Learning 3D Photography Videos via Self-supervised Diffusion on Single Images

Xiaodong Wang[1*] , **Chenfei Wu**[2*] , **Shengming Yin**[2] , **Minheng Ni**[2] , **Jianfeng Wang**[2] , **Linjie Li**[2] , **Zhengyuan Yang**[2] , **Fan Yang**[2] , **Lijuan Wang**[2] , **Zicheng Liu**[2] , **Yuejian Fang**[1†] and **Nan Duan**[2†]

[1]Peking University
[2]Microsoft

{wangxiaodong21s@stu, fangyj@ss}.pku.edu.cn, {chewu, v-sheyin, t-mni, jianfw, Lindsey.Li, zhengyang, fanyang, lijuanw, zliu, nanduan}@microsoft.com

Abstract

3D photography renders a static image into a video with appealing 3D visual effects. Existing approaches typically first conduct monocular depth estimation, then render the input frame to subsequent frames with various viewpoints, and finally use an inpainting model to fill those missing/occluded regions. The inpainting model plays a crucial role in rendering quality, but it is normally trained on out-of-domain data. To reduce the training and inference gap, we propose a novel self-supervised diffusion model as the inpainting module. Given a single input image, we automatically construct a training pair of the masked occluded image and the ground-truth image with random cycle rendering. The constructed training samples are closely aligned to the testing instances, without the need for data annotation. To make full use of the masked images, we designed a Masked Enhanced Block (MEB), which can be easily plugged into the UNet and enhance the semantic conditions. Towards real-world animation, we present a novel task: out-animation, which extends the space and time of input objects. Extensive experiments on real datasets show that our method achieves competitive results with existing SOTA methods.

1 Introduction

Recent advances in generation have delivered impressive photorealistic visual content, such as images [Rombach *et al.*, 2022; Wu *et al.*, 2022; Ramesh *et al.*, 2022] and videos [Wu *et al.*, 2021; Villegas *et al.*, 2022; Singer *et al.*, 2022]. 3D photography is a special case of video generation that converts a static image into a 3D video, with applications in virtual reality and animation, attracting wide attention [Shih *et al.*, 2020; Wiles *et al.*, 2020; Tucker and Snavely, 2020; Jampani *et al.*, 2021]. It requires the model to generate a sequence of *consistent* and *reasonable* subsequent frames (novel views) from a starting image only by interactively changing the camera

---

Supplementary Material can be found in https://wang-xiaodong1899.github.io/assets/IJCAI23_supple.pdf

viewpoint, thus putting forward higher requirements to training data and model design.

Existing 3D photography methods can be summarized into two categories: *multi-view methods* and *single-view methods*, as shown in the first two rows in Fig. 1, respectively. Multi-view methods [Hu *et al.*, 2021; Lai *et al.*, 2021; Rockwell *et al.*, 2021; Li *et al.*, 2021; Li and Kalantari, 2020] usually take frames from video, multi-view images, or fake multi-view images [Han *et al.*, 2022] as input and produce 3D representations of scenes, such as point clouds [Wiles *et al.*, 2020] and MPIs (MultiPlane Images) [Tucker and Snavely, 2020], and then train an encoder-decoder model projecting the source view V^i to target view V^{i+1}. As shown in the first row in Fig. 1, such methods need image pairs in training, but there is a lack of large-scale multi-view datasets.

To alleviate the dependency on multi-view datasets, single-view methods [Jampani *et al.*, 2021; Shih *et al.*, 2020] try to train with only RGB image data, first estimate the monocular depth [Ranftl *et al.*, 2022], and then use MPIs or LDIs (Layered Depth Images) representations to better inpaint the occluded regions of target frames. Early method [Shih *et al.*, 2020] searches the occluded regions by hard layering the discontinuities of estimated depth and inpaints the depth and RGB images separately. Further, [Jampani *et al.*, 2021] extends a soft depth layering to decompose an image into the foregrounds and backgrounds, and it prefers inpainting the backgrounds for preserving foreground details. These single-view methods convert the 3D photography into an inpainting task by predefining the masked regions on image data, and these predefined masks by priors are more suitable for generation than random masks, therefore, they can generate good 3D videos only trained on single images.

However, facing complex scenes in 3D photography, the predefined masks in single-view methods are not consistent with real occluded regions caused by 3D rendering. As shown in the second row in Fig. 1, single-view methods first predefine the masks according to edge or depth priors (the yellow masked regions surrounding the foregrounds as in [Jampani *et al.*, 2021]), and then train GAN-based models to inpaint these regions towards original images. During inference, they first render the source frame V^i into a masked V^{i+1} at the target viewpoint via the 3D renderer W, as the blue masked image shows in the second row in Fig. 1, and then use the trained model to inpaint those occluded regions to obtain the novel
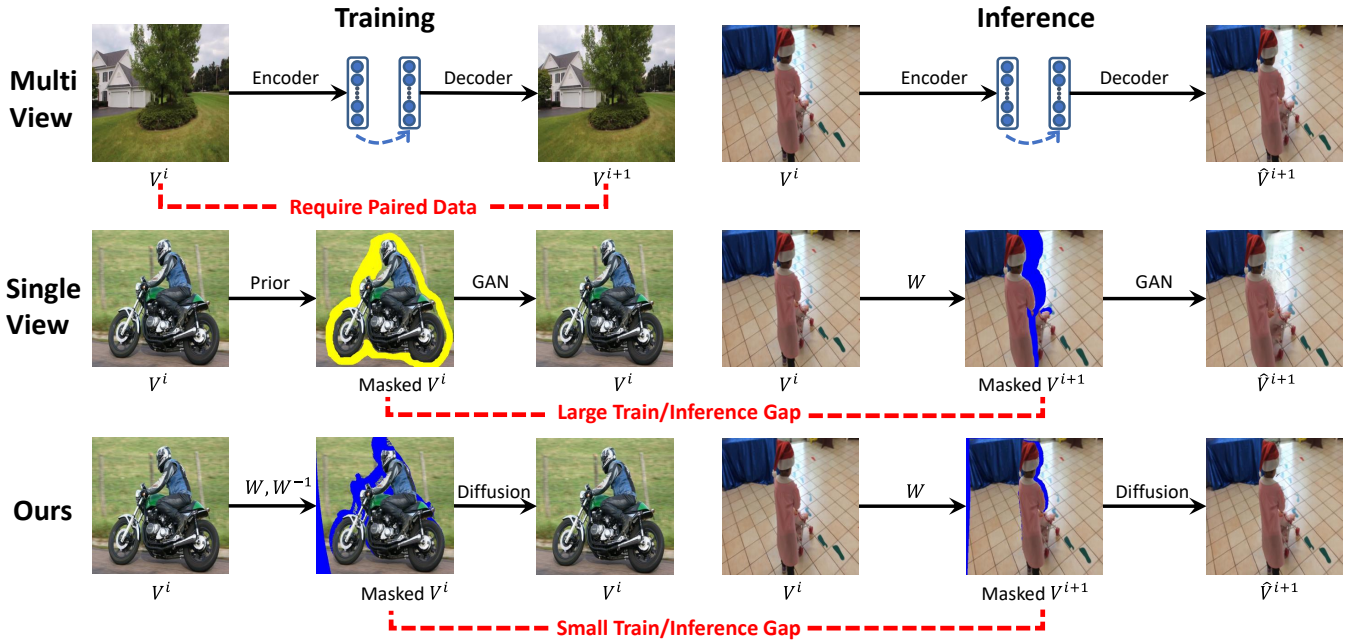
Figure 1: Comparison of different methods. Our proposed self-supervised diffusion belongs to single-view methods but is different from existing works. We incorporate a pair of 3D rendering function W, W^{-1} to reduce the train/inference gap of existing single-view methods.

view $\hat{V}^{i+1}$. Since the rendered masks are different from pre-defined masks, there is a large gap between training and inference, leading to obvious visual distortions.

To this end, we propose a novel self-supervised diffusion model, which only trains on single images but can generate high-quality 3D videos shown in Fig. 1. To align inpainting masks towards real occluded masks, we use the same 3D renderer W to compose a cycle rendering (W, W^{-1}). In cycle rendering, we first use W to randomly render the image V^i to a virtual image at a nearby viewpoint, and then we use W^{-1} to render the virtual image back to the viewpoint of V^i. As a result, we obtain the masked V^i with occluded regions. Further, we utilize a conditional diffusion model that denoises cycle-rendered masked images into original images. We regard the masked images as the conditions, and the original images as the ground-truth images of our diffusion model, respectively. During inference, we can inpaint the masked V^{i+1} via our trained diffusion model to generate the novel view $\hat{V}^{i+1}$. This self-supervised way effectively reduces the gap between training and inference, guaranteeing the high quality of generated 3D videos. To fully leverage the semantic information of masked images, we present a Masked Enhanced Block (MEB) to better embed them into the denoising UNet. Specifically, the masked images and occluded masks are fed into the MEBs via two stacked spatially adaptive normalization layers [Park *et al.*, 2019]. By leveraging the enhanced module to inpaint masked images, our model outputs higher fidelity results with fewer visual distortions. To further promote the visual quality and diversity of novel views, we can transfer the text-to-image knowledge into our self-supervised diffusion model, relieving the pressure of inpainting.

Towards the real application of animation, we further present a novel task: *out-animation*, which requires models to generate videos that extend the space and time dimensions of input objects (or selected parts of an image). We propose a two-stage pipeline: we first perform the image outpainting with the same denoising network to generate consistent and reasonable scenes for input objects according to their content and text prompts, and then sample the 3D videos via the trained self-supervised diffusion model. Experiments on novel view synthesis and image outpainting of real datasets validate the effectiveness of our method.

We summarize our contributions as follows:

- We propose a novel self-supervised diffusion model, which trains only on single images but can generate high-quality 3D photography videos.

- We propose MEB, a Masked Enhanced Block that leverages the unmasked image conditions for the denoising process of our diffusion model.

- We present a novel task: out-animation, and adapt the proposed diffusion models to the new task. Experiments on real datasets validate the effectiveness of our models.

2 Related Work

2.1 3D Photography

Multi-view methods. Many methods [Han *et al.*, 2022; Hu *et al.*, 2021; Lai *et al.*, 2021; Rockwell *et al.*, 2021; Li *et al.*, 2021; Li and Kalantari, 2020; Wang *et al.*, 2022a; Mildenhall *et al.*, 2021] learn to predict the 3D representations (such as NeRf [Mildenhall *et al.*, 2021], point clouds [Wiles *et al.*, 2020], and MPIs [Tucker and Snavely, 2020] using multi-view supervision, so they assume many views such as related two views or multi-views can be used

to train the models. [Tucker and Snavely, 2020] first applies the MPI representations from single image input and renders them to novel views, and then predicts the target views via multi-view supervision. SynSin [Wiles *et al.*, 2020] proposes a novel point cloud renderer that transfers the latent 3D point cloud features into the target views, and rendered features are passed a refinement network to generate target predictions. To alleviate the lack of large-scale in-the-wild multi-view datasets, AdaMPI [Han *et al.*, 2022] trains an inpainting network via warp-back strategy to construct fake multi-view datasets and then trains an MPI-based model.

Single-view methods. Single-view methods [Tucker and Snavely, 2020; Shih *et al.*, 2020; Jampani *et al.*, 2021; Kopf *et al.*, 2020; Niklaus *et al.*, 2019] only require single view images with LDIs (Layered Depth Images) [Shih *et al.*, 2020] representations. Most single-view methods estimate the dense monocular depths and fill in the predefined occluded regions. 3d photo [Shih *et al.*, 2020] makes full use of the three inpainting modules that separately inpaint the edges, depth, and color images to predict the impressive novel views. SLIDE [Jampani *et al.*, 2021] proposes a soft layering to separate an image into foregrounds and backgrounds, and then it prefers inpainting the color and depth of backgrounds, therefore, it better preserves the foreground details.

2.2 Image Synthesis

Text-conditional Image Synthesis

Text-to-image recently becomes popular due to the remarkable success of autoregressive models and diffusion models. Parti [Yu *et al.*, 2022] and NUWA-Infinity [Wu *et al.*, 2022] are two-stage approaches, which first compress an image into discretized latent space, and autoregressively predict discrete image tokens based on text tokens. Diffusion-based methods generate the images via a denoising network conditioned on text representations, such as DALL·E2 [Ramesh *et al.*, 2022], Imagen [Saharia *et al.*, 2022], and Stable-Diffusion [Rombach *et al.*, 2022].

Semantic Image Synthesis

This task aims to create images from semantic segmentations, and the challenge is to generate better images in terms of visual fidelity and spatial alignment. SPADE [Park *et al.*, 2019] is the most popular method to achieve promising images by easily introducing a spatially adaptive normalization layer. From this conditional normalization perspective, methods [Liu *et al.*, 2019; Tan *et al.*, 2021; Wang *et al.*, 2021; Zhu *et al.*, 2020] have designed more tailored ways to embed the semantic masks.

Recently, researchers also pay more attention to *Image Inpainting* and *Image Outpainting*. Inpainting aims to fill missing regions in images, while outpainting tries to extend the images. Some text-conditional image synthesis methods could directly adapt the outpainting or inpainting. Stable-Diffusion and DALL·E2 can be applied on both outpainting and inpainting tasks. The core of 3D photography is similar to image-inpainting, but inpaints the occluded regions for moved 3D target views. The MEB in our UNet is inspired by the various embedding ways of semantic masks, but we utilize the diffusion models rather than GAN-based models.

3 Methodology

In this paper, we propose a novel self-supervised diffusion model for learning 3D photography. We are the first to leverage the diffusion model for 3D photography, and the quality of novel views will benefit from the randomness continuously involved by noise at each step. We organize the rest of this section as follows: we first review the conditional diffusion models. Then, we present our self-supervised diffusion model. After that, we will present the out-animation task, and introduce our pipeline for achieving this task.

3.1 Preliminaries

We briefly review the theory of conditional diffusion models [Ho *et al.*, 2020]. Conditional diffusion models are latent variable models that aim to learn the form of $p_\theta(\mathbf{z}_0|\mathbf{c})$ while the conditional data follows $q(\mathbf{z}_0|\mathbf{c})$, where c is the condition. The joint distribution $p_\theta(\mathbf{z}_{0:T}|\mathbf{c})$ is called the *reverse process*, which is a Markov chain with learned Gaussian transitions starting at $p(\mathbf{z}_T) = \mathcal{N}(\mathbf{z}_T; \mathbf{0}, \mathbf{I})$, as shown below:

$$p_\theta(\mathbf{z}_{0:T}|\mathbf{c}) := p(\mathbf{z}_T) \prod_{t=1}^{T} p_\theta(\mathbf{z}_{t-1}|\mathbf{z}_t, \mathbf{c}), \quad (1)$$

$$p_\theta(\mathbf{z}_{t-1}|\mathbf{z}_t, \mathbf{c}) := \mathcal{N}(\mathbf{z}_{t-1}; \mu_\theta(\mathbf{z}_t, \mathbf{c}, t), \mathbf{\Sigma}_\theta(\mathbf{z}_t, \mathbf{c}, t)) \quad (2)$$

where t is an arbitrary timestep. The approximate posteriors $q(\mathbf{z}_{1:T}|\mathbf{z}_0)$, called the *forward process*, is fixed as a Markov chain that gradually adds Gaussian noise to the data obeying a variance scheduler $\beta_1, \ldots, \beta_T$:

$$q(\mathbf{z}_{1:T}|\mathbf{z}_0) := \prod_{t=1}^{T} q(\mathbf{z}_t|\mathbf{z}_{t-1}), \quad (3)$$

$$q(\mathbf{z}_t|\mathbf{z}_{t-1}) := \mathcal{N}(\mathbf{z}_t; \sqrt{1-\beta_t}\mathbf{z}_{t-1}, \beta_t\mathbf{I}) \quad (4)$$

We define $\alpha_t := 1 - \beta_t$ and $\overline{\alpha}_t := \prod_{s=1}^{t} \alpha_s$. Then we sample $\mathbf{z}_t$ at the timestep t in closed form as follows:

$$q(\mathbf{z}_t|\mathbf{z}_0) = \mathcal{N}(\mathbf{z}_t; \sqrt{\overline{\alpha}_t}\mathbf{z}_0, (1 - \overline{\alpha}_t)\mathbf{I}) \quad (5)$$

The conditional diffusion models are trained to minimize the variational lower bound (VLB), and the objective is equivalent to a denoising process as follows:

$$\mathcal{L}_{t-1} = \mathbb{E}_{\mathbf{z}_0, \mathbf{c}, \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), t} \left[\|\epsilon - \epsilon_\theta(\mathbf{z}_t, t, \mathbf{c})\|_2^2 \right] \quad (6)$$

where $\mathcal{L}_{t-1}$ is the loss function at the timestep $t-1$, and ϵ_θ is the denoising network.

3.2 Self-supervised Diffusion Model

To address the lack of large-scale multi-view data and generate high-quality 3D videos, we present the self-supervised diffusion model. The two crucial parts of our model are the self-supervised strategy for denoising and the M-UNet (a modified UNet with several masked enhanced blocks).

Self-supervised strategy. Previous works [Li *et al.*, 2022; Han *et al.*, 2022] attempted to construct the training pairs from images, i.e., occluded images and ground-truth images, to get rid of the requirement for multi-view datasets. But they only trained the traditional inpainting models to inpaint the

(a) Training Phase of Self-supervised Diffusion Model
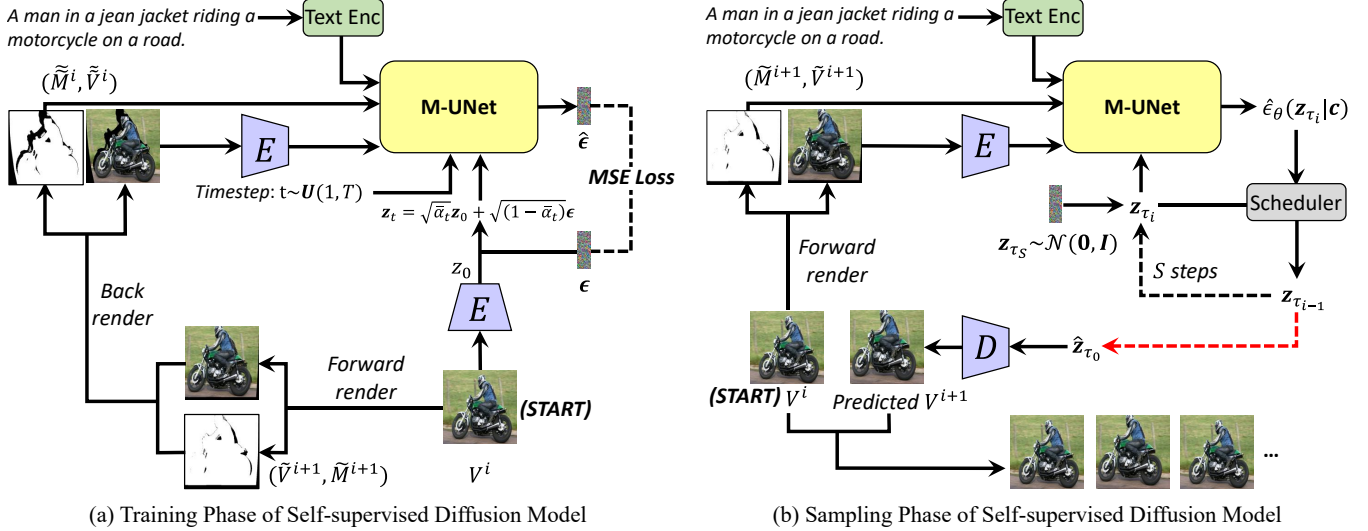
(b) Sampling Phase of Self-supervised Diffusion Model

Figure 2: The overall framework. (a) A self-supervised way to train the diffusion model to inpaint the rendered masked regions which align closely to real occlusions, and (b) inferences a consistent and realistic 3D video by our effective model with the small train-inference gap.

RGBD images, which are prone to predict inconsistent results in occluded regions. To this end, we take the best of the strong synthesis capability of the diffusion model to facilitate the high visual quality of these occluded regions.

Inspired by [Li *et al.*, 2022], we use a cycle rendering way to achieve the self-supervised diffusion process in our model. The overall self-supervised denoising process is shown in Fig. 2 (a). The cycle rendering includes a forward rendering and a back rendering. Given an input image, we regard it as the i-th frame V^i. In forward rendering, we first estimate the depth image D^i by monocular depth estimation (such as [Ranftl *et al.*, 2022]). Then, we randomly sample a nearby viewpoint, which is a virtual viewpoint with a relative camera pose $T = [R|t]$, where R is the rotation matrix and t is the translation matrix. We render the frame V^i to a pseudo next frame $\tilde{V}^{i+1}$ at that virtual viewpoint as follows:

$$(\tilde{V}^{i+1}, \tilde{D}^{i+1}) = \mathcal{W}(V^i, D^i, T) \qquad (7)$$

where $\mathcal{W}$ is the same 3D render process as [Liu *et al.*, 2021]. In forward rendering, we can find some occluded regions deriving from the depth image $\tilde{D}^{i+1}$, marked by $\tilde{M}^{i+1}$, so we obtain the pair $(\tilde{V}^{i+1}, \tilde{M}^{i+1})$ as shown in Fig. 2 (a). However, we do not have the ground truth of this frame. So, we render this occluded next frame back to the original viewpoint via a back rendering as follows:

$$(\tilde{\tilde{V}}^i, \tilde{\tilde{D}}^i) = \mathcal{W}((\tilde{V}^{i+1} \cdot \tilde{M}^{i+1}, \tilde{D}^{i+1} \cdot \tilde{M}^{i+1}), T^{-1}) \quad (8)$$

where T^{-1} is the inverse camera pose, and mask $\tilde{M}^{i+1}$ is element-wise multiplied with the RGBD frame $(\tilde{V}^{i+1}, \tilde{D}^{i+1})$, and then the render way is same as the forward rendering. After that, we can obtain a masked image with mask $(\tilde{\tilde{V}}^i, \tilde{\tilde{M}}^i)$ at the same viewpoint with original frame V^i. These pairs $\{(\tilde{\tilde{V}}^i, \tilde{\tilde{M}}^i), V^i\}$ enable the diffusion model to inpaint the occluded regions only on single images by a self-supervised way. We regard the masked images $\tilde{\tilde{V}}^i$ as image

conditions, and the image V^i is the ground truth of our diffusion model. We will encode the masked images to latent features, and input them into M-UNet with masks and text prompts as conditions. We encode the ground truth to latent features $\mathbf{z}_0$, and we add the noise to latent features as follows:

$$\mathbf{z}_t = \sqrt{\overline{\alpha}_t}\mathbf{z}_0 + \sqrt{(1 - \overline{\alpha}_t)}\epsilon \qquad (9)$$

The noisy latent features $\mathbf{z}_t$ are fed into M-UNet with the timestep t to predict the added noise $\hat{\epsilon}$. At every timestep, the denoising objective is the MSE loss between $\hat{\epsilon}$ and ϵ.

M-UNet. Based on UNet, we propose M-UNet with a set of Masked Enhanced Blocks (MEBs). Traditional inpainting networks struggle to predict consistency content for occluded regions, and one reason is that unmasked spatial information suffers loss in image-to-image networks, as well as the concatenation in diffusion models [Wang *et al.*, 2022b]. The diffusion model is very flexible for receiving conditions, so we design a masked enhanced block to fully leverage the unmasked regions. As shown in Fig. 3 (a), we take the same architecture of UNet as in Stable-diffusion [Rombach *et al.*, 2022], but we design several MEBs for the downblocks of the UNet, since image conditions would suffer less loss when embedded into downblocks [Wang *et al.*, 2022b]. We utilize two stacked spatially normalization layers to embed the masked image features $\tilde{\mathbf{z}}^i$ and the mask $\tilde{M}^i$ for enhancing the spatial information, and add the timestep t embedding as normal, as shown in Fig. 3 (b). We formulate the stacked spatially normalization layers as follows:

$$\mathbf{f}^i = \gamma_M(\tilde{M}) \cdot (\gamma_{\mathbf{z}}(\tilde{\mathbf{z}}) \cdot \text{Norm}(\mathbf{f}^{i-1}) + \beta_{\mathbf{z}}(\tilde{\mathbf{z}})) + \beta_{\mathbf{M}}(\tilde{M}) \ (10)$$

where the $\mathbf{f}^{i-1}$ and $\mathbf{f}^i$ are the input and output features. $\gamma(\cdot)$ and $\beta(\cdot)$ are the convolution layers to map latent features or masks into high-level spatially-adaptive features and add them with input features. The enhancement denotes that the embedded mask further enhances the spatial information of
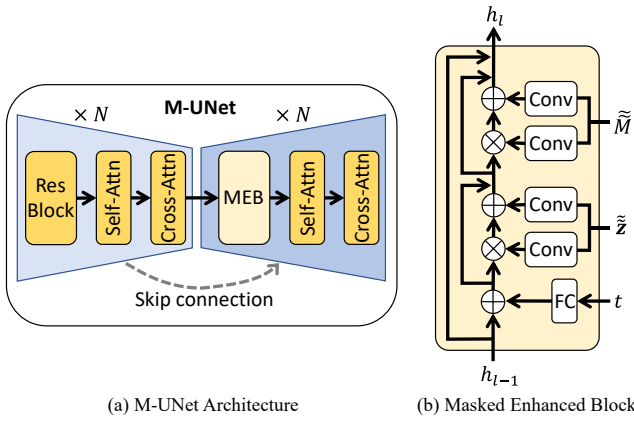
(a) M-UNet Architecture   (b) Masked Enhanced Block

Figure 3: Illustrations of M-UNet (a) and Masked Enhanced Block (MEB) (b). MEB is a key component of M-UNet.

Figure 4: Two-stage pipeline for out-animation.

unmasked regions. We add the skip-connection at the start of MEB, the start of the first embed layer, and the start of second embed layer, as shown in Fig. 3 (b), and h_{l-1} and h_l are the input and output features of MEB. We insert the MEB into each downblock to fully leverage the image conditions. The stacked normalization layers are crucial parts of M-UNet, which performs better than other embedding ways.

Training. The training process is shown in Fig. 2 (a). We denote the M-UNet as ϵ_θ, the noise added from Gaussian distribution as ϵ, the mask as $\tilde{M}^i$, and the occluded image features as $\tilde{\tilde{z}}^i = E(\tilde{\tilde{V}}^i)$ encoded by KL-VAE encoder [Rombach *et al.*, 2022]. The text prompt features are denoted as **p** encoded by CLIP text encoder [Radford *et al.*, 2021]. We regard conditions of M-UNet as the set of $\mathbf{c} = \{\tilde{M}^i, \tilde{\tilde{z}}^i, \mathbf{p}\}$, and then we learn the conditional self-supervised diffusion model via:

$$\mathcal{L}(V^i) = \mathbb{E}_{E(V^i),\mathbf{c},\epsilon\sim\mathcal{N}(\mathbf{0},\mathbf{I}),t}\left[\|\epsilon - \epsilon_\theta(\mathbf{z}_t,t,\mathbf{c})\|_2^2\right] \quad (11)$$

with t uniformly sampled from $\{1,\ldots,T\}$, ϵ_θ is optimized during training. To support the classifier-free guidance sampling, we randomly drop 10% text prompts during training.

Sampling. We use the classifier-free guidance for sampling videos with our diffusion model as shown in Fig. 2 (b). Given the ith frame V^i, we first forward render it to obtain an occluded image and mask $(\tilde{V}^{i+1}, \tilde{M}^{i+1})$. The occluded image is encoded to latent features $\tilde{z}^{i+1}$. We use the blank text $\emptyset$ with the mask and occluded features as the extra conditions, and normal conditions $\mathbf{c} = \{\tilde{M}^{i+1}, \tilde{z}^{i+1}, \mathbf{p}\}$ consists of the text prompt, the mask, and occluded image features. The guidance way in the sampling procedure can be formulated:

$$\hat{\epsilon}_\theta(\mathbf{z}_t,\mathbf{c}) = \epsilon_\theta(\mathbf{z}_t,\mathbf{c}) + s \cdot (\epsilon_\theta(\mathbf{z}_t,\mathbf{c}) - \epsilon_\theta(\mathbf{z}_t,\tilde{M}^{i+1},\tilde{z}^{i+1},\emptyset)) \quad (12)$$

where s is the guidance scale. We sample the latent features on a subset $\{\mathbf{z}_{\tau_0},\ldots,\mathbf{z}_{\tau_S}\}$, where τ is an increasing sub-sequence of $[0,\ldots,T]$ of length $S+1$. We denoise each $\mathbf{z}_{\tau_i}$ to $\mathbf{z}_{\tau_{i-1}}$ as follows:

$$\mathbf{z}_{\tau_{i-1}} = \sqrt{\overline{\alpha}_{\tau_{i-1}}}\left(\sqrt{\frac{1}{\overline{\alpha}_{\tau_i}}}\mathbf{z}_{\tau_i} - \sqrt{\frac{1}{\overline{\alpha}_{\tau_i}}-1}\hat{\epsilon}_\theta(\mathbf{z}_{\tau_i},\mathbf{c})\right) + \sqrt{1-\overline{\alpha}_{\tau_{i-1}}}\hat{\epsilon}_\theta(\mathbf{z}_{\tau_i},\mathbf{c}) \quad (13)$$
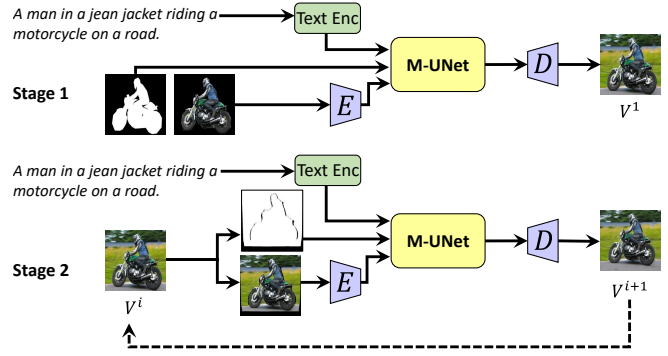
This is known as the DDIM sampler. At the start of sampling, $\mathbf{z}_{\tau_S}$ is the pure Gaussian noise, after S steps, we obtain the predicted features $\hat{\mathbf{z}}_{\tau_0}$ as the red dotted arrow shows in Fig. 2 (b), and $\hat{\mathbf{z}}_{\tau_0}$ can be decoded to a predicted $\hat{V}^{i+1}$ frame with a single pass through KL-VAE decoder D. Then, we can randomly sample a 3D video iteratively.

3.3 Out-animation

Towards real-world animation, we propose a new task named *Out-animation*, which generates complete videos with reference to given input objects. The challenge is how to generate high-quality and diverse videos which contain the input objects and retain their inherent properties.

The definition of **Out-animation** is: Given input objects (or selected parts of an image), out-animation aims to extend input objects along both space and time dimensions and outputs an animated video.

In this work, we present a simple two-stage pipeline. Our pipeline consists of two stages as shown in Fig. 4, and both only depend on the denoising M-UNet.

For stage-1, input objects may have arbitrary shapes. To generate desired complete images, we can train our diffusion model based on M-UNet to denoise input objects to the ground truth images. Training data can be collected from semantic segmentation datasets (such as COCO [Caesar *et al.*, 2018]). Similar to our strategy for inpainting 3D photography, we regard source images as the ground truth and segmented objects and masks as image conditions. Also, text prompts are conditions for this diffusion model. For inference of a new scene, we input the masks, objects, and text prompts to the trained diffusion model. Then, after S steps denoising process, we can obtain the final latent features, and decode them to the first frame V^1 via KL-VAE decoder. The stage-2 pipeline regards results that came from stage-1 as the starting frame, and iteratively renders the current frame V^i and then generates the next frame V^{i+1} as shown in Fig. 4.

Our method is different from the 3d photography methods [Shih *et al.*, 2020; Jampani *et al.*, 2021; Li *et al.*, 2022] based on depth-inpainting, since we only utilize depth images in our rendering processes. We aim to leverage the diffusion model to predict better novel views, rather than training a complex model or extra model for depth-inpainting.

4 Experiments

For a fair comparison with state-of-the-art methods, we evaluate synthesis results on two datasets, RealEstate10k (RE10K) [Zhou *et al.*, 2018] which provides about 10 million frames derived from about 80k video clips of static scenes, and MannequinChallenge (MC) [Li *et al.*, 2019] with more than 170k frames derived from about 2k YouTube videos. To validate the effectiveness of our M-UNet, we also evaluate the image outpainting in COCO [Caesar *et al.*, 2018].

Baselines and Metrics. For **novel view synthesis**, we quantitatively and qualitatively compared our method with recent state-of-the-art methods for which code is released: SynSin [Wiles *et al.*, 2020], Single-image MPI (SMPI) [Tucker and Snavely, 2020], 3d-photo [Shih *et al.*, 2020], and AdaMPI [Han *et al.*, 2022]. In our experiments, we evaluate the released pretrained models. SynSin and SMPI models were trained on RealEstate10K. 3d-photo, AdaMPI, and ours were trained on COCO. For a fair comparison, 3d-photo, AdaMPI, and our method all use MiDaS [Ranftl *et al.*, 2022] for depth estimation. SynSin and SMPI models are trained on RE10K, which randomly samples source and target frames from 57K training clips. Our model loaded the weights from Stable-diffusion, and only trained the 3D photography on COCO. We use the same intrinsic matrices, source camera poses and target camera poses for all methods. Following 3d-photo [Shih *et al.*, 2020], we measure three metrics including LPIPS [Zhang *et al.*, 2018], PSNR, and SSIM. For **image outpainting**, we quantitatively and qualitatively compared with the pretrained model SDM-v1.4 [Rombach *et al.*, 2022] and the finetuned version on COCO by our implement. We measure the FID, IS, and CLIP-SIM of outpainted images with ground-truth images. Metric details can be found in the supplementary.

Quantitative Results on RealEstate10k. RealEstate10k (RE10K) [Zhou *et al.*, 2018] is a video dataset consisting of 10K YouTube videos of static scenes. We randomly sample 1K video clips from test set for evaluation. We use the given camera intrinsics and extrinsics following RE10K. Specifically, we choose the first frame (t=1) from each test clip as the source view and consider the fifth (t=5) frame and tenth (t=10) frames as target views. We compute the evaluation metrics between predictions and ground-truth views. The results are shown in Tab. 1. We report our results of the trained model loaded from Stable-diffusion and the model only trained on COCO from scratch. 3d-photo and our method do not need video datasets and perform zero-shot on RE10K dataset because both are only trained on COCO. Compare with multi-view methods (SynSin and SMPI), our method outperforms SynSin on almost all results and shows better LPIPS than SMPI, indicating that we can achieve high realistic results than others. However, since we will generate new better content for occluded regions, the structural similarity with GT target views would slightly drop, but show better visual effects. Compared zero-shot performance with 3d-photo and AdaMPI, we perform better results than 3d-photo by a large margin and competitive results with AdaMPI.

Quantitative Results on MannequinChallenge (MC). MC is a video dataset of video clips of people freezing in

Method	Video	LPIPS↓		PSNR↑		SSIM↑	
		$t=5$	$t=10$	$t=5$	$t=10$	$t=5$	$t=10$
3d-photo*		0.209	0.266	16.00	15.27	0.43	0.41
SynSin	✓	0.063	0.097	24.42	21.73	0.81	0.71
SMPI	✓	0.055	0.106	**26.90**	**23.32**	**0.87**	**0.78**
AdaMPI*		0.056	0.100	25.83	22.01	0.84	0.73
Ours (Scratch)*		0.052	0.099	25.11	21.18	0.83	0.72
Ours*		**0.049**	**0.095**	25.35	21.36	0.84	0.72

Table 1: Results on RealEstate10K. ∗ denotes zero-shot performance. Underline masks the single-view methods. We evaluate the target views at two timesteps t=5 and t=10, size=256×256.

Method	Video	LPIPS↓		PSNR↑		SSIM↑	
		$t=3$	$t=5$	$t=3$	$t=5$	$t=3$	$t=5$
3d-photo		0.495	0.590	11.88	11.23	0.29	0.27
SynSin	✓	0.304	0.404	15.53	13.83	0.44	0.37
SMPI	✓	0.349	0.453	**16.26**	**14.56**	0.47	**0.41**
AdaMPI		0.298	0.393	16.21	14.35	**0.49**	**0.41**
Ours (Scratch)		0.277	0.372	15.87	13.87	0.47	0.37
Ours		**0.272**	**0.367**	16.01	13.99	0.47	0.38

Table 2: Zero-shot Results on MannequinChallenge, underline masks the single-view methods. Two timesteps t=3 and t=5.

diverse and natural poses, which is collected and processed similarly to RE10K. Similar to [Jampani *et al.*, 2021], we randomly sample 200 video clips from the test set, and we choose the first frame (t=1) as the source view, and the third frame (t=3) and fifth frame (t=5) as target views. The results are shown in Tab. 2. Our method outperforms 3d-photo and SynSin on all results by a large margin. Our method achieved the best LPIPS and comparable results for PSNR and SSIM with SMPI and AdaMPI, indicating the best perceptual similarity and good structure similarity.

Quantitative Results on COCO. We filter COCO-2017 [Caesar *et al.*, 2018] to an outpainting dataset containing 117266 training images and 4952 test images, and details can be found in the supplementary. We evaluate the pretrained Stable-diffusion V1.4 (SDM), a finetuned model (SDM-FT) on COCO for the text-to-image task. And we finetune a model (w/ concat) via concatenation in the input layer. Evaluation results are shown in Tab. 3. SDM-FT achieves a better FID than SDM, but degrades on IS and CLIP-SIM. It is reasonable that the generated images from the finetuned model are closer to the original images, while the quality, diversity, and text relevance are slightly worse. Our model sig-

Method	FID↓	IS↑	CLIP-SIM↑
SDM (T2I)	16.77	35.13	0.3212
SDM-FT (T2I)	13.01	34.09	0.3112
SDM w/ cross-fusion	11.33	38.40	0.3214
SDM w/ concat	11.22	36.08	0.3223
Ours w/ MEB, full (Scratch)	16.51	35.95	0.3218
Ours w/ MEB, w/o mask	10.66	37.59	0.3216
Ours w/ MEB, full	**10.65**	**38.61**	**0.3226**

Table 3: Outpainting results on COCO.

(a) Qualitative Results for 3D Photography on Benchmarks (RE10K, MC)

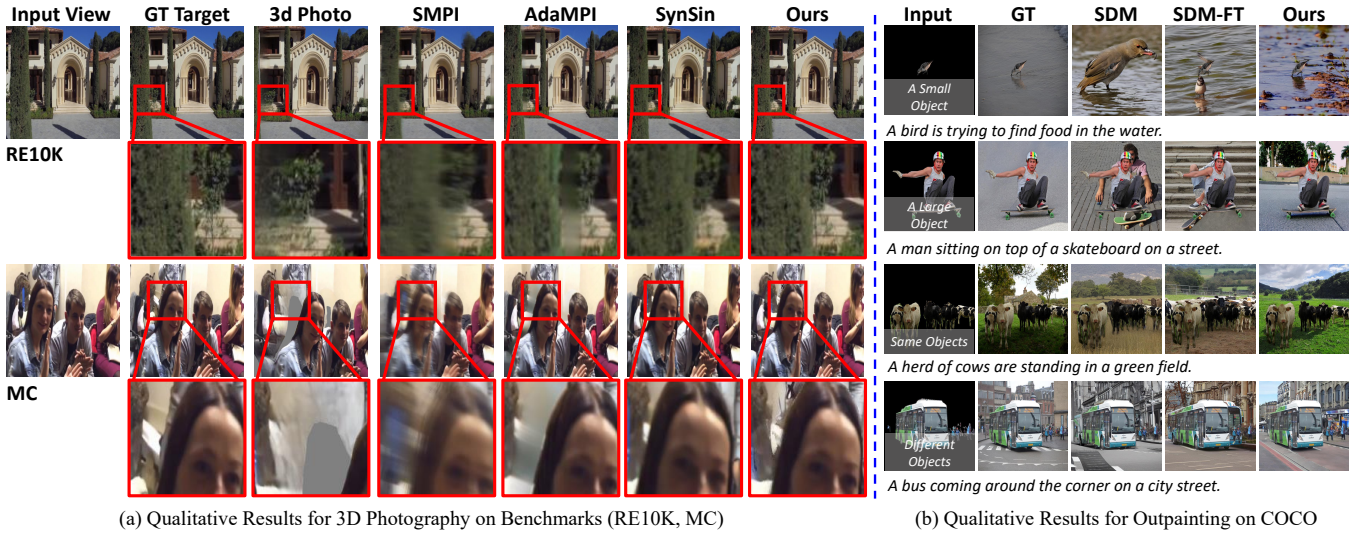(b) Qualitative Results for Outpainting on COCO

Figure 5: Qualitative results on benchmarks. (a) 3D Photography results, our method generates clear visual content and preserves details better than others. (b) Image Outpainting results, our method is able to handle various input objects and generate realistic images.

nificantly reduces the FID score to 10.65, and achieves better IS and CLIP-SIM than Stable-diffusion models. Without the pretrained weights, our model can still achieve better outpainting results than SDM on all metrics. We also compare with the cross fusion, introduced two learnable weights [Zhu *et al.*, 2020] to embed masks, and a variant of our method without masks. Our embedding way shows better quality, diversity, and text relevance, and also shows the improvement by the enhancement of masks.

Qualitative Results. We qualitatively compare the novel view synthesis in RE10K and MC. In Fig. 5. (a), the goal is to generate a better target view given the input view when the camera moves. The first two rows show results on RE10K, and the second row shows the details in the same regions that need to be generated. Such as the right occlusion regions of trees, 3d-photo produces results with large visual distortions and other methods generate blurrier results, while our method generates the best realistic results and preserves the structures better. We show another sample in a more challenging benchmark MC in the last two rows. 3d-photo generates worse results with more artifacts and the wrong position, and other methods also generate blurrier results in occlusion regions, while our method generates clear content in these regions and preserves the clear details for other regions. We compare various outpainted sample results in COCO in Fig. 5. (b). We use the same captions to generate the scenes for all methods. Compared with SDM and SDM-FT, our method generates more consistent and high-fidelity results.

Model Analysis. We vary the guidance strength for outpainting task on COCO in Fig. 6. The experiment reveals that increasing guidance strength leads to improvements in both FID and IS. However, it also indicates that once guidance strength surpasses 7.5, the gains in both FID and IS become less significant. We compare the inference time for generating one frame in Table. 4. Our method shows significant improvement compared to 3d-photo and is comparable
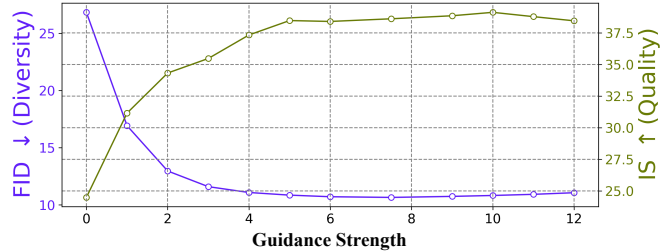

Figure 6: Diversity and quality by varying guidance strength.

Method	Backbone	Time (second)
3d-photo [Shih *et al.*, 2020]	CNN	120
SynSin [Wiles *et al.*, 2020]	GAN	1.8
SMPI [Tucker and Snavely, 2020]	GAN	1.8
AdaMPI [Han *et al.*, 2022]	GAN	0.6
Ours	Diffusion	4

Table 4: Inference time for 1 frame on NVIDIA's V100 GPU.

to other GAN-based methods. Tab. 4 illustrates a 20-step diffusion process. However, it is possible to improve the speed by sacrificing performance with fewer steps.

5 Conclusion

To reduce the train-inference gap in 3D photography training on single images, a novel self-supervised diffusion model is proposed that can generate high-quality 3D videos from single images. We first generate the masked regions that are closely aligned to real occluded regions in 3D rendering and then train a diffusion model with masked enhanced modules to inpaint these regions. Towards the real application of animation, we present the out-animation, which extends the space and time of input objects. Experimental results on real datasets validate the effectiveness of our method.

Ethical Statement

This work can be used to generate backgrounds of images, which may include people, products, etc. Therefore, this technology may be wrongly used to infringe copyright or personal privacy. Out-animation can generate new content and carry risks related to deceptive and otherwise harmful content. As technology improves, it will mistake generated images for authentic ones. Meanwhile, more research needs to be done to change the societal biases in training data.

Acknowledgments

Contribution Statement

Xiaodong Wang and Chenfei Wu contributed equally to this work. Xiaodong Wang was responsible for the design and implementation, while Chenfei Wu oversaw this work. The other authors were involved in method discussions, experiments, and paper writing of this work. Nan Duan led this work. Yuejian Fang and Nan Duan are the corresponding authors of this work.

References

[Caesar *et al.*, 2018] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1209–1218, 2018.

[Han *et al.*, 2022] Yuxuan Han, Ruicheng Wang, and Jiaolong Yang. Single-view view synthesis in the wild with learned adaptive multiplane images. *arXiv preprint arXiv:2205.11733*, 2022.

[Ho *et al.*, 2020] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

[Hu *et al.*, 2021] Ronghang Hu, Nikhila Ravi, Alexander C. Berg, and Deepak Pathak. Worldsheet: Wrapping the world in a 3d sheet for view synthesis from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12528–12537, 2021.

[Jampani *et al.*, 2021] Varun Jampani, Huiwen Chang, Kyle Sargent, Abhishek Kar, Richard Tucker, Michael Krainin, Dominik Kaeser, William T. Freeman, David Salesin, and Brian Curless. SLIDE: Single image 3d photography with soft layering and depth-aware inpainting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12518–12527, 2021.

[Kopf *et al.*, 2020] 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 Transactions on Graphics (TOG)*, 39(4):76–1, 2020.

[Lai *et al.*, 2021] Zihang Lai, Sifei Liu, Alexei A. Efros, and Xiaolong Wang. Video Autoencoder: Self-supervised disentanglement of static 3D structure and motion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9730–9740, 2021.

[Li and Kalantari, 2020] 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.

[Li *et al.*, 2019] Zhengqi Li, Tali Dekel, Forrester Cole, Richard Tucker, Noah Snavely, Ce Liu, and William T. Freeman. Learning the depths of moving people by watching frozen people. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4521–4530, 2019.

[Li *et al.*, 2021] Jiaxin Li, Zijian Feng, Qi She, Henghui Ding, Changhu Wang, and Gim Hee Lee. Mine: Towards continuous depth mpi with nerf for novel view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12578–12588, 2021.

[Li *et al.*, 2022] Zhengqi Li, Qianqian Wang, Noah Snavely, and Angjoo Kanazawa. Infinitenature-zero: Learning perpetual view generation of natural scenes from single images. In *European Conference on Computer Vision*, pages 515–534. Springer, 2022.

[Liu *et al.*, 2019] Xihui Liu, Guojun Yin, Jing Shao, Xiaogang Wang, et al. Learning to predict layout-to-image conditional convolutions for semantic image synthesis. *Advances in Neural Information Processing Systems*, 32, 2019.

[Liu *et al.*, 2021] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14458–14467, 2021.

[Mildenhall *et al.*, 2021] 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.

[Niklaus *et al.*, 2019] Simon Niklaus, Long Mai, Jimei Yang, and Feng Liu. 3d ken burns effect from a single image. *ACM Transactions on Graphics (ToG)*, 38(6):1–15, 2019.

[Park *et al.*, 2019] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2337–2346, 2019.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, and Jack Clark. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[Ramesh *et al.*, 2022] 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.

[Ranftl *et al.*, 2022] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards Robust Monocular Depth Estimation: Mixing Datasets for Zero-Shot Cross-Dataset Transfer. 44(3):1623–1637, 2022.

[Rockwell *et al.*, 2021] Chris Rockwell, David F. Fouhey, and Justin Johnson. Pixelsynth: Generating a 3d-consistent experience from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14104–14113, 2021.

[Rombach *et al.*, 2022] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

[Saharia *et al.*, 2022] 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.

[Shih *et al.*, 2020] Meng-Li Shih, Shih-Yang Su, Johannes Kopf, and Jia-Bin Huang. 3d photography using context-aware layered depth inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8028–8038, 2020.

[Singer *et al.*, 2022] 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.

[Tan *et al.*, 2021] Zhentao Tan, Dongdong Chen, Qi Chu, Menglei Chai, Jing Liao, Mingming He, Lu Yuan, Gang Hua, and Nenghai Yu. Efficient semantic image synthesis via class-adaptive normalization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[Tucker and Snavely, 2020] Richard Tucker and Noah Snavely. Single-view view synthesis with multiplane images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 551–560, 2020.

[Villegas *et al.*, 2022] Ruben Villegas, Mohammad Babaeizadeh, Pieter-Jan Kindermans, Hernan Moraldo, Han Zhang, Mohammad Taghi Saffar, Santiago Castro, Julius Kunze, and Dumitru Erhan. Phenaki: Variable length video generation from open domain textual description. *arXiv preprint arXiv:2210.02399*, 2022.

[Wang *et al.*, 2021] Yi Wang, Lu Qi, Ying-Cong Chen, Xiangyu Zhang, and Jiaya Jia. Image synthesis via semantic composition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13749–13758, 2021.

[Wang *et al.*, 2022a] Qianqian Wang, Zhengqi Li, David Salesin, Noah Snavely, Brian Curless, and Janne Kontkanen. 3d moments from near-duplicate photos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3906–3915, 2022.

[Wang *et al.*, 2022b] Weilun Wang, Jianmin Bao, Wengang Zhou, Dongdong Chen, Dong Chen, Lu Yuan, and Houqiang Li. Semantic image synthesis via diffusion models. *arXiv preprint arXiv:2207.00050*, 2022.

[Wiles *et al.*, 2020] Olivia Wiles, Georgia Gkioxari, Richard Szeliski, and Justin Johnson. Synsin: End-to-end view synthesis from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7467–7477, 2020.

[Wu *et al.*, 2021] Chenfei Wu, Lun Huang, Qianxi Zhang, Binyang Li, Lei Ji, Fan Yang, Guillermo Sapiro, and Nan Duan. Godiva: Generating open-domain videos from natural descriptions. *arXiv preprint arXiv:2104.14806*, 2021.

[Wu *et al.*, 2022] Chenfei Wu, Jian Liang, Xiaowei Hu, Zhe Gan, Jianfeng Wang, Lijuan Wang, Zicheng Liu, Yuejian Fang, and Nan Duan. Nuwa-infinity: Autoregressive over autoregressive generation for infinite visual synthesis. *arXiv preprint arXiv:2207.09814*, 2022.

[Yu *et al.*, 2022] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2022.

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

[Zhou *et al.*, 2018] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *arXiv preprint arXiv:1805.09817*, 2018.

[Zhu *et al.*, 2020] Peihao Zhu, Rameen Abdal, Yipeng Qin, and Peter Wonka. Sean: Image synthesis with semantic region-adaptive normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5104–5113, 2020.