MVDiffusion++: A Dense High-resolution Multi-view Diffusion Model for Single or Sparse-view 3D Object Reconstruction

Shitao Tang[1*], Jiacheng Chen[1*], Dilin Wang[2*], Chengzhou Tang[2], Fuyang Zhang[1], Yuchen Fan[2], Vikas Chandra[2], Yasutaka Furukawa[1†], and Rakesh Ranjan[2†]

[1] Simon Fraser University
{shitaot, jca348, fuyangz, furukawa}@sfu.ca
[2] Meta Reality Labs
{wdilin, chengzhout, ycfan, vchandra, rakeshr}@meta.com

Abstract. This paper presents a neural architecture MVDiffusion++ for 3D object reconstruction that synthesizes dense and high-resolution views of an object given one or a few images without camera poses. MVDiffusion++ achieves superior flexibility and scalability with two surprisingly simple ideas: 1) A "pose-free architecture" where standard self-attention among 2D latent features learns 3D consistency across an arbitrary number of conditional and generation views without explicitly using camera pose information; and 2) A "view dropout strategy" that discards a substantial number of output views during training, which reduces the training-time memory footprint and enables dense and high-resolution view synthesis at test time. We use 3D objects for training and the Google Scanned Objects for evaluation with standard novel view synthesis and 3D reconstruction metrics, where MVDiffusion++ significantly outperforms the current state of the arts. We also demonstrate a text-to-3D application example by combining MVDiffusion++ with a text-to-image generative model.

1 Introduction

Human vision demonstrates remarkable flexibility. Look at the images of objects at the left in Figure 1. While unable to create millimeter-accurate 3D models, our visual system can combine information from a few images to form a coherent 3D representation in our minds, including intricate facial features of a tiger or the arrangement of blocks forming a toy train, even parts that are fully obscured.

3D reconstruction technology [1, 6, 29, 42] has evolved over the last fifteen years in a fundamentally different way. Unlike the human ability to infer 3D shapes from a few images, the technology takes hundreds of images of an object, estimates their precise camera parameters, and reconstructs high-fidelity 3D geometry at a sub-millimeter accuracy.

This paper explores a new paradigm of 3D reconstruction that combines the high-fidelity of computational methods and the flexibility of human visual systems.

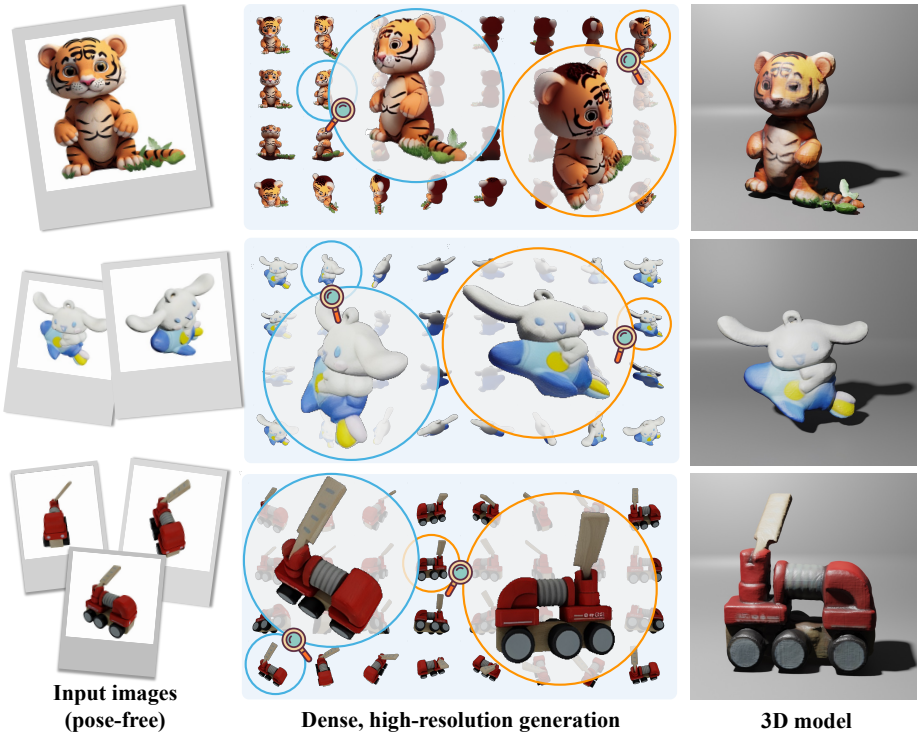

Fig. 1: MVDiffusion++ generates dense(32) and high-resolution(512×512) images of an object from a single or multiple unposed images. The input images of the three examples are from a latent diffusion model, OmniObject3D[38], and Google Scanned Objects[4], respectively.

Our inspiration comes from exciting recent developments in multi-view image generative models [14, 17, 18, 27, 28, 30, 34]. MVDiffusion [30] is an early attempt to extend pre-trained image diffusion models to a multi-view generative system, when pixel correspondences across views are available (e.g., generating perspective images to form a panorama). MVDream [28] and Wonder3D [18] further extend to more general settings where generated images yield 3D reconstruction via techniques such as NeRF [20] or NeuS [33].

This paper pushes the frontier of multi-view diffusion models towards flexible and high-fidelity 3D reconstruction systems. Concretely, the paper presents MVDiffusion++, a novel approach to generate dense (32) and high-resolution (512×512) images of an object, conditioned with single or sparse input views without camera poses, whose reliable estimation is difficult due to minimal or no visual overlaps. Standard 3D reconstruction techniques turn generated images into a 3D model. Two simple ideas are at the heart of our method. First, we leverage a latent diffusion inpainting model with conditional and generation branches, where self-attention among 2D features learns 3D consistency without

using camera poses or image projection formula. Second, we introduce "view dropout" training strategy, which randomly excludes generation views in each batch, enabling the use of high-resolution images during training. During testing, this simple approach surprisingly generates high-quality, dense views for all the images simultaneously.

MVDiffusion++ achieves state-of-the-art performance on the task of novel view synthesis, single-view reconstruction, and sparse-view reconstruction. For single-view reconstruction, our method achieves 0.6973 IoU and 0.0165 Chamfer distance on the Google Scanned Objects dataset, higher than SyncDreamer [17] by 0.1552 in terms of Vol. IOU. For novel view synthesis in sparse view setting, MVDiffusion++ improves the PSNR by 8.19 compared with a recent pose-free view synthesis method, LEAP [11]. Lastly, we demonstrate applications in text-to-3D by combining MVDiffusion++ with a text-to-image generative model.

2 Related work

This paper presents a multi-view image generative model for object reconstruction, given one or a few condition images. The section reviews related work on multi-view image generation and single to sparse-view 3D reconstruction techniques.

Multi-view image generation. The evolution of text-to-image diffusion models has paved the way for multi-view image generation. MVDiffusion [30] introduces an innovative multi-branch Unet architecture for denoising multi-view images simultaneously. This approach, however, is constrained to cases with one-to-one image correspondences. Syncdreamer [17] uses 3D volumes and depth-wise attention for maintaining multi-view consistency. MVDream [28] takes a different path, incorporating 3D self-attention to extend the work to more general cases. Similarly, Wonder3D [18] and Zero123++ [27] apply 3D self-attention to single-image conditioned multi-view image generation. These methods, while innovative, tend to produce sparse, low-resolution images due to the computational intensity of the attention mechanism. In contrast, our framework represents a more versatile solution capable of generating dense, high-resolution multi-view images conditioned on an arbitrary number of images.

Single view reconstruction. Single View Image Reconstruction is an active research area [17, 18, 21, 36, 39, 40], driven by the advancements of generative models [17, 18, 21, 36]. Large reconstruction model [10] and DMV3D [39] predict triplanes from a single image, but the 3D volume limits its resolutions. The other method, Syncdreamer [17] generates multi-view images with a latent diffusion model by constructing a cost volume. These images are then used to recover 3D structures using conventional reconstruction methods like Neus. However, this process requires substantial GPU memory, limiting it to low resolutions. Similarly, Wonder3D faces challenges due to the computational demands of self-attention, leading to similar restrictions. In contrast, our approach introduces a "view dropout" technique, which randomly samples a limited number of views for training in each iteration. This enables our model to generate a variable

number of high-resolution images while employing full 3D self-attention, effectively addressing the limitations faced by existing methods.

Sparse view reconstruction. Sparse View Image Reconstruction (SVIR) [11, 41] is a challenging task where only a limited number of images, typically two to ten, are given. Traditional 3D reconstruction methods estimate camera poses first, then perform dense reconstruction using techniques such as multi-view stereo [29, 42] or NeRF [33]. However, camera pose estimation is difficult for SVIR, where visual overlaps are none to minimal. To address this, FvOR [41] optimizes camera poses and shapes jointly. LEAP [11] along with PF-LRM [35] highlight the issues of noisy camera poses and suggest a pose-free approach. However, they are not based on generative models, lacking generative priors, and suffer from low-resolution outputs due to the use of volume rendering. In contrast, our method employs a diffusion model to generate high-resolution multi-view images directly, then a reconstruction system Neus [33] to recover a mesh model.

3 Preliminary: Multi-view latent diffusion models

MVDiffusion [30] is a multi-view latent diffusion model [17, 27, 28, 30], generating multiple images given a text or an image, when pixel-wise correspondences are available across views. MVDiffusion is the foundation of the proposed approach, where the section reviews its architecture and introduces notations (See Figure 2).

For generating eight perspective views forming a panorama, eight latent diffusion models (LDM) denoise eight noisy latent images $\{Z_1(t), Z_2(t), \cdots Z_8(t)\}$ simultaneously. A UNet is the core of a LDM model, consisting of a sequence of blocks through the four levels of the feature pyramid.

Let U_b^i denote the feature image of i-th image at b-th block. A CNN initializes an input U_i^0 from $Z_i(t)$ at the first block. Each UNet block has four network modules. The first is a novel correspondence-aware attention (CAA), enforcing consistency across views with visual overlaps: The left/right neighboring images (U_{i-1}^b, U_{i+1}^b) for panorama. The remaining three modules are from the original: 1) Self-attention (SA) layers; 2) Cross-attention (CA) layers from the condition with the CLIP embedding; and 3) CNN layers with the pixel-wise concatenation of a positional encoding of time $\tau(t)$. At test time, a standard DDPM sampler [9] updates all noisy latents with the predicted noise from the last CNN layer. The training objective is defined as follows by omitting the conditions for notation simplicity, where $\boldsymbol{\epsilon}^i$ is a Gaussian and ϵ_θ denotes the UNet output.

$$L_{MVLDM} := \mathbb{E}_{\{Z_i(0)\}_{i=1}^N, \{\boldsymbol{\epsilon}^i \sim \mathcal{N}(0,I)\}_{i=1}^N, t}\Big[\sum_{i=1}^{N} \|\boldsymbol{\epsilon}^i - \epsilon_\theta^i(\{Z_i(t)\}, \tau(t))\|_2^2\Big]. \quad (1)$$

4 MVDiffusion++

MVDiffusion++ pushes the frontier of multi-view diffusion models for 3D modeling in their *flexibility* and *scalability* by generating dense and higher-resolution images given an arbitrary number of un-posed condition views. With the prevalence

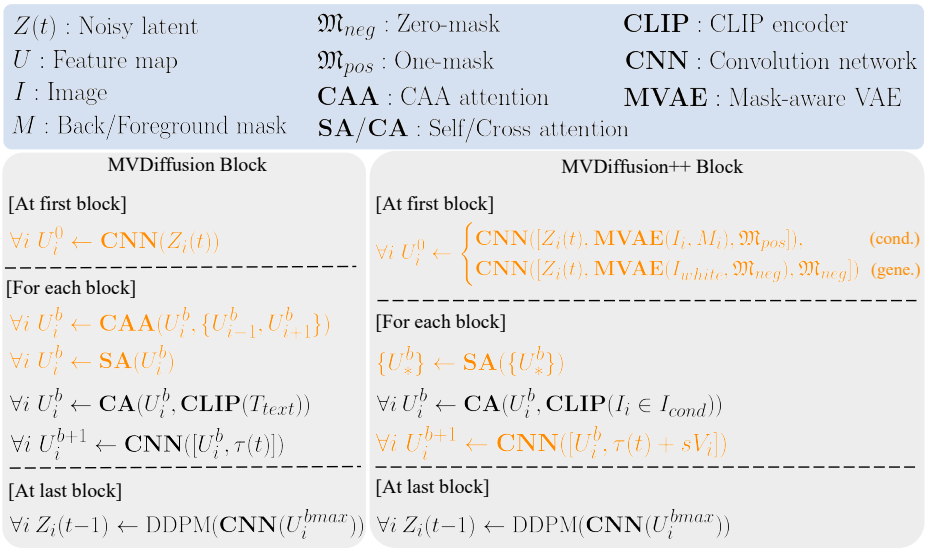

Fig. 2: The denoising architectures for MVDiffusion and MVDiffusion++for sampling multi-view images. The order of the MVDiffusion network modules is rearranged to highlight the differences (in orange) with MVDiffusion++.

of Transformer models [31], high-fidelity 3D modeling would require large-scale attention over dense and high-resolution image features, potentially with volumetric ones. Furthermore, 3D consistency learning is at the heart of the task, which would usually require precise image projection models and/or camera parameters. Our surprising discovery is that self-attention among 2D latent image features is all we need for 3D learning without projection models or camera parameters, and a simple training strategy would further achieve dense and high-resolution multi-view image generation. The section defines the task (i.e., input condition and output target images), then explains the two key ideas: 1) pose-free multi-view conditional diffusion model for flexibility and 2) view dropout training strategy for scalability. §5 provides the remaining system details.

4.1 Task: Input condition images and output target images

The generation target is a set of dense (32) and high-resolution (512×512) images, positioned at uniform 2D grid points on a sphere. Specifically, there are eight azimuth angles (every 45°) and four elevation angles (every 30° in the range $[-30°, 60°]$). Camera up-vectors are aligned with gravity, and their optical axes pass through the sphere center. Our input condition is one or a few images without camera poses, where visual overlaps are too minimal or possibly none for Structure from Motion algorithms to work reliably. The number of condition images is up to a pre-determined number, which is 10 in our experiments but

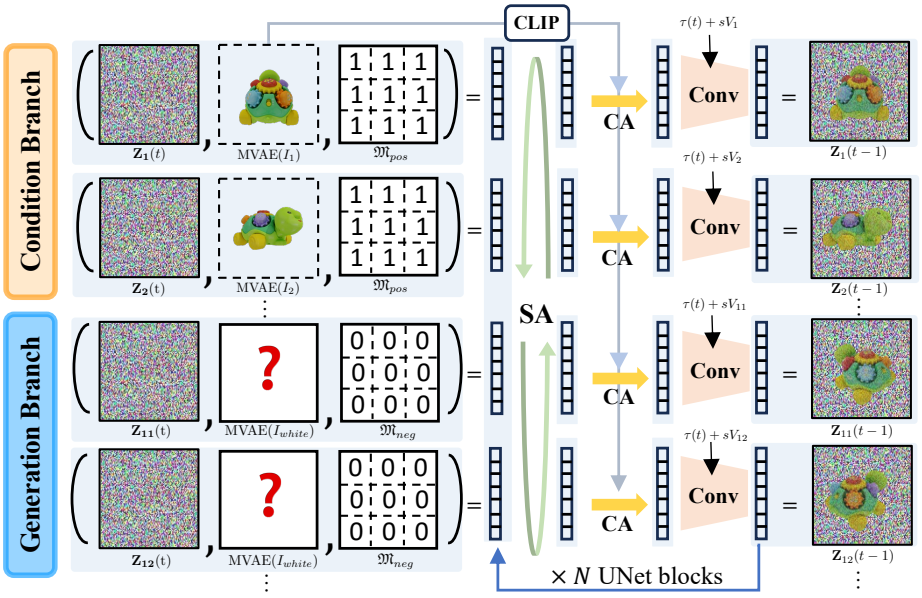

Fig. 3: Illustration of the pose-free multi-view conditional diffusion model of MVDiffusion++. The model takes any number of input images and generates images at fixed viewpoints. The condition branch and generation branch have different input configurations but share the same structure and weights.

can easily change. The input image resolution is 512×512. The horizontal and vertical field-of-view of both the input and output views is 60°.

We use synthetic rendered images from 3D object databases for training and evaluations. The task settings vary slightly between datasets, with details provided in §5. Here, we explain one preprocessing step that removes ambiguity in the training task. 3D object databases and Google Scanned Object [4] align the Z-axis with the object up-vectors. However, the azimuth of the ground-truth object pose is ambiguous without camera poses of the condition images. Therefore, we rotate the output views to align the azimuth of the first condition and the first output image.

4.2 Pose-free multi-view conditional diffusion model

MVDiffusion++ is a multi-view latent diffusion model as defined in §3, comprising of a *condition branch* for single or sparse-view input images and a *generation branch* for output images (See Figure 2 and Figure 3). Note that the condition branch shares the same architecture and is tasked to generate the condition images that are also given as guidance (i.e., a trivial task).

Diffusion process. The forward diffusion process is the same as MVDiffusion, except for the image resolution and the pre-trained VAE. Concretely, it 1) converts

all $512 \times 512 \times 3$ input/output image (I_i) with foreground masks to $64 \times 64 \times 4$ latent images (Z_i) by a fine-tuned latent diffusion VAE (denoted as MVAE, see §5 for the fine-tuning process); and 2) adds a Gaussian noise with a linear schedule, as suggested by zero-123++ [27] to each feature of Z_i.

Denoising process. The denoising process is highlighted in Figure 2, where a latent diffusion UNet with a few modifications processes a noisy latent $Z_i(t)$ at each denoising step t. The UNet consists of 9 blocks of network modules over the four levels of feature pyramids on either side of the encoder/decoder. The details are explained as follows.

[At first block] The UNet feature U_i^0 at the first block is initialized with the concatenation of 1) the noisy latents Z_i; 2) a constant binary mask of either 1 or 0, denoted by $\mathfrak{M}_{pos}$ or $\mathfrak{M}_{neg}$ to indicate the branch type (condition or generation); and 3) the condition latents (MVAE(I_i, M_i)) where we use the conditonal VAE from latent diffusion to encode the condition image (I_i) with its segmentation mask (M_i). Note that this concatenation has $9 = (4 + 4 + 1)$ channels, and a 1×1 final convolution layer reduces the channel dimension to 4. For a generation branch, we pass a white image as I_i and a binary image of 1 (i.e., $\mathfrak{M}_{pos}$) as M_i. For training 3D objects and Google Scaned Object datasets, we use the masks provided by the datasets. Otherwise, we run segmentation to generate the masks.

[For each block] Three network modules process the input: 1) Global self-attention mechanism among the UNet features across all the images, learning 3D consistency; 2) Cross-attention mechanism, injecting the CLIP embedding of the condition images to all the other images through the CLIP embedding; and 3) CNN layers, process per-image features while injecting the timestep frequency encoding $\tau(t)$ and the learnable embedding of an image index V_i. For the self-attention module, we copy the network architecture and model weights and apply it across all the views. This module is inspired by MVDream [28], while the key differences in our work are 1) Scalability deployment via the view-drop training strategy in §4.3; and 2) Handling of multiple condition images without camera poses via the network design. $42 = (32 + 10)$ learnable embedding vectors $\{V_i\}$ are trained for 32 generation and 10 condition images, each of which is multiplied with a zero-initialized trainable scale s to avoid model disruption at initailization.

[At last block] The output of the last UNet block yields the noise estimation, and a standard DDPM sampler [9] takes it to produce the noisy latent of the next timestep $Z_i(t-1)$ for each sampling step. The loss function is the same as MVDiffusion. Note that the model is first trained with ϵ-prediction and then with v-prediction (See §5), where Equation 1 is the loss function for the ϵ-prediction model. The velocity [26], $\boldsymbol{v}^i(t) = \alpha_t \boldsymbol{\epsilon}^i - \gamma_t Z_i(0)$, becomes the prediction target for the v-prediction model, while α_t and γ_t are predefined angular parameters.

4.3 View dropout training strategy

MVDiffusion++ training would face a scalability challenge. $42 (= 32 + 10)$ copies of UNet features yield more than 130k tokens, where the global self-attention

mechanism becomes infeasible even with the latest memory efficient transformers for large language models [2, 3]. We propose a simple yet surprisingly effective *view dropout* training strategy, which completely discards a set of views across all layers during training. Specifically, we randomly drop 24 out of 32 views for each object at each training iteration, significantly reducing memory consumption at training. At test time, we run the entire architecture and generate 32 views.

5 Remaining system details

This section explains the remaining system details on the data preparations, the mesh extraction process, the MVAE pre-fine-tuning, and the three-stage training strategy.

5.1 Training data preparation

We use 180k models whose aesthetic scores [22] are at least 5 for training. For each object 3D model, we translate the bounding box center to the origin and apply uniform scaling so that the longest dimension matches $[-1, 1]$. The output camera centers are placed at a distance of 1.5 from the origin. Input condition views are chosen in a similar way as Zero-123 [16]. Concretely, an azimuth angle is randomly chosen from one of the eight discrete angles of the output cameras (also see §4.1). The elevation angle is set randomly from [-10°, 45°]. The distance of the camera center from the origin is set randomly from [1.5, 2.2]. We use Blender to render images.

5.2 Testing data preparation

Single-view cases. Google Scanned Object (GSO) [4] is our testing dataset, where we borrow the rendered images and the evaluation pipeline from SyncDreamer [17]. Concretely, the test set consists of 30 objects. Each object has 16 images with a fixed elevation of 30° and every 22.5° for azimuth. SyncDreamer selected condition images by "visual plausibility", which we copy. The details are provided in the supplementary. Since the azimuth angles in our training setting are every 45°, eight images (starting from and including the condition image) are used for evaluation. The resolution of the rendered images is 256x256, while the image resolution of our architecture is 512x512. We upscale the condition images to 512x512 for our system inputs. The ground-truth images are 256x256 and we downscale our generated images to 256x256 for evaluation, while 512x512 images are used for the mesh reconstruction. The Chamfer Distances (CD) and volume IoU between the ground-truth and reconstructed shapes are reported for single-view 3D reconstruction. The PSNR, SSIM [37], and LPIPS [44] are reported for novel view synthesis (NVS) by averaging over the eight images.

Sparse-view cases. Sparse-view un-posed condition is a new setup (except the work of LEAP [11] and PF-LRM [35] to our knowledge). We use a process

similar to the single-view setting to render images. Concretely, we first render 10 condition images for each of the 30 GSO objects. The azimuth and the elevation angles are chosen randomly from [0, 360) and [-10, 45] respectively. We render 32 ground-truth target images while aligning the azimuth of the first target view and the first input view (See §4.1). The same evaluation metrics are used, while we vary the number of condition images to be 1, 2, 4 and 10.

5.3 Mesh extraction from generated images

After generating 32 images, a neural implicit reconstruction method recovers a mesh model, similar to SyncDreamer [17] and Wonder3D [18]. Specifically, we use grid-based NeuS [7, 13], where the foreground masks are decoded from the latent images $\{Z_i(0)\}$ by MVAE. Since our generated images have high resolution and quality, we directly run the monocular normal estimator released by Omnidata [5] to obtain additional normal supervisions for NeuS without a normal generation module like Wonder3D. We borrow the NeuS implementation from Wonder3D's official codebase but do not use their ranking-based loss. With a single Nvidia 2080 Ti, it takes around 3 minutes to reconstruct a textured mesh model. The mesh could directly use the exported vertex color or be re-textured with the generated images.

5.4 Mask-aware VAE pre-fine-tuning

We copy the network architecture and model weights of the default VAE and add additional input and output channels to handle the mask. We found that fine-tuning Mask-aware VAE (M-VAE) only with object images improves performance. Concretely, we use approximately 3 million RGBA images rendered from a collection of 3D objects to fine-tune M-VAE as a pre-processing. We follow the original VAE hyperparameters with a base learning rate of 4.5e-6 and a batch size of 64. The training runs for 60,000 iterations. The binary cross entropy loss is used for the mask channel. The process improves PSNR from 36.6 to 41.2.

6 Experiments

We train the model with a batch size of 1024 using 128 Nvidia H100 GPUs for about a week. At test time, we use DDPM [9] sampler with 75 steps to sample the multi-view images, and it takes our model 30s, 77s, 123s, and 181s to generate 8, 16, 24, and 32 images, respectively. The section presents the single view experiments in §6.1, the sparse view experiments in §6.2, and text-to-3D application experiments in §6.3.

6.1 Single-view object modeling

Three state-of-the-art single-view object modeling methods are our main baselines: SyncDreamer [17], Wonder3D [18], and Open-LRM [8]. Since the evaluation

Table 1: Single-view object modeling results, evaluating reconstructed meshes (left) and generated images (right). The ground-truth meshes and images are prepared by SyncDreamer [17] based on the Google Scanned Object [4] dataset. ICP is necessary to align reconstructed meshes for methods marked with *.

Task →	3D reconstruction		Novel view synthesis		
Method	Chamfer Dist.↓	Vol. IoU↑	PSNR↑	SSIM↑	LPIPS↓
Realfusion [19]	0.0819	0.2741	15.26	0.722	0.283
Magic123 [25]	0.0516	0.4528	-	-	-
One-2-3-45 [15]	0.0629	0.4086	-	-	-
Point-E [23]	0.0426	0.2875	-	-	-
Shap-E [12]	0.0436	0.3584	-	-	-
Zero123 [16]	0.0339	0.5035	18.93	0.779	0.166
SyncDreamer [17]	0.0261	0.5421	20.05	0.798	0.146
Wonder3D [18]*	0.0329	0.5768	-	-	-
Open-LRM [8]*	0.0285	0.5945	-	-	-
Ours	**0.0165**	**0.6973**	**21.45**	**0.844**	**0.129**

pipeline is the same as SyncDreamer, we copy numbers of other baselines in their paper for comparison, which includes Zero123 [16], RealFusion [19], Magic123 [25], One-2-3-45 [15], Point-E [23], and Shap-E [12]. The following introduces the three main baselines and how we reproduce their systems:

• *SyncDreamer* generates 16 images from fixed viewpoints given a single input image. The image resolution is 256x256. Their denoising network ϵ_θ initializes from Zero123 and leverages 3D feature volumes and depth-wise attention to learn multi-view consistency. It requires users to provide the elevation of the input image.

• *Wonder3D* takes a single input image as the canonical view and generates 6 images as well as the normal maps. The image resolution is 256×256. Multi-view self-attention and an extra cross-domain attention ensure the consistency of generation results, while the views are sparser than ours. We run the official codebase on the GSO input images to get the results. However, the released model assumes orthographic cameras and we cannot use the same test set to evaluate the NVS performance. ICP aligns the reconstructed mesh with the ground truth before computing the metrics.

• *Open-LRM* is an open-source implementation of Large Reconstruction Model (LRM) [10], a generalized reconstruction model that predicts a triplane NeRF from a single input image using a feed-forward transformer-based network. ICP aligns the reconstructed mesh with the ground truth before computing the CD and volume IoU.

Results. Table 1 presents the quantitative evaluations of the reconstructed 3D meshes and the generated images. MVDiffusion++ consistently outperforms all the competing methods with clear margins. Note that the evaluation is not completely fair for Wonder3D that assumes orthographic camera projections,

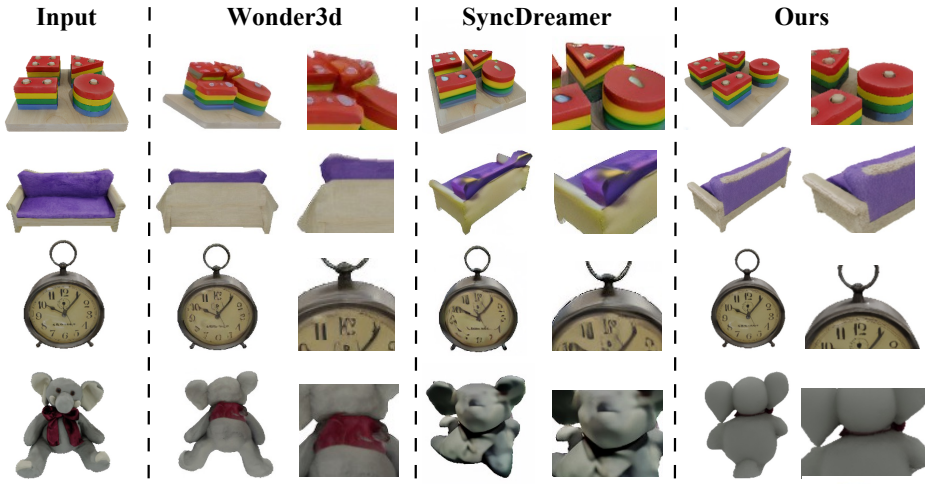

Fig. 4: Single-view object modeling results of generated images. The input image and the generated images by Wonder3D and SyncDreamer are in 256×256. Our rendered images are in 512×512, showing higher fidelity and richer details.

where perspective images are used in the experiments. However, we believe the clear performance gaps suffice to demonstrate the strength of our method.

Figure 4 and Figure 5 show generated images and reconstructed mesh models. In Figure 4, our method clearly shows the number on the clock (row 3), while others exhibit blurry numbers. In Figure 5, our method can recover a plausible and detailed shape of the turtle example (row 1), while Wonder3D and OpenLRM fail to recognize it as a turtle and exhibit significant artifacts.

6.2 Sparse-view object modeling

Sparse-view un-posed input images is a challenging setting, where we are aware of only a few existing approaches such as LEAP [11] and PF-LRM [35], a sparse-view pose-free extension of LRM [10]. There is no public implementation of PF-LRM, and we pick LEAP as the first baseline. The literature on multi-view 3D reconstruction is extensive. It would be valuable to contrast our approach, even though they require camera poses as input. As a compromise, we have selected NeuS [33] as our second benchmark by providing the ground-truth camera poses as their input.

• *LEAP* leverages a transformer to predict neural volumes of radiance fields from a sparse number of views and is also pose-free. LEAP employs DINOv2 [24] as the feature extractor and has reasonable generalization capacity.

• *NeuS* is a 3D reconstruction method, where we provide the ground-truth camera poses of the condition images as well as surface normals estimated by Omnidata's monocular normal estimator [5]. We use the public grid-based NeuS implementation [7]. This baseline is similar to MonoSDF [43] or NeurIS [32]

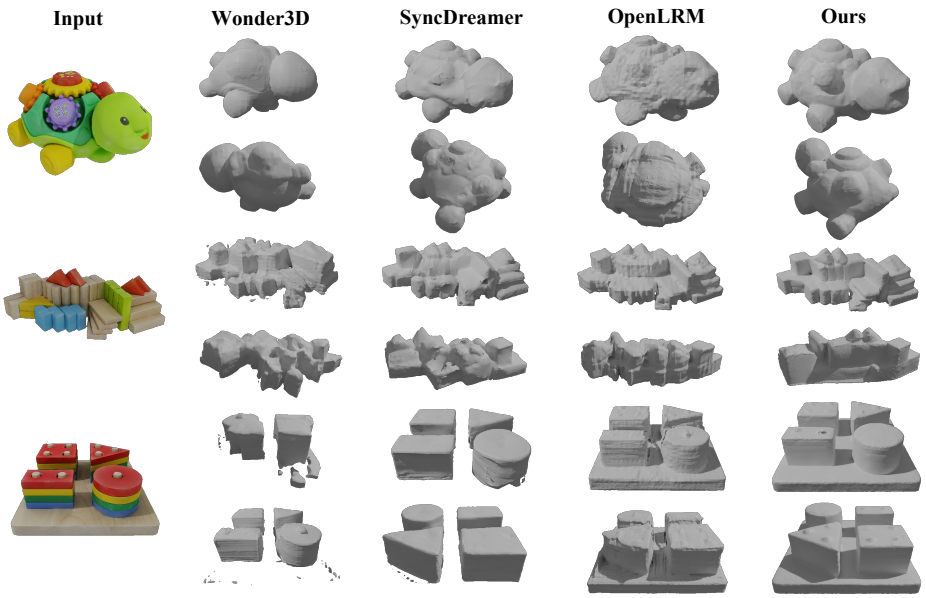

Fig. 5: Single-view object modeling results of reconstructed mesh models. Our meshes are exported from dense (32) and high-resolution (512×512) generated images, demonstrating finer details.

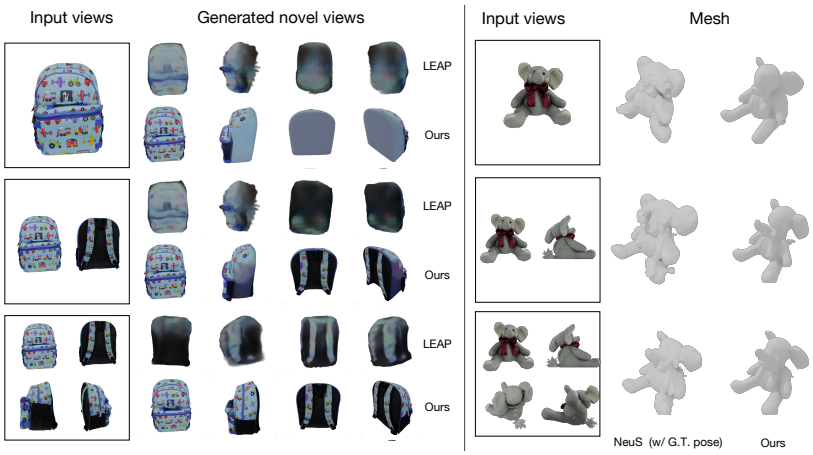

Fig. 6: Novel view synthesis and 3D reconstruction with sparse-view input images. **Left**: a qualitative example of novel view synthesis, comparing LEAP [11] and MVDiffusion++ with different numbers of unposed input images. **Right**: qualitative comparison of reconstructed meshes between NeuS [33] with ground-truth relative poses and our pose-free MVDiffusion++.

Table 2: Sparse-view object modeling results, evaluating reconstructed meshes (left) and generated images (right), based on the GSO [4] dataset.

Method	Views	Chamfer Dist.↓	Vol. IoU↑
Sync-Dreamer	1	0.0318	0.5610
NeuS[33] (G.T. pose)	1	0.0536	0.4400
	2	0.0307	0.5884
	4	0.0158	0.7323
	10	0.0096	0.8092
Ours	1	0.0208	0.6689
	2	0.0158	0.7260
	4	0.0122	0.7737
	10	0.0101	0.8046

Method	Views	PSNR↑	SSIM↑	LPIPS↓
Sync-Dreamer	1	19.46	0.847	0.188
LEAP[11]	1	14.66	0.47	0.43
	2	16.22	0.59	0.36
	4	16.54	0.61	0.35
	10	16.84	0.64	0.34
Ours	1	20.25	0.862	0.157
	2	21.73	0.872	0.137
	4	23.44	0.886	0.117
	10	25.03	0.899	0.102

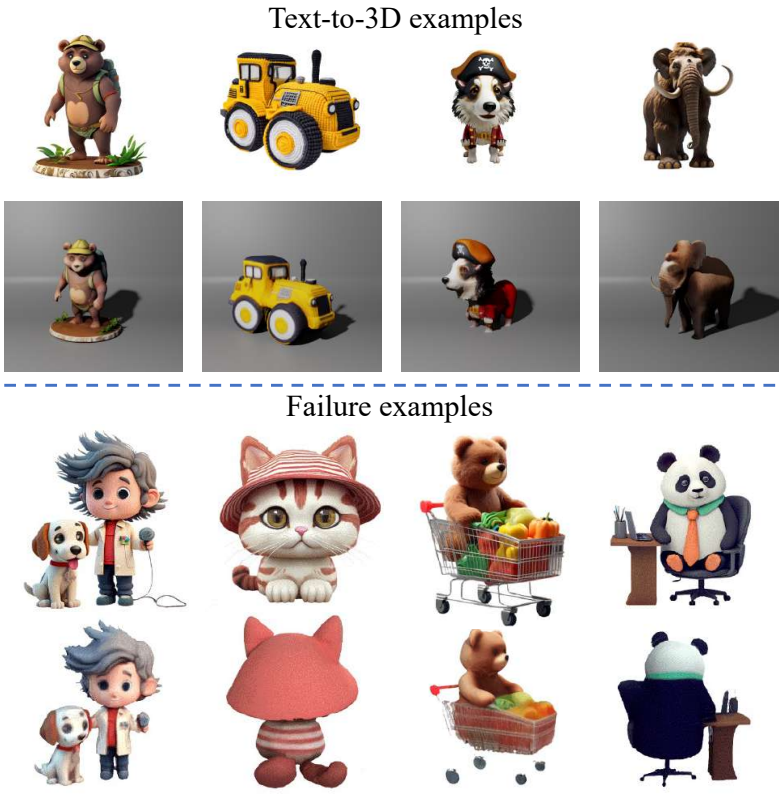

Fig. 7: Text-to-3D application examples. (Top) A text-to-image model generates an image given a text prompt. (Bottom) MVDiffusion++ turns the generated image into a 3D model. We also show some failure examples at the bottom.

equipped with ground-truth foreground masks and camera poses, thus sets a performance upper bound for methods without generative priors.

Results. Table 2 and Figure 6 present the quantitative and qualitative comparison results, respectively. Compared to LEAP, MVDiffusion++ generates images with much better quality. LEAP and our method both exploit multi-view self-attention to establish global 3D consistency. Therefore, we attribute our better performance to the strong image priors inherited from the pre-trained latent diffusion models. Our reconstructed meshes outperform NeuS in most settings, a notable achievement considering that NeuS uses ground-truth camera poses. This comparison highlights the practicality of our method, enabling users to achieve high-quality 3D models from just a few object snapshots.

6.3 Text-to-3D application

MVDiffusion++ shows consistent performance with minimal errors on the GSO dataset and achieves remarkable generalization capabilities. To further challenge the system, we demonstrate a text-to-3D application, where a text-to-image model prepares an input condition image. MVDiffusion++ turns the condition image into a 3D model. Figure 7 has four examples demonstrating the power of our approach.

7 Limitations and future challenges

This paper presents a pose-free technique for reconstructing objects using an arbitrary number of images. Central to this approach is a sophisticated multi-branch, multi-view diffusion model. This model processes any number of conditional images to produce dense, consistent views from fixed perspectives. This capability significantly enhances the performance of existing reconstruction algorithms, enabling them to generate high-quality 3D models. Our results show that MVDiffusion++ sets a new standard in performance for both single-view and sparse-view object reconstruction.

Figure 7 presents typical failure modes and the limitations of our approach. Our method struggles with thin structures as in the leftmost example, which fails to reconstruct a cable. Our method occasionally generates implausible images for views occluded in the input, a notable instance being the depiction of a cat with two tails. These shortcomings are predominantly attributed to the lack of training data, where one future work will expand the framework to incorporate videos, which offer richer contextual and spatial information, potentially enabling dynamic video generation.

Acknowledgements. This research is partially supported by NSERC Discovery Grants, NSERC Alliance Grants, and John R. Evans Leaders Fund (JELF). We thank the Digital Research Alliance of Canada and BC DRI Group for providing computational resources.

Bibliography

[1] Agarwal, S., Furukawa, Y., Snavely, N., Simon, I., Curless, B., Seitz, S.M., Szeliski, R.: Building rome in a day. Communications of the ACM **54**(10), 105–112 (2011)

[2] Dao, T.: Flashattention-2: Faster attention with better parallelism and work partitioning. arXiv preprint arXiv:2307.08691 (2023)

[3] Dao, T., Fu, D., Ermon, S., Rudra, A., Ré, C.: Flashattention: Fast and memory-efficient exact attention with io-awareness. Advances in Neural Information Processing Systems **35**, 16344–16359 (2022)

[4] Downs, L., Francis, A., Koenig, N., Kinman, B., Hickman, R., Reymann, K., McHugh, T.B., Vanhoucke, V.: Google scanned objects: A high-quality dataset of 3d scanned household items. In: ICRA (2022)

[5] Eftekhar, A., Sax, A., Malik, J., Zamir, A.: Omnidata: A scalable pipeline for making multi-task mid-level vision datasets from 3d scans. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10786–10796 (2021)

[6] Furukawa, Y., Curless, B., Seitz, S.M., Szeliski, R.: Towards internet-scale multi-view stereo. In: 2010 IEEE computer society conference on computer vision and pattern recognition. pp. 1434–1441. IEEE (2010)

[7] Guo, Y.C.: Instant neural surface reconstruction (2022), https://github.com/bennyguo/instant-nsr-pl

[8] He, Z., Wang, T.: Openlrm: Open-source large reconstruction models. `https://github.com/3DTopia/OpenLRM` (2023)

[9] Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)

[10] Hong, Y., Zhang, K., Gu, J., Bi, S., Zhou, Y., Liu, D., Liu, F., Sunkavalli, K., Bui, T., Tan, H.: Lrm: Large reconstruction model for single image to 3d. arXiv preprint arXiv:2311.04400 (2023)

[11] Jiang, H., Jiang, Z., Zhao, Y., Huang, Q.: Leap: Liberate sparse-view 3d modeling from camera poses. arXiv preprint arXiv:2310.01410 (2023)

[12] Jun, H., Nichol, A.: Shap-e: Generating conditional 3d implicit functions. arXiv preprint arXiv:2305.02463 (2023)

[13] Li, Z., Müller, T., Evans, A., Taylor, R.H., Unberath, M., Liu, M.Y., Lin, C.H.: Neuralangelo: High-fidelity neural surface reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8456–8465 (2023)

[14] Liu, M., Shi, R., Chen, L., Zhang, Z., Xu, C., Wei, X., Chen, H., Zeng, C., Gu, J., Su, H.: One-2-3-45++: Fast single image to 3d objects with consistent multi-view generation and 3d diffusion. arXiv preprint arXiv:2311.07885 (2023)

[15] Liu, M., Xu, C., Jin, H., Chen, L., Xu, Z., Su, H.: One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization. arXiv preprint arXiv:2306.16928 (2023)

[16] Liu, R., Wu, R., Van Hoorick, B., Tokmakov, P., Zakharov, S., Vondrick, C.: Zero-1-to-3: Zero-shot one image to 3d object. In: ICCV (2023)
[17] Liu, Y., Lin, C., Zeng, Z., Long, X., Liu, L., Komura, T., Wang, W.: Syncdreamer: Generating multiview-consistent images from a single-view image. arXiv preprint arXiv:2309.03453 (2023)
[18] Long, X., Guo, Y.C., Lin, C., Liu, Y., Dou, Z., Liu, L., Ma, Y., Zhang, S.H., Habermann, M., Theobalt, C., et al.: Wonder3d: Single image to 3d using cross-domain diffusion. arXiv preprint arXiv:2310.15008 (2023)
[19] Melas-Kyriazi, L., Laina, I., Rupprecht, C., Vedaldi, A.: Realfusion: 360deg reconstruction of any object from a single image. In: CVPR (2023)
[20] Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. In: ECCV (2020)
[21] Mittal, P., Cheng, Y.C., Singh, M., Tulsiani, S.: Autosdf: Shape priors for 3d completion, reconstruction and generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 306–315 (2022)
[22] Murray, N., Marchesotti, L., Perronnin, F.: Ava: A large-scale database for aesthetic visual analysis. In: 2012 IEEE conference on computer vision and pattern recognition. pp. 2408–2415. IEEE (2012)
[23] Nichol, A., Jun, H., Dhariwal, P., Mishkin, P., Chen, M.: Point-e: A system for generating 3d point clouds from complex prompts. arXiv preprint arXiv:2212.08751 (2022)
[24] Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
[25] Qian, G., Mai, J., Hamdi, A., Ren, J., Siarohin, A., Li, B., Lee, H.Y., Skorokhodov, I., Wonka, P., Tulyakov, S., et al.: Magic123: One image to high-quality 3d object generation using both 2d and 3d diffusion priors. arXiv preprint arXiv:2306.17843 (2023)
[26] Salimans, T., Ho, J.: Progressive distillation for fast sampling of diffusion models. arXiv preprint arXiv:2202.00512 (2022)
[27] Shi, R., Chen, H., Zhang, Z., Liu, M., Xu, C., Wei, X., Chen, L., Zeng, C., Su, H.: Zero123++: a single image to consistent multi-view diffusion base model. arXiv preprint arXiv:2310.15110 (2023)
[28] Shi, Y., Wang, P., Ye, J., Long, M., Li, K., Yang, X.: Mvdream: Multi-view diffusion for 3d generation. arXiv preprint arXiv:2308.16512 (2023)
[29] Stereopsis, R.M.: Accurate, dense, and robust multiview stereopsis. IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE **32**(8) (2010)
[30] Tang, S., Zhang, F., Chen, J., Wang, P., Furukawa, Y.: Mvdiffusion: Enabling holistic multi-view image generation with correspondence-aware diffusion. arXiv preprint arXiv:2307.01097 (2023)
[31] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)

[32] Wang, J., Wang, P., Long, X., Theobalt, C., Komura, T., Liu, L., Wang, W.: Neuris: Neural reconstruction of indoor scenes using normal priors. In: European Conference on Computer Vision. pp. 139–155. Springer (2022)
[33] Wang, P., Liu, L., Liu, Y., Theobalt, C., Komura, T., Wang, W.: Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. In: NeurIPS (2021)
[34] Wang, P., Shi, Y.: Imagedream: Image-prompt multi-view diffusion for 3d generation. arXiv preprint arXiv:2312.02201 (2023)
[35] Wang, P., Tan, H., Bi, S., Xu, Y., Luan, F., Sunkavalli, K., Wang, W., Xu, Z., Zhang, K.: Pf-lrm: Pose-free large reconstruction model for joint pose and shape prediction. arXiv preprint arXiv:2311.12024 (2023)
[36] Wang, Y., Lira, W., Wang, W., Mahdavi-Amiri, A., Zhang, H.: Slice3d: Multi-slice, occlusion-revealing, single view 3d reconstruction. arXiv preprint arXiv:2312.02221 (2023)
[37] Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. TIP (2004)
[38] Wu, T., Zhang, J., Fu, X., Wang, Y., Ren, J., Pan, L., Wu, W., Yang, L., Wang, J., Qian, C., et al.: Omniobject3d: Large-vocabulary 3d object dataset for realistic perception, reconstruction and generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 803–814 (2023)
[39] Xu, Y., Tan, H., Luan, F., Bi, S., Wang, P., Li, J., Shi, Z., Sunkavalli, K., Wetzstein, G., Xu, Z., et al.: Dmv3d: Denoising multi-view diffusion using 3d large reconstruction model. arXiv preprint arXiv:2311.09217 (2023)
[40] Yan, X., Yang, J., Yumer, E., Guo, Y., Lee, H.: Perspective transformer nets: Learning single-view 3d object reconstruction without 3d supervision. Advances in neural information processing systems **29** (2016)
[41] Yang, Z., Ren, Z., Bautista, M.A., Zhang, Z., Shan, Q., Huang, Q.: Fvor: Robust joint shape and pose optimization for few-view object reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2497–2507 (2022)
[42] Yao, Y., Luo, Z., Li, S., Fang, T., Quan, L.: Mvsnet: Depth inference for unstructured multi-view stereo. In: Proceedings of the European conference on computer vision (ECCV). pp. 767–783 (2018)
[43] Yu, Z., Peng, S., Niemeyer, M., Sattler, T., Geiger, A.: Monosdf: Exploring monocular geometric cues for neural implicit surface reconstruction. Advances in neural information processing systems **35**, 25018–25032 (2022)
[44] Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR (2018)