Ray Conditioning: Trading Photo-consistency for Photo-realism in Multi-view Image Generation

Eric Ming Chen[1] Sidhanth Holalkere[1] Ruyu Yan[1] Kai Zhang[2] Abe Davis[1]
[1]Cornell University [2]Adobe Research

https://ray-cond.github.io

Abstract

Multi-view image generation attracts particular attention these days due to its promising 3D-related applications, e.g., image viewpoint editing. Most existing methods follow a paradigm where a 3D representation is first synthesized, and then rendered into 2D images to ensure photo-consistency across viewpoints. However, such explicit bias for photo-consistency sacrifices photo-realism, causing geometry artifacts and loss of fine-scale details when these methods are applied to edit real images. To address this issue, we propose ray conditioning, a geometry-free alternative that relaxes the photo-consistency constraint. Our method generates multi-view images by conditioning a 2D GAN on a light field prior. With explicit viewpoint control, state-of-the-art photo-realism and identity consistency, our method is particularly suited for the viewpoint editing task.

1. Introduction

Modeling the distributions of natural images has long been an important problem that is extensively studied. Generative adversarial networks (GANs) and diffusion models are two types of generative models that have successfully shown impressive capabilities of learning image distributions—the generated samples are almost indistinguishable from real photos [11, 33, 16, 34].

While optimizing for the photo-realism of individual samples, these generative models rarely allow for multi-view image generation, where photo-consistency matters. Recently, many multi-view image synthesizers have been proposed that try to optimize for both photo-realism and photo-consistency [6, 41, 13, 26]. They generally follow a "synthesize-3D-then-render" paradigm: 3D representations are synthesized, and then images are rendered (at specified camera poses). Such 3D-aware generative models, especially EG3D [6], achieve high-quality multi-view image generation results, despite being trained only on single-view image data with poses.

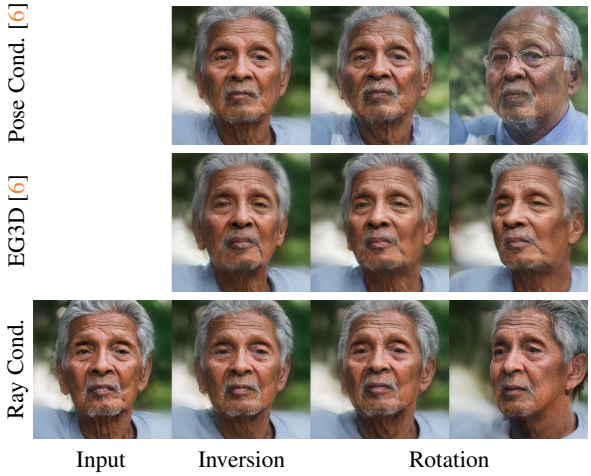

Figure 1. **Challenges With Viewpoint Editing.** Conditioning a 2D GAN's latent space on pose does not ensure that an identity remains consistent across views. 3D-aware GANs such as EG3D [6] struggle to reconstruct high-frequency details such as wrinkles and hair. Our ray conditioning method most faithfully reproduces the input image, and preserves identity when editing the viewpoint.

However, photo-realism and photo-consistency are oftentimes two competing goals: photo-realism very explicitly favors image quality over control, while photo-consistency more implicitly favors control over quality. A few examples of this conflict include fine-scale detail and view-dependent appearance in the multi-view 3D reconstruction and 3D generation problems. Details are either filtered from 3D representations (leaving them smoother and more diffuse than real ones) or misinterpreted as geometric artifacts [39]. As shown in Figure 1, although EG3D allows explicit camera control and generates photo-consistent images at different viewpoints, it fails to reproduce the subtle details, e.g., the wrinkles and hairs, in the input image. Conditioning a 2D GAN's latent space on camera pose does not ensure that the identity remains consistent across views.

Our work is motivated by the observation that, for cer-

Figure 2. **Viewpoint Editing with Ray Conditioning.** Ray conditioning enables photo-realistic multi-view image editing on natural photos via GAN inversion. The left half shows headshots of four individuals and their corresponding synthesized results from another viewpoint. The right half shows a portrait of two individuals (top row), the GAN inversion results of their faces (top row corners), and the resulting image (bottom row), in which their faces are replaced with synthesized faces looking in a different direction (bottom row corners). To produce the latter, we used Photoshop to blend the synthesized faces with the original image.

tain classes of images with shared canonical structure, e.g. faces, it is possible to achieve viewpoint control without optimizing explicitly for 3D structure. The result is a modified 2D GAN that offers precise control over generated viewpoints without sacrificing photo-realism. Furthermore, we are able to train on data that does not contain multiple viewpoints of any single subject, letting us leverage the same diverse and abundant data used for regular GANs. Our method combines the photo-realism of existing GANs with the control offered by geometric models, outperforming related methods in both generation and inversion quality. This makes our method particularly well-suited for viewpoint editing in static images.

Key to our method is the proposed ***ray conditioning*** mechanism that enables explicit viewpoint control. The method is simple. Rather than using a 3D model, our method conditions each pixel in a generated image on the ray through it—a technique inspired by the light field [24, 12]. The spatial priors of ray conditioning enables the image synthesizer to learn multi-view consistency from only single-view image collections and their estimated poses. By choosing a geometry-free approach, we relax the 3D photo-consistency constraints in exchange for increased photo-realism. Despite this, our approach still offers competitive identity preservation capability when editing viewpoints. Figure 2 represents the quality and control we can achieve with ray conditioning. Evaluation on both single-view and multi-view data shows that our method is a significant improvement in image quality over Light Field Networks (LFNs), another geometry-free image synthesizer [31], demonstrating the promising potential of this line of research.

In summary, our contributions are as follows.

1. We propose a simple yet effective geometry-free generative model named ray conditioning for multi-view image generation, and show that it achieves greater photo-realism than geometry-based baselines while maintaining explicit control of viewpoints.
2. We demonstrate the advantages of our method in the downstream application of editing real images' viewpoints where photo-realism is favored over photo-consistency.
3. Our ray conditioning method is also the first geometry-free *multi-view* image synthesizer that can generate highly realistic images in high resolution (1024×1024) given only single-view posed image collections.

2. Related Work

Image Synthesis and Latent Space Editing. In the past decade, GANs [11] have revolutionized image synthesis. They are trained by optimizing two neural networks, a generator and a discriminator, at the same time. The generator tries to fool the discriminator by generating fake images that look as real as possible. The discriminator tries to distinguish the synthetic images from the real ones. Once training is complete, the generator is able to generate images that are almost indistinguishable from natural images. Among all the proposed GAN architectures, for its high image quality, StyleGAN [21, 22, 20] is perhaps the most

widely adopted. We base our method on StyleGAN2, and enable explicit viewpoint control via ray conditioning.

Several works have found that the latent spaces of StyleGAN are remarkably linear [18, 30, 17, 37], disentangling attributes such as facial expression, hair color, and pose. Pose disentanglement is of particular interest to the domain of this work, because it serves as a proxy for 3D information. Related work have utilized this property of 2D GANs to discover visual correspondences between images [27]. However, properly harnessing this ability can be non-trivial. In a latent space, editing directions often have to be found ad-hoc for each dataset, and lack a intuitive interpretation [30]. In contrast, our method allows for interpretable and explicit control of viewpoints.

Multi-view Image Synthesis. Our work is closely related to the direction of multi-view image synthesis, where prior work can be roughly classified into two categories: *geometry-based* and *geometry-free*.

Geometry-based multi-view image generation approaches typically adapt 2D GAN-based image synthesizers in a way that a neural 3D representation is first generated, from which 2D images are then rendered using neural rendering. Representative work in this category include EG3D [6], GMPI [41], StyleSDF [26], StyleNeRF [13], GIRAFFE [25], π-GAN [5], etc. These methods mainly differ from each other in choice of the 3D representation and rendering algorithm. For instance, EG3D adapts the StyleGAN2 generator to predict a feature volume in a compact triplane representation. They then use volume rendering to render a feature map, which is later decoded into a high-resolution image through a convolutional decoder; on the other hand, GMPI generates a multiplane image RGBA representation [42] and renders it using homography warping and alpha compositing. These *geometry-based* methods have demonstrated impressive multi-view image synthesis quality given only single-view posed data; the generated images are very view-consistent and detailed. However, we observe that the synthesize-3D-then-render approach they adopt indeed trades some photo-realism for photo-consistency, as shown in Figure 1. Moreover, as noted by concurrent work [38], the geometry prior in an image synthesizer also increases the difficulty of inverting a real posed image. Both issues sacrifice the methods' performance in viewpoint editing for real posed images. We seek to circumvent these issues by optimizing for photo-realism and easy invertibility.

Geometry-free methods traditionally learn view consistency priors by training image synthesizers on large *multi-view* datasets rather than using a 3D representation. LFNs [31] and 3DiM [35] are two successful methods that have inspired our approach. LFNs represents each scene as a light field parametrized by a multilayer perceptron (MLP) which maps a ray to a color. For generation, the MLP is also conditioned on a randomly-sampled latent code through meta-learning [14]. However, compared to a convolutional neural network (CNN), a MLP cannot effectively utilize the inductive bias of spatial smoothness in natural images. The generated results are oftentimes blurry compared to CNN-based image synthesizers. 3DiM uses a more powerful image generator—a diffusion model [33, 16, 34], and achieve state-of-the-art single-view novel view synthesis results on the ShapeNet dataset [7]. At their core is a pose-conditional image-to-image diffusion model trained using ground-truth multi-view images as supervision. Being a conditional image synthesizer, 3DiM cannot perform unconditional multi-view image generation. While both works have made strong methodological contributions, neither has been proven to generate results at a resolution higher than 128×128, to learn without multi-view datasets, nor to learn over photo-realistic images. In comparison, we show that our method (ray conditioning and a CNN synthesizer) can be trained with only *single-view* posed data at 1024×1024 resolution, and can perform even better than *geometry-based* methods on practical viewpoint-editing applications.

3. Method

Like prior work on 3D-aware GANs [6], we focus on unstructured single-view image collections of objects from the same category, e.g., human faces, with labeled camera poses. Namely, we focus on a set of $(\mathbf{I}, \mathbf{K}, \mathbf{E})$ triples, where $\mathbf{I}$ is an image, $\mathbf{K}$ are its corresponding intrinsics, and $\mathbf{E} = [\mathbf{R} \mid \mathbf{t}]$ are its corresponding camera extrinsics (camera-to-world transformations). For the sake of downstream applications such as portrait reposing, we seek to train a GAN that allows us to explicitly control the viewpoints of the synthesized images—without explicitly modeling the geometry. We accomplish this by adding our proposed ray-conditioning mechanism to an off-the-shelf image generator: StyleGAN2 [22]. Our method requires minimal modifications to the image generator's architecture, while achieving higher photo-realism than methods that use a 3D representation [6, 41, 26, 13].

3.1. Photo Collections as Unstructured Light Fields

Images are often regarded as a sample of a scene's 5D plenoptic function, $L : (\mathbf{p}, \mathbf{d}) \mapsto \mathbf{c}$, which describes the light intensity $\mathbf{c}$ in an arbitrary direction $\mathbf{d} \in \mathbb{S}^2$, and at arbitrary location $\mathbf{p} \in \mathbb{R}^3$ [2]. As described in the classic light field works [24, 12], if we choose to sample images outside a convex hull surrounding the object, the 5D plenoptic function becomes 4D. In this case, the 4D plenoptic function is also called the 4D light field. A pixel (u, v) of an image $\mathbf{I}$ can be interpreted as a sample of the light through a ray

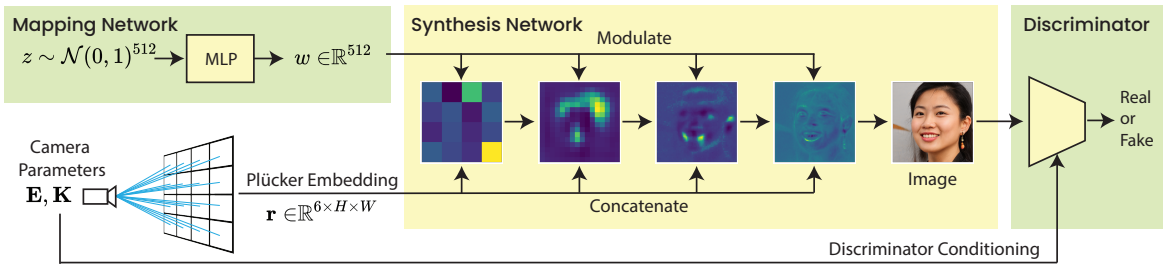

Figure 3. **The Ray Conditioning Method.** The StyleGAN synthesis network progressively convolves and upscales a low-resolution feature map into a high-resolution one. To condition the generator on a camera, we concatenate these feature maps with an appropriately downsampled Plücker embedding of the sampled camera parameters. By doing so, the GAN learns to associate camera rays with appearance.

direction:

$$\mathbf{d}_{u,v} = \mathbf{R}\mathbf{K}^{-1}\left[u, v, 1\right]^T, \tag{1}$$

$$\mathbf{d}_{u,v} = \mathbf{d}_{u,v} / \|\mathbf{d}_{u,v}\|_2, \tag{2}$$

$$u, v \in [0, W) \times [0, H), \tag{3}$$

where W, H are $\mathbf{I}$'s width and height. A perspective image is then a measurement of the 4D light field at the location of the camera origin, i.e., $\mathbf{p} = \mathbf{o}$, from a bundle of ray directions $\mathbf{d}_{u,v}$ falling inside the camera's field-of-view.

Typically, the light field of a single scene is measured using a dense grid of synchronized cameras [36]. Novel views can then be synthesized by interpolating these densely captured images. However, for a photo collection with only one image per scene, such as one of faces, the light field measurements are highly unstructured. Due to the varying identities, expressions etc, each image can be thought of as being a single-shot sampling of its scene's light field.

The task of generative modeling is thus to model a distribution of light field observations over an entire photo collection to picture what missing view points may have looked like.

3.2. Ray Conditioning for Image Synthesis

Given a posed image collection $\{(\mathbf{I}, \mathbf{K}, \mathbf{E})\}$ that measures light fields in a highly unstructured way, we aim to learn the distribution of a light field L defined on ray bundles $\mathbf{r} \in \mathbb{R}^{6\times H\times W}$. $\mathbf{r}$ is a 2D feature map which assigns each pixel (u, v) in $\mathbf{I}$ to the camera ray through that pixel, $\mathbf{r}_{u,v}$, as illustrated in Figure 3. We only consider ray bundles that follow the same distribution as all the input ray bundles; hence L is not well-defined for out-of-distribution ray bundles, as shown in Figure 8.

We add ray conditioning to a GAN G to model the distribution of a light field L implicitly: the generator maps a Gaussian-distributed noise code $\mathbf{z}$ and ray bundle $\mathbf{r}$ to an image sample: $G(\mathbf{z}, \mathbf{r}) = \mathbf{I}$. By fixing the noise code $\mathbf{z}$, and using different ray bundles $\mathbf{r}$, we can sample images at different viewpoints from the same learnt light field. Concretely, we use the StyleGAN2 backbone, as does EG3D [6]. As shown in Figure 3, we add ray conditioning to each level of the progressively-growing synthesis network. At each level, we first compute the spatial ray embedding to the same resolution as the feature map, and then concatenate the ray embedding and feature map together along the channel dimension.

The ray embedding needs to be carefully chosen: the standard 5D ray parametrization $\mathbf{r}_{u,v} = (\mathbf{o}, \mathbf{d}_{u,v})$ is redundant for a 4D light field as it fails to consider the assumption of zero decay in empty space: $L(\mathbf{o}, \mathbf{d}) = L(\mathbf{o} + t\mathbf{d}, \mathbf{d})$. Inspired by LFNs [31], we remove this redundancy through the Plücker parametrization $\mathbf{r}_{u,v} = (\mathbf{o} \times \mathbf{d}_{u,v}, \mathbf{d}_{u,v})$, where $\times$ is the cross product. With this parametrization, we have:

$$(\mathbf{o} + t\mathbf{d}) \times \mathbf{d} = \mathbf{o} \times \mathbf{d} + t\mathbf{d} \times \mathbf{d} = \mathbf{o} \times \mathbf{d}. \tag{4}$$

We require minimal modifications to the backbone StyleGAN2 architecture: each convolution kernel just needs to accept extra ray embedding inputs. Hence the induced computational overhead is almost negligible. Moreover, we can start from a pretrained StyleGAN2 model, and finetune it to make it amenable to explicit viewpoint control. Unlike prior works that prioritize photo-consistency over photo-realism, our method maintains the high image generation fidelity of StyleGAN2, as shown in Figure 5.

3.3. Viewpoint Editing for Real Posed Images

For generated images, new viewpoints can be achieved by changing the ray bundles $\mathbf{r}$ in our generated light field $G(\mathbf{z}, \mathbf{r})$. Like prior work in StyleGAN-based real image editing, we can invert a real posed image $(\mathbf{I}, \mathbf{K}, \mathbf{E})$ into a latent space of StyleGAN first, and then modify the ray bundles to edit the viewpoint.

As our method closely resembles the backbone StyleGAN2, we can directly use off-the-shelf GAN inversion methods [28] for high-quality inversions and viewpoint edits. To invert the image's camera parameters, we use Deep 3D Face Reconstruction [10]. This is in stark contrast to geometry-based methods like EG3D that require more dedicated inversion methods, as shown by concurrent work [38] and Figure 4.

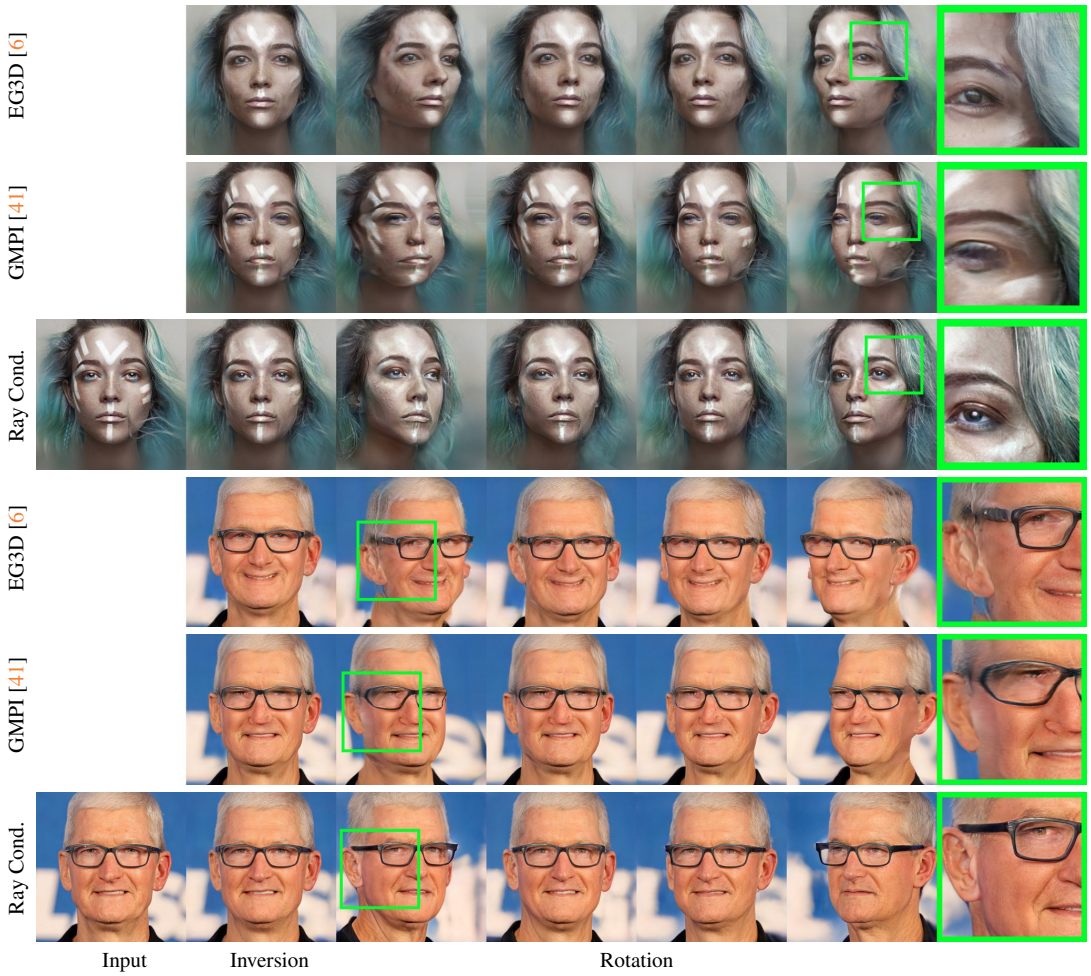

Figure 4. **Viewpoint Editing via GAN Inversion.** We invert an image into each GAN using PTI [28]. Ray conditioning is able to best preserve the details of the input images, as shown in the rich detail of the skin, eyes, and hair. Because radiance fields are biased towards low frequency results, the EG3D [6] inversions lack the detail that ray conditioning can offer. In the second example, there are also geometry artifacts near the ears and border of the face. GMPI [41] struggles to invert input images, causing distortion in the novel views.

We also differ from prior latent-space viewpoint editing work in terms of offering intuitive explicit viewpoint control and increased viewpoint change range. We provide results in the supplementary material. In addition, those latent-space editing directions require paired training data [30, 1]. One such method, InterfaceGAN [30], relies on an external binary classifier to determine whether a generated face is facing left or right. InterfaceGAN also does not allow for explicit control of pose, and relies on manual tuning to achieve the desired pose.

4. Experiments

We validate our approach by testing its ability to generate multi-view images on two single-view posed datasets: Flickr-Faces-HQ (FFHQ) [21] and AFHQv2 Cat Faces [8]. We show that our geometry-free approach outperforms geometry-based baselines in terms of photo-realism when generating multi-view images. We also demonstrate our model's strength for downstream viewpoint editing of real images. Finally, we compare our method with a prior geometry-free method, LFNs [31], on a multi-view posed dataset: SRN Cars [32, 7], and show significant improvement in multi-view generation quality.

4.1. Metrics

We adopt the same metrics used in EG3D [6] for quantitative evaluation.

Image Quality. To compare image generation quality, we report the FID score [15] and KID score$\times$100 [3] between 50k generated images and the entire training set.

Identity Consistency. As a proxy for view consistency, we use ArcFace [9, 29], a facial recognition model, to compute identity consistency between two random views of one individual, and average it over 1024 samples.

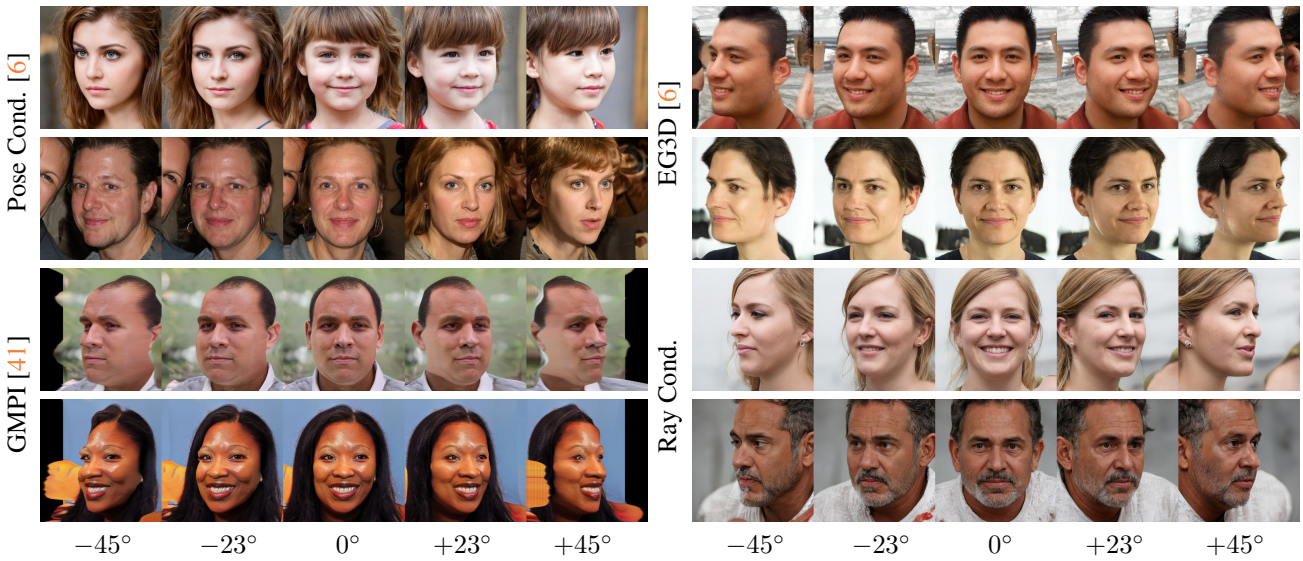

Figure 5. **Unconditional Multi-view Image Generation.** As reflected in the photos, our method is able to maintain identity consistency and photorealism across angles. Pose conditioning fails to be view consistent. By foregoing a 3D model, we do not have the geometric artifacts that GMPI [41] and EG3D [6] may have. Although there are some view-dependent changes between images, we achieve the highest image quality at steep angles. All results are generated with $\psi_{\text{trunc}} = 0.7$.

Pose Accuracy. We sample one camera pose for 1024 individuals and estimate the camera pose of the synthesized image with Deep 3D Face Reconstruction [10]. We report the mean squared error between camera angles.

Generation Speed. We benchmark the time for generating an image from a latent vector and report the generation speed in frames per second (FPS). Generation speed is critical for GAN inversion, as inversion methods may require hundreds of queries to the GAN to invert a real image.

4.2. Results and Discussion

We report quantitative results comparing ray conditioning to competitive 3D-aware baselines in Table 1, and show samples of generated images in Figure 5. All baseline results except for FPS are quoted from StyleGAN2-ADA [19], GMPI [41] and EG3D [6]. We also report standard deviations for the ID metric and pose metric, which were previously not reported. Notably, we are able to achieve strong results with only a 2D GAN backbone. Because our image quality is not limited by the resolution of a geometric model, ray conditioning is able to achieve the highest FID and KID scores on faces, which are most similar with the scores of StyleGAN2. We are also able to achieve competitive identity accuracies and pose accuracies, demonstrating our model's ability to maintain view consistency and model camera pose.

These metrics corroborate what we see in Figure 5. Compared to EG3D and GMPI, ray conditioning maintains the highest visual quality across yaw changes. GMPI appears to have lower quality images than EG3D and ray conditioning. As reported in the GMPI ablation study, we believe that this is due to GMPI's synthetic shading, which is necessary to learn accurate depth information. At yaws of $\pm 23°$, all methods do a good job at representing rotation. However, beyond that, EG3D and GMPI appear to show artifacts. For EG3D, there is noise near the ears in both examples. For the second individual, there is an unnatural border near the ears as well. This is a known failure mode of EG3D and other radiance fields called billboarding, described in Section 1.3 of their appendix. For GMPI, the multiplane images cause noticeable quality issues in the rendered images. Images appear to be blurry. While ray conditioning sacrifices some view consistency, such as changes in smile, the results remain realistic even at difficult yaws. These results provide an explanation for why ray conditioning is able to achieve the best FID and KID scores. Forfeiting a 3D representation allows for high quality image synthesis across angles. The fact that all of the most competitive methods are built upon the StyleGAN architecture underscores its superior ability to disentangle pose and appearance. Ray conditioning is a natural extension of StyleGAN for generating multi-view images.

In Figure 6, we compare ray conditioning to Light Field Networks [31] (LFNs), a model designed for geometry-free view synthesis. LFNs use an autodecoder to condition the color of a ray on a normally distributed latent vector. Because LFNs are trained with a L2 reconstruction loss instead of an adversarial loss, output images tend to be blurry. Relative to the training dataset, LFNs achieves an FID score of 41.8 while ray conditioning achieves an FID score of 3.39. The samples from ray conditioning are much sharper, and show more diversity. We provide videos in the supple-

	FID↓	KID↓	FFHQ ID ↑	Pose↓	FPS↑	AFHQv2 Cats FID↓	AFHQv2 Cats KID ↓
Image synth.							
StyleGAN2 [19] 512^2	-	-	-	-	69	$3.55^\dagger$	$0.066^\dagger$
StyleGAN2 [22] 1024^2	2.70	0.048	-	-	62	-	-
Geometry-based MV synth.							
StyleSDF [26] 256^2	11.5	0.370	-	-	-	12.8^*	0.447^*
StyleNeRF [13] 1024^2	8.10	0.240	-	-	-	14.0^*	0.350^*
EG3D [6] 512^2	4.70	0.132	**0.77**±0.15	**0.005**±0.005	33	$\mathbf{2.77}^\dagger$	$\mathbf{0.041}^\dagger$
GMPI [41] 512^2	8.29	0.454	**0.74**±0.16	**0.006**±0.009	13	7.79	0.474
GMPI [41] 1024^2	7.50	0.407	**0.75**±0.16	**0.007**±0.010	6	-	-
Geometry-free MV synth.							
Ray Conditioning 512^2	3.50	0.076	**0.75**±0.15	**0.006**±0.007	**48**	3.44	0.103
Ray Conditioning 1024^2	**3.28**	**0.066**	**0.76**±0.14	**0.006**±0.007	**38**	-	-

Table 1. **Multi-view Image Generation Metrics.** Ray conditioning enables multi-view (MV) image synthesis by conditioning a 2D GAN on a ray embedding of a camera. It achieves high degrees of photorealism, identity consistency, and pose accuracy. We compare each multi-view GAN method to a StyleGAN2 baseline, showing the loss of fildelity due to geometric inductive biases. All metrics except for FPS are quoted from StyleGAN2-ADA [19], EG3D [6] and GMPI [41]. We also compute and report standard deviations for the ID and pose scores, which were previously not reported. We bold the best statistically significant results. *Trained on all of AFHQ instead of the cats subset. †Trained with adaptive discriminator augmentation [19].

mentary material. As shown in Figure 7, on FFHQ, LFNs struggles to reconstruct the input data. It is also not able to synthesize novel views when trained on FFHQ, a single-view dataset. Ray conditioning demonstrates that light field conditioning concept introduced in LFNs is capable of synthesizing compelling results on only single-view data.

4.3. Ablation Study

We compare ray conditioning to a simpler alternative: pose conditioning. Similar to StyleGAN2-ADA [19] and EG3D [6], we first flatten the camera extrinsics $\mathbf{E} \in \mathbb{R}^{4\times4}$ and intrinsics $\mathbf{K} \in \mathbb{R}^{3\times3}$ into a conditioning vector $\mathbf{c} \in \mathbb{R}^{25}$. We then input both a randomly sampled $\mathbf{z} \sim \mathcal{N}(0,1)^{512}$ and the conditioning vector $\mathbf{c}$ into the mapping network for predicting a $\mathbf{w}$ code, as shown in Figure 3. Pose conditioning also provides explicit viewpoint control; however, it encodes a viewpoint as a 1D vector, rather than 2D feature map as we do for ray conditioning. We show that the lack of spatial inductive bias in pose conditioning causes the identity of generated people to vary wildly with small rotations in Figure 5. At a resolution of 512×512, pose conditioning achieves an average ID similarity score of 0.68 ± 0.20, while ray conditioning achieves 0.75 ± 0.15. Ray conditioning produces much more view-consistent results than pose conditioning.

4.4. Viewpoint Editing for Real Posed Images

We compare our ray conditioning against prior work for the application of editing viewpoints of real posed images, as described in Section 3.3. Since our method and baselines are all variants of StyleGAN, we invert images using Pivotal Tuning Inversion (PTI) [28], and synthesize images of the same individual from different viewpoints. We show in Figure 4 that our ray conditioning method achieves the same explicit viewpoint control as the geometry-based methods while preserving a much higher degree of photo-realism. In the first example, the detail is incredibly noticeable in the eyes and hair, which closely resemble the input image. Eyes are especially important for human perception of identity and familiarity, but are often difficult to invert for geometry-based methods due to their specularity.

Moreover, geometry-based methods such as EG3D and GMPI can introduce geometric artifacts in synthesized images. When fitting a geometry-based representation to a single image, there is often ambiguity on whether to modify the geometry or texture. Incorrect geometry can create seemingly correct images. This is only realized after a shift of viewpoint. For radiance fields, this has been coined as

Figure 6. **Unconditional Generation of Cars**. We observe that ray conditioning, a GAN-based method, has sharper images and more diverse samples than the baseline LFNs [31]. In particular, ray conditioning achieves an FID of 3.39 whereas LFNs achieves an FID of 41.8. (FID is computed from 50k random samples.)

Figure 7. **Face Reconstruction Comparison**. LFNs [31] struggle to reconstruct the training data of FFHQ, making it unsuitable for generating light fields from a *single-view* dataset of natural images. Moreover, it is not able to synthesize novel views. Ray conditioning is able to successfully reconstruct the training images with GAN inversion.

shape radiance ambiguity [39], and still a challenging problem for many 3D representations. In the bottom individual of Figure 4, we see that although EG3D is able to reproduce the input image, the disoccluded parts around the ears exhibit strong geometry artifacts when viewed at a different yaw angle. GMPI is more severely hurt by the ambiguity between geometry and appearance when fitting to the input image, which leads to distortion in the novel views. This is most noticeable in the bottom individual's glasses, which appear to be glued to the face. Additionally, since geometry-based methods tend to smooth textures in favor of photo-consistency, they lack the level of details that our ray conditioning can offer. Rich details in hair, skin, and eyes are inherently view-dependent, and are best captured by relaxing constraints on photo-consistency.

Furthermore, we quantitatively evaluate ray conditioning and EG3D on GAN inversion and viewpoint editing with the CelebA-HQ [23] dataset. Neither method was trained on this dataset, allowing for a measure of cross-dataset generalization. To evaluate the similarity between input images and inversions, we calculate PSNR, SSIM, LPIPS [40], and ID scores. To evaluate the image quality of synthesized novel views, we compare the FID and KID$\times 100$ against the original images. Both metrics were computed from one novel viewpoint for 100 images at a resolution of 512×512. The results in Table 2 show that ray conditioning can achieve higher image quality and detail preservation in both input inversions and after viewpoint editing.

5. Conclusion

We propose ray conditioning, a method for multi-view image generation with explicit viewpoint control.

Our key insight is that we do not need to generate consistent 3D geometry to control the viewpoint of generated images. Instead, 4D ray conditioning lets us generate different viewpoints individually, placing fewer constraints on the generator, which leaves it freer to optimize for photo-realism. However, through our experiments, we find that this comes with a trade-off. While ray conditioning creates realistic static images, it may introduce aliasing in videos. It also does not generalize well to out-of-distribution camera poses, as shown in Figure 8.

The difference between EG3D and ray conditioning echoes that of 3D geometry-based representations and light fields. If a subject is perfectly photo-consistent, then all views of the subject can be perfectly encoded in a 3D set of RGBA points. However, view-dependent effects such as specularities violate this assumption, as does high-frequency geometry when 3D resolution is finite [4]. The 4D light field accommodates such features by representing rays individually, which lets light reflected from a shared 3D point vary with angle.

By conditioning a 2D GAN on a light field prior, as opposed to using a 3D representation, we achieve the best photo-realism among all existing multi-view image synthesizers, with competitive identity consistency across viewpoints. We believe that, our method pushes forward the boundary of geometry-free generative models, and hope our conclusions can inspire a variety of work in new scene representations.

	Inversion				Novel Views	
	PSNR↑	SSIM↑	LPIPS ↓	ID↑	FID↓	KID ↓
EG3D [6]	26.56	**0.78**	0.12	0.75	65.0	0.0180
Ray Cond.	**27.49**	**0.78**	**0.10**	**0.85**	**58.5**	**0.0036**

Table 2. **GAN Inversion Metrics.** We measure reconstruction quality between input images and inversions on four metrics. We then compare the FID and KID$\times 100$ between the original images and images with random viewpoint changes. Ray conditioning can achieve higher image quality and detail in both input inversions, and after viewpoint change.

Figure 8. **Limitations on Camera Control**. Ray conditioning does not generalize well to out-of-distribution camera poses. The first row shows that ± 1.0 of x-axis camera translation leads to viewpoint distortion and identity shift. The second row shows that an out of distribution rotation of $\pm 75^\circ$ leads to identity shift.

References

[1] Rameen Abdal, Peihao Zhu, Niloy J. Mitra, and Peter Wonka. Styleflow: Attribute-conditioned exploration of stylegan-generated images using conditional continuous normalizing flows. *ACM Trans. Graph.*, 40(3), May 2021. 5

[2] Edward H. Adelson and James R. Bergen. The plenoptic function and the elements of early vision. 1991. 3

[3] Mikolaj Binkowski, Danica J. Sutherland, Michal Arbel, and Arthur Gretton. Demystifying mmd gans. *ArXiv*, abs/1801.01401, 2018. 5

[4] Jin-Xiang Chai, Xin Tong, Shing-Chow Chan, and Heung-Yeung Shum. Plenoptic sampling. In *ACM Trans. Graph.*, 2000. 8

[5] Eric Chan, Marco Monteiro, Petr Kellnhofer, Jiajun Wu, and Gordon Wetzstein. pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In *arXiv*, 2020. 3

[6] Eric R. Chan, Connor Z. Lin, Matthew A. Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas Guibas, Jonathan Tremblay, Sameh Khamis, Tero Karras, and Gordon Wetzstein. Efficient geometry-aware 3D generative adversarial networks. In *arXiv*, 2021. 1, 3, 4, 5, 6, 7, 8

[7] Angel X. Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. ShapeNet: An Information-Rich 3D Model Repository. Technical Report arXiv:1512.03012 [cs.GR], Stanford University — Princeton University — Toyota Technological Institute at Chicago, 2015. 3, 5

[8] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 5

[9] Jiankang Deng, Jia Guo, Xue Niannan, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, 2019. 5

[10] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In *IEEE Computer Vision and Pattern Recognition Workshops*, 2019. 4, 6

[11] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, 2014. 1, 2

[12] Steven J. Gortler, Radek Grzeszczuk, Richard Szeliski, and Michael F. Cohen. The lumigraph. *ACM Trans. Graph.*, 1996. 2, 3

[13] Jiatao Gu, Lingjie Liu, Peng Wang, and Christian Theobalt. Stylenerf: A style-based 3d aware generator for high-resolution image synthesis. In *Int. Conf. Learn. Represent.*, 2022. 1, 3, 7

[14] David Ha, Andrew M. Dai, and Quoc V. Le. Hypernetworks. In *Int. Conf. Learn. Represent.*, 2017. 3

[15] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NIPS*, 2017. 5

[16] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *arXiv preprint arxiv:2006.11239*, 2020. 1, 3

[17] Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. Ganspace: Discovering interpretable gan controls. In *Adv. Neural Inform. Process. Syst.*, 2020. 3

[18] Ali Jahanian*, Lucy Chai*, and Phillip Isola. On the "steerability" of generative adversarial networks. In *International Conference on Learning Representations*, 2020. 3

[19] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. In *Adv. Neural Inform. Process. Syst.*, 2020. 6, 7

[20] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. In *Neural Information Processing Systems*, 2021. 2

[21] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4396–4405, 2018. 2, 5

[22] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8107–8116, 2019. 2, 3, 7

[23] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020. 8

[24] Marc Levoy and Pat Hanrahan. Light field rendering. *ACM Trans. Graph.*, 1996. 2, 3

[25] Michael Niemeyer and Andreas Geiger. Giraffe: Representing scenes as compositional generative neural feature fields. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021. 3

[26] Roy Or-El, Xuan Luo, Mengyi Shan, Eli Shechtman, Jeong Joon Park, and Ira Kemelmacher-Shlizerman. StyleSDF: High-Resolution 3D-Consistent Image and Geometry Generation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 13503–13513, June 2022. 1, 3, 7

[27] William Peebles, Jun-Yan Zhu, Richard Zhang, Antonio Torralba, Alexei Efros, and Eli Shechtman. Gan-supervised dense visual alignment. In *CVPR*, 2022. 3

[28] Daniel Roich, Ron Mokady, Amit H Bermano, and Daniel Cohen-Or. Pivotal tuning for latent-based editing of real images. *ACM Trans. Graph.*, 2021. 4, 5, 7

[29] Sefik Ilkin Serengil and Alper Ozpinar. Lightface: A hybrid deep face recognition framework. In *2020 Innovations in Intelligent Systems and Applications Conference (ASYU)*, pages 23–27. IEEE, 2020. 5

[30] Yujun Shen, Ceyuan Yang, Xiaoou Tang, and Bolei Zhou. Interfacegan: Interpreting the disentangled face representation learned by gans. *TPAMI*, 2020. 3, 5

[31] Vincent Sitzmann, Semon Rezchikov, William T. Freeman, Joshua B. Tenenbaum, and Fredo Durand. Light field networks: Neural scene representations with single-evaluation

rendering. In *Adv. Neural Inform. Process. Syst.*, 2021. 2, 3, 4, 5, 6, 7, 8

[32] Vincent Sitzmann, Michael Zollhöfer, and Gordon Wetzstein. Scene representation networks: Continuous 3d-structure-aware neural scene representations. In *Advances in Neural Information Processing Systems*, 2019. 5

[33] Jascha Narain Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. *ArXiv*, abs/1503.03585, 2015. 1, 3

[34] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *Int. Conf. Learn. Represent.*, 2021. 1, 3

[35] Daniel Watson, William Chan, Ricardo Martin-Brualla, Jonathan Ho, Andrea Tagliasacchi, and Mohammad Norouzi. Novel view synthesis with diffusion models. *ArXiv*, abs/2210.04628, 2022. 3

[36] Bennett Wilburn, Neel Joshi, Vaibhav Vaish, Eino-Ville Talvala, Emilio Antunez, Adam Barth, Andrew Adams, Mark Horowitz, and Marc Levoy. High performance imaging using large camera arrays. *ACM Trans. Graph.*, 24(3):765–776, jul 2005. 4

[37] Zongze Wu, Dani Lischinski, and Eli Shechtman. Stylespace analysis: Disentangled controls for stylegan image generation. *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 12858–12867, 2021. 3

[38] Jiaxin Xie, Hao Ouyang, Jingtan Piao, Chenyang Lei, and Qifeng Chen. High-fidelity 3d gan inversion by pseudo-multi-view optimization. *arXiv preprint arXiv:2211.15662*, 2022. 3, 4

[39] Kai Zhang, Gernot Riegler, Noah Snavely, and Vladlen Koltun. Nerf++: Analyzing and improving neural radiance fields. *arXiv:2010.07492*, 2020. 1, 8

[40] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018. 8

[41] Xiaoming Zhao, Fangchang Ma, David Güera, Zhile Ren, Alexander G. Schwing, and Alex Colburn. Generative multiplane images: Making a 2d gan 3d-aware. In *Eur. Conf. Comput. Vis.*, 2022. 1, 3, 5, 6, 7

[42] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *ArXiv*, abs/1805.09817, 2018. 3