Appendix

In this document, we present implementation details and additional results that are supplemental to the main paper. In Appendix A, we outline key design details for the ray conditioning method. In Appendix B, we include more results of viewpoint editing. We also compare to concurrent work in 3D-aware image inversion, and show that these methods still have challenges with respect to realism. Finally, in Appendix C, we include more details about the datasets used.

A. Design Details

We outline some key design details in this section.

A.1. Pretraining and Weight Initialization

Recall that the ray embedding is a $6 \times H \times W$ feature map which is concatenated to each intermediate representation of StyleGAN. To accommodate these extra features, we add 6 channels to each convolutional layer of StyleGAN. To begin training, we initialize all prior weights to be those of a pretrained StyleGAN model. The extra 6 channels for each layer is then initialized normally, with the default StyleGAN2 gain parameters [11, 6]. The discriminator also uses pretrained discriminator weights. The pose conditioninal discriminator module is initialized as an MLP in the same way done in EG3D and StyleGAN2-ADA [2, 8]. In Figure 2, we demonstrate the effect that ray conditioning has on the intermediate layers of StyleGAN. Consistent with prior work [18, 7], subject pose is generally a coarse-level feature which takes form around resolutions 16×16 to 64×64. Ray conditioning also converges very fast. In Figure 1, we show that ray conditioning takes about 160kimgs, or 1.5 hours on 2 Nvidia A6000s, to learn camera pose.

Init 80kimgs 160kimgs 240kimgs 320kimgs

Figure 1. **Training Convergence Speed.** For faces, we initialize our ray conditioning model from a pretrained StyleGAN2 model. Through training, it learns to properly generate images from a target pose. Ray conditioning converges quickly. After 160kimgs (1.5 hours on 2 A6000s), ray conditioning is already able to properly generate an image at a target pose.

A.2. Effect on Latent Space

Many prior work have found directions in StyleGAN's $\mathcal{W}$ latent space which correspond to subject pose [7, 14]. Surprisingly, ray condition nullifies these directions in the latent space. When we try to modify an image's pose using InterfaceGAN directions [14] after ray conditioning, the image stays the same. This implies that ray conditioning "moves" the pose information previously embedded in the $\mathcal{W}$ latent space into the convolutional weights assigned to the ray embedding.

A.3. StyleGAN2 vs. StyleGAN3.

We have implemented ray conditioning successfully for both StyleGAN2 [11] and StyleGAN3 [9]. StyleGAN3 is better suited for video generation tasks because of its antialiasing abilities. However, for multi-view image generation, we have found no advantage for using StyleGAN3. Following EG3D [2] and GMPI [20], we choose to use StyleGAN2 because of its slightly higher image quality and faster model.

A.4. Training Details

FFHQ and AFHQ models were trained starting from official StyleGAN checkpoints. They were trained on $2\times$ Nvidia A6000 GPUs for 1040kimgs - 1440kimgs. For SRN Cars, we train from scratch for 13,520kimgs. Hyperparameters are set to the same as those of StyleGAN2.

B. Additional Results

B.1. Viewpoint Editing Examples

We believe that ray conditioning is a natural choice for portrait editing over 3D-aware GANs. In Figure 5, we recreate Figure 2 in the main paper with EG3D. We see that the resulting images from EG3D appear more cartoonish than human-like. There are also geometry distortions. In the top right individual (yellow background), we see geometry artifacts near his ear. In the bottom example (pink background), the challenges are also apparent when we try to blend faces into a preexisting image. EG3D fails to achieve our intended effect of photorealistic viewpoint editing. In the supplementary material, we also include a video demonstrating the viewpoint control we have over input samples.

B.2. 3D-Aware GAN Inversion

Many have recognized the issues with using a 2D GAN inversion method such as PTI [13] with a 3D-aware GAN such as EG3D [2]. Although PTI can successfully invert an input image, it can cause geometry artifacts that are only realized after a change in viewpoint. Several work have created dedicated 3D GAN inversion methods for EG3D. However, we find that these methods can still cause aliasing in the inverted images, creating a loss of quality. In

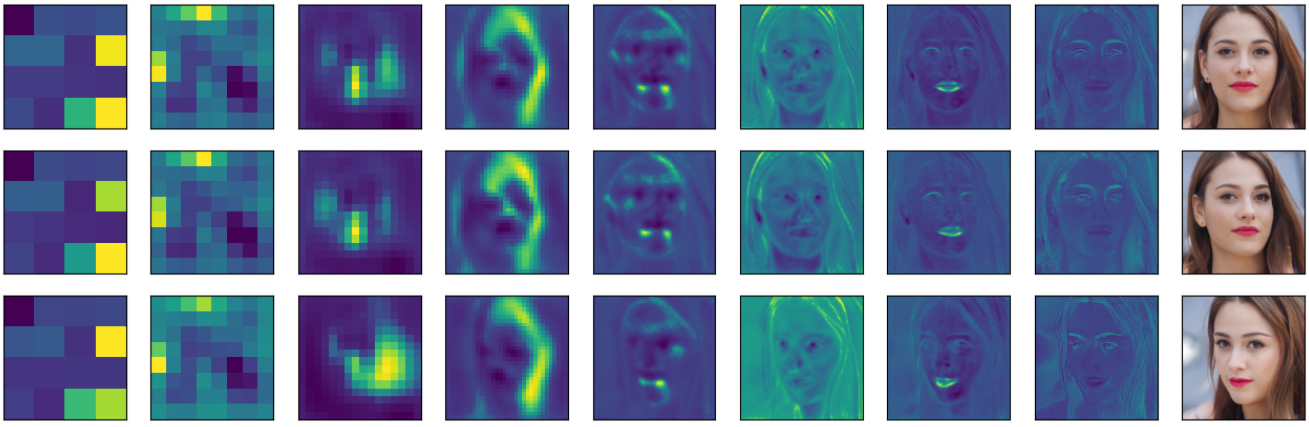


Figure 2. **Ray Conditioning Affects Intermediate Features.** To learn viewpoint control, we condition each intermediate layer of StyleGAN on a ray embedding. We picture coarse to fine resolution feature maps of one latent seed from different poses. Consistent with prior work [18, 7], we find that subject pose is a coarse-level feature in the StyleGAN latent space. It begins to take form around resolutions 16×16 to 64×64.

addition, they can struggle when inverting side-facing images. The occluded regions can have incorrect geometry. The eyes can also lose their specularity, which is important for human perception of identity. In Figure 6, and Figure 7 we demonstrate these issues. We compare ray conditioning to two related work. 3D GAN Inversion with Pose Estimation [12] is a recent work designed for EG3D inversion. HFGI3D [19] is a concurrent work also designed for EG3D inversion. Inversion with dedicated 3D-aware methods can also take much longer than with PTI. For instance, 3D GAN Inversion with Pose Estimation [12] takes on average 3 minutes per image. HFGI3D [19] can take 8 minutes. Meanwhile, PTI with ray conditioning only takes 1 minute per image.

B.3. Latent Space Pose Editing

InterfaceGAN and similar work [14, 7, 1] allow for viewpoint change by finding directions in the StyleGAN latent space which correspond to pose. However, these methods can only do binary changes such as left facing or right facing, instead of explicit viewpoint control. Because they do not operate per-pixel like how ray conditioning does, they also lack the spatial inductive bias which makes ray conditioning effective. We show an example of the differences between latent space editing and ray conditioning in Figure 8.

B.4. Latent Space Samples

In Figure 11, we show uncurated latent space samples from StyleGAN2 with ray conditioning. Even without a 3D representation, ray conditioning is still able change the viewpoint of generated samples. We picture latent seeds 0-31. To show our image quality, we also present larger images in Figure 10.

B.5. Additional Results on Cars

In Figure 3, we demonstrate that our model can enable 360° viewpoint editing when trained on a dataset with 360° of views. The car stays consistent as we rotate the camera. We also include videos of smooth trajectories in the supplementary material. We use StyleGAN3 [9] because of its antialiasing properties.

Figure 3. **360° View Consistency from Ray Conditioning.** When trained on a multi-view dataset, ray conditioning is able to generate view consistent results with 360° of rotation. Please see the accompanying videos for continuous results.

B.6. More Experiments on Light Field Networks

In terms of image quality, ray conditioning is a large improvement over Light Field Networks (LFNs) [16]. We pro-

vide more results from LFNs on FFHQ in Figure 4. As discussed in the main paper, LFNs have two main challenges. First, LFNs struggle to reconstruct high frequency details on a photo-realistic dataset such as FFHQ. We also attempted to train LFNs with SIREN [15] activations instead of ReLU activations, but the model struggled to converge. Second, LFNs are unable to generate novel views when trained on a dataset with only one image per instance. Our work demonstrates that light field priors introduced in LFNs can be naturally extended from MLPs to more powerful CNN-based image synthesizers.

Figure 4. **Light Field Networks on FFHQ.** Light Field Networks (LFNs) [16] have two key challenges. First, they struggle to reconstruct high frequency details from input images. Second, when trained on a dataset with only one image per face, LFNs [16] struggle to construct novel views (NVs). When combined with a more powerful generative model such as a GAN, ray conditioning helps to address both of these problems.

C. Evaluation Details

C.1. Datasets

Similar to prior work [2, 20], our method requires a dataset of images and estimated camera poses. We outline the datasets used below.

FFHQ Human Faces. FFHQ [10] is a dataset of $\sim$ 70k 1024×1024 images of front-facing faces. We use camera poses provided by EG3D, which are estimated by a deep face pose estimator [5]; camera poses are assumed to be distributed on a sphere, all facing a shared center. As previously reported by EG3D, this dataset contains bias which may affect the resulting generations. For instance, people in front-facing images are more likely to smile. People who appear to be lower than the camera tend to be children. In Figure 9, we present the distribution of subject pose in terms of yaw (horizontal rotation), and pitch (vertical rotation).

AFHQv2 Cat Faces. AFHQv2 [4] is a dataset consisting of many animal faces. We train our model on the cat subset using the camera poses provided by EG3D. This subset only consists of $\sim$ 5k 512×512 images, which is much smaller than FFHQ. Some pretraining is expected for good results.

SRN Cars. The SRN Cars [17] training set is a collection of $\sim$2.5k ShapeNet [3] cars, each rendered from 250 cameras distributed on a sphere at a resolution of 128×128. Because it contains multiple images per object, it is commonly used for evaluating geometry-free view synthesis models. We demonstrate our method's ability to generate 360° light fields, and compare to the LFNs [16] baseline on this dataset. Unlike for FFHQ or AFHQ, we start training from scratch.

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. 2

[2] 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

[3] 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

[4] 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. 3

[5] 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. 3

[6] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. *Int. Conf. Comput. Vis.*, pages 1026–1034, 2015. 1

[7] Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. Ganspace: Discovering interpretable gan controls. In *Adv. Neural Inform. Process. Syst.*, 2020. 1, 2

[8] 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. 1

[9] 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. 1, 2

Ray Conditioning + PTI [13] EG3D [2] + PTI [13]

Figure 5. **Photo Editing Comparison.** We believe that ray conditioning is a natural choice for portrait editing over 3D-aware GANs. To illustrate, we compare the ray conditioning results from Figure 2 in the main paper to EG3D [2]. We see that EG3D cannot create a change in viewpoint without sacrificing some realism. There are noticeable geometry issues, and the edited individuals seem more cartoonish than human-like.

[10] 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. 3

[11] 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. 1

[12] Jaehoon Ko, Kyusun Cho, Daewon Choi, Kwangrok Ryoo, and Seungryong Kim. 3d gan inversion with pose optimization. *WACV*, 2023. 2, 5, 6

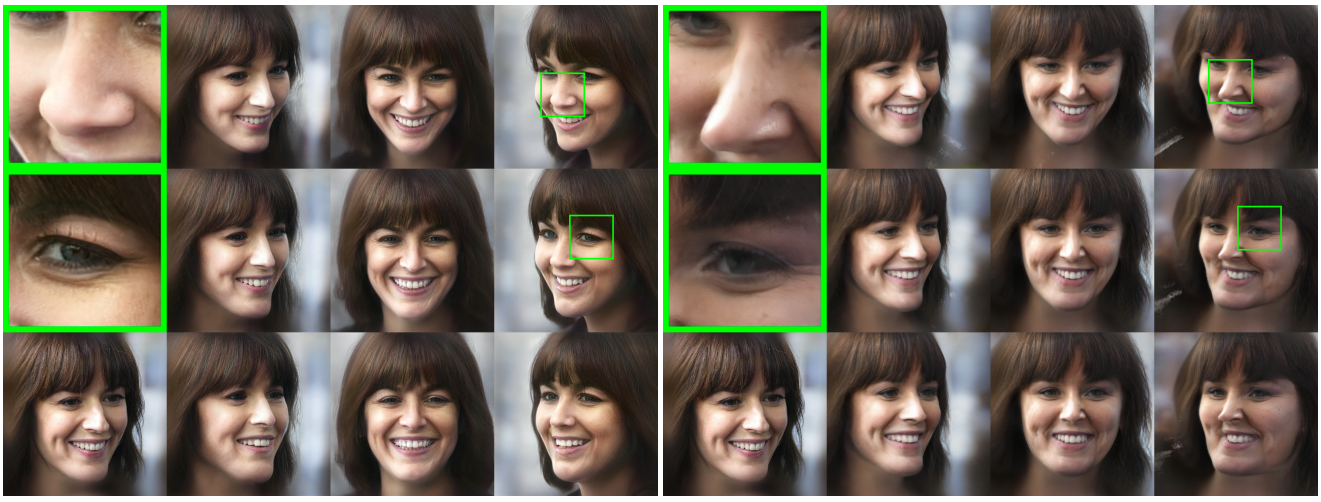
[13] Daniel Roich, Ron Mokady, Amit H Bermano, and Daniel Cohen-Or. Pivotal tuning for latent-based editing of real images. *ACM Trans. Graph.*, 2021. 1, 4, 5, 6

[14] Yujun Shen, Ceyuan Yang, Xiaoou Tang, and Bolei Zhou. Interfacegan: Interpreting the disentangled face representation learned by gans. *TPAMI*, 2020. 1, 2, 7

[15] Vincent Sitzmann, Julien N.P. Martel, Alexander W. Bergman, David B. Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. In *Adv. Neural Inform. Process. Syst.*, 2020. 3

[16] 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

[17] 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. 3



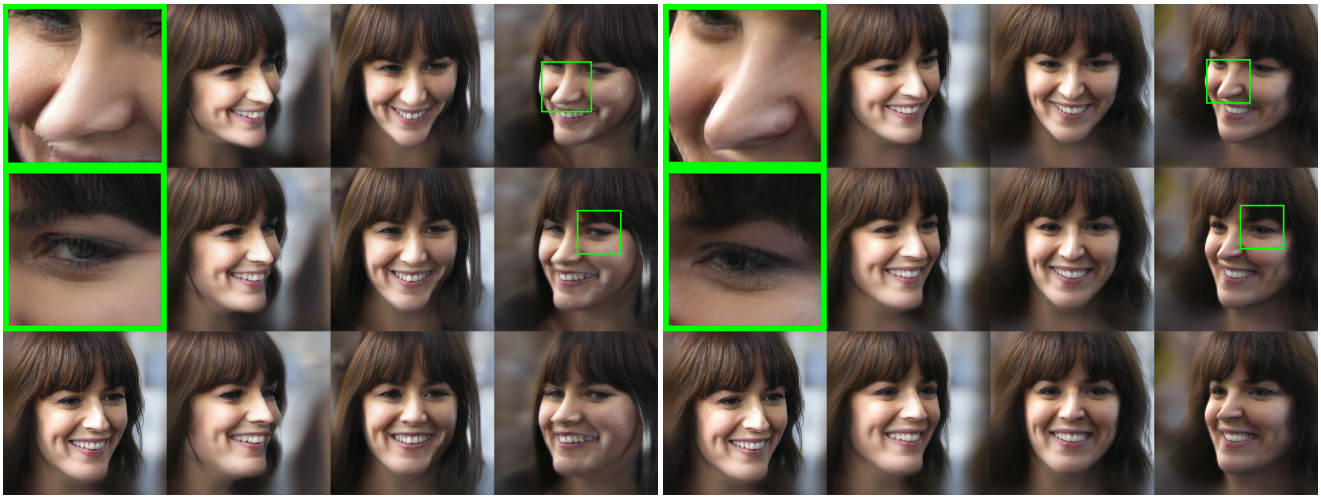
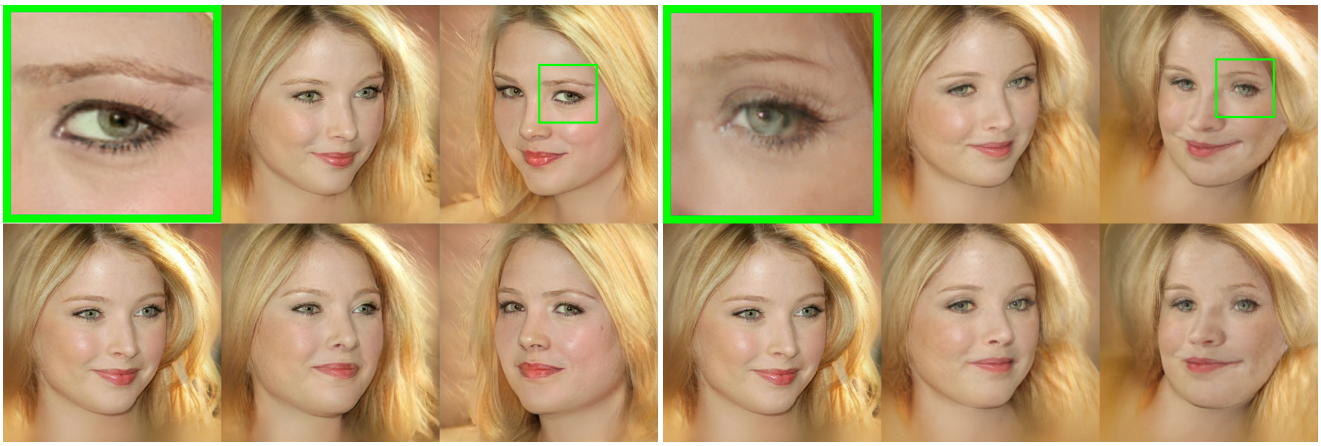


Figure 6. **Comparison to 3D-Aware GAN Inversion Methods.** Even dedicated 3D GAN inversion methods for EG3D [12, 19] can struggle to generate realistic novel views. It is especially challenging for 3D-aware GAN inversion methods when the input image is not front facing. It tends to distort the geometry of the input individual. For instance, when combined with EG3D, 3D GAN inversion [12] and PTI [13] appear to widen the individual's face. PTI [13], 3D GAN Inversion [12], and HFGI3D [19] all appear to make the individual's nose more pointy.

[18] 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. 1, 2

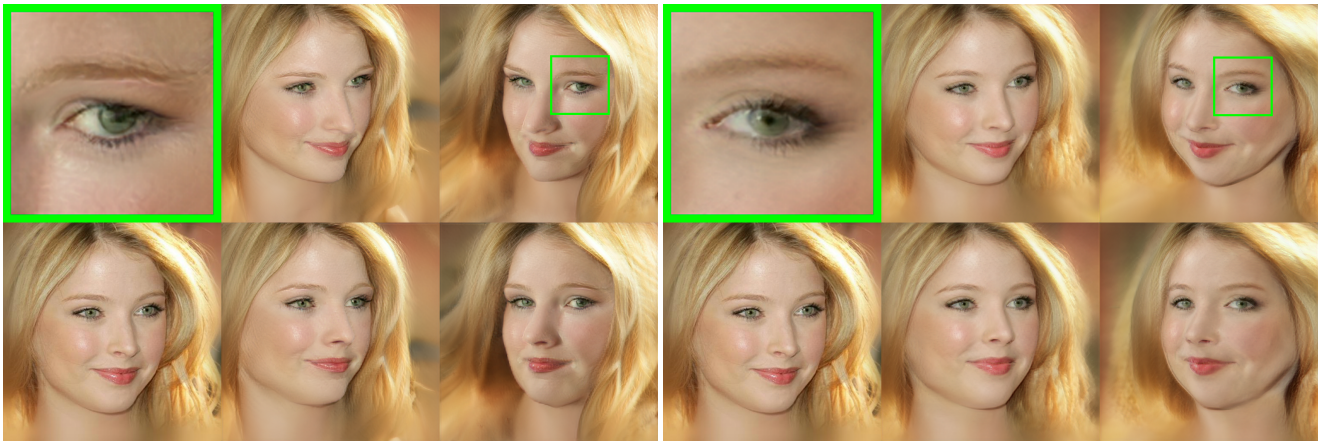
[19] Jiaxin Xie, Hao Ouyang, Jingtan Piao, Chenyang Lei, and Qifeng Chen. High-fidelity 3d gan inversion by pseudo-multi-view optimization. *arXiv preprint arXiv:2211.15662*, 2022. 2, 5, 6

[20] 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



Input Novel Views
Ray Conditioning + PTI [13]

Input Novel Views
EG3D [2] + PTI [13]



Input Novel Views
EG3D [2] + 3D GAN Inversion [12]

Input Novel Views
EG3D [2] + HFGI3D [19]

Figure 7. **Comparison to 3D-Aware GAN Inversion Methods.** Ray conditioning provides the best image quality compared to geometry-based GANs and inversion methods. The loss of image quality is noticeable. For 3D GAN Inversion [12] and HFGI3D [19], there are streaks across the face, hinting at spatial aliasing issues. 3D GAN Inversion and HFGI3D also cannot reconstruct the specularity of the eyes, which is very apparent in the ray conditioning example.

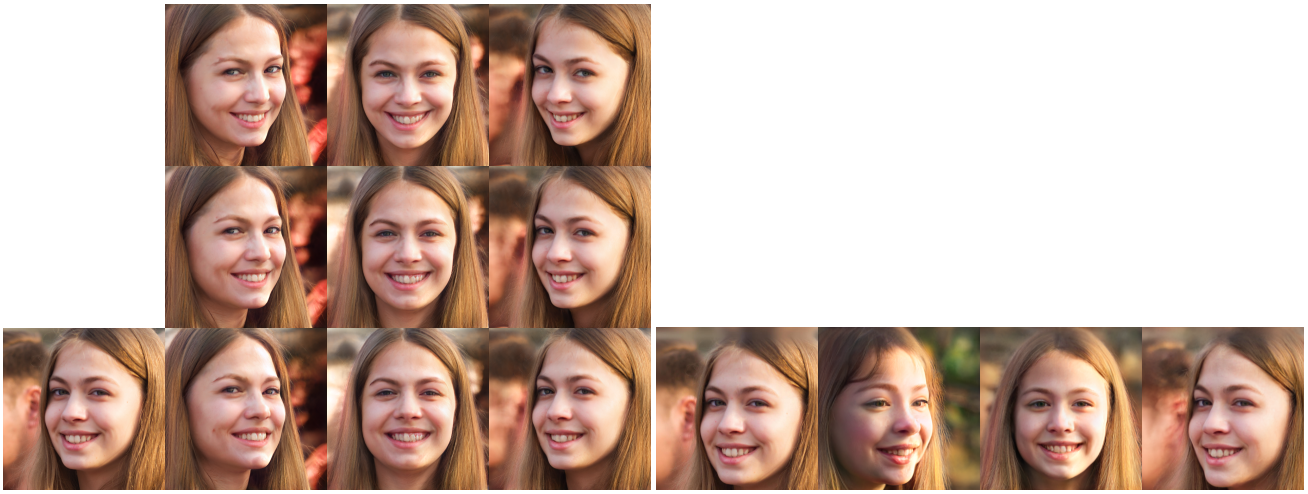


Figure 8. **Comparison to InterfaceGAN.** Latent space editing techniques such as InterfaceGAN [14] can generate binary changes to images to make them left facing or right facing. However, these models lack the same level of control that ray conditioning and geometry-based generative models do. We can achieve free viewpoint control with ray conditioning, while latent space editing is restricted to one dimension. InterfaceGAN also lacks the same amount of disentanglement as we do, causing identity shift.

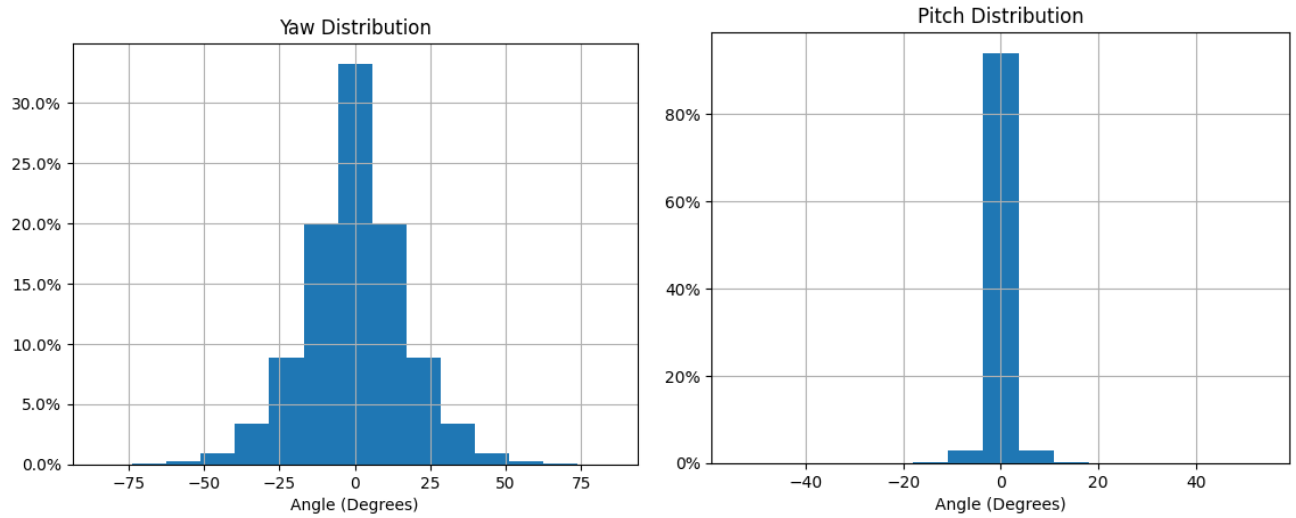


Figure 9. **FFHQ Pose Distribution.** FFHQ consists mainly of front facing photographs. Its standard deviation for yaw (horizontal rotation) is 16°. Its standard deviation for pitch (vertical rotation) is 2° degrees. As noted by EG3D [2], this unbalanced distribution of subject pose is a challenge for all multi-view generative models trained on FFHQ. Proper data augmentation to reduce distribution bias is still an open and important problem.

Figure 10. **Curated Latent Space Samples.** Given the 12 front facing images, we show that we can map them to a different viewpoint while maintaining the quality of the generated faces.

Figure 11. **Uncurated Latent Samples.** We picture two views from latent vectors of seeds 0-31 to demonstrate the quality of our GAN. Even without a 3D representation, ray conditioning successfully generates images of the same individual from different view points. Results were generated with a truncation of $\psi = 0.7$.