Overcoming Camera-Facing Gaze Bias in EG3D Scene Generation

Kinam Kim, Juhyun Lee, Youngdo Lee
KAIST
{kimkinam1118, dhwndudkaps2, lyd0531}@kaist.ac.kr

Abstract

Notwithstanding that EG3D [3] achieved realistic image quality with high computational efficiency, the gaze following problem hinders the real-world application of 3D GANs such as virtual reality and human-computer interaction. Due to the data bias when rotating the face, all existing 3D GANs cannot maintain the direction of eye when the camera is rotating and gazes the camera. In this work, we tackle this problem by fusing 3D-aware image synthesis with gaze estimation for the first time. We subjoin a pre-trained state-of-the-art gaze estimation model called L2CS-Net [1] into training pipeline of EG3D, successfully addressed the inherent gaze following issue on FFHQ dataset. Code is available at: https://github.com/3D-eye-centric-bias/Gaze-Corrected-EG3D

Figure 1. Comparison of generated sample for same individual from ours (up), and EG3D (down). Our framework enables EG3D, controlling the eyes to gaze the straight ahead, not following the camera when rotating face.

1. Introduction

Recently, remarkable advancement in deep learning triggered creation application such as generating high-fidelity images or 3D objects. In particular, deep generative 3D-aware image synthesis [10, 18], which generates high-resolution photorealistic images, capturing 3D-consistency and intricate geometry of 3D objects from 2D image collections, draws wide attention.

Recent progress [3, 4, 17, 18, 21] is mainly driven by blending a Neural Radiance Field (NeRF) [16] into a Generative Adversarial Network (GAN) [9]. Although prior works have achieved great success, these works, more concretely EG3D [3], suffer from the gaze following problem [3], degrading the model performance. Due to the data bias when rotating faces, the generator consistently creates an incorrect interpretation of eyes that gaze the camera. In the real world, the human gaze serves as a vital cue in a plethora of applications, including virtual reality and human-robot interaction. Therefore, addressing the gaze following problem holds practical significance in computer vision and graphics applications and is possible to enhance the model's generalization capability.

In this paper, we propose the first approach to effectively preserve eye's orientation when rotating face images. Our strategy involves the integration of a gaze estimation [5, 19] model, designed to extract the direction vector of the gaze, in parallel with the discriminator of EG3D. And we aim to modify the training algorithm to ensure that the model generates a face from the same latent code while enabling it to rotate. We provide the details of our approach in Section 3. Also, Fig. 3 summarizes our training framework. And through our new framework, we obtained the successful results on the FFHQ dataset, solving the gaze following problem without harming the multi-view consistency and pose accuracy of EG3D (see Fig. 1).

In summary, our contributions are the following:

- We tackle the gaze following problem for the first time in the field of 3D-aware image synthesis, by integrating pre-trained gaze estimation model into the training pipeline of EG3D.
- We successfully solved the gaze following problem, coming at a price of image quality. However, in terms of view consistency and pose accuracy, our model achieve better than 90% of the performance of EG3D.

2. Related work

Generative 3D-aware image synthesis. GANs have recently achieved the great success in deep generative 3D-aware image synthesis. Although state-of-the-art GANs are viable in 2D only and lack knowledge of 3D structure, groundbreaking prior works integrate NeRF models into 2D GANs to overcome the problem [3, 4, 17, 18, 21]. While such methods obtain 3D consistency with fair quality, they encounter challenges when training on high-resolution images, arising from the expensive rendering process involved in radiance fields. GIRAFFE [17] discards this inefficiency by convolutional network renderer, but harming the 3D consistency. And EG3D [3] improves both efficiency and image quality by introducing a tri-plane-based framework with dual-discrimination strategy, and have demonstrated state-of-the-art results on the FFHQ [14] dataset.

All aforementioned works, however, ignored the gaze following problem in spite of its importance in real-world applications, focusing on efficiency and image quality only. Hence, in this work we advocate to achieve better disentanglement and fine-grained control of eye gaze.

Gaze estimation. Mainly due to the advances of deep learning, CNN-based methods are the most successful approaches for gaze estimation. Without any dedicated hardwares, it captures human eye appearance and regresses the gaze direction from the images captured by cameras that are inexpensive and off-the-shelf such as webcams. And most of these works have focused on developing novel network architectures and framework. [1,6,8,15,20] For example, Gaze360 model [15] integrated LSTM [13] into ResNet [11] to predict gaze angles from webcam video. Despite achieving high accuracy, these models exhibit deficiencies in robustness and generalization, particularly when exposed to unconstrained factors including varying lighting conditions, diverse head poses. L2CS-Net [1] demonstrated the state-of-the-art results by gaze bin classification; estimating the neighborhood of the gaze angle in a robust manner instead of directly regressing the gaze direction.

The indispensable properties of gaze estimation model to combine in 3D-aware image synthesis are efficiency and generalization. Definitely, gaze estimation model pretrained on gaze datasets must be robust on image generation dataset including FFHQ. And L2CS-Net obtained robustness with simple ResNet-based networks, not hampering the efficiency and end-to-end pipeline of EG3D. In this regime, we adopt L2CS-Net to guide EG3D for more controllable synthesis of eye gazes.

3. Method

3.1. Data preprocessing

The main problem that this paper aims to address is that the eyes of the images generated by the EG3D model appear to be looking at the camera. The FFHQ [14] dataset comprises a significant number of frontal photographs. The camera poses employed by the EG3D generator during training are derived from the actual dataset, FFHQ. Consequently, an abundance of front-facing data results in a proportional increase in frontal perspectives within the generated images. However, the issue addressed in this paper, namely the problem of iris bias, is not prominently evident in frontal photographs. Frontal images, characterized by a direct gaze towards the camera, pose challenges in distinguishing whether iris bias is manifest or if it is an inherent characteristic of facing forward. In essence, a dataset rich in front-facing images hinders the extraction of an appropriate camera position for learning iris bias, leading to a decrease in overall model performance.

To address this challenge, gaze angles were extracted for FFHQ dataset images, and only those with a yaw absolute angle exceeding 0.4 were selectively curated for inclusion in the training dataset, effectively addressing the impact of biased iris positioning.

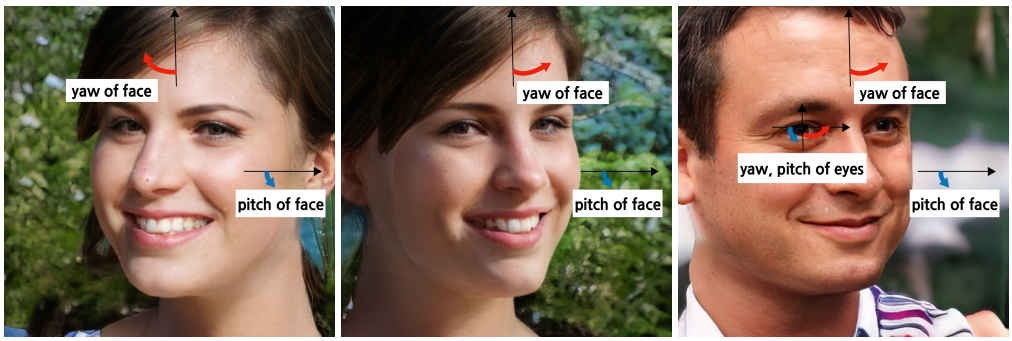
3.2. Gaze corrected EG3D architecture

Fine-tuning with adjusted gaze angles is essential to generate images where the human looks forward rather than at the camera. To achieve the generation of images with a forward-facing gaze, it is necessary to redefine the loss function. The ideal images we seek are those where the gaze aligns with the face's direction, looking straight ahead (Fig. 2c). However, the images generated by the existing EG3D model consistently exhibit a gaze directed at the camera, irrespective of the face's orientation (Fig. 2a and Fig. 2b). In other words, the desired outcome occurs when the yaw and pitch values representing the face's direction align with the yaw and pitch values of the eyes.

To address this, we extract the face angle from the camera position and input the generated image into the gaze estimation model, L2CS, to extract the angle of the eyes. The loss value (Sec. 3.3), calculated using these two angles, is incorporated into the generator's loss function to facilitate training, ensuring alignment between the face angle and the gaze angle of the eyes (see Fig. 3).

3.3. Loss function

In order to align the face angle and eye angle through fine-tuning, it is essential to adjust the loss function to minimize the difference between these angles during training. We modified the existing generator loss function of the EG3D model by incorporating Mean Squared Error (MSE)



(a) EG3D model output (right face) (b) EG3D model output (right face) (c) ideal generated image

Figure 2. The images generated by the EG3D exhibit a discrepancy between the yaw and pitch values of the eyes and those of the body, consistently biased towards facing the camera. (a) In images generated by EG3D, the body is oriented to the left, while the eyes are facing forward. (b) In EG3D-generated images, the body is oriented to the right, but the eyes are facing forward. (c) An ideal image would be one where the subject is looking straight ahead without facing the camera.

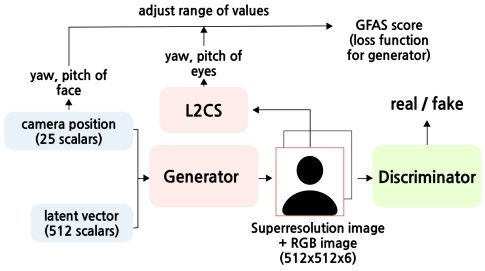


Figure 3. Overall of corrected gaze EG3D architecture.

loss between the yaw and pitch values of the face and gaze, defining a new loss function for this purpose (Eq. (1)). We denote pitch of face as p_f, pitch of gaze as p_g, yaw of face as y_f, and yaw of gaze as y_g.

$$MSE(f,g) = \frac{1}{N}\sum^{N}\sqrt{(p_f - p_g)^2 + (y_f - y_g)^2} \quad (1)$$

4. Experiments and results

4.1. Experimental Setups

Datasets. We compare our results with EG3D on commonly-use dataset in 3D-aware image synthesis named FFHQ [14], a high-quality collection of human face images. It consists of 70,000 images at 1024×1024 resolution, spanning a diverse range of ages, ethnicities, and facial features. We note that there exist some images that resisted face detection [3], so that we discard them for our purpose. Therefore, our final dataset contain 69,957 images and are preprocessed as described in Section 3.1.

Computing Resources and Time. Our experimental process encompassed several key stages, each demanding specific computational resources and time. Initially, data preprocessing, including dataset download and preparation, took about 6 hours. For the training phase, we utilized four NVIDIA GeForce RTX 3090 GPUs, dedicating approximately 10 hours to train the model effectively. Finally, the evaluation process, which involved comprehensive performance assessments of our model, the baseline, and various ablation study configurations, required an additional 6 hours. This comprehensive outline of the resources and time investment provides insights into the computational intensity of our experimental work.

Evaluation metrics. To evaluate our result and compare with EG3D effectively, we inherit evaluation metrics from their works. For image quality, we computed Fréchet Inception Distance (FID) [12] and Kernel Inception Distance (KID) [2] between 50k generated images and real images. Also, for multi-view facial identity consistency, we calculate the mean of ArcFace [7] cosine similarity score between the same synthesized face rendered from random camera poses. In addition, to assess the alignment of face direction and gaze direction, we develop new metrics named GAFS (Gaze-Face Alignment Score), the average of Eq. (1) across randomly generated 1024 images.

4.2. Comparisons

Baseline. As we described, we compare our approaches against EG3D [3], the state-of-the-art methods for deep generative 3D-aware image synthesis, on FFHQ [14].

Qualitative results. We provide the best result synthesized by our method with FFHQ in Fig. 4 to highlight the power of our approach. While EG3D suffer from the gaze following problem, our model rotates not only the face, but also its gaze following the face. Furthermore, the generated images maintain higher quality and view-consistency of EG3D.

Figure 4. Qualitative comparison between EG3D and ours. Although the details such as glasses and hair style are also influenced by the model, our model controls the direction of eye gaze to be consistent with facial direction.

	FFHQ				
	GFAS ↓	FID ↓	KID ↓	ArcFace ↑	Pose ↓
EG3D	0.230	**4.1**	**0.12**	0.45	0.038
Ours	**0.215**	5.8	0.19	**0.50**	**0.037**

Table 1. Quantitative evaluation using GFAS, FID, KID×100, ArcFace and Pose×10.

	FFHQ			
	GFAS ↓	FID ↓	ArcFace ↑	Pose ↓
Baseline	0.233	6.3	0.45	**0.031**
+GL	0.225	**5.3**	0.45	0.038
+DP	0.228	7.8	0.47	0.039
+GL, DP (Ours)	**0.215**	5.8	**0.50**	0.037

Table 2. Quantitative evaluation of ablation study. The table compares the baseline model with models having added gaze loss (+GL), data preprocessing (+DP), and both (+GL, DP). GFAS, FID, Pose × 10, and ArcFace metrics are presented.

Quantitative results. We provide our quantitative evaluations in Tab. 1. For image quality, multi-view facial identity consistency, and eye gaze direction assessment, we computed various metrics as described in Section 4.1. As a result, due to an additional constraint on eye gaze of images, our model shows significant degradation in image quality, FID and KID. But in terms of view consistency and pose accuracy, we demonstrate that our model succeeded to maintain the state-of-the-art performance of EG3D.

4.3. Ablation Study

Our ablation study, detailed in Tab. 2, focuses on the impact of Gaze Loss (GL) and Data Preprocessing (DP) on model performance. Both GL and DP independently improve the Gaze-Face Alignment Score (GFAS), but their combined use leads to a greater improvement, demonstrating their synergistic effect. DP enhances identity features as reflected by ArcFace scores, though at the cost of image quality (FID score). In contrast, GL improves FID score. Together, they optimize both GFAS and balance FID and ArcFace scores, indicating a comprehensive improvement in image synthesis.

5. Discussion

Limitations and future work. Our approach has limitations, such as unintended transformations of facial features when using L2CS-Net, due to gradients affecting the entire face image. Also, Adjusting latent variables for specific features like the eyes can inadvertently alter other aspects, as shown in Figure 1 (clothing changes) and Figure 4 (glasses changes). This highlights the challenge of disentangling facial features in gaze correction models. To enhance our model, we propose some possible future works:

- **Discriminator Aware of Gaze Angle:** Integrate gaze information into the discriminator for more accurate gaze alignment.
- **Freezing Early Layers of Generator:** Focus training on later layers to adjust eye details without changing overall facial features.
- **Using 3D Masks for Targeted Modifications:** Apply 3D masks for precise adjustments to specific facial features, addressing latent variable entanglement.

Conclusions. Our work represents a significant stride in 3D-aware image synthesis, focusing on the critical aspect of gaze correction in generated images. By innovatively integrating gaze loss and data preprocessing into the 3D generative framework, we have successfully addressed the inherent gaze following issue prevalent in previous models. Our method enhances the gaze-face alignment, as indicated by the reduced GFAS.

One of the key strengths of our approach is its independence from the specific architecture of EG3D. By defining a universal loss function and applying data preprocessing, our method gains versatility, suitable for various 3D generative models to enhance gaze alignment training. This adaptability opens up broader possibilities in 3D image synthesis. The effectiveness of our model in maintaining accurate gaze direction and identity features signifies a notable advancement in generative image synthesis, paving the way for more realistic and tailored applications. Future research can build on our work to expand the capabilities of 3D-aware generative models.

Acknowledgements

This project was conducted by Kinam Kim, Juhyun Lee, and Youngdo Lee. Specifically, Kinam Kim, the leader of the project, implemented the main training and evaluation codes including data preprocessing and spent a lot of time monitoring the training of models. Also, Juhyun Lee put in hard work to setup the experiment environment without dependency conflict and was in charge of the documentation of the project. She implemented loss function codes including gaze loss and also participated for training the model. And Youngdo Lee implemented gaze estimation model and connected with EG3D model for the project. He built the visualization code for qualitative comparisons and also participated in evaluating the model. For implementation, we mainly borrowed dedicated implementations of EG3D and L2CS-Net. Their official implementations are available at https://github.com/NVlabs/eg3d and https://github.com/Ahmednull/L2CS-Net, respectively. But we autonomously combined two codes into one pipeline and also modified or newly implemented training and evaluation codes for our experimental purpose. Also, We borrowed code from Deep3DFaceReconPytorch https://github.com/3D-eye-centric-bias/Deep3DFaceRecon_pytorch for the extraction of camera parameters from the images. Please see our codes at a link provided in the abstract.

References

[1] Ahmed A Abdelrahman, Thorsten Hempel, Aly Khalifa, and Ayoub Al-Hamadi. L2CS-Net: fine-grained gaze estimation in unconstrained environments. *arXiv preprint arXiv:2203.03339*, 2022. 1, 2

[2] Mikolaj Binkowski, Danica J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. 3

[3] 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 *CVPR*, 2022. 1, 2, 3

[4] Eric R. Chan, Marco Monteiro, Petr Kellnhofer, Jiajun Wu, and Gordon Wetzstein. 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 (CVPR)*, pages 5799–5809, June 2021. 1, 2

[5] Yihua Cheng, Haofei Wang, Yiwei Bao, and Feng Lu. Appearance-based gaze estimation with deep learning: A review and benchmark, 2021. 1

[6] Murthy L R D and Pradipta Biswas. Appearance-Based Gaze Estimation Using Attention and Difference Mechanism. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 3143–3152, June 2021. 2

[7] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. ArcFace: Additive Angular Margin Loss for Deep Face Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 3

[8] Tobias Fischer, Hyung Jin Chang, and Yiannis Demiris. RT-GENE: Real-Time Eye Gaze Estimation in Natural Environments. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018. 2

[9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. 1

[10] Jiatao Gu, Lingjie Liu, Peng Wang, and Christian Theobalt. StyleNeRF: A Style-based 3D Aware Generator for High-resolution Image Synthesis. In *International Conference on Learning Representations*, 2022. 1

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2

[12] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. 3

[13] Sepp Hochreiter and Jürgen Schmidhuber. Long Short-Term Memory. *Neural Comput.*, 9(8):1735–1780, nov 1997. 2

[14] Tero Karras, Samuli Laine, and Timo Aila. A Style-Based Generator Architecture for Generative Adversarial Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 2, 3

[15] Petr Kellnhofer, Adria Recasens, Simon Stent, Wojciech Matusik, and Antonio Torralba. Gaze360: Physically Unconstrained Gaze Estimation in the Wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 2

[16] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 1

[17] Michael Niemeyer and Andreas Geiger. GIRAFFE: Representing scenes as compositional generative neural feature fields. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1, 2

[18] Katja Schwarz, Yiyi Liao, Michael Niemeyer, and Andreas Geiger. GRAF: Generative Radiance Fields for 3D-Aware Image Synthesis. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 20154–20166. Curran Associates, Inc., 2020. 1, 2

[19] Xinming Wang, Jianhua Zhang, Hanlin Zhang, Shuwen Zhao, and Honghai Liu. Vision-Based GazeEstimation: A Review. *IEEE Transactions on Cognitive and Developmental Systems*, 14(2):316–332, 2022. 1

[20] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. MPIIGaze: Real-World Dataset and Deep Appearance-Based Gaze Estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(1):162–175, 2019. 2

[21] Peng Zhou, Lingxi Xie, Bingbing Ni, and Qi Tian. CIPS-3D: A 3D-Aware Generator of GANs Based on Conditionally-Independent Pixel Synthesis, 2021. 1, 2