Pixel-Aligned Multi-View Generation with Depth Guided Decoder

Zhenggang Tang,[1] Peiye Zhuang,[2] Chaoyang Wang,[2] Aliaksandr Siarohin,[2] Yash Kant[3]
Alexander Schwing,[1] Sergey Tulyakov[2], Hsin-Ying Lee[2]
[1]University of Illinois Urbana-Champaign [3]Snap Inc. [2]University of Toronto

1. Introduction

Multi-view images that show an object from a small number of different viewpoints, have emerged as a commonly used auxiliary representation in 3D generation.

Despite promising results, it is challenging for current multi-view generation methods to achieve pixel-level image alignment across views. The coarse alignment achieved with current methods introduces ambiguity in subsequently employed reconstruction methods. as shown in Fig. 1, irrespective of whether per-instance optimization [1] or feed-forward methods [2] are used for 3D generation, they get blurred results due to pixel-level misalignment. Pixel-level alignment issues arise because existing multi-view diffusion models are mostly fine-tuned from an image diffusion model with additional multi-view attention [3, 4] or an intermediate implicit 3D representation [5, 6]. Notably, the diffusion process occurs in a latent space with limited resolution, and decoding is performed independently for each frame without cross-view communication, making pixel-level alignment difficult. To improve, some multi-view diffusion models are fine-tuned from video diffusion models with camera trajectory control [7, 8]. Although the multi-view latents are jointly decoded using a video decoder, achieving pixel-level alignment remains challenging due to the sparsity of adjacent multi-view frames.

In this work, we propose to address the pixel-level alignment issue by improving existing VAE decoders. Following prior multi-view generation works, We adopt the VAE decoder from Stable Video Diffusion [9] as our backbone. Differently, to enable cross-view attention at higher-resolution and and achieve better pixel-level multiview alignment, we modify the VAE decoder in two ways: First, we propose a depth-truncated epipolar attention mechanism applied to high-resolution layers. This attention mechanism extracts cross view features that are crucial for better feature alignment. However, the depth information is not available during inference. Moreover, the multi-view latents are often not accurately aligned. Second, to solve this, we augment data with structured-noise depth to mitigate the domain gap between training and inference. We propose to augment data with structured-noise depth, appending both high- and low-frequency noise to the ground-truth depth. Then during inference, we simply employ depth predicted by an off-the-shelf multi-view 3D reconstruction method [1]. This is feasible as we obtain a model that is more robust to imperfect predictions.

We conduct extensive qualitative and quantitative experiments against baseline methods. We visually compare with other multi-view generation methods by adopting the same 3D reconstruction methods [1] and quantitatively measure PSNR, SSIM, LPIPS, and the number of correspondences, on the reconstructed 3D objects. The proposed method performs favorably against existing state-of-the-art multi-view generation methods.

2. Related work

3D generation. To utilize pre-trained image diffusion models for 3D generation, Score Distillation Sampling (SDS) [10] and its variants [11, 12] have been proposed to distill knowledge from 2D models in a per-instance optimization manner, taking minutes to hours for each generation. Recently, to avoid time-consuming optimization, feed-forward methods [13, 14] have emerged. They use multi-view images as an auxiliary representation followed by 3D reconstruction. In this work, we focus on pixel-aligned multi-view image generation to facilitate better 3D reconstruction, ultimately leading to better 3D generation.

Multi-view image generation. To adapt from large-scale image [15] or video [7] diffusion models by finetuning, multi-view diffusion models incorporate multi-view cross attention [3, 16] or adopt intermediate 3D representations like voxels [5] or a triplane-based neural radiance field (NeRF) [6]. However, irrespective of the approach, existing efforts still struggle to synthesize pixel-level aligned multi-view images.

3. Method

3.1. Overview

We aim to generate multi-view images with better pixel-level alignment (Fig. 2). For this, we focus on improving the decoder of a latent diffusion model, Stable Video Diffusion

Figure 1. **Visualization of our method**. Comparing to the baseline methods (column 4-7, 10-11), our proposed method enables to generate pixel-aligned multi-view images, which can lead to better 3D reconstruction quality.

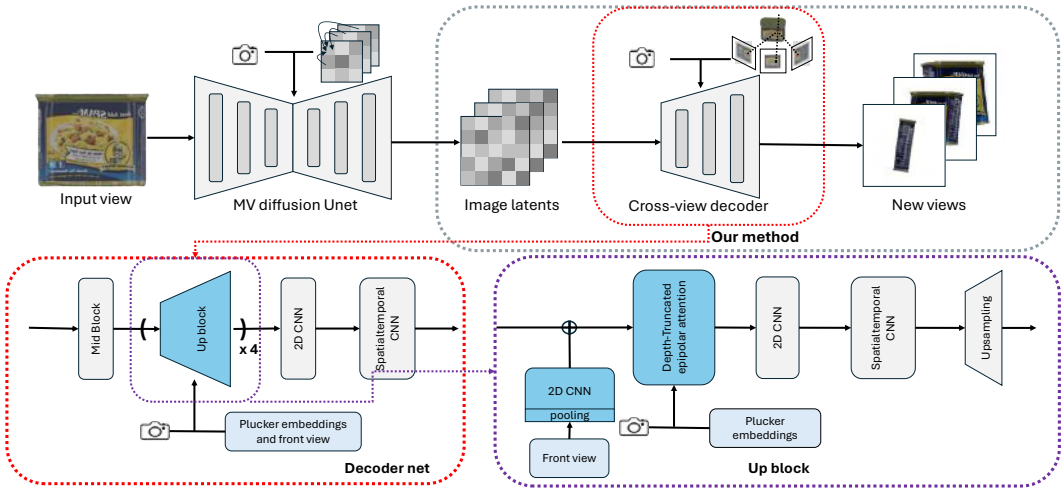

Figure 2. **Overview.** (top) We focus on improving the decoder for pixel-aligned multi-view image generation. (bottom-left) The decoder contains four Up-blocks to upsample the resolution from 32 to 256. (bottom-right) We propose several additions, highlighted with blue color: a condition from the input front-view image and a depth-truncated epipolar attention mechanism.

(SVD) [9].

First we propose a depth-truncated epipolar attention mechanism (Section 3.2, Fig.3). It aggregates features from multi-view latents by making use of depth information. To further mitigate the domain gap between the ground-truth depth used in training and the predicted depth used in inference, we propose a structured-noise depth augmentation strategy (Section 3.3). The strategy can also help handle the imperfect generated multi-view latents during inference. Implementation details are in Section 3.4.

3.2. Depth-truncated epipolar attention

To generate pixel-level aligned multi-view images, the decoder needs to gather and process information from multi-view latents. An epipolar attention mechanism is an excellent candidate for this task, because it permits to combine information from corresponding points across views. Importantly, to attain a more accurate pixel-level information exchange, the attention is preferably applied to any resolution, particularly also on higher resolution latents. However, a vanilla epipolar attention mechanism often spreads too much attention on irrelevant parts, which makes it difficult for the network to learn to extract the correct adjacent features. Moreover, it also consumes a lot of memory and easily leads to out-of-memory errors given current hardware memory constraints, even on high-end equipment. To address this, we propose a depth-truncated epipolar attention mechanism. This approach not only aggregates multi-view information at higher resolutions, but also further improves the quality by enabling the model to focus on crucial regions.

Concretely, consider a feature map from a referenced view $F^{\mathbf{ref}}$, and feature maps from N_v other views $\{F^j\}_{j \neq \mathbf{ref}}$. For a source point s on $F^{\mathbf{ref}}$, we can get the

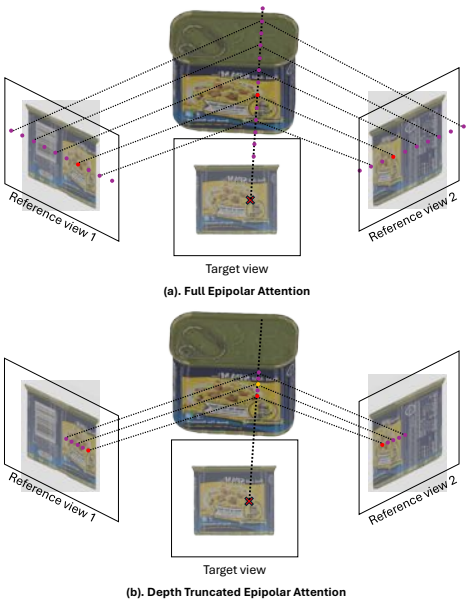

Figure 3. **Epipolar Attention.** (a) Full epipolar attention aggregates information along the whole epipolar line, covering unnecessary ranges (only the red dot is the correct position), which limits applicability to lower resolution representations due to memory constraints. (b) Depth-truncated epipolar attention samples only points near the 3D location of that pixel (the red dot). It enables epipolar attention on higher-resolution representations and improves information aggregation.

epipolar lines $\{l^j\}_{j\neq\mathbf{ref}}$ on the other views. Instead of using all points on the epipolar lines, we only sample points around the regions of interest. Specifically, given a known or estimated depth value z and the camera intrinsics, we can unproject the point to 3D space $s_{\mathbf{3D}}$. We the sample N_p points $\{p_i\}$ around $s_{\mathbf{3D}}$ in range $[-r, r]$ in a stratified way along the 3D line formed by $s_{\mathbf{3D}}$ and $T_{\mathbf{c2w}} \cdot s$, where $T_{\mathbf{c2w}}$ is a camera-to-world transformation. We then project these points to the epiplor lines $\{l^j\}$ on the other views to extract features.

To compute the cross attention among views, we first aggregate features across views. For the N_p sampled points, we get features $\{f_i^j\}_{i=1,\ldots,N_p,\ j=1,\ldots,N_v}$ after projecting the points onto the epipoloar line on the j^{th} feature map. For each point, we aggregate these features by a concatenation operation followed by a 2-layer MLP,

$$\begin{aligned} f_i^{\mathbf{mv}} &= \mathbf{MLP}(\mathbf{concat}(\{f_i^j\}_j)), \\ \mathbf{MLP} &: \mathbb{R}^{N\times \text{dim}} \to \mathbb{R}^{\text{dim}}, f_i^{\mathbf{mv}} \in \mathbb{R}^d. \end{aligned} \tag{1}$$

Then, we aggregate the features across N_p points by stacking along the feature dimension and get $F^{\mathbf{mv}} \in \mathbb{R}^{N_p\times d}$.

We use this feature map to compute the keys and values of classic attention while the queries are computed from the reference view, i.e.,

$$\begin{aligned} Q &= W_Q \cdot F^{\mathbf{ref}}, W_Q \in \mathbb{R}^{d\times HW} \\ K &= W_K \cdot F^{\mathbf{mv}}, W_K \in \mathbb{R}^{d\times N_p} \\ V &= W_V \cdot F^{\mathbf{mv}}, W_V \in \mathbb{R}^{d\times N_p} \end{aligned} \tag{2}$$

We apply the depth-truncated epipolar attention on all latent resolutions (from 32 to 256) in all Up-blocks of the decoder.

3.3. Structured-noise depth augmentation

The proposed depth-truncated epipolar attention mechanism requires access to depth information. During training, we leverage 3D data to obtain ground-truth depth. During inference, we can predict depth using off-the-shelf depth predictors. However, the predicted depth is usually imperfect. Furthermore, the multi-view latents are encoded from ground-truth 3D assets during training, but are generated by the diffusion process during inference. That is, the multi-view latents we are decoding might not be accurately aligned. Therefore, we need a strategy to mitigate the domain gap.

Rather than warping the ground truth latents to simulate the misalignment of generated latents, we warp the ground truth depth, which can be regarded as equivalently warping the latents. For this, we propose structured-noise depth augmentation as the noising process. During training, we uniformly sample noise in lower resolution $(3, 64, 128)$ hierarchically. We then upsample these noises to the 256 resolution, and add them to the 256 resolution ground-truth depth map D. We formulate the process as follows:

$$\begin{aligned} Z_i &\sim \mathcal{U}(-s_i, s_i)^{i\times i}, i \in \{3, 64, 128\} \\ \{Z_i'\} &= \mathbf{Upsample}(\{Z_i\}, 256), Z_i' \in \mathbb{R}^{256\times 256} \\ D' &= D + Z_3' + Z_{64}' + Z_{128}' \end{aligned} \tag{3}$$

The noisy depths D' now contain both high and low frequency noise. Note that the depth map will be pooled to different resolution for different hierarchies of Up-blocks. Note, compared to the naive strategy which perturbs each depth pixel with independent Gaussian or uniform noise, our strategy does not cause the noise to be cancelled out in lower resolutions. This makes our method more robust.

During inference, we simply use the predicted depth (we use Neus [1] in this work), as we find our model to be robust to inaccurate predictions.

3.4. Implementation details

We train our model on a subset of the Objaverse [17] dataset, which includes around 23k objects with high-quality geometry and texture. To render the dataset, our procedure is similar to wonder3D [3]: we render RGB images from six fixed views—front, front right, right, back,

Method	PSNR ↑	SSIM ↑	LPIPS ↓
Realfusion	15.26	0.722	0.283
Zero123	18.93	0.779	0.166
SyncDreamer	20.06	0.798	**0.146**
Wonder3D	20.55	0.845	0.166
Ours	**20.74**	**0.847**	0.164

Table 1. **Image-conditioned novel view synthesis on Google Scanned Objects.** We report PSNR, SSIM, and LPIPS on the generated novel view images of GSO objects.

left, and front left. Then the images are encoded by the VAE encoder of Stable Diffusion [15]. During inference, the image latents as the decoder's input are generated by wonder3D [3] given the input image. See appendix for more implementation details.

4. Experiments

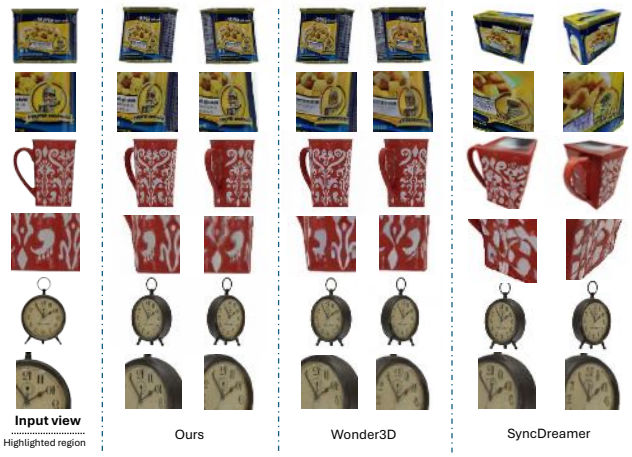

Figure 4. **Qualitative comparisons with baselines.** See appendix for more visualizations.

Methods	Ours	Zero123	Wonder3D	w/o depth aug.	Indep. aug.	Full epi.	w/o epi.
No. of corr.	**458.87**	54.28	329.56	291.59	259.03	254.20	245.94

Table 2. **Evaluating pixel-level alignment.** To better understand the necessity of pixel-aligned multi-view images, we measure the number of correspondences using AspanFormer.

4.1. Multi-view consistency

Following prior works, we evaluate baselines and our method on a subset of the Google Scanned Object (GSO) dataset [18], which includes a variety of objects in common life. The subset matches what is used in SyncDreamer [5] and wonder3D [3], including 30 objects from humans and animals to everyday objects. For each object, we render its front view in a 256 resolution and use it as the input to all methods. Moreover, we use the photometric

Figure 5. **Qualitative comparisons after 3D rendering.** To better understand the impact of pixel-level aligned multi-view images in the 3D generation pipeline, we reconstruct the 3D object using generated multi-view images. We can clearly observe that inconsistent multi-view images lead to reconstructed 3D objects which are blurry.

PSNR, SSIM [19], and LPIPS [20] as evaluation metrics. The quantitative results are summarized in Tab. 1. Note that Wonder3D's performance in our evaluation is lower than reported in the original paper. We tried our best to re-implement their evaluation. Results still improve upon those of other methods. Our method performs favorably to Wonder3D in PSNR and SSIM.

Qualitatively, the multi-view images generated by our method are more consistent, as shown in Fig. 4. We provide zoomed-in illustrations to highlight complex textures. Notably, our method generates textures that are more faithful to the input view, while Wonder3D and SyncDreamer both yield blurred textures. This is due to their diffusion process occurring in latent space with limited resolution. Moreover, their decoder doesn't consider other views.

Next, we measure the number of correspondences between adjacent views using the off-the-shelf dense matching method AspanFormer [21]. As shown in Tab. 2, the proposed method outperforms Wonder3D by 40%. This result highlights the improved pixel-alignment.

4.2. Rerendering from 3D generation

Next, we show that consistent multi-view generation is beneficial for 3D asset generation. As shown in Fig. 5, we optimize NeuS [1] again using the images decoded by our method, and re-render the NeuS results from the fixed views. The Wonder3D [3] baseline reconstruction follows its procedure. We observe the baseline's re-rendering to be much more blurred due to pixel-level misalignment. In contrast, our method's re-rendering remains consistent and maintains the fine details observed in the front view.

References

[1] Wang P. Liu L. Liu Y. Theobalt C. Komura T. and Wang W., "Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction," 2021. 1, 3, 4

[2] Li J. Tan H. Zhang K. Xu Z. Luan F. Xu Y. Hong Y. Sunkavalli K. Shakhnarovich G. and Bi S., "Instant3d: Fast text-to-3d with sparse-view generation and large reconstruction model," in *ICLR*, 2024. 1

[3] Long X. Guo Y. Lin C. Liu Y. Dou Z. Liu L. Ma Y. Zhang S. Habermann M. and Theobalt C., "Wonder3d: Single image to 3d using cross-domain diffusion," *arXiv preprint arXiv:2310.15008*, 2023. 1, 3, 4

[4] Shi Y. Wang P. Ye J. Long M. Li K. and Yang X., "Mvdream: Multi-view diffusion for 3d generation," in *ICLR*, 2024. 1

[5] X. Liu L. Komura T. Liu Y. Lin C. Zeng Z. Long and Wang W., "Syncdreamer: Generating multiview-consistent images from a single-view image," in *ICLR*, 2024. 1, 4

[6] Xu Y. Tan H. Luan F. Bi S. Wang P. Li J. Shi Z. Sunkavalli K. Wetzstein G. and Xu Z., "Dmv3d: Denoising multi-view diffusion using 3d large reconstruction model," in *ICLR*, 2024. 1

[7] C. Rombach R. Voleti V. Yao C. Boss M. Letts A. Pankratz D. Tochilkin D. Laforte and Jampani V., "Sv3d: Novel multi-view synthesis and 3d generation from a single image using latent video diffusion," *arXiv preprint arXiv:2403.12008*, 2024. 1

[8] Chen Z. Wang Y. Wang F. Wang Z. and Liu H., "V3d: Video diffusion models are effective 3d generators," *arXiv preprint arXiv:2403.06738*, 2024. 1

[9] Blattmann A. Dockhorn T. Kulal S. Mendelevitch D. Kilian M. Lorenz D. Levi Y. English Z. Voleti V. and Letts A., "Stable video diffusion: Scaling latent video diffusion models to large datasets," *arXiv preprint arXiv:2311.15127*, 2023. 1, 2

[10] Poole B. Jain A. Barron J. and Mildenhall B., "Dreamfusion: Text-to-3d using 2d diffusion," in *ICLR*, 2023. 1

[11] Wang Z. Lu C. Wang Y. Bao F. Li C. Su H. and Zhu J., "Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation," *arXiv preprint arXiv:2305.16213*, 2023. 1

[12] J. Zhu and Zhuang P., "Hifa: High-fidelity text-to-3d with advanced diffusion guidance," *arXiv preprint arXiv:2305.18766*, 2023. 1

[13] Hong Y. Zhang K. Gu J. Bi S. Zhou Y. Liu D. Liu F. Sunkavalli K. Bui T. and Tan H., "Lrm: Large reconstruction model for single image to 3d," in *ICLR*, 2024. 1

[14] Tang J. Chen Z. Chen X. Wang T. Zeng G. and Liu Z., "Lgm: Large multi-view gaussian model for high-resolution 3d content creation," *arXiv preprint arXiv:2402.05054*, 2024. 1

[15] Rombach R. Blattmann A. Lorenz D. Esser P. and Ommer B., "High-resolution image synthesis with latent diffusion models," in *CVPR*, 2022. 1, 4

[16] Tang Z. Fan Y. Wang D. Xu H. Ranjan R. Schwing A. and Yan Z., "Mv-dust3r+: Single-stage scene reconstruction from sparse views in 2 seconds," *arXiv preprint arXiv:2412.06974*, 2024. 1

[17] Deitke M. Schwenk D. Salvador J. Weihs L. Michel O. VanderBilt E. Schmidt L. Ehsani K. Kembhavi A. and Farhadi A., "Objaverse: A universe of annotated 3d objects," in *CVPR*, 2023. 3

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

[19] Wang Z. Bovik A. Sheikh H. and Simoncelli E., "Image quality assessment: from error visibility to structural similarity," *IEEE TIP*, 2004. 4

[20] Zhang R. Isola P. Efros A. Shechtman E. and Wang O., "The unreasonable effectiveness of deep features as a perceptual metric," in *CVPR*, 2018. 4

[21] Chen H., Luo Z. Zhou L. Tian Y. Zhen M. Fang T. Mckinnon D. Tsin Y., and Quan L., "Aspanformer: Detector-free image matching with adaptive span transformer," in *ECCV*, 2022. 4