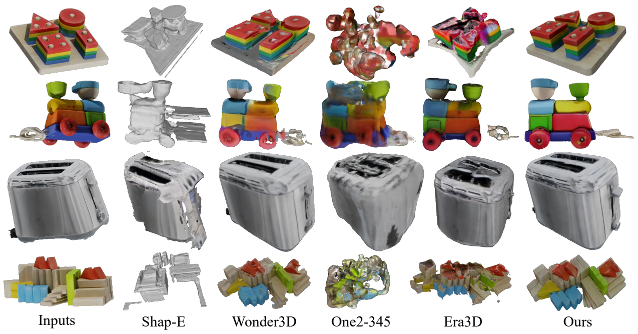

Figure 5. Qualitative comparison of 3D reconstruction results on the GSO dataset. CTR3D generates the best results with more views and high resolution, compared with existing methods.

Methods	256 resolution		512 resolution	
	CD ↓	IoU ↑	CD ↓	IoU ↑
6 views	0.0211	0.511	0.0217	0.530
8 views	0.0209	0.518	0.0201	0.572
10 views	0.0206	0.514	0.0187	0.578
12 views	0.0199	0.520	0.0180	0.585

Table 2. Ablation experiment of the number and resolution of views in CTR3D. We use CD and IoU as metrics.

4.2. Ablation study

In this section, we conduct a series of experiments to demonstrate the effectiveness of our key design. Due to limited computational resources, we used only a subset of our training set containing about 8,000 data pairs for model training and validated on 20 samples outside the training set. In Table 2, we first experimented with training our model to generate different numbers of views and resolutions. The results show that higher image resolution and more views lead to lower Chamfer distance in the generated models, supporting our idea that increasing view density and resolution improves 3D generation quality.

We further validate our design for the multi-view reduc-

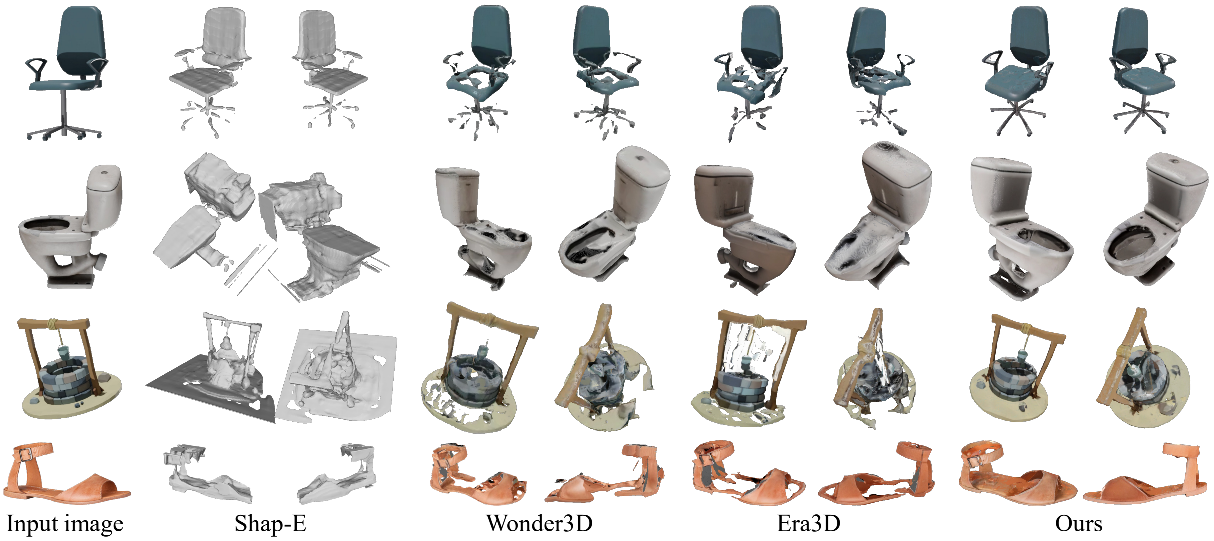

Figure 6. Reconstruction results on in the wild data. CTR3D can produce more complete results compared with previous methods.

Methods	CD ↓	IoU ↑	Run Time (ms)		
			256*6	512*6	512*12
Dense-att	0.0218	0.432	2.902	23.97	47.49
ToMe (50%)	0.0229	0.413	4.467	11.29	21.95
ToMe (10%)	0.0245	0.401	3.965	6.751	13.02
ours(no IN)	0.0222	0.428	1.602	1.805	3.126
ours(no RW)	0.0232	0.418	1.482	1.785	3.130
ours	0.0211	0.511	1.678	1.805	3.132

Table 3. Ablation of the multi-view reduction and recovery block in CTR3D. We fix the resolution as 256 and the view number as 6 for generation evaluation. The running time under different settings are also reported.

tion scheme. Specifically, we compare our method with the baseline dense attention (Dense-att), ToMe [31] and our method with our instant normalization (IN) and re-weight (RW) block disabled. For fair comparison, we implement ToMe in the same network as our method and finetune it by the same training steps. The results are presented in Table 3, which indicates that our design achieves the best performance. While using the ToMe method for multi-view token reduction can improve computational efficiency, it results in lower quality outputs. Note that the running speed of ToMe is slower than our method even at a lower reduction rate of 10%, primarily due to the time-consuming matching process. Although dense attention can obtain impressive performance in the setting of 6 low-resolution views generation, it struggles to work efficiently when view and resolution increase. In contrast, our proposed method not only produces high reconstruction metrics but also presents excellent running time. Finally, in Table 4, we reported the performance comparison between our reduced cross-view attention and existing cross-view attention layers [22, 27].

Methods	N-view	Memory usage (G)		Running time (ms)	
		256	512	256	512
Dense	6	1.03	2.07	2.902	23.927
Epipolar	6	1.06	2.58	2.997	20.87
Era3D	6	1.01	1.13	1.886	2.327
Ours	6	0.39	0.58	1.602	1.805
Ours	12	0.76	1.14	1.620	3.132

Table 4. Memory usage and running time of CTR3D in different multi-view generation settings.

We measure the average memory consumption and running time of each attention layer under the same framework, with Xformers consistently enabled. Our approach offers faster speeds and lower memory consumption, enabling the generation of more high-resolution multi-view images.

5. Conclusions

In this paper, we propose a novel method for high-quality multi-view image generation called CTR3D. We design a multi-view token reduced attention layer to facilitate dense and high-resolution multi-view generation. Our method consists of three steps: redundancy removal to enhance computational efficiency, attention interaction to ensure multi-view consistency, and token recovery to maintain the overall token count. By eliminating redundant tokens, our method consistently generates more high-resolution views compared to existing approaches, outperforming state-of-the-art methods both quantitatively and qualitatively.

Acknowledgements. This work was supported by HKUST Project 24251090T019 and Natural Science Foundation of China (NSFC) No.62302481.

References

[1] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 3

[2] Daniel Bolya and Judy Hoffman. Token merging for fast stable diffusion. In *CVPR Workshop*, pages 4598–4602, 2023. 2, 4

[3] Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your vit but faster. In *ICLR*, 2023. 2, 3

[4] Shuning Chang, Pichao Wang, Ming Lin, Fan Wang, David Junhao Zhang, Rong Jin, and Mike Zheng Shou. Making vision transformers efficient from a token sparsification view. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6195–6205, 2023. 3

[5] Anpei Chen, Haofei Xu, Stefano Esposito, Siyu Tang, and Andreas Geiger. Lara: Efficient large-baseline radiance fields. In *European Conference on Computer Vision*, pages 338–355. Springer, 2024. 1

[6] 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

[7] Zilong Chen, Yikai Wang, Feng Wang, Zhengyi Wang, and Huaping Liu. V3d: Video diffusion models are effective 3d generators. *arXiv preprint arXiv:2403.06738*, 2024. 3

[8] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *CVPR*, pages 13142–13153, 2023. 6

[9] Laura Downs, Anthony Francis, Nate Koenig, Brandon Kinman, Ryan Hickman, Krista Reymann, Thomas B McHugh, and Vincent Vanhoucke. Google scanned objects: A high-quality dataset of 3d scanned household items. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 2553–2560. IEEE, 2022. 6

[10] Yiftach Edelstein, Or Patashnik, Dana Cohen-Bar, and Lihi Zelnik-Manor. Sharp-it: A multi-view to multi-view diffusion model for 3d synthesis and manipulation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 21458–21468, 2025. 3

[11] Mohsen Fayyaz, Soroush Abbasi Koohpayegani, Farnoush Rezaei Jafari, Sunando Sengupta, Hamid Reza Vaezi Joze, Eric Sommerlade, Hamed Pirsiavash, and Jürgen Gall. Adaptive token sampling for efficient vision transformers. In *ECCV*, pages 396–414, 2022. 2, 3

[12] Ryan Grainger, Thomas Paniagua, Xi Song, Naresh Cuntoor, Mun Wai Lee, and Tianfu Wu. Paca-vit: learning patch-to-cluster attention in vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18568–18578, 2023. 3

[13] Jiatao Gu, Alex Trevithick, Kai-En Lin, Joshua M Susskind, Christian Theobalt, Lingjie Liu, and Ravi Ramamoorthi. Nerfdiff: Single-image view synthesis with nerf-guided distillation from 3d-aware diffusion. 2023. 3

[14] Joakim Bruslund Haurum, Sergio Escalera, Graham W Taylor, and Thomas B Moeslund. Which tokens to use? investigating token reduction in vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 773–783, 2023. 3

[15] 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. 2

[16] Hanzhe Hu, Zhizhuo Zhou, Varun Jampani, and Shubham Tulsiani. Mvd-fusion: Single-view 3d via depth-consistent multi-view generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9698–9707, 2024. 3

[17] Ka-Hei Hui, Aditya Sanghi, Arianna Rampini, Kamal Rahimi Malekshan, Zhengzhe Liu, Hooman Shayani, and Chi-Wing Fu. Make-a-shape: a ten-million-scale 3d shape model. In *Forty-first International Conference on Machine Learning*, 2024. 3

[18] Team Hunyuan3D, Bowen Zhang, Chunchao Guo, Haolin Liu, Hongyu Yan, Huiwen Shi, Jingwei Huang, Junlin Yu, Kunhong Li, Penghao Wang, et al. Hunyuan3d-omni: A unified framework for controllable generation of 3d assets. *arXiv preprint arXiv:2509.21245*, 2025. 3

[19] Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, et al. Perceiver io: A general architecture for structured inputs & outputs. arxiv 2021. *arXiv preprint arXiv:2107.14795*, 3. 1

[20] Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *International conference on machine learning*, pages 4651–4664. PMLR, 2021. 1

[21] Heewoo Jun and Alex Nichol. Shap-e: Generating conditional 3d implicit functions. *arXiv preprint arXiv:2305.02463*, 2023. 7

[22] Yash Kant, Ziyi Wu, Michael Vasilkovsky, Guocheng Qian, Jian Ren, Riza Alp Guler, Bernard Ghanem, Sergey Tulyakov, Igor Gilitschenski, and Aliaksandr Siarohin. Spad: Spatially aware multiview diffusers. *arXiv preprint arXiv:2402.05235*, 2024. 8

[23] Minchul Kim, Shangqian Gao, Yen-Chang Hsu, Yilin Shen, and Hongxia Jin. Token fusion: Bridging the gap between token pruning and token merging. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1383–1392, 2024. 3

[24] Zhenglun Kong, Peiyan Dong, Xiaoliang Ma, Xin Meng, Mengshu Sun, Wei Niu, Xuan Shen, Geng Yuan, Bin Ren, Minghai Qin, Hao Tang, and Yanzhi Wang. Spvit: Enabling faster vision transformers via soft token pruning. In *arxiv*, 2021. 3

[25] Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A frame-

work for attention-based permutation-invariant neural networks. In *International conference on machine learning*, pages 3744–3753. PMLR, 2019. 1

[26] Jiahao Li, Hao Tan, Kai Zhang, Zexiang Xu, Fujun Luan, Yinghao Xu, Yicong Hong, Kalyan Sunkavalli, Greg Shakhnarovich, and Sai Bi. Instant3d: Fast text-to-3d with sparse-view generation and large reconstruction model. *arXiv preprint arXiv:2311.06214*, 2023. 3

[27] Peng Li, Yuan Liu, Xiaoxiao Long, Feihu Zhang, Cheng Lin, Mengfei Li, Xingqun Qi, Shanghang Zhang, Wenhan Luo, Ping Tan, et al. Era3d: High-resolution multiview diffusion using efficient row-wise attention. *arXiv preprint arXiv:2405.11616*, 2024. 2, 3, 6, 7, 8, 1

[28] Weiyu Li, Rui Chen, Xuelin Chen, and Ping Tan. Sweetdreamer: Aligning geometric priors in 2d diffusion for consistent text-to-3d. *arXiv preprint arXiv:2310.02596*, 2023. 3

[29] Weiyu Li, Jiarui Liu, Rui Chen, Yixun Liang, Xuelin Chen, Ping Tan, and Xiaoxiao Long. Craftsman: High-fidelity mesh generation with 3d native generation and interactive geometry refiner. *arXiv preprint arXiv:2405.14979*, 2024. 3, 1, 2

[30] Xinhai Li, Huaibin Wang, and Kuo-Kun Tseng. Gaussiandiffusion: 3d gaussian splatting for denoising diffusion probabilistic models with structured noise. *arXiv preprint arXiv:2311.11221*, 2023. 3

[31] Xirui Li, Chao Ma, Xiaokang Yang, and Ming-Hsuan Yang. Vidtome: Video token merging for zero-shot video editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7486–7495, 2024. 2, 4, 8

[32] Yixun Liang, Kunming Luo, Xiao Chen, Rui Chen, Hongyu Yan, Weiyu Li, Jiarui Liu, and Ping Tan. Unitex: Universal high fidelity generative texturing for 3d shapes. *arXiv preprint arXiv:2505.23253*, 2025. 3

[33] 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

[34] 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 *CVPR*, pages 300–309, 2023. 3

[35] 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, 2023. 7

[36] Minghua Liu, Chao Xu, Haian Jin, Linghao Chen, Zexiang Xu, Hao Su, et al. One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization. *arXiv preprint arXiv:2306.16928*, 2023. 3

[37] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9298–9309, 2023. 3

[38] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9298–9309, 2023. 3, 7

[39] Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. Syncdreamer: Generating multiview-consistent images from a single-view image. *arXiv preprint arXiv:2309.03453*, 2023. 2, 3, 7

[40] Zexiang Liu, Yangguang Li, Youtian Lin, Xin Yu, Sida Peng, Yan-Pei Cao, Xiaojuan Qi, Xiaoshui Huang, Ding Liang, and Wanli Ouyang. Unidream: Unifying diffusion priors for relightable text-to-3d generation. *arXiv preprint arXiv:2312.08754*, 2023. 3

[41] Sifan Long, Zhen Zhao, Jimin Pi, Shengsheng Wang, and Jingdong Wang. Beyond attentive tokens: Incorporating token importance and diversity for efficient vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10334–10343, 2023. 3

[42] Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, et al. Wonder3d: Single image to 3d using cross-domain diffusion. *arXiv preprint arXiv:2310.15008*, 2023. 2, 3, 6, 7, 1

[43] Luke Melas-Kyriazi, Christian Rupprecht, Iro Laina, and Andrea Vedaldi. Realfusion: 360° reconstruction of any object from a single image. In *Arxiv*, 2023. 7

[44] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 2, 3

[45] Guocheng Qian, Jinjie Mai, Abdullah Hamdi, Jian Ren, Aliaksandr Siarohin, Bing Li, Hsin-Ying Lee, Ivan Skorokhodov, Peter Wonka, Sergey Tulyakov, et al. Magic123: One image to high-quality 3d object generation using both 2d and 3d diffusion priors. *arXiv preprint arXiv:2306.17843*, 2023. 7

[46] Yongming Rao, Wenliang Zhao, Benlin Liu, Jiwen Lu, Jie Zhou, and Cho-Jui Hsieh. Dynamicvit: Efficient vision transformers with dynamic token sparsification. In *NeurIPS*, 2021. 2, 3

[47] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. 3

[48] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 3, 4, 6

[49] Michael Ryoo, AJ Piergiovanni, Anurag Arnab, Mostafa Dehghani, and Anelia Angelova. Tokenlearner: Adaptive space-time tokenization for videos. *Advances in neural information processing systems*, 34:12786–12797, 2021. 3, 1

[50] Kyle Sargent, Zizhang Li, Tanmay Shah, Charles Herrmann, Hong-Xing Yu, Yunzhi Zhang, Eric Ryan Chan, Dmitry Lagun, Li Fei-Fei, Deqing Sun, et al. Zeronvs: Zero-shot

360-degree view synthesis from a single real image. *arXiv preprint arXiv:2310.17994*, 2023. 3

[51] Ruoxi Shi, Hansheng Chen, Zhuoyang Zhang, Minghua Liu, Chao Xu, Xinyue Wei, Linghao Chen, Chong Zeng, and Hao Su. Zero123++: a single image to consistent multi-view diffusion base model. *arXiv preprint arXiv:2310.15110*, 2023. 2, 1

[52] Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. *arXiv preprint arXiv:2308.16512*, 2023. 2, 3

[53] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 6

[54] Jiaxiang Tang, Jiawei Ren, Hang Zhou, Ziwei Liu, and Gang Zeng. Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. *arXiv preprint arXiv:2309.16653*, 2023. 3

[55] Shitao Tang, Fuayng Zhang, Jiacheng Chen, Peng Wang, and Furukawa Yasutaka. Mvdiffusion: Enabling holistic multi-view image generation with correspondence-aware diffusion. *arXiv preprint 2307.01097*, 2023. 3

[56] Shitao Tang, Jiacheng Chen, Dilin Wang, Chengzhou Tang, Fuyang Zhang, Yuchen Fan, Vikas Chandra, Yasutaka Furukawa, and Rakesh Ranjan. Mvdiffusion++: A dense high-resolution multi-view diffusion model for single or sparse-view 3d object reconstruction. *arXiv preprint arXiv:2402.12712*, 2024.

[57] Shitao Tang, Jiacheng Chen, Dilin Wang, Chengzhou Tang, Fuyang Zhang, Yuchen Fan, Vikas Chandra, Yasutaka Furukawa, and Rakesh Ranjan. Mvdiffusion++: A dense high-resolution multi-view diffusion model for single or sparse-view 3d object reconstruction. In *European Conference on Computer Vision*, pages 175–191. Springer, 2025. 3

[58] 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, 2025. 3

[59] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *arXiv preprint arXiv:2106.10689*, 2021. 4, 7, 1

[60] Peng Wang, Hao Tan, Sai Bi, Yinghao Xu, Fujun Luan, Kalyan Sunkavalli, Wenping Wang, Zexiang Xu, and Kai Zhang. Pf-lrm: Pose-free large reconstruction model for joint pose and shape prediction. *arXiv preprint arXiv:2311.12024*, 2023. 2, 3

[61] 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. *arXiv preprint arXiv:2305.16213*, 2023. 2, 3

[62] Shuang Wu, Youtian Lin, Feihu Zhang, Yifei Zeng, Jingxi Xu, Philip Torr, Xun Cao, and Yao Yao. Direct3d: Scalable image-to-3d generation via 3d latent diffusion transformer. *arXiv preprint arXiv:2405.14832*, 2024. 3, 1

[63] 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.

[64] 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

[65] Yinghao Xu, Hao Tan, Fujun Luan, Sai Bi, Peng Wang, Jiahao Li, Zifan Shi, Kalyan Sunkavalli, Gordon Wetzstein, Zexiang Xu, et al. Dmv3d: Denoising multi-view diffusion using 3d large reconstruction model. *arXiv preprint arXiv:2311.09217*, 2023. 3

[66] Hongyu Yan, Kunming Luo, Weiyu Li, Yixun Liang, Shengming Li, Jingwei Huang, Chunchao Guo, and Ping Tan. Posemaster: Generating 3d characters in arbitrary poses from a single image. *arXiv preprint arXiv:2506.21076*, 2025. 3

[67] Haibo Yang, Yang Chen, Yingwei Pan, Ting Yao, Zhineng Chen, Chong-Wah Ngo, and Tao Mei. Hi3d: Pursuing high-resolution image-to-3d generation with video diffusion models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 6870–6879, 2024. 3

[68] Xianghui Yang, Huiwen Shi, Bowen Zhang, Fan Yang, Jiacheng Wang, Hongxu Zhao, Xinhai Liu, Xinzhou Wang, Qingxiang Lin, Jiaao Yu, et al. Hunyuan3d 1.0: A unified framework for text-to-3d and image-to-3d generation. *arXiv preprint arXiv:2411.02293*, 2024. 3

[69] Taoran Yi, Jiemin Fang, Guanjun Wu, Lingxi Xie, Xiaopeng Zhang, Wenyu Liu, Qi Tian, and Xinggang Wang. Gaussiandreamer: Fast generation from text to 3d gaussian splatting with point cloud priors. *arXiv preprint arXiv:2310.08529*, 2023. 3

[70] Hongxu Yin, Arash Vahdat, Jose M Alvarez, Arun Mallya, Jan Kautz, and Pavlo Molchanov. A-vit: Adaptive tokens for efficient vision transformer. In *CVPR*, pages 10809–10818, 2022. 2, 3

[71] 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, 1, 2

[72] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 6

[73] Zhuofan Zong, Kunchang Li, Guanglu Song, Yali Wang, Yu Qiao, Biao Leng, and Yu Liu. Self-slimmed vision transformer. In *European Conference on Computer Vision*, pages 432–448. Springer, 2022. 3, 1