InfiniteNature-Zero: Learning Perpetual View Generation of Natural Scenes from Single Images

Zhengqi Li[1], Qianqian Wang[1,2], Noah Snavely[1], and Angjoo Kanazawa[3]

[1] Google Research
[2] Cornell Tech, Cornell University
[3] UC Berkeley

Abstract. We present a method for learning to generate unbounded flythrough videos of natural scenes starting from a single view, where this capability is learned from a collection of *single photographs*, without requiring camera poses or even multiple views of each scene. To achieve this, we propose a novel self-supervised view generation training paradigm, where we sample and rendering virtual camera trajectories, including cyclic ones, allowing our model to learn stable view generation from a collection of single views. At test time, despite never seeing a video during training, our approach can take a single image and generate long camera trajectories comprised of hundreds of new views with realistic and diverse contents. We compare our approach with recent state-of-the-art supervised view generation methods that require posed multi-view videos and demonstrate superior performance and synthesis quality.

1 Introduction

There are millions of photos of natural landscapes on the Internet, capturing breathtaking scenery across the world. Recent advances in vision and graphics have led to the ability to turn such images into compelling 3D photos [38,69,30]. However, most prior work can only extrapolate scene content within a limited range of views corresponding to a small head movement. What if, instead, we could step into the picture and fly through the scene like a bird and explore the 3D world, where diverse elements like mountain, lakes, and forests emerge naturally as we move? This challenging new task was recently proposed by Liu *et al.* [43], who called it *perpetual view generation*: given a single RGB image, the goal is to synthesize a video depicting a scene captured from a moving camera with an arbitrary long camera trajectory. Methods that tackle this problem can have applications in content creation and virtual reality.

However, perceptual view generation is an extremely challenging problem: as the camera travels through the world, the model needs to fill in unseen missing regions in a harmonious manner, and must add new details as new scene content approaches the camera, all the while maintaining photo-realism and diversity. Liu *et al.* [43] proposed a supervised solution that generates sequences of views in an auto-regressive manner. In order to train the model, Liu *et al.* (which we will refer to as *Infinite Nature*), require a large dataset of posed video clips of

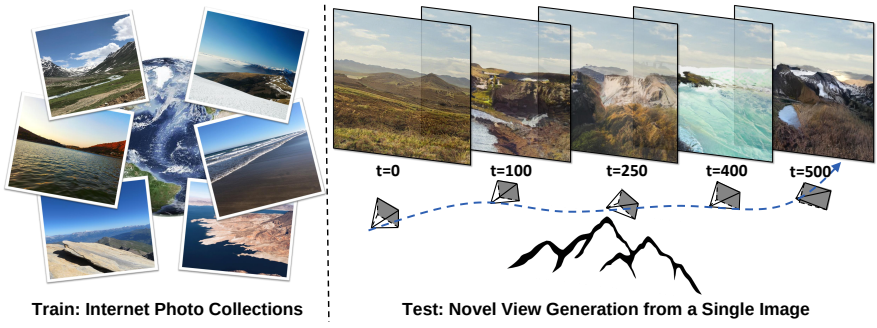

Fig. 1: **Learning perpetual view generation from single images.** Given a single RGB image input, our approach generates novel views corresponding to a continuous long camera trajectory, without ever seeing a video during training.

nature scenes to supervise their model. In essence, perpetual view generation is a video synthesis task, but the requirement of *posed* video makes data collection a big challenge. Obtaining large amounts of diverse, high-quality, and long videos of nature scenes is challenging enough, let alone estimating accurate camera poses on these videos at scale. In contrast, Internet photos of nature landscapes are much easier to collect and have spurred many research problems such as panorama generation [73,42], image extrapolation [10,62], image editing [56], and multi-model image synthesis [17,29].

How can we use existing single-image datasets for the 3D view generation task? In other words, can we learn view generation by simply observing many photos, without requiring video or camera poses? Training with less powerful supervision would seemingly make this already challenging synthesis task even more challenging. And doing so is not a straightforward application of prior methods. For instance, prior single-image view synthesis methods either require posed multi-view data [86,60,36], or can only extrapolate within a limited range of viewpoints [38,69,30,28]. Other methods for video synthesis [1,78,91,40] require videos spanning multiple views as training data, and can only generate a limited number of novel frames with no ability to control camera motion at runtime.

In this work, we present a novel method for learning perpetual view generation from only a collection of single photos, without requiring multiple views of each scene or camera information. Despite using much less information, our approach improves upon the visual quality of prior methods that require multi-view data. We do so by utilizing *virtual* camera trajectories and computing losses that enable high-quality perpetual view generation results. Specifically, we introduce a self-supervised view synthesis strategy via cyclic virtual camera trajectory, which provides the network signals for learning to generate a single step of view-synthesis without multi-view data. In addition, to learn to generate a long sequence of novel views, we employ an adversarial perpetual view generation training technique, encouraging views along a long virtual camera trajectory to be realistic and generation to be stable. The only requirement for our approach

is an off-the-shelf monocular depth network to obtain disparity for the initial frame, but this depth network does not need to be trained on our data. In this sense, our method is self-supervised, leveraging underlying pixel statistics from single-image collections. Because we train with no video data whatsoever, we call our approach *InfiniteNature-Zero*.

We show that naively training our model based on prior video/view generation methods leads to training divergence or mode collapse. We therefore introduce balanced GAN sampling and progressive trajectory growing strategies that stabilize model training. In addition, to prevent artifacts and drift during inference, we propose a global sky correction technique that yields more consistent and realistic synthesis results along long camera trajectories.

We evaluate our method on two public nature scene datasets, and compare with recent supervised video synthesis and view generation methods. We demonstrate superior performance compared to state-of-the-art baselines trained on multi-view collections, even though our model only requires single-view photos during training. To our knowledge, our work is the first to tackle unbounded 3D view generation for natural scenes trained on 2D image collections, and believe this capability will enable new methods for generative 3D synthesis that leverage more limited supervision. We encourage viewing the supplemental video for animated comparisons.

2 Related Work

Image extrapolation. An inspiring early approach to infinite view extrapolation was proposed by Kaneva *et al.* [32]. That method continually retrieves, transforms, and blends imagery to create an infinite 2D landscape. We revisit this idea in the 3D context, which requires inpainting, i.e., filling missing content within an image [25,89,90,44,94], as well as *outpainting*, extending the image and inferring unseen content outside the image boundaries [84,87,74,4,60,62] in order to generate the novel view. Super-resolution [21,39] is also an important aspect of perpetual view generation, as approaching a distant object requires synthesizing additional high-resolution detail. Image-specific GAN methods demonstrate super-resolution of textures and natural images as a form of image extrapolation [96,72,66,71]. In contrast to the above methods that address these problems individually, our methods handles all three sub-problems jointly.

Generative view synthesis. View synthesis is the problem of generating novel views of a scene from existing views. Many view synthesis methods require multiple views of a scene as input [41,7,95,49,19,11,47,59,50,83,45], though recent work also can generate novel views from a single image [9,77,55,76,68,86,31,70,37,61]. These methods often require multi-view posed datasets such as RealEstate10k [95]. However, empowered by advances in neural rendering, recent work shows that one can unconditionally generate 3D scene representations like neural radiance fields for 3D-aware image synthesis [63,53,16,52,23,5]. Many of these methods only require unstructured 2D images for training. When GAN inversion is possible, these methods can also be used for single-image view synthesis, although they

have only been demonstrated on specific object categories like faces [6,5]. All of the work mentioned above allows for a limited range of output viewpoints. In contrast, our method can generate new views perpetually given a single input image. Most related to our work is Liu *et al.* [43], which also performs perpetual view generation. However, Liu *et al.* require posed videos during training. Our method can be trained with unstructured 2D images, and also experimentally achieves better view generation diversity and quality.

Video synthesis. Our problem is also related to the problem of video synthesis [13,75], which can be roughly divided into the three categories: 1) unconditional video generation [78,51,20,46], which generates a video sequence given input random variables; 2) video prediction [81,82,85,80,27,40], which generates a video sequence from one or more initial observations; and 3) video-to-video synthesis, which maps a video from a source domain to a target domain. Most video prediction methods focus on generating videos of dynamic objects under a static camera [81,18,82,15,88,92,40], e.g., human motion [3] or the movement of robot arms [18]. In contrast, we focus on generating new views of static nature scenes with a moving camera. Several video prediction methods can also simulate moving cameras [14,79,1,40], but unlike our approach, they require long video sequences for training, do not reason about the underlying 3D scene geometry, and do not allow for explicit control over camera viewpoint. More recently, Koh *et al.* [36] propose a method to navigate and synthesize indoor environments with controllable camera motion. However, they require ground truth RGBD panoramas as supervision and can only generate novel frames up to 6 steps. Many prior methods in this vein also require 3D input, such as voxel grids [24] or dense point clouds [48], whereas we require only a single RGB image.

3 Learning view generation from single-image collections

We formulate the task of perpetual view generation as follows: given an starting RGB image I_0, generate an image sequence $(\hat{I}_1, \hat{I}_2, ..., \hat{I}_t, ...)$ corresponding to an arbitrary camera trajectory $(c_1, c_2, ..., c_t, ...)$ starting from I_0, where the camera viewpoints c_t can be specified either algorithmically or via user input.

The prior Infinite Nature method tackles this problem by decomposing it into three phases: **render**, **refine** and **repeat** [43]. Given an RGBD image $(\hat{I}_{t-1}, \hat{D}_{t-1})$ at camera c_{t-1}, the **render** phase renders a new view $(\tilde{I}_t, \tilde{D}_t)$ at c_t by transforming and warping $(\hat{I}_{t-1}, \hat{D}_{t-1})$ using a differentiable 3D renderer $\mathcal{W}$. This yields a warped view $(\tilde{I}_t, \tilde{D}_t) = \mathcal{W}\left((I_{t-1}, D_{t-1}), T^t_{t-1}\right)$, where T^t_{t-1} is an $SE(3)$ transformation from c_{t-1} to c_t. In the **refine** phase, the warped RGBD image $(\tilde{I}_t, \tilde{D}_t)$ is fed into a refinement network F_θ to fill in missing content and add details: $(\hat{I}_t, \hat{D}_t) = F_\theta(\tilde{I}_t, \tilde{D}_t)$. The refined outputs $(\hat{I}_t, \hat{D}_t)$ are then treated as a starting view for the next iteration of the **repeat** step, from which the process can be repeated. We refer readers to the original work for more details [43].

To supervise the view generation model, Infinite Nature trains on video clips of natural scenes, where each video frame has camera pose dervied from structure from motion (SfM) [95]. During training, it randomly chooses one frame in a

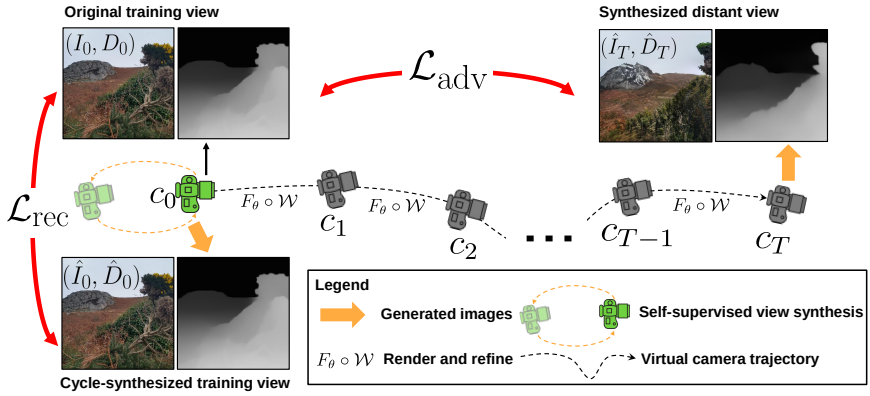

Fig. 2: **Self-supervised view generation via virtual cameras.** Given a starting RGBD image (I_0, D_0) at viewpoint c_0, our training procedure samples two virtual camera trajectories: 1) a cycle to and back from a single virtual view (orange paths), creating a *self-supervised view synthesis* signal enforced by the reconstruction loss $\mathcal{L}_{\text{rec}}$. 2) a path of virtual cameras for which we generate corresponding images via render-refine-repeat process (black path). An adversarial loss $\mathcal{L}_{\text{adv}}$ between the final view $(\hat{I}_T, \hat{D}_T)$ and the real image (I_0, D_0) enables the network to learn long-range view generation.

video clip as the starting view I_0, and performs the render-refine-repeat process along the provided SfM camera trajectory. At a camera viewpoint c_t along the trajectory, a reconstruction loss and an adversarial loss are computed between the image predicted by the network $(\hat{I}_t, \hat{D}_t)$ and the corresponding real RGBD frame (I_t, D_t). However, obtaining long nature videos with accurate camera poses is difficult due to potential distant or non-Lambertian contents of landscapes (e.g., sea, mountain, and sky). In contrast, our method does not require videos at all, whether with camera poses or not.

We show that 2D photo collections alone provide sufficient supervision signals to learn perceptual view generation, given an off-the-shelf monocular depth prediction network. Our key idea is to sample and render *virtual* camera trajectories starting from the training image, using the refined depth to warp to the next view. In particular, we generate two kinds of camera trajectories, illustrated in Fig. 2: First, we generate *cyclic* camera trajectories that start and end at the training image, from which the image needs to be reconstructed in a self-supervised manner (Sec. 3.1). This self-supervision trains our network to learn geometry-aware view refinement during view generation. Second, we synthesize longer virtual camera trajectories from which we compute an adversarial loss $\mathcal{L}_{\text{adv}}$ on the rendered image (Sec. 3.2). This signal trains our network to learn stable view generation for long camera trajectories. The rest of this section describes the two training signals in detail, as well as a sky correction component (Sec. 3.3) that prevents

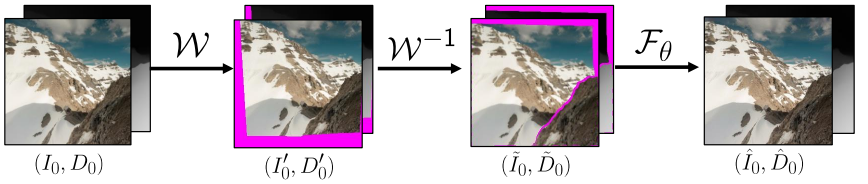

Fig. 3: **Self-supervised view synthesis.** From a real RGBD image (I_0, D_0), we synthesize an input $(\tilde{I}_0, \tilde{D}_0)$ to a refinement model by cycle-rendering through a virtual viewpoint. From left to right: input image; input rendered to a virtual "previous" view; virtual view rendered *back* to the starting viewpoint; final image $(\hat{I}_0, \hat{D}_0)$ refined with refinement network $\mathcal{F}_\theta$, trained to match the starting image.

drift in sky regions during test-time, yielding more realistic and stable long-range trajectories for nature scenes.

3.1 Self-supervised view synthesis

In Infinite Nature's supervised learning framework, a reconstruction loss is applied between predicted and corresponding real RGBD images to train the network to learn to refine the inputs rendered from a previous viewpoint. Note that unlike the task of free-form image inpainting [94], this next-view supervision provides crucial signals for the network to learn to add suitable details and to fill in missing regions around disocclusions using background context, while preserving 3D perspective—in other words, we can't fully simulate the necessary signal using standard inpainting supervision. Instead, our idea is to treat the known real image as the held-out "next" view, and simulate a rendered image input from a virtual "previous" viewpoint. We implement this idea in an elegant way by rendering a *cyclic* virtual camera trajectory starting and ending at the known input training view, then comparing the final rendered image at the end of the cycle to the known ground truth input view. In practice, we find that a cycle including just one other virtual view (i.e., warping to a sampled viewpoint, then rendering back to the input viewpoint) is sufficient. Fig. 3 shows an example sequence of views produced in such a cyclic rendering step.

In particular, we first estimate the depth D_0 from a real image I_0 using the off-the-shelf mono-depth network [57]. We randomly sample a nearby viewpoint with a relative camera pose T within a set of maximum values for each camera parameter. We then synthesize the view at virtual pose T by rendering (I_0, D_0) to a new image $(I'_0, D'_0) = \mathcal{W}((I_0, D_0), T)$. Next, to encourage the network to learn to fill in missing background contents at disocclusions, we create a per-pixel binary mask M'_0 derived from the rendered disparity D'_0 at the virtual viewpoint [43,30]. Lastly, we render this virtual view with mask (I'_0, D'_0, M'_0) back to the starting viewpoint with T^{-1}: $(\tilde{I}_0, \tilde{D}_0, \tilde{M}_0) = \mathcal{W}\left((I'_0, D'_0, M'_0), T^{-1}\right)$ where the rendered mask is element-wise multiplied with the rendered RGBD image. Intuitively, this strategy constructs inputs whose pixel statistics, including blurriness and missing

content, are similar to those produced by warping one view forward to a next viewpoint, forming naturalistic input to view refinement.

The cycle-rendered images $(\tilde{I}_0, \tilde{D}_0)$ are then fed into the refinement network F_θ, whose outputs $(\hat{I}_0, \hat{D}_0) = F_\theta(\tilde{I}_0, \tilde{D}_0)$ are compared to the original RGBD image (I_0, D_0) to yield a reconstruction loss $\mathcal{L}_{\text{rec}}$. Because this method does not require actual multiple views or SfM camera poses, we can generate an effectively infinite set of virtual camera motions during training. Because the target view is always an input training view we seek to reconstruct, this approach can be thought of as a self-supervised way of training view synthesis.

3.2 Adversarial perpetual view generation

Although the insight above enables the network to learn to refine a rendered image, directly applying such a network iteratively during inference over multiple steps quickly degenerates (see third row of Fig. 4). As observed by prior work [43], we must train a synthesis model through multiple recurrently-generated camera viewpoints in order for learned view generation to be stable. Therefore, in addition to the self-supervised training in Sec. 3.1, we also train on longer virtual camera trajectories. In particular, during training, for a given input RGBD image (I_0, D_0), we randomly sample a virtual camera trajectory $(c_1, c_2, ..., c_T)$ starting from (I_0, D_0) by iteratively performing render-refine-repeat T times, yielding a sequence of generated views $(\hat{I}_1, \hat{I}_2, ..., \hat{I}_T)$. To avoid the camera flying into out-of-distribution viewpoints (e.g., crashing into mountains or water) we adopt the auto-pilot algorithm from [43] to sample the camera path. The auto-pilot algorithm determines the pose of the next view based on the proportion of sky and foreground elements as determined by the estimated disparity map at the current viewpoint (see supplemental material for more details). Next, we discuss how we train our model using such sampled virtual camera trajectories.

Balanced GAN sampling. We now have a generated sequence of views along a virtual camera trajectory from the input image, but we do not have the ground truth sequence corresponding to these views. How can we train the model without such ground truth? We find that it is sufficient to compute an adversarial loss that trains a discriminator to distinguish between real images and the synthesized "fake" images along the virtual camera path. One straightforward implementation of this idea is to treat all T predictions $\{\hat{I}_t, \hat{D}_t\}_{t=1}^{T}$, along the virtual path as fake samples, and sample T real images randomly from the dataset. However, this strategy leads to unstable training, because there is a significant discrepancy in pixel statistics between the generated view sequence and the set of sampled real photos: a generated sequence along a camera trajectory has frames with similar content with smoothly changing viewpoints, whereas randomly sampled real images from the dataset exhibit completely different content and viewpoints. This vast difference in the distribution of images that the discriminator observes leads to unstable training in conditional GAN settings [24]. To address this issue, we propose a simple but effective technique to stabilize the training. Specifically, for a generated sequence, we only feed the discriminator the generated image

$(\hat{I}_T, \hat{D}_T)$ at the last camera c_T as the fake sample, and use its corresponding input image (I_0, D_0) at the starting view as the real sample, as shown in Fig. 2. In this case, the real and fake sample in each batch will exhibit similar content and viewpoint variations. Further, during each training iteration, we randomly sample the length of virtual camera trajectory T between 1 and a predefined maximum length $T_{\max}$, so that the prediction at any viewpoint and step will be sufficiently trained.

Progressive trajectory growing. We observe that without the guidance of ground truth sequences, the discriminator quickly gain an overwhelming advantage over the generator at the beginning of training. Similarly to the issues explored in prior 2D GAN works [34,33,67], it would take longer time for the network to predict plausible views at more distant viewpoints. As a result, the discriminator will easily distinguish real images from fake ones generated at distant views, and offer meaningless gradients to the generator. To address this issue, we propose to grow the length of virtual camera trajectory progressively. In particular, we begin with self-supervised view synthesis as described in Sec. 3.1 and pretrain the model for 200K steps. We then increase the maximum length of the virtual camera trajectory T by 1 every 25K iterations until reaching the predefined maximum length $T_{\max}$. This progressive growing strategy ensures that the images rendered at a previous viewpoint c_{t-1} has been sufficiently initialized before being fed into the refinement network to generate view at the next viewpoint c_t.

3.3 Global sky correction

The sky is an indispensable visual element of nature scenes with unique characteristics — it should change much more slowly than the foreground content, due to the sky being at infinity. However, we found that the sky synthesized by Infinite Nature can contain unrealistic artifacts after multiple steps. We also found that mono-depth predictions can be inaccurate in sky regions, leading to sky contents to quickly approach the camera in an unrealistic manner.

Therefore, at test time we devise a method to correct the sky regions of refined RGBD images at each step by leveraging the sky content from the starting view. In particular, we use an off-the-shelf semantic segmentation [8] and the predicted disparity map to determine soft sky masks for the starting and for each generated view, which we found to be effective in identifying sky pixels. We then correct the sky texture and disparity at every step by alpha blending the homography-warped sky content from the starting view (warped according to the camera rotation's effect on the plane at infinity) with the foreground content in the current generated view. To avoid redundantly outpainting the same sky regions, we expand the input image and disparity through GAN inversion [12,10] to seamlessly create a canvas of higher resolution and field of view. We refer readers to supplementary material for more details. As shown in the penultimate column of Fig. 4, by applying global sky correction at test time, the sky regions exhibit significantly fewer artifacts, resulting in more realistic generated views.

Fig. 4: **Generated views after 50 steps with different settings.** Each row shows results for a different input image. From left to right: input view; results without balanced GAN sampling; without adversarial perpetual view generation strategy; without progressive trajectory growing; without self-supervised view synthesis; without global sky correction; full approach.

3.4 Network and supervision losses

We adopt a variant of conditional StyleGAN model, CoMod-GAN [94], as the backbone refinement module F_θ. Specifically, F_θ consists of a global encoder and a StyleGAN generator, where encoder produces a global latent code z_0 from the input view. At each refine step, we co-modulate intermediate feature layers of the StyleGAN generator through concatenation of z_0 and a latent code z mapped from a Gaussian noise. The training loss for generator and discriminator is:

$$\mathcal{L}^F = \mathcal{L}^F_{\text{adv}} + \lambda_1 \mathcal{L}_{\text{rec}}, \quad \mathcal{L}^D = \mathcal{L}^D_{\text{adv}} + \lambda_2 \mathcal{L}_{R_1} \tag{1}$$

where $\mathcal{L}^F_{\text{adv}}$ and $\mathcal{L}^D_{\text{adv}}$ are non-saturated GAN losses [22], applied on the last step of the camera trajectory and the corresponding training image. $\mathcal{L}_{\text{rec}}$ is a reconstruction loss between real images and cycle-synthesized views described in Sec 3.1: $\mathcal{L}_{\text{rec}} = \sum_l ||\phi^l(\hat{I}_0) - \phi^l(I_0)||_1 + ||\hat{D}_0 - D_0||_1$, where ϕ^l is a feature outputs at scale l from the different layer of a pretrained VGG network [64]. $\mathcal{L}_{R_1}$ is a gradient regularization term that is applied to discriminator during training [35].

4 Experiments

4.1 Datasets and baselines

We evaluate our approach on two public datasets of nature scenes: the Landscape High Quality dataset (LHQ) [73], a collection of 90K landscapes photos collected from the Internet, and the Aerial Coastline Imagery Dataset (ACID) [43], a video dataset of nature scenes with SfM camera poses.

On the ACID dataset, where posed video data is available, we compare with several state-of-the-art supervised learning methods. Our main baseline is Infinite Nature, a recent state-of-the-art view generation method designed for natural

Method	MV?	View Synthesis PSNR↑	SSIM↑	LPIPS↓	View Generation FID ↓	FID_{sw} ↓	KID↓	Style↓
GFVS [61]	Yes	11.3/11.9	0.68/0.69	0.33/0.34	109	117	0.87	14.6
PixelSynth [60]	Yes	20.0/19.7	0.73/0.70	0.19/0.20	111	119	1.12	10.54
SLAMP [1]	Yes	-	-	-	114	138	1.91	15.2
DIGAN [91]	Yes	-	-	-	53.4	57.6	0.43	5.85
Liu *et al.* [43]	Yes	23.0/**21.1**	**0.83/0.74**	0.14/0.18	32.4	37.2	0.22	9.37
Ours	No	**23.5/21.1**	0.81/0.71	**0.10/0.15**	**19.3**	**25.1**	**0.11**	**5.63**

Table 1: **Quantitative comparisons on the ACID test set.** "MV?" indicates whether a method requires (posed) multi-view data for training. We report view synthesis results with two different GTs (shown as X/Y): sequences rendered with 3D Photos [70] (left), and real sequences (right). KID and Style are scaled by 10 and 10^5 respectively. See Sec. 4.4 for descriptions of baselines.

scenes [43]. We also compare with other recent view and video synthesis methods, including geometry-free view synthesis [61] (GFVS) and PixelSynth [60], both of which are based on VQ-VAE [58,17] for long-range view synthesis. Additionally, we compare with two recent video synthesis methods, SLAMP [1] and DIGAN [91]. Following their original protocols, we train both methods with video clips of 16 frames from the ACID dataset until convergence.

For the LHQ dataset, since there is no multi-view training data and we are unaware of prior methods that can train on single images, we show results from our approach with different configurations, described in more detail in Sec. 4.5.

4.2 Metrics

We evaluate synthesis quality on two tasks that we refer to as *short-range view synthesis* and *long-range view generation*. By view synthesis, we mean the ability to render views near a source view with high fidelity, and we report standard error metrics between predicted and ground truth views, including PSNR, SSIM and LPIPS [93]. Since there is no multi-view data on the LHQ dataset, we create pseudo ground truth images over a trajectory of length 5 from a global LDI mesh [65], computed with 3D Photos [70]; we refer to the supplementary material for more details. On the ACID dataset, we report error on real video sequences where we use SfM-aligned depth maps to render images from each method. We also report results from ground truth sequences created with 3D Photos, since we observe that in real video sequences, pixel misalignments can also be caused by factors like scene motion and errors in mono-depth and camera poses.

For the task of view generation, following previous work [43], we adopt the Fréchet Inception Distance (FID), sliding window FID (FID_{sw}) of window size $\omega = 20$, and Kernel Inception Distance (KID) [2] to measure synthesis quality of different approaches. We also introduce a style consistency metric that computes an average style loss between the starting image and all the generated views along a camera trajectory. This metric reflects how much the style of a generated

Method	Configurations					View Synthesis			View Generation			
	$\mathcal{L}_{\text{rec}}$	$\mathcal{L}_{\text{adv}}$	PTG	BGS	Sky	PSNR↑	SSIM↑	LPIPS↓	FID ↓	FID_{sw} ↓	KID ↓	Style↓
Naive	✓	✓				28.0	0.87	0.07	38.1	52.1	0.25	6.36
w/o BGS	✓	✓	✓		✓	28.0	0.89	0.08	34.9	41.1	0.20	6.45
w/o PTG	✓	✓		✓	✓	28.1	0.90	0.07	35.3	42.6	0.21	6.04
w/o repeat	✓				✓	26.8	0.86	0.15	61.3	85.5	0.40	8.15
w/o SVS		✓	✓	✓	✓	26.6	0.85	0.08	23.4	30.2	0.12	6.37
w/o sky	✓	✓	✓	✓		28.3	0.90	0.07	24.8	31.3	0.11	6.43
Ours (full)	✓	✓	✓	✓	✓	**28.4**	**0.91**	**0.06**	**19.4**	**25.8**	**0.09**	**5.91**

Table 2: **Ablation study on the LHQ test set.** KID and Style are scaled by 10 and 10^5 respectively. See Sec. 4.5 for a description of each baseline.

sequence deviates from the original image; we evaluate it over a trajectory of length 50. For FID and KID calculations, we compute real statistics from 50K images randomly sampled from each of the dataset, and calculate fake statistics from 70K and 100K genereated images on ACID and LHQ respectively, where 700 and 1000 test images are used as starting images evaluated over 100 steps. Note that since SLAMP and DIGAN do not support camera viewpoint control, we only evaluate them on view generation metrics.

4.3 Implementation details

We set the maximum camera trajectory length $T_{\text{max}} = 10$. The weight of R_1 regularization λ_2 is set to 0.15 and 0.004 for LHQ and ACID datasets, respectively. During training, we found that treating a predicted view along a long virtual trajectory as ground truth and adding a small self-supervised view synthesis loss over these predictions yield more stable view generation results. Therefore we set reconstruction weight $\lambda_1 = 1$ for input training image at starting viewpoint, and $\lambda_1 = 0.05$ for the predicted frame along a long camera trajectory. Following [35], we apply lazy regularization to the discriminator gradient regularization every 16 training steps and adopt gradient clipping and exponential moving averaging to update the parameters of refinement network. For all experiments, we train on centrally cropped images of 128×128 for 1.8M steps with batch size 32 using 8 NVIDIA A100 GPUs, which takes ∼6 days to converge. At each rendering stage, we use softmax splatting [54] to 3D render images through depth. Our method can also generate higher resolution of 512×512 views. Instead of directly training the model at high resolution, which would take an estimated 3 weeks, we train an extra super-resolution module that takes one day to converge using the same self-supervised learning idea. We refer readers to the supplementary material for more details and high-resolution results.

4.4 Quantitative comparisons

Table 1 shows quantitative comparisons between our approach and other baselines on the ACID test set. Although the model only observes single images, our

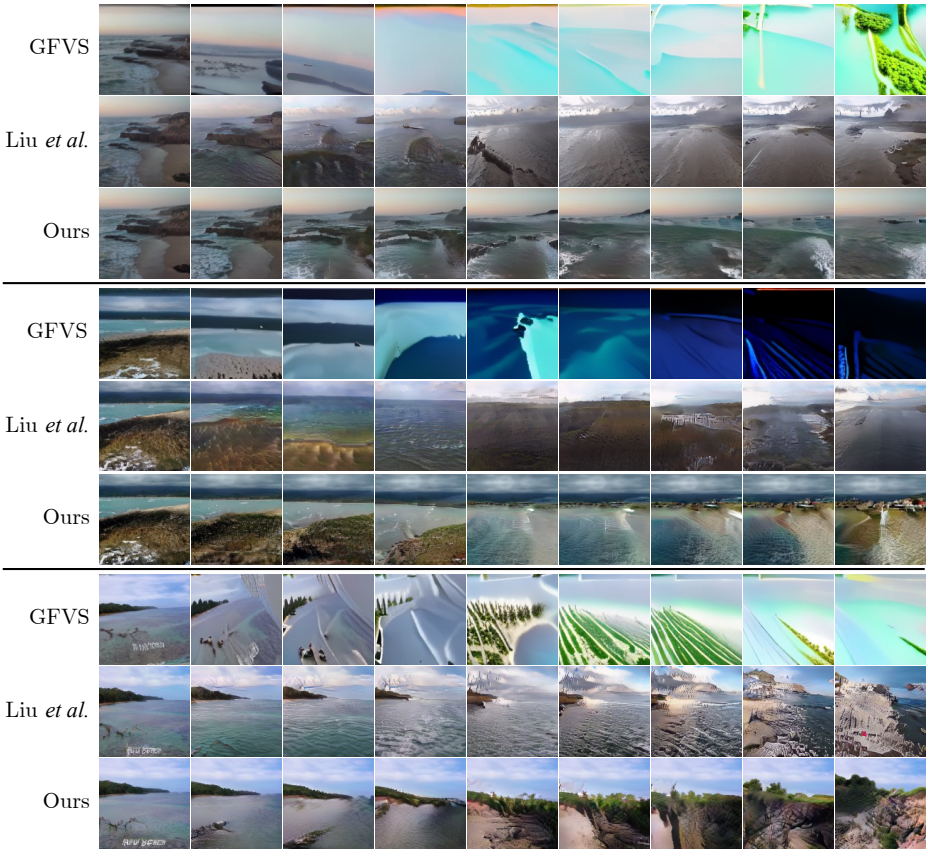

Fig. 5: **Qualitative comparisons on the ACID test set.** From left to right, we show generated views over trajectories of length 100 on three methods: GFVS [61], Liu *et al.* [43] and Ours.

approach outperforms the other baselines in view generation on all error metrics, while achieving competitive performance on the view synthesis task. Specifically, our approach demonstrates the best FID and KID scores, indicating better realism and diversity of our generated views. Our method also achieves the best style consistency, suggesting better style preservation. For the view synthesis task, we achieve the best LPIPS over the baselines, suggesting higher perceptual quality for our rendered images. We also obtain competitive low-level PSNR and SSIM with the supervised learning methods from Infinite Nature on the ACID test set, which applied explicit reconstruction loss over real sequences.

4.5 Ablation study

We perform an ablation study on the LHQ test set to analyze the effectiveness of components in the proposed system. We ablate our system with different

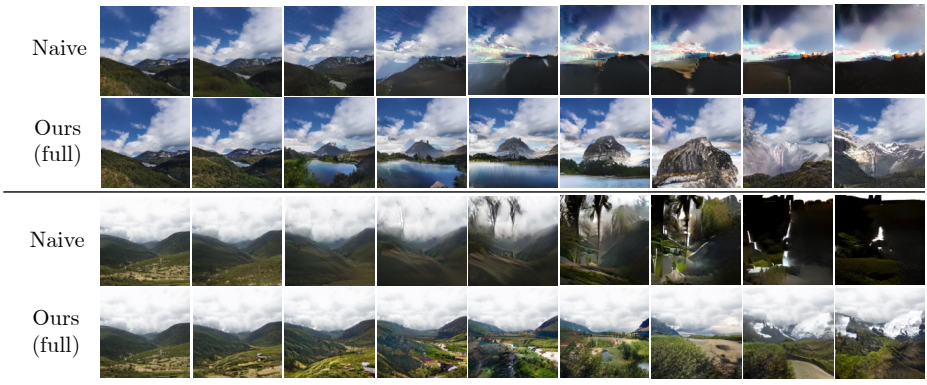

Fig. 6: **Qualitative comparisons on the LHQ test set.** On three starting views, from left to right, we show generated views over trajectories of length 100 from a naive baseline and our full approach. See Sec. 4.5 for more details.

configurations: (1) a naive baseline where we apply an adversarial loss between all the predictions along a camera trajectory and a set of randomly sampled real photos, and apply geometry re-grounding introduced in Infinite Nature [43] during testing (Naive); without (2) using balanced GAN sampling (w/o BGS); (3) progressive trajectory growing (w/o PTG), (4) GAN training via long camera trajectories (w/o repeat), (5) applying self-supervised view synthesis (w/o SVS), and (6) employing global sky correction (w/o sky). Quantitative and qualitative comparisons are shown in Table 2 and Fig. 4 respectively. Our full system achieves the best view synthesis and view generation performance compared with other alternatives. In particular, adding self-supervised view synthesis significantly improves view synthesis performance. Training via virtual camera trajectories, adopting introduced GAN sampling/training strategies, and applying global sky correction all improve view generation performance by a large margin.

4.6 Qualitative comparisons

Fig. 5 shows visual comparisons between our approach, Infinite Nature [43], and GFVS [61] on the ACID test set. GFVS quickly degenerates due to the large distance between the input and generated viewpoints. Infinite Nature can generate plausible views over multiple steps, but the content and style of generated views quickly transform into an unrelated unimodal scene. Our approach, in contrast, not only generates more consistent views with respect to starting images, but demonstrates significantly improved synthesis quality and realism.

Fig. 6 shows visual comparisons between the naive baseline described in Sec. 4.5 and our full approach. The generated views from the baseline quickly deviate from realism due to ineffective training/inference strategies. In contrast, our full approach can generate much more realistic, consistent, and diverse results

Fig. 7: **Perpetual view generation.** Given a single RGB image, we show the results of our method generating sequences of 500 realistic new views of natural scenes without suffering significant drift. Please see video for animated results.

over long camera trajectories. For example, the views generated by our approach cover diverse and realistic natural elements such as lakes, trees, and mountains.

4.7 Single-image perpetual view generation

Finally, we visualize our model's ability to generate long view trajectories from a single RGB image in Fig. 7. Although our approach only sees single images during training, it learns to generate long sequences of 500 new views depicting realistic natural landscapes, without suffering significant drift or degeneration. We refer readers to the supplementary material for the full effect and results generated from different types of camera trajectories.

5 Discussion

Limitations and future directions. Our method inherits some limitations from prior video and view generation methods. For example, although our method produces globally consistent backgrounds, it does not ensure global consistency of foreground contents. Addressing this issue requires generating an entire 3D world model, which is an exciting direction to explore. In addition, as with Infinite Nature, our method can generate unrealistic views if the desired camera trajectory is not seen during training such as in-place rotation. Alternative generative methods such as VQ-VAE [58] and diffusion models [26] may provide promising paths towards addressing this limitation.

Conclusion. We presented a method for learning perpetual view generation of natural scenes solely from single-view photos, without requiring camera poses and multi-view data. At test time, given a single RGB image, our approach allows for generating hundreds of new views covering realistic natural scenes along a long camera trajectory. We conduct extensive experiments and demonstrate the improved performance and synthesis quality of our approach over prior supervised approaches. We hope this work demonstrates a new step towards unbounded generative view synthesis of nature scenes from Internet photo collections.

References

1. Akan, A.K., Erdem, E., Erdem, A., Guney, F.: Slamp: Stochastic latent appearance and motion prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 14728–14737 (October 2021)
2. Bińkowski, M., Sutherland, D.J., Arbel, M., Gretton, A.: Demystifying mmd gans. arXiv preprint arXiv:1801.01401 (2018)
3. Blank, M., Gorelick, L., Shechtman, E., Irani, M., Basri, R.: Actions as space-time shapes. In: Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1. vol. 2, pp. 1395–1402. IEEE (2005)
4. Bowen, R.S., Chang, H., Herrmann, C., Teterwak, P., Liu, C., Zabih, R.: Oconet: Image extrapolation by object completion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2307–2317 (2021)
5. Chan, E.R., Lin, C.Z., Chan, M.A., Nagano, K., Pan, B., Mello, S.D., Gallo, O., Guibas, L., Tremblay, J., Khamis, S., Karras, T., Wetzstein, G.: Efficient geometry-aware 3D generative adversarial networks. In: arXiv (2021)
6. Chan, E.R., Monteiro, M., Kellnhofer, P., Wu, J., Wetzstein, G.: pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5799–5809 (2021)
7. Chaurasia, G., Duchene, S., Sorkine-Hornung, O., Drettakis, G.: Depth synthesis and local warps for plausible image-based navigation. ACM Transactions on Graphics (TOG) **32**(3), 1–12 (2013)
8. Chen, L.C., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587 (2017)
9. Chen, X., Song, J., Hilliges, O.: Monocular neural image based rendering with continuous view control. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4090–4100 (2019)
10. Cheng, Y.C., Lin, C.H., Lee, H.Y., Ren, J., Tulyakov, S., Yang, M.H.: In&out: Diverse image outpainting via gan inversion. arXiv preprint arXiv:2104.00675 (2021)
11. Choi, I., Gallo, O., Troccoli, A., Kim, M.H., Kautz, J.: Extreme view synthesis. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 7781–7790 (2019)
12. Chong, M.J., Lee, H.Y., Forsyth, D.: Stylegan of all trades: Image manipulation with only pretrained stylegan. arXiv preprint arXiv:2111.01619 (2021)
13. Clark, A., Donahue, J., Simonyan, K.: Adversarial video generation on complex datasets. arXiv preprint arXiv:1907.06571 (2019)
14. Clark, A., Donahue, J., Simonyan, K.: Efficient video generation on complex datasets. ArXiv **abs/1907.06571** (2019)
15. Denton, E., Fergus, R.: Stochastic video generation with a learned prior. In: International conference on machine learning. pp. 1174–1183. PMLR (2018)
16. DeVries, T., Bautista, M.A., Srivastava, N., Taylor, G.W., Susskind, J.M.: Unconstrained scene generation with locally conditioned radiance fields. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14304–14313 (2021)
17. Esser, P., Rombach, R., Ommer, B.: Taming transformers for high-resolution image synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12873–12883 (2021)
18. Finn, C., Goodfellow, I., Levine, S.: Unsupervised learning for physical interaction through video prediction. Advances in neural information processing systems **29** (2016)

19. Flynn, J., Broxton, M., Debevec, P., DuVall, M., Fyffe, G., Overbeck, R., Snavely, N., Tucker, R.: Deepview: View synthesis with learned gradient descent. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2367–2376 (2019)
20. Fox, G., Tewari, A., Elgharib, M., Theobalt, C.: Stylevideogan: A temporal generative model using a pretrained stylegan. arXiv preprint arXiv:2107.07224 (2021)
21. Glasner, D., Bagon, S., Irani, M.: Super-resolution from a single image. In: 2009 IEEE 12th international conference on computer vision. pp. 349–356. IEEE (2009)
22. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. Advances in neural information processing systems **27** (2014)
23. Gu, J., Liu, L., Wang, P., Theobalt, C.: StyleNeRF: A style-based 3d-aware generator for high-resolution image synthesis. arXiv preprint arXiv:2110.08985 (2021)
24. Hao, Z., Mallya, A., Belongie, S., Liu, M.Y.: Gancraft: Unsupervised 3d neural rendering of minecraft worlds. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14072–14082 (2021)
25. Hays, J., Efros, A.A.: Scene completion using millions of photographs. ACM Transactions on Graphics (ToG) **26**(3), 4–es (2007)
26. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems **33**, 6840–6851 (2020)
27. Hsieh, J.T., Liu, B., Huang, D.A., Fei-Fei, L.F., Niebles, J.C.: Learning to decompose and disentangle representations for video prediction. Advances in neural information processing systems **31** (2018)
28. Hu, R., Ravi, N., Berg, A.C., Pathak, D.: Worldsheet: Wrapping the world in a 3d sheet for view synthesis from a single image. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2021)
29. Huang, X., Mallya, A., Wang, T.C., Liu, M.Y.: Multimodal conditional image synthesis with product-of-experts gans. arXiv preprint arXiv:2112.05130 (2021)
30. Jampani, V., Chang, H., Sargent, K., Kar, A., Tucker, R., Krainin, M., Kaeser, D., Freeman, W.T., Salesin, D., Curless, B., et al.: Slide: Single image 3d photography with soft layering and depth-aware inpainting. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12518–12527 (2021)
31. Jang, W., Agapito, L.: Codenerf: Disentangled neural radiance fields for object categories. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12949–12958 (2021)
32. Kaneva, B., Sivic, J., Torralba, A., Avidan, S., Freeman, W.T.: Infinite images: Creating and exploring a large photorealistic virtual space. In: Proceedings of the IEEE (2010)
33. Karnewar, A., Wang, O.: Msg-gan: Multi-scale gradients for generative adversarial networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7799–7808 (2020)
34. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of gans for improved quality, stability, and variation. In: International Conference on Learning Representations (2018)
35. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of stylegan. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8110–8119 (2020)
36. Koh, J.Y., Lee, H., Yang, Y., Baldridge, J., Anderson, P.: Pathdreamer: A world model for indoor navigation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14738–14748 (2021)

37. Kopf, J., Matzen, K., Alsisan, S., Quigley, O., Ge, F., Chong, Y., Patterson, J., Frahm, J.M., Wu, S., Yu, M., Zhang, P., He, Z., Vajda, P., Saraf, A., Cohen, M.: One shot 3d photography. ACM Transactions on Graphics (Proceedings of ACM SIGGRAPH) **39**(4) (2020)
38. Kopf, J., Matzen, K., Alsisan, S., Quigley, O., Ge, F., Chong, Y., Patterson, J., Frahm, J.M., Wu, S., Yu, M., et al.: One shot 3d photography. ACM Transactions on Graphics (TOG) **39**(4), 76–1 (2020)
39. Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., et al.: Photo-realistic single image super-resolution using a generative adversarial network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4681–4690 (2017)
40. Lee, W., Jung, W., Zhang, H., Chen, T., Koh, J.Y., Huang, T., Yoon, H., Lee, H., Hong, S.: Revisiting hierarchical approach for persistent long-term video prediction. arXiv preprint arXiv:2104.06697 (2021)
41. Levoy, M., Hanrahan, P.: Light field rendering. In: Proceedings of the 23rd annual conference on Computer graphics and interactive techniques. pp. 31–42 (1996)
42. Lin, C.H., Cheng, Y.C., Lee, H.Y., Tulyakov, S., Yang, M.H.: InfinityGAN: Towards infinite-pixel image synthesis. In: International Conference on Learning Representations (2022)
43. Liu, A., Tucker, R., Jampani, V., Makadia, A., Snavely, N., Kanazawa, A.: Infinite nature: Perpetual view generation of natural scenes from a single image. In: Proc. Int. Conf. on Computer Vision (ICCV). pp. 14458–14467 (2021)
44. Liu, H., Wan, Z., Huang, W., Song, Y., Han, X., Liao, J.: Pd-gan: Probabilistic diverse gan for image inpainting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9371–9381 (2021)
45. Liu, L., Gu, J., Lin, K.Z., Chua, T.S., Theobalt, C.: Neural sparse voxel fields. NeurIPS (2020)
46. Liu, Y., Shu, Z., Li, Y., Lin, Z., Perazzi, F., Kung, S.Y.: Content-aware gan compression. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12156–12166 (2021)
47. Lombardi, S., Simon, T., Saragih, J., Schwartz, G., Lehrmann, A., Sheikh, Y.: Neural volumes: Learning dynamic renderable volumes from images. ACM Trans. Graph. **38**(4), 65:1–65:14 (Jul 2019)
48. Mallya, A., Wang, T.C., Sapra, K., Liu, M.Y.: World-consistent video-to-video synthesis. In: European Conference on Computer Vision. pp. 359–378. Springer (2020)
49. Mildenhall, B., Srinivasan, P.P., Ortiz-Cayon, R., Kalantari, N.K., Ramamoorthi, R., Ng, R., Kar, A.: Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. ACM Transactions on Graphics (TOG) **38**(4), 1–14 (2019)
50. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. In: European conference on computer vision. pp. 405–421. Springer (2020)
51. Munoz, A., Zolfaghari, M., Argus, M., Brox, T.: Temporal shift gan for large scale video generation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 3179–3188 (2021)
52. Niemeyer, M., Geiger, A.: Campari: Camera-aware decomposed generative neural radiance fields. In: 2021 International Conference on 3D Vision (3DV). pp. 951–961. IEEE (2021)
53. Niemeyer, M., Geiger, A.: Giraffe: Representing scenes as compositional generative neural feature fields. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11453–11464 (2021)

54. Niklaus, S., Liu, F.: Softmax splatting for video frame interpolation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5437–5446 (2020)
55. Niklaus, S., Mai, L., Yang, J., Liu, F.: 3d ken burns effect from a single image. ACM Transactions on Graphics (ToG) **38**(6), 1–15 (2019)
56. Park, T., Zhu, J.Y., Wang, O., Lu, J., Shechtman, E., Efros, A., Zhang, R.: Swapping autoencoder for deep image manipulation. Advances in Neural Information Processing Systems **33**, 7198–7211 (2020)
57. Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., Koltun, V.: Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. IEEE transactions on pattern analysis and machine intelligence (2020)
58. Razavi, A., Van den Oord, A., Vinyals, O.: Generating diverse high-fidelity images with vq-vae-2. Advances in neural information processing systems **32** (2019)
59. Riegler, G., Koltun, V.: Free view synthesis. In: European Conference on Computer Vision (2020)
60. Rockwell, C., Fouhey, D.F., Johnson, J.: Pixelsynth: Generating a 3d-consistent experience from a single image. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14104–14113 (2021)
61. Rombach, R., Esser, P., Ommer, B.: Geometry-free view synthesis: Transformers and no 3d priors. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14356–14366 (2021)
62. Saharia, C., Chan, W., Chang, H., Lee, C.A., Ho, J., Salimans, T., Fleet, D.J., Norouzi, M.: Palette: Image-to-image diffusion models. arXiv preprint arXiv:2111.05826 (2021)
63. Schwarz, K., Liao, Y., Niemeyer, M., Geiger, A.: Graf: Generative radiance fields for 3d-aware image synthesis. Advances in Neural Information Processing Systems **33**, 20154–20166 (2020)
64. Sengupta, A., Ye, Y., Wang, R., Liu, C., Roy, K.: Going deeper in spiking neural networks: Vgg and residual architectures. Frontiers in neuroscience **13**, 95 (2019)
65. Shade, J., Gortler, S., He, L.w., Szeliski, R.: Layered depth images. In: Proceedings of the 25th annual conference on Computer graphics and interactive techniques. pp. 231–242 (1998)
66. Shaham, T.R., Dekel, T., Michaeli, T.: Singan: Learning a generative model from a single natural image. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4570–4580 (2019)
67. Shaham, T.R., Dekel, T., Michaeli, T.: Singan: Learning a generative model from a single natural image. 2019 IEEE/CVF International Conference on Computer Vision (ICCV) pp. 4569–4579 (2019)
68. Shi, L., Hassanieh, H., Davis, A., Katabi, D., Durand, F.: Light field reconstruction using sparsity in the continuous fourier domain. ACM Transactions on Graphics (TOG) **34**(1), 1–13 (2014)
69. Shih, M.L., Su, S.Y., Kopf, J., Huang, J.B.: 3d photography using context-aware layered depth inpainting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8028–8038 (2020)
70. Shih, M.L., Su, S.Y., Kopf, J., Huang, J.B.: 3d photography using context-aware layered depth inpainting. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
71. Shocher, A., Bagon, S., Isola, P., Irani, M.: Ingan: Capturing and remapping the" dna" of a natural image. arXiv preprint arXiv:1812.00231 (2018)

72. Shocher, A., Cohen, N., Irani, M.: "zero-shot" super-resolution using deep internal learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3118–3126 (2018)
73. Skorokhodov, I., Sotnikov, G., Elhoseiny, M.: Aligning latent and image spaces to connect the unconnectable. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14144–14153 (2021)
74. Teterwak, P., Sarna, A., Krishnan, D., Maschinot, A., Belanger, D., Liu, C., Freeman, W.T.: Boundless: Generative adversarial networks for image extension. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10521–10530 (2019)
75. Tian, Y., Ren, J., Chai, M., Olszewski, K., Peng, X., Metaxas, D.N., Tulyakov, S.: A good image generator is what you need for high-resolution video synthesis. arXiv preprint arXiv:2104.15069 (2021)
76. Tucker, R., Snavely, N.: Single-view view synthesis with multiplane images. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020)
77. Tulsiani, S., Tucker, R., Snavely, N.: Layer-structured 3d scene inference via view synthesis. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 302–317 (2018)
78. Tulyakov, S., Liu, M.Y., Yang, X., Kautz, J.: Mocogan: Decomposing motion and content for video generation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1526–1535 (2018)
79. Villegas, R., Pathak, A., Kannan, H., Erhan, D., Le, Q.V., Lee, H.: High fidelity video prediction with large stochastic recurrent neural networks. Advances in Neural Information Processing Systems **32** (2019)
80. Villegas, R., Yang, J., Hong, S., Lin, X., Lee, H.: Decomposing motion and content for natural video sequence prediction. arXiv preprint arXiv:1706.08033 (2017)
81. Vondrick, C., Pirsiavash, H., Torralba, A.: Generating videos with scene dynamics. Advances in neural information processing systems **29** (2016)
82. Vondrick, C., Torralba, A.: Generating the future with adversarial transformers. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1020–1028 (2017)
83. Wang, Q., Wang, Z., Genova, K., Srinivasan, P.P., Zhou, H., Barron, J.T., Martin-Brualla, R., Snavely, N., Funkhouser, T.: Ibrnet: Learning multi-view image-based rendering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4690–4699 (2021)
84. Wang, Y., Tao, X., Shen, X., Jia, J.: Wide-context semantic image extrapolation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1399–1408 (2019)
85. Wang, Y., Long, M., Wang, J., Gao, Z., Yu, P.S.: PredRNN: Recurrent neural networks for predictive learning using spatiotemporal LSTMs. In: Advances in Neural Information Processing Systems. pp. 879–888 (2017)
86. Wiles, O., Gkioxari, G., Szeliski, R., Johnson, J.: Synsin: End-to-end view synthesis from a single image. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7467–7477 (2020)
87. Yang, Z., Dong, J., Liu, P., Yang, Y., Yan, S.: Very long natural scenery image prediction by outpainting. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10561–10570 (2019)
88. Ye, Y., Singh, M., Gupta, A., Tulsiani, S.: Compositional video prediction. In: International Conference on Computer Vision (ICCV) (2019)

89. Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., Huang, T.S.: Generative image inpainting with contextual attention. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5505–5514 (2018)
90. Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., Huang, T.S.: Free-form image inpainting with gated convolution. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4471–4480 (2019)
91. Yu, S., Tack, J., Mo, S., Kim, H., Kim, J., Ha, J.W., Shin, J.: Generating videos with dynamics-aware implicit generative adversarial networks. In: The Tenth International Conference on Learning Representations (2022)
92. Yu, S., Tack, J., Mo, S., Kim, H., Kim, J., Ha, J.W., Shin, J.: Generating videos with dynamics-aware implicit generative adversarial networks. In: The Tenth International Conference on Learning Representations (2022)
93. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR (2018)
94. Zhao, S., Cui, J., Sheng, Y., Dong, Y., Liang, X., Chang, E.I., Xu, Y.: Large scale image completion via co-modulated generative adversarial networks. In: International Conference on Learning Representations (2021)
95. Zhou, T., Tucker, R., Flynn, J., Fyffe, G., Snavely, N.: Stereo magnification: Learning view synthesis using multiplane images. In: SIGGRAPH (2018)
96. Zhou, Y., Zhu, Z., Bai, X., Lischinski, D., Cohen-Or, D., Huang, H.: Non-stationary texture synthesis by adversarial expansion. arXiv preprint arXiv:1805.04487 (2018)