Consistent View Synthesis with Pose-Guided Diffusion Models

Hung-Yu Tseng[1] Qinbo Li[1] Changil Kim[1] Suhib Alsisan[1] Jia-Bin Huang[1,2] Johannes Kopf[1]
[1]Meta [2]University of Maryland, College Park

Figure 1. **Consistent view synthesis via pose-guided diffusion model.** (*top*) Given an input image and a sequence of camera poses, we present a pose-guided diffusion model to generate a sequence of frames that are photorealistic and 3D consistent. (*bottom*) Our proposed method can synthesize diverse sequences from the same set of inputs.

Abstract

Novel view synthesis from a single image has been a cornerstone problem for many Virtual Reality applications that provide immersive experiences. However, most existing techniques can only synthesize novel views within a limited range of camera motion or fail to generate consistent and high-quality novel views under significant camera movement. In this work, we propose a pose-guided diffusion model to generate a consistent long-term video of novel views from a single image. We design an attention layer that uses epipolar lines as constraints to facilitate the association between different viewpoints. Experimental results on synthetic and real-world datasets demonstrate the effectiveness of the proposed diffusion model against state-of-the-art transformer-based and GAN-based approaches. More qualitative results are available at https://poseguided-diffusion.github.io/.

1. Introduction

Offering immersive 3D experiences from daily photos has attracted considerable attention. It is a cornerstone technique for a wide range of applications such as 3D photo [18, 49], 3D asset generation [35], and 3D scene navigation [4]. Notably, rapid progress has been made in addressing the *single-image view synthesis* [40, 50, 61, 69] issue. Given an arbitrarily narrow field-of-view image, these frameworks can produce high-quality images from novel viewpoints. However, these methods are limited to viewpoints that are within a small range of the camera motion.

The *long-term single-image view synthesis* task is recently proposed to address the limitation of small camera motion range. As demonstrated in Figure 1, the task attempts to generate a *video* from a single image and a sequence of camera poses. Note that different from the single-image view synthesis problem, the viewpoints of the last few video frames produced under this setting may be far away from the original viewpoint. Take the results shown in Figure 1, for instance, the cameras are moving into different rooms that were not observed in the input images.

Generating long-term view synthesis results from a single image is challenging for two main reasons. First, due to the large range of the camera motion, e.g., moving into a new room, a massive amount of new content needs to be hallucinated for the regions that are not observed in the input image. Second, the view synthesis results should be *consistent* across viewpoints, particularly in the regions observed in the input viewpoint or previously hallucinated in the other views.

Both explicit- and implicit-based solutions are proposed to handle these issues. Explicit-based approaches [17, 24, 25, 40] use a "warp and refine" strategy. Specifically, the image is first warped from the input to novel viewpoints according to some 3D priors, i.e., monocular depth estimation [37, 38]. Then a transformer or GAN-based generative model is designed to refine the warped image. However, the success of the explicit-based schemes hinges on the accuracy of the monocular depth estimation. To address this limitation, Rombach *et al*. [42] designed a geometry-free transformer to implicitly learn the 3D correspondences between the input and output viewpoints. Although reasonable new content is generated, the method fails to produce coherent results across viewpoints. The LoR [39] framework leverages the auto-regressive transformer to further improve the consistency. Nevertheless, generating consistent, high-quality long-term view synthesis results remains challenging.

In this paper, we propose a framework based on diffusion models for consistent and realistic long-term novel view synthesis. Diffusion models [14, 52, 54] have achieved impressive performance on many content creation applications, such as image-to-image translation [44] and text-to-image generation [2, 36, 45]. However, these methods only work on 2D images and lack 3D controllability. To this end, we develop a *pose-guided* diffusion model with the epipolar attention layers. Specifically, in the UNet [43] network of the proposed diffusion model, we design the epipolar attention layer to associate the input view and output view features. According to the camera pose information, we estimate the epipolar line on the input view feature map for each pixel on the output view feature map. Since these epipolar lines indicate the candidate correspondences, we use the lines as the constraint to compute the attention weight between the input and output views.

We conduct extensive quantitative and qualitative studies on real-world Realestate10K [76] and synthetic Matterport3D [7] datasets to evaluate the proposed approach. With the epipolar attention layer, our pose-guided diffusion model is capable of synthesizing long-term novel views that 1) have realistic new content in unseen regions and 2) are consistent with the other viewpoints. We summarize the contributions as follows:

- We propose a pose-guided diffusion model for the long-term single-image view synthesis task.
- We consider the epipolar line as the constraint and design an epipolar attention to associate pixels in the images at input and output views for the UNet network in the diffusion model.
- We validate that the proposed method synthesizes realistic and consistent long-term view synthesis results on the Realestate10K and Matterport3D datasets.

2. Related Work

Novel view synthesis. Novel view synthesis aims to generate high-quality images at arbitrary viewpoints given a set of posed images of a particular scene. With the emergence of deep learning, early approaches [9, 11, 59] use Convolutional Neural Networks to synthesize novel views. Instead of generating novel views directly, several methods [34, 56, 57, 78] predict the appearance flow for producing images of new viewpoints. Recently, various 3D representations are leveraged for this task, including 3D point clouds [1, 17, 29, 33, 69], and layered representations such as layered depth images [18, 49] as well as multiplane images [77]. These representations are used in 3D photo [18], light fields [23, 30] and many other novel view synthesis applications [10, 55, 61]. Very recently, neural radiance field (NeRF) [31] methods reconstruct the target scene implicitly with multi-layer perceptrons and demonstrate impressive novel view synthesis results in various scenarios, including 360-degree [3, 72] or city-scale [58] 3D scenes.

Nevertheless, the approaches mentioned above can only 1) *interpolate* between multiple input views or 2) *extrapolate* from single/multiple views within a limited range of camera movement. To synthesize a realistic novel view along camera trajectories that are far away from the input viewpoint, the PixelSynth [40] and SE3DS [17] schemes progressively construct a 3D point cloud from the input viewpoint according to the estimated depth, then repeatedly apply the "warp and refine" strategy to produce novel views. On the other hand, the GeoGPT [42] framework uses a geometry-free transformer that does not rely on monocular depth estimation. The LoR [39] approach improves the transformer model to reduce the temporal flickering gen-

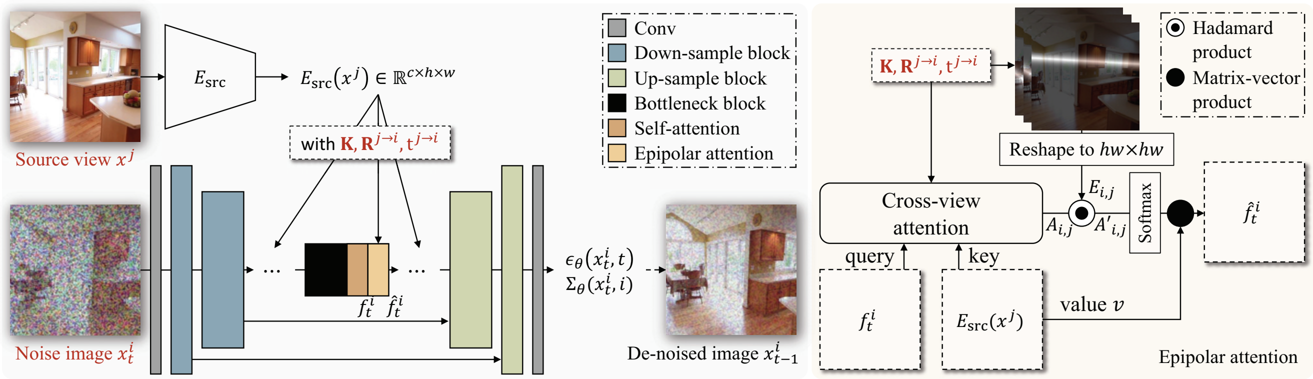

Figure 2. **Method overview.** (*left*) The core component of our pose-guided diffusion model is the UNet that takes the source view image and camera poses as the input (red font), and de-noises the image at the target viewpoint. We use an encoder to extract features from the source view features. We design an epipolar attention to associate the target view with the source view features, and add the epipolar attention layer after *each* self-attention layer in the UNet network. The UNet model takes as input the source view features as well as the camera parameters via the epipolar attention layers, and predicts the de-noised target view image. (*right*) According to the input camera parameters, we compute the epipolar line as the constraint to estimate the attention between the source view and target view features.

erated along a camera trajectory. Nonetheless, it remains challenging to produce high-quality novel views. In this work, we propose a pose-guided diffusion model that synthesizes a consistent and realistic sequence of novel views.

Diffusion models. De-noising diffusion models [14, 52, 54] are generative models that learn to generate data samples from Gaussian noise through a series of de-noising processes. Recently, diffusion models have demonstrated remarkable performance on a variety of 2D content creation tasks, including image super-resolution [22, 41, 46, 64], image in-painting [27, 44], image de-blurring [21, 68], and text-to-image [2, 36, 45]. In addition to working on 2D images, diffusion models are also emerging in the video generation [13, 16, 51, 63] or 3D shape generation [28, 70, 75] tasks. As these methods lack 3D camera pose controllability, they cannot be directly applied to the view synthesis problem. We also build upon diffusion models for our task. In contrast to existing diffusion models for image synthesis, our approach offers full controllability of the viewpoints.

Concurrent with our work, 3DiM [67] also leverages diffusion models for view synthesis tasks. Our work differs in two aspects. First, 3DiM focuses on *object-centric* synthetic scenes (e.g., ShapeNet dataset). In contrast, we focus on long-term view generation of *scene-centric* realistic scenes with complex appearances. Second, we exploit the epipolar constraints across views explicitly with the proposed epipolar cross-view attention layer. We demonstrate that integrating these geometric constraints leads to substantial quality improvements.

Attention. Attention aims to capture the long-range dependencies, e.g., the relationship between two distant image pixels. Attention mechanisms are widely used in deep learning tasks such as image recognition [26] and image generation [71]. In particular, self-attention layers [62, 65] capture the dependencies within the *same* data. On the other hand, cross-attention [74] models the relationships between instances of *different* data, e.g., two images, or an image vs. a text sequence. The proposed epipolar attention can be considered as a type of cross-attention, where the epipolar lines are introduced as geometric constraints to compute the dependencies between the source view and target view image pixels.

3. Methodology

Our goal is to synthesize a sequence of images $\{x^i\}_{i=2}^n$ given an input image x^1, and a sequence of camera poses $\{\mathbf{K}^i, \mathbf{R}^i, \mathbf{t}^i\}_{i=2}^n$, i.e., intrinsics, rotation, and translation, respectively. We design a pose-guided diffusion model to auto-regressively generate the image at each viewpoint i to produce the final sequence. In this section, we first introduce diffusion models in Section 3.1. We then illustrate the proposed pose-guided diffusion model in Section 3.2. Finally, we describe how we produce the consistent novel view video in Section 3.3.

3.1. Diffusion Model

Diffusion models [14, 52, 54] learn to convert an empirical (i.e., isotropic Gaussian) distribution into the target data (i.e., real image) distribution through a series of de-noising operations. A forward process is derived to gradually add noise to the real image so the image becomes indistinguishable from the Gaussian noise. On the other hand, a backward process is learned to reverse the forward process, i.e., map from noises to real images.

Forward and backward process. Given an image x_0 sampled from the real image distribution $\mathcal{P}(x)$, the forward pro-

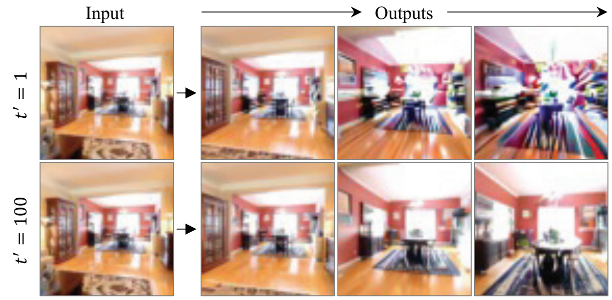

Figure 3. **Artifacts from fixing noises in the backward process.** To improve the consistency across different views in the same sequence, we use the same set of initialization noise x_T and diffusion noise $\{\epsilon_t\}_{t=T}^{t'}$ described in (3) to generate all views in the same video. (*top*) However, we find that fixing noises in all backward steps, i.e., $\{\epsilon_t\}_{t=T}^{1}$, creates obvious artifacts. (*bottom*) We address this by using the fixed noises in the early backward steps only, i.e., $\{\epsilon_t\}_{t=T}^{100}$, and re-sample the noises in the last few backward steps. This helps improve consistency while maintaining realism.

cess converts the image to noise by a T-steps process that gradually adds Gaussian noise to x_0, namely

$$x_t = \sqrt{\alpha_t} x_{t-1} + (1-\alpha_t)\epsilon_t \qquad \epsilon_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{1}$$

where $t = [1, \cdots, T]$. The notation α_t is computed from the noise schedule, which is pre-determined such that $x_T \approx \mathcal{N}(\mathbf{0}, \mathbf{I})$. We can further marginalize the forward process to

$$x_t = \sqrt{\bar{\alpha}_t} x_0 + (1-\bar{\alpha}_t)\epsilon_t \qquad \epsilon_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{2}$$

where $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$. The backward process can then be formulated as

$$x_{t-1} = \mu_\theta(x_t, t) + \Sigma_\theta(x_t, t)\epsilon_t \qquad \epsilon_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{3}$$

where $t = [T, \cdots, 1]$. Typically, an UNet [8, 43] model parameterized by θ is used to learn the backward process.

Training. We use the DDPM [14] strategy that trains the UNet model to predict $\epsilon_\theta(x_t, t)$ instead of $\mu_\theta(x_t)$ in (3), such that

$$\mu_\theta(x_t, t) = \big(x_t - (\frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}})\epsilon_\theta(x_t, t)\big)/\sqrt{\alpha_t}. \tag{4}$$

The UNet model is trained using the mean square loss:

$$L_{\text{diffusion}} = \mathbb{E}_{x,t}\big[\|\epsilon_t - \epsilon_\theta(x_t, t)\|_2^2\big]. \tag{5}$$

As for the term $\Sigma_\theta(x_t, t)$, we follow the improved DDPM [32] approach that uses an additional objective L_{vlb} for training the UNet model to make the prediction.

3.2. Pose-Guided Diffusion Model

We present an overview of the proposed pose-guided diffusion model in Figure 2. Given the source view image x^j at the j-th viewpoint, where $j \in [1, \cdots, n]$, the goal is to de-noise the target view image x_t^i at the diffusion time step t. We first use the source view encoder to extract the feature maps from the source view image x^j. Combining the feature maps using the proposed epipolar attention layer, the UNet model predicts the $\epsilon_\theta(x_t^i, t)$ and $\Sigma_\theta(x_t^i, t)$ terms to estimate the de-noised image x_{t-1}^i. We obtain the final target view image x_0^i by iterating through the backward process.

Source view encoder. Given the source view image x^j, we use a deep convolutional neural network E_{src} to extract the feature map $E_{\text{src}}(x^j) \in \mathbb{R}^{c \times h \times w}$, where $c \times w$ matches the resolution of the attention layer in the UNet network. In practice, we use the pre-trained MiDaS [38] model as the source view encoder. Our early experiments show that such a strategy facilitates faster training of the pose-guided diffusion model. Note that we extract multiple intermediate feature maps from the MiDas model according to the resolutions of the attention layers used in the UNet model.

UNet network. We modify the commonly-used UNet architecture in diffusion models [8] as our UNet network. As demonstrated in the left part of Figure 2, we add the proposed epipolar attention layer after *each* of the self-attention layers in the UNet network.

Epipolar attention. The proposed epipolar attention aims to *associate* the target view with the source view. The core idea is to leverage the epipolar line as the *constraint* to reduce the number of candidate source view pixels corresponding to a particular target view pixel. We present the epipolar attention in the right-hand side of Figure 2. Given the query calculated from intermediate UNet feature f_t^i and the key computed from the source view feature $E_{\text{src}}(x^j)$, we first use the cross-view attention [74] to compute the affinity matrix $A_{i,j} \in \mathbb{R}^{hw \times hw}$. The term $h \times w$ indicates the resolution of the epipolar attention layer. Second, for each pixel position on the intermediate UNet feature map f_t^i, we compute the epipolar line on the source view feature map E_{src} according to the camera parameters $\mathbf{K}$, $\mathbf{R}^{j\to i}$, and $\mathbf{t}^{j\to i}$. The line is then converted to a weight map of shape $h \times w$ where the values indicate the inverse distance to the epipolar line. We estimate the weight maps for all positions in f_t^i, stack these maps, and reshape to get the epipolar weight matrix $E_{i,j} \in \mathbb{R}^{hw \times hw}$. We re-weight the affinity matrix by $A'_{i,j} = A_{i,j} \odot E_{i,j}$, where $\odot$ denotes the Hadamard product. Finally, the output of the epipolar attention layer $\hat{f}_t^i \in \mathbb{R}^{c \times h \times w}$ is computed as

$$\hat{f}_t^i = \text{reshape}\big(\text{softmax}(A'_{i,j}) \cdot v\big), \tag{6}$$

where v is the value term calculated from the source view feature map $E_{\text{src}}(x^j)$. We detail the computation of the epipolar line in the supplementary document.

Super-resolution. We use the cascaded diffusion [15, 36, 45] strategy to obtain the final spatial resolution. Specif-

Table 1. **Quantitative evaluation on short-term view synthesis.** We report the average PSNR (↑), SSIM (↑), and LPIPS (↓) scores between the first five generated and ground-truth frames in the videos. The best performance is in **bold**.

Methods	Re10K			MP3D		
	PSNR (↑)	SSIM (↑)	LPIPS (↓)	PSNR (↑)	SSIM (↑)	LPIPS (↓)
GeoGPT [42]	20.90	0.61	2.53	16.87	**0.63**	3.46
LoR [39]	20.93	0.61	2.35	19.64	0.61	3.30
SE3DS [17]	18.24	0.59	3.20	-	-	-
Ours	**22.64**	**0.68**	**2.19**	**20.59**	**0.63**	**2.90**

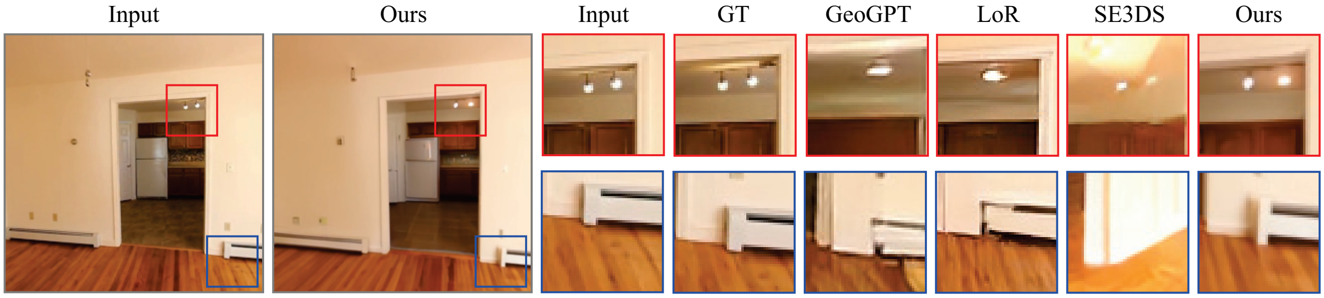

Figure 4. **Qualitative comparisons of short-term view synthesis.** We present the short-term single-image view synthesis results generated by different methods. The patches are all cropped from the same location of the patches in the second image (i.e., Ours).

ically, we use a base pose-guided diffusion model to produce the sequence of resolution 64×64. Then another pose-guided super-resolution diffusion model, detailed in the supplementary document, is used to generate the final 256×256 video.

3.3. Consistent Long-Term View Synthesis

Our goal is to synthesize *a sequence of* novel views given the input image. Although the proposed pose-guided diffusion model learns to generate a single novel view during the training time, we can use the auto-regressive inference to produce long-term view synthesis in the test time. A simple way is to consider the target view x^{i-1} generated at the previous step as the source view x^j to generate the novel view in the current step, i.e., $j = i - 1$. Nevertheless, this approach produces temporal flickering in the final video due to the frame-by-frame processing strategy. We use the following two solutions to address the issue.

Stochastic conditioning. We find that using stochastic conditioning [67] slightly improves the temporal flickering. Specifically, at each step in the backward process described in (3), instead of using the previous frame x^{i-1}, we randomly sample the source view image x^j from the set of prior frames $x^j \sim \text{Uniform}(\{x^k, \cdots, x^{i-1}\})$. Such a strategy encourages the diffusion model to be guided by all the previous frames, thus improving the temporal consistency.

Fixing noises in the backward process. The noises introduced during the backward process illustrated in (3) also contribute to the temporal inconsistency. To reduce the variance of the backward process across different views, we use the same initialization noise x_T and diffusion noises $\{\epsilon_t\}_{t=T}^{1}$ to generate all images in the same video. Nevertheless, we observe noticeable artifacts if we fix all diffusion noises $\{\epsilon_t\}_{t=T}^{1}$ during the backward process, as demonstrated in Figure 3. In practice, fixing the diffusion noises $\{\epsilon_t\}_{t=T}^{t'}$ to a certain backward step t' alleviates the issue and improves the temporal consistency.[1]

4. Experimental Results

4.1. Experimental Setup

Datasets. We focus on two multi-view datasets for all experiments: real-world RealEstate10K (Re10K) [76] and synthetic Matterport 3D (MP3D) [7]. We use $61,986$ video clips in the Re10K dataset for training and randomly sample 500 sequences from the testing split for the evaluation. As for the MP3D dataset, we follow the common protocol [20, 39, 40, 69] to use the Habitat agent [48] to render $6,000$ training videos and 500 testing videos. For both datasets, we resize and center-crop the video to the spatial resolution of 256×256.

Compared methods. We compare our method with several state-of-the-art methods: two recent transformer-based approaches GeoGPT [42] and LoR [39], as well as a very recent GAN-based scheme SE3DS [17].

Evaluation setting. We evaluate the short-term and long-term view synthesis results. We generate a 20-frames video for each testing image, and consider the first 5 frames as the short-term views:

[1] We set t' to be 100 in all experiments, indicating that we re-sample the noise ϵ in the last 100 backward steps.

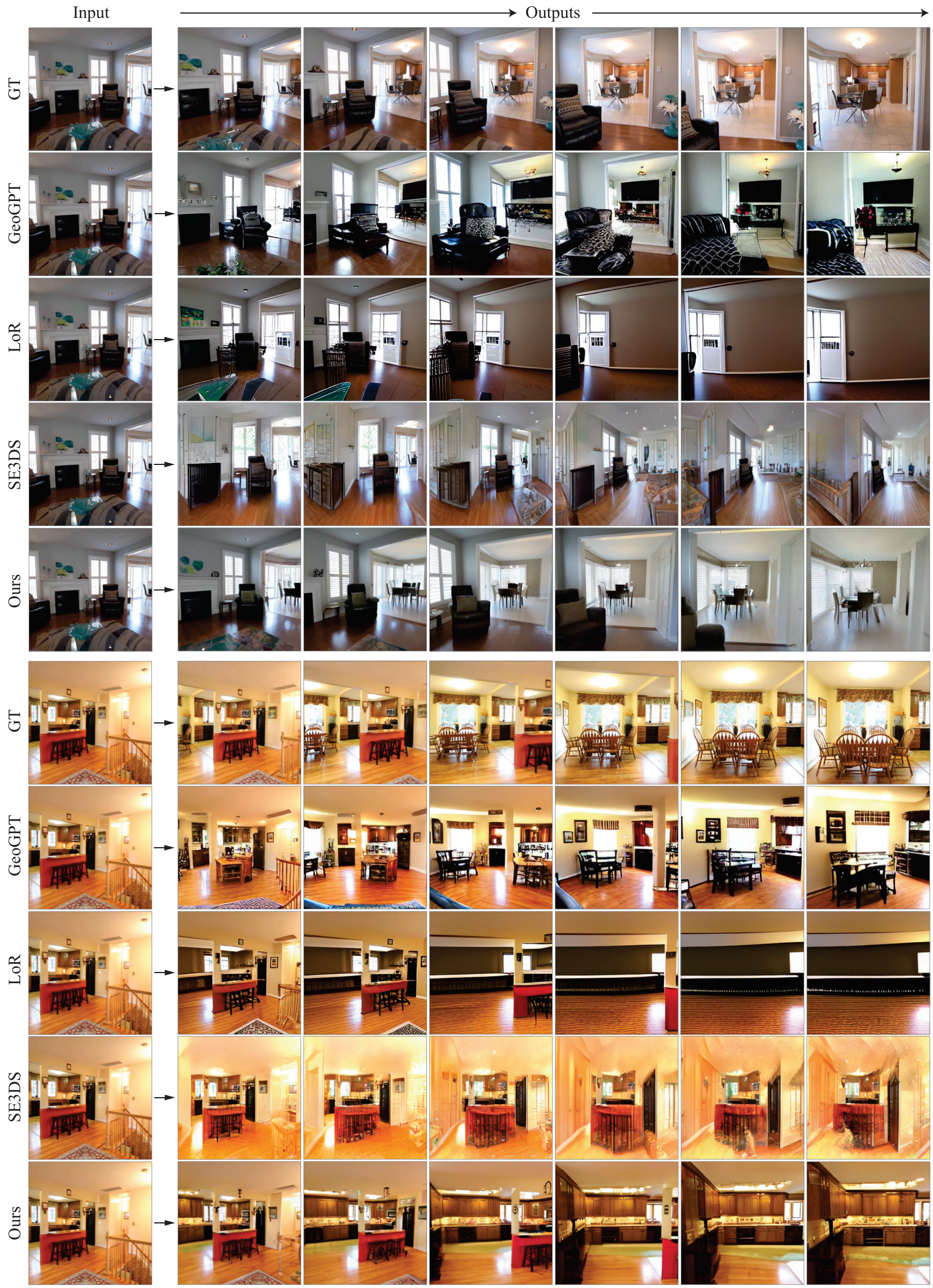

Figure 5. **Qualitative comparisons.** We present the long-term single-image view synthesis results generated by different methods.

- Short-term: We use pairwise metrics PSNR, SSIM, and LPIPS [73] to measure the difference between the generated and ground-truth images.
- Long-term: We measure generated image quality and temporal consistency. For image quality, we use the FID [12] and KID [5] scores to estimate the realism of the last (i.e., 20-th) generated frame. We use the flow warping error (E_{warp}) [19] to quantify the temporal consistency. Specifically, we use the RAFT [60] model to compute optical flow between two consecutive generated frames. Then the error is computed as

$$E_{\text{warp}} = \sum_{i=2}^{20} M^{(i-1)\to i} \|x^i - \hat{x}^{i-1}\|_1, \quad (7)$$

where M is the visibility mask, and $\hat{x}^{i-1}$ is warped from the output frame x^{i-1} using to the optical flow.

More details are provided in the supplementary materials.

4.2. Short-term View Synthesis

We present the quantitative comparisons in Table 1 and qualitative results in Figure 4. While the SE3DS method struggles to produce realistic results, the GeoGPT and LoR frameworks have similar performance on producing short-term novel views. However, the details generated by these two transformer-based methods are slightly inconsistent with the input view. In contrast, the proposed approach synthesizes 1) details that are consistent with the input view and 2) accurate parallax that corresponds to the camera motion.

4.3. Long-Term View Synthesis

We measure the last frame FID and KID scores to evaluate the per-frame quality, and calculate the flow warping error E_{warp} to access the temporal consistency of the generated 20-frames videos. We demonstrate the quantitative comparisons in Table 2, and show example qualitative results in Figure 5. Similar to the short-term view synthesis setting, the SE3DS scheme struggles to generate appealing results, especially under large camera motion, e.g., the bottom example in Figure 5. On the other hand, the GeoGPT model synthesizes realistic novel views. Nevertheless, the results are not consistent across different viewpoints, i.e., the scene changes drastically frame-by-frame. In contrast to the GeoGPT approach, the novel views produced by the LoR method are more consistent. Nonetheless, we observe a quality degradation in the last few generated frames. Compared to these existing approaches, our model generates novel view sequences that 1) maintain the image quality over time and 2) contain less temporal flickering.

Per-frame quality vs. temporal consistency. It is challenging to assess the overall long-term view synthesis performance since there are two perspectives: per-frame quality (FID, KID) and temporal consistency (E_{warp}). Therefore, we plot the FID vs. E_{warp} curves of videos with different lengths generated by various methods in Figure 6. Consistent with the observation we have from Table 2 and Figure 5, the GeoGPT model fails to generate consistent images, while the LoR approach struggles to maintain the generated image quality over time. In contrast, the proposed pose-guided diffusion model synthesizes novel views that are consistent and remain realistic over time.

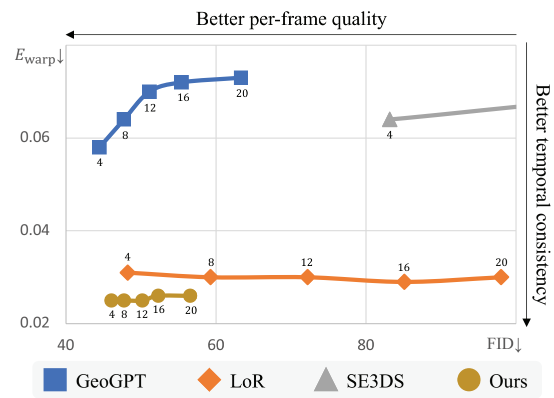

Figure 6. **Last frame quality vs. temporal flickering.** We show the FID (↓) of the last frames and flow-warping errors E_{warp} (↓) given different generated video lengths $\{4, 8, \cdots, 20\}$. Our method generates not only realistic but also consistent long-term single-image view synthesis results.

4.4. Ablation Study

We conduct ablation studies using the Re10K dataset to further analyze the proposed approach.

Epipolar attention. In order to understand the effectiveness of the proposed epipolar attention, we make a comparison to two baselines: Concat and Cross-view attention. In Concat baseline, we use the commonly-used UNet [8] structure with two modifications. First, we concatenate the source view x^j with the noise image x_t^i at the target view as the input to the UNet model. Second, we flatten the input camera pose parameters, compute the embedding vector and add the vector to the diffusion time-step embedding for the UNet network.[2] As for the Cross-view attention baseline, we simply remove the epipolar constraint (i.e., use $A_{i,j}$ instead of $A'_{i,j}$ in (6)) in the proposed epipolar attention layer. To ensure a fair comparison, we use identical hyper-parameters to train different models to generate 64×64 sequences, then use a third-party video super-resolution [6] model to get the final results of resolution 256×256. Furthermore, all the compared methods use stochastic conditioning and noise-fixing.

We present the results in Table 3. Compared to the Cross-view attention baseline, the Concat baseline fails to

[2]The strategy is similar to adding class conditioning embedding [8], or adding text embedding [45] to the diffusion time-step embedding.

Table 2. **Quantitative evaluation on long-term view synthesis.** Given the $20-frames$ videos, we report the average FID (↓) and KID (↓) scores of the last generated frames, and use all generated frames to compute the flow warping error E_{warp} (↓). The best performance is in **bold**. We also report the score of real testing videos for reference.

Methods	Re10K			MP3D		
	FID (↓)	KID (↓)	E_{warp} (↓)	FID (↓)	KID (↓)	E_{warp} (↓)
Real	41.09	0.011	0.018	58.83	0.011	0.019
GeoGPT [42]	63.30	**0.016**	0.073	213.14	0.046	0.057
LoR [39]	98.01	0.034	0.030	113.50	0.048	0.036
SE3DS [17]	235.8	0.153	0.060	-	-	-
Ours	**56.33**	**0.016**	**0.023**	**72.48**	**0.019**	**0.035**

Table 3. **Impact of epipolar attention.** We report the FID (↓) and KID (↓) scores of the last generated video frames. We use different diffusion models to generate the 64×64 sequences, then use the same video super-resolution [6] model to get the 256×256 videos for fair comparison. The best performance is in **bold**.

Methods	Re10K	
	FID (↓)	KID (↓)
Source/target views concatenation	87.22	0.034
Cross-view attention	81.37	0.033
Epipolar attention (Ours)	**69.63**	**0.025**

generate high-quality novel views in long-term, since it is challenging to learn the correspondence between source and target views via concatenated inputs. On the other hand, our approach synthesizes realistic novel views as the proposed attention leverages epipolar lines as the constraint to estimate the dependency between the source and target views.

Super-resolution. In this study, we compare different super-resolution approaches: monocular image super-resolution (ESRGAN) [66], video super-resolution (RealBasicVSR) [6], and our pose-guided super-resolution diffusion model. For a fair comparison, we use the same 64×64 sequences generated by the low-resolution pose-guided diffusion model as the input. The results are shown in Table 4. The videos super-resolved by the RealBasicVSR method contain less flickering compared to the other methods since they process the low-resolution sequences frame-by-frame. On the other hand, the pose-guided diffusion model generates much more high-quality novel views. Therefore, we use the pose-guided diffusion model to super-resolve the low-resolution novel view videos in all experiments. Nevertheless, we argue that video super-resolution diffusion models may be critical to further reduce the temporal flickering while maintaining the visual quality.

5. Limitations and Future Works

The proposed method has the following limitations. First, our approach cannot handle the case where scene scales vary dramatically across different videos, e.g., landscape videos explored in [24, 25]. Take Figure 7, for instance, the scale of the scene is significantly larger than

Table 4. **Super-resolution models.** We report the average LPIPS (↓) scores for the short-term, FID (↓) and E_{warp} scores for the long-term novel view synthesis results. We use the same 64×64 results and different super-resolution methods to get the 256×256 videos. The best performance is in **bold**.

Methods	Re10K		
	LPIPS (↓)	FID (↓)	E_{warp} (↓)
Real-ESRGAN [66]	2.32	75.05	0.021
RealBasicVSR [6]	2.28	69.63	**0.014**
Pose-guided diffusion model	**2.19**	**56.33**	0.023

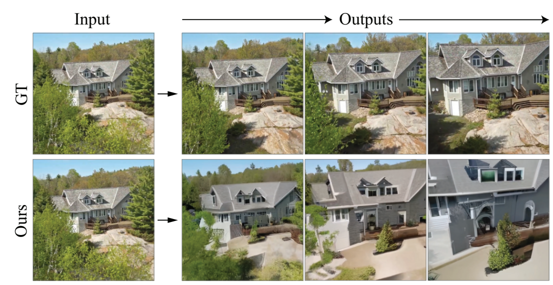

Figure 7. **Failure case.** Our proposed method fails to generate realistic novel views if the scale of the scene is significantly different from those in the training data.

those in the Re10K training data. We believe that handling such cases requires proper scale normalization or data augmentation. We leave this exploration to future work. Second, the inference is time-consuming as it involves multiple steps (i.e., 250 in practice) in the backward process to predict one single novel view. As many recent efforts [47, 53] are made to accelerate the inference speed of the diffusion model, we plan to explore these solutions in the future.

6. Conclusions

In this work, we introduce a pose-guided diffusion model to synthesize a novel view video under massive camera motion from a single image. The core of our diffusion model is the epipolar attention that estimates the dependencies between images of two camera viewpoints. Qualitative and quantitative results show that the proposed pose-guided diffusion model generates novel views that are 1) realistic, even the viewpoints far away from the input view, and 2) consistent across various viewpoints.

References

[1] Kara-Ali Aliev, Artem Sevastopolsky, Maria Kolos, Dmitry Ulyanov, and Victor Lempitsky. Neural point-based graphics. In *ECCV*, 2020. 2

[2] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, et al. ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022. 2, 3

[3] Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *ICCV*, 2021. 2

[4] Miguel Angel Bautista, Pengsheng Guo, Samira Abnar, Walter Talbott, Alexander Toshev, Zhuoyuan Chen, Laurent Dinh, Shuangfei Zhai, Hanlin Goh, Daniel Ulbricht, et al. Gaudi: A neural architect for immersive 3d scene generation. In *NeurIPS*, 2022. 1

[5] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. In *ICLR*, 2018. 7

[6] Kelvin C.K. Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. Investigating tradeoffs in real-world video super-resolution. In *CVPR*, 2022. 7, 8

[7] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. In *International Conference on 3D Vision*, 2017. 2, 5

[8] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *NeurIPS*, 2021. 4, 7

[9] Alexey Dosovitskiy, Jost Tobias Springenberg, Maxim Tatarchenko, and Thomas Brox. Learning to generate chairs, tables and cars with convolutional networks. *IEEE TPAMI*, 39(4):692–705, 2016. 2

[10] John Flynn, Michael Broxton, Paul Debevec, Matthew DuVall, Graham Fyffe, Ryan Overbeck, Noah Snavely, and Richard Tucker. Deepview: View synthesis with learned gradient descent. In *CVPR*, 2019. 2

[11] John Flynn, Ivan Neulander, James Philbin, and Noah Snavely. Deepstereo: Learning to predict new views from the world's imagery. In *CVPR*, 2016. 2

[12] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, 2017. 7

[13] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022. 3

[14] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 2, 3, 4

[15] Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *JMLR*, 23:47–1, 2022. 4

[16] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *arXiv preprint arXiv:2204.03458*, 2022. 3

[17] Jing Yu Koh, Harsh Agrawal, Dhruv Batra, Richard Tucker, Austin Waters, Honglak Lee, Yinfei Yang, Jason Baldridge, and Peter Anderson. Simple and effective synthesis of indoor 3d scenes. *arXiv preprint arXiv:2204.02960*, 2022. 2, 5, 8

[18] Johannes Kopf, Kevin Matzen, Suhib Alsisan, Ocean Quigley, Francis Ge, Yangming Chong, Josh Patterson, Jan-Michael Frahm, Shu Wu, Matthew Yu, et al. One shot 3d photography. *ACM TOG*, 39(4):76–1, 2020. 1, 2

[19] Wei-Sheng Lai, Jia-Bin Huang, Oliver Wang, Eli Shechtman, Ersin Yumer, and Ming-Hsuan Yang. Learning blind video temporal consistency. In *ECCV*, 2018. 7

[20] Zihang Lai, Sifei Liu, Alexei A Efros, and Xiaolong Wang. Video autoencoder: self-supervised disentanglement of static 3d structure and motion. In *ICCV*, 2021. 5

[21] Sangyun Lee, Hyungjin Chung, Jaehyeon Kim, and Jong Chul Ye. Progressive deblurring of diffusion models for coarse-to-fine image synthesis. *arXiv preprint arXiv:2207.11192*, 2022. 3

[22] Haoying Li, Yifan Yang, Meng Chang, Shiqi Chen, Huajun Feng, Zhihai Xu, Qi Li, and Yueting Chen. Srdiff: Single image super-resolution with diffusion probabilistic models. *Neurocomputing*, 479:47–59, 2022. 3

[23] Qinbo Li and Nima Khademi Kalantari. Synthesizing light field from a single image with variable mpi and two network fusion. *ACM Trans. Graph.*, 39(6):229–1, 2020. 2

[24] Zhengqi Li, Qianqian Wang, Noah Snavely, and Angjoo Kanazawa. Infinitenature-zero: Learning perpetual view generation of natural scenes from single images. In *ECCV*, 2022. 2, 8

[25] Andrew Liu, Richard Tucker, Varun Jampani, Ameesh Makadia, Noah Snavely, and Angjoo Kanazawa. Infinite nature: Perpetual view generation of natural scenes from a single image. In *ICCV*, 2021. 2, 8

[26] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021. 3

[27] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *CVPR*, 2022. 3

[28] Shitong Luo and Wei Hu. Diffusion probabilistic models for 3d point cloud generation. In *CVPR*, 2021. 3

[29] Moustafa Meshry, Dan B Goldman, Sameh Khamis, Hugues Hoppe, Rohit Pandey, Noah Snavely, and Ricardo Martin-Brualla. Neural rerendering in the wild. In *CVPR*, 2019. 2

[30] Ben Mildenhall, Pratul P Srinivasan, Rodrigo Ortiz-Cayon, Nima Khademi Kalantari, Ravi Ramamoorthi, Ren Ng, and Abhishek Kar. Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. *ACM TOG*, 38(4):1–14, 2019. 2

[31] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf:

Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 2

[32] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *ICML*, 2021. 4

[33] Simon Niklaus, Long Mai, Jimei Yang, and Feng Liu. 3d ken burns effect from a single image. *ACM TOG*, 38(6):1–15, 2019. 2

[34] Eunbyung Park, Jimei Yang, Ersin Yumer, Duygu Ceylan, and Alexander C Berg. Transformation-grounded image generation network for novel 3d view synthesis. In *CVPR*, 2017. 2

[35] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 1

[36] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 2, 3, 4

[37] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *ICCV*, 2021. 2

[38] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE TPAMI*, 2020. 2, 4

[39] Xuanchi Ren and Xiaolong Wang. Look outside the room: Synthesizing a consistent long-term 3d scene video from a single image. In *CVPR*, 2022. 2, 5, 8

[40] Chris Rockwell, David F Fouhey, and Justin Johnson. Pixelsynth: Generating a 3d-consistent experience from a single image. In *ICCV*, 2021. 1, 2, 5

[41] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 3

[42] Robin Rombach, Patrick Esser, and Björn Ommer. Geometry-free view synthesis: Transformers and no 3d priors. In *ICCV*, 2021. 2, 5, 8

[43] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, 2015. 2, 4

[44] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 Conference Proceedings*, 2022. 2, 3

[45] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022. 2, 3, 4, 7

[46] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE TPAMI*, 2022. 3

[47] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *ICLR*, 2021. 8

[48] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, et al. Habitat: A platform for embodied ai research. In *ICCV*, 2019. 5

[49] Meng-Li Shih, Shih-Yang Su, Johannes Kopf, and Jia-Bin Huang. 3d photography using context-aware layered depth inpainting. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1, 2

[50] Meng-Li Shih, Shih-Yang Su, Johannes Kopf, and Jia-Bin Huang. 3d photography using context-aware layered depth inpainting. In *CVPR*, 2020. 1

[51] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022. 3

[52] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, 2015. 2, 3

[53] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2020. 8

[54] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *NeurIPS*, 2019. 2, 3

[55] Pratul P Srinivasan, Richard Tucker, Jonathan T Barron, Ravi Ramamoorthi, Ren Ng, and Noah Snavely. Pushing the boundaries of view extrapolation with multiplane images. In *CVPR*, 2019. 2

[56] Pratul P Srinivasan, Tongzhou Wang, Ashwin Sreelal, Ravi Ramamoorthi, and Ren Ng. Learning to synthesize a 4d rgbd light field from a single image. In *ICCV*, 2017. 2

[57] Shao-Hua Sun, Minyoung Huh, Yuan-Hong Liao, Ning Zhang, and Joseph J Lim. Multi-view to novel view: Synthesizing novel views with self-learned confidence. In *ECCV*, 2018. 2

[58] Matthew Tancik, Vincent Casser, Xinchen Yan, Sabeek Pradhan, Ben Mildenhall, Pratul P Srinivasan, Jonathan T Barron, and Henrik Kretzschmar. Block-nerf: Scalable large scene neural view synthesis. In *CVPR*, 2022. 2

[59] Maxim Tatarchenko, Alexey Dosovitskiy, and Thomas Brox. Multi-view 3d models from single images with a convolutional network. In *ECCV*, 2016. 2

[60] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *ECCV*, 2020. 7

[61] Richard Tucker and Noah Snavely. Single-view view synthesis with multiplane images. In *CVPR*, 2020. 1, 2

[62] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 3

[63] Vikram Voleti, Alexia Jolicoeur-Martineau, and Christopher Pal. Masked conditional video diffusion for prediction, generation, and interpolation. *arXiv preprint arXiv:2205.09853*, 2022. 3

[64] Chulin Wang, Kyongmin Yeo, Xiao Jin, Andres Codas, Levente J Klein, and Bruce Elmegreen. S3rp: Self-supervised super-resolution and prediction for advection-diffusion process. *arXiv preprint arXiv:2111.04639*, 2021. 3

[65] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, 2018. 3

[66] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *ICCV*, 2021. 8

[67] Daniel Watson, William Chan, Ricardo Martin-Brualla, Jonathan Ho, Andrea Tagliasacchi, and Mohammad Norouzi. Novel view synthesis with diffusion models. *arXiv preprint arXiv:2210.04628*, 2022. 3, 5

[68] Jay Whang, Mauricio Delbracio, Hossein Talebi, Chitwan Saharia, Alexandros G Dimakis, and Peyman Milanfar. Deblurring via stochastic refinement. In *CVPR*, 2022. 3

[69] Olivia Wiles, Georgia Gkioxari, Richard Szeliski, and Justin Johnson. Synsin: End-to-end view synthesis from a single image. In *CVPR*, 2020. 1, 2, 5

[70] Xiaohui Zeng, Arash Vahdat, Francis Williams, Zan Gojcic, Or Litany, Sanja Fidler, and Karsten Kreis. Lion: Latent point diffusion models for 3d shape generation. *arXiv preprint arXiv:2210.06978*, 2022. 3

[71] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In *ICML*, 2019. 3

[72] Kai Zhang, Gernot Riegler, Noah Snavely, and Vladlen Koltun. Nerf++: Analyzing and improving neural radiance fields. *arXiv preprint arXiv:2010.07492*, 2020. 2

[73] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 7

[74] Brady Zhou and Philipp Krähenbühl. Cross-view transformers for real-time map-view semantic segmentation. In *CVPR*, 2022. 3, 4

[75] Linqi Zhou, Yilun Du, and Jiajun Wu. 3d shape generation and completion through point-voxel diffusion. In *ICCV*, 2021. 3

[76] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: learning view synthesis using multiplane images. *ACM TOG*, 37(4):1–12, 2018. 2, 5

[77] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: learning view synthesis using multiplane images. *ACM TOG*, 37(4):1–12, 2018. 2

[78] Tinghui Zhou, Shubham Tulsiani, Weilun Sun, Jitendra Malik, and Alexei A Efros. View synthesis by appearance flow. In *ECCV*, 2016. 2