High-Fidelity Eye Animatable Neural Radiance Fields for Human Face

Hengfei Wang
hxw080@student.bham.ac.uk
Zhongqun Zhang
zxz064@student.bham.ac.uk
Yihua Cheng✉
y.cheng.2@bham.ac.uk
Hyung Jin Chang
h.j.chang@bham.ac.uk

School of Computer Science
University of Birmingham
Birmingham, UK

Abstract

Face rendering using neural radiance fields (NeRF) is a rapidly developing research area in computer vision. While recent methods primarily focus on controlling facial attributes such as identity and expression, they often overlook the crucial aspect of modeling eyeball rotation, which holds importance for various downstream tasks. In this paper, we aim to learn a face NeRF model that is sensitive to eye movements from multi-view images. We address two key challenges in eye-aware face NeRF learning: *how to effectively capture eyeball rotation for training* and *how to construct a manifold for representing eyeball rotation*. To accomplish this, we first fit FLAME, a well-established parametric face model, to the multi-view images considering multi-view consistency. Subsequently, we introduce a new Dynamic Eye-aware NeRF (DeNeRF). DeNeRF transforms 3D points from different views into a canonical space to learn a unified face NeRF model. We design an eye deformation field for the transformation, including rigid transformation, e.g., eyeball rotation, and non-rigid transformation. Through experiments conducted on the ETH-XGaze dataset, we demonstrate that our model is capable of generating high-fidelity images with accurate eyeball rotation and non-rigid periocular deformation, even under novel viewing angles. Furthermore, we show that utilizing the rendered images can effectively enhance gaze estimation performance.

1 Introduction

Face rendering is an important task in computer vision and computer graphics. It is widely demanded by applications such as virtual reality [17, 24, 26, 52, 53], digital human [6, 19, 22, 35] and CG film-making [2, 36, 44]. Conventional methods fit a parametric face model based on a face template mesh [8, 23]. Generative adversarial networks (GAN) directly render photo-realistic images with deep neural networks [1, 13, 14]. Recent research has incorporated the Neural Radiance Fields (NeRF) [25] for face rendering. NeRF models encode 3D geometry and exhibit great multi-view consistency that enables face rendering

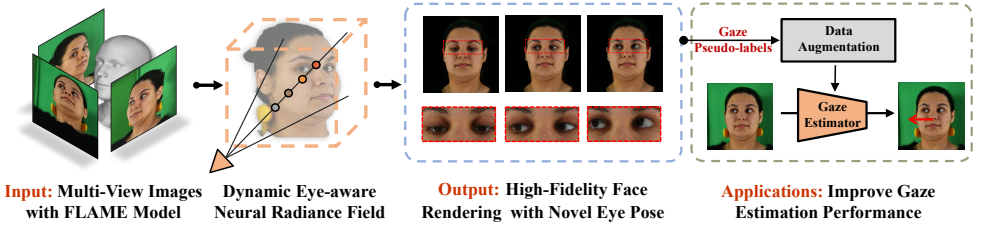

Figure 1: The Dynamic Eye-aware Neural Radiance Field (DeNeRF) is designed to render high-fidelity faces with animatable eyes using a set of multi-view images. It enables face rending under novel view and eye pose. By leveraging DeNeRF, we are able to obtain pseudo gaze labels from the model, which can be utilized to enhance gaze estimation methods.

under novel viewpoints. Additionally, various previous works explore parametric face NeRF models [18], which enable the control of facial attributes in rendered images such as identity and expression.

While most face NeRF models focus on facial semantic attributes, they tend to overlook the importance of controllable eye movement in face rendering. Eye movement can enhance image realism and is critical for multiple downstream tasks. However, the eyeball is located inside the face and incompletely visible in face images. It is difficult to determine the precise 3D eyeball rotation and position based on only face images. Furthermore, the eyeball rotation is related to geometric architecture, which means the model should be rotation-aware. Eyeball rotation also leads to non-rigid deformation such as the periocular deformation. Previous methods perform the rotation in feature space with gaze directions [32, 54]. However, it is also non-trivial to obtain accurate gaze directions from images and such approaches often result in low-quality rendered images.

In this paper, we propose a novel approach called Dynamic Eye-aware NeRF (DeNeRF). DeNeRF learns a dynamic face NeRF model from multi-view images, enabling face rendering with unseen eyeball and head poses. We address two challenges in eye-aware NeRF learning: *effectively capturing eyeball rotation* and *constructing a suitable manifold for representing such rotation*. To capture eyeball rotation, we begin by fitting a well-established parametric face model, FLAME [23], to the multi-view face images. FLAME is originally designed for face tracking from a single face image, and we modify the fitting process to account for consistency across multi-view images. This allows us to obtain parameters such as eyeball and head pose, which we then use for DeNeRF learning.

We further define a unified canonical space in DeNeRF. Given a pixel in the observed images, we sample 3D positions based on view directions in the observation space. We transform the 3D positions from the observation space into the canonical space with given poses. We design an eye deformation field for the transformation, including rigid transformation (*e.g.*, eyeball rotation and head rotation) and non-rigid transformation. We input the 3D positions in the canonical space into a NeRF model and render the pixel for alignment. To reduce the computational costs, we adopt a patch-based sampling approach [33]. We sample patches in images for alignment in each iteration. To enforce realistic eye region, we generate an eye mask for each image and enlarge the sampling ratio in the eye region.

Overall, our contributions are three-fold:

- We propose DeNeRF which learns a dynamic face NeRF model from multi-view images. To capture the eyeball pose accurately, we design a new fitting process for the

FLAME model, ensuring consistency across multiple views.
- We define a unified canonical space to construct a rotation-aware manifold. We transform 3D positions in the observation space into a canonical space based on an eye deformation field, including both rigid and non-rigid transformations. The DeNeRF is learned in the canonical space.
- DeNeRF enables high-fidelity face rendering under novel eyeball poses and head poses. Extensive experiments prove the rendered images can effectively enhance the performance of the downstream gaze estimation task.

2 Related Work

Neural Radiance Field. NeRF [25] proposed to learn implicit neural representations of a static scene from multi-view images, showing high-quality novel view synthesis. It models the continuous radiance field of a static scene by utilizing a mapping function that takes both a 3D spatial point $\mathbf{x}$ and view direction $\mathbf{d}$ as input, and outputs the corresponding RGB color $\mathbf{c}$ and volume density σ values. A standard NeRF is parameterized with a Multi-Layer Perceptron (MLP) as

$$H_\theta : (\gamma(\mathbf{x}), \gamma(\mathbf{d})) \rightarrow (\mathbf{c}, \sigma), \tag{1}$$

where θ represents the parameters of the network and γ refers to a positional encoding function [25, 37] that transforms $\mathbf{x}$ and $\mathbf{d}$ into a high-dimensional space. Different from conventional generative models, NeRF is a 3D-aware model and represents 3D object/scene via implicit neural representations. Novel views are generated from the implicit neural representation with volume rendering. Further explorations [10, 17, 18, 27, 29, 30] adapt NeRF to represent dynamic scenes. Hong *et al.* [18] acquire the latent codes of disentangled facial attributes from 3D morphable model. They control the code to render images with different poses and identities. Some methods also use NeRF for talking face generation [7, 15, 45]. They usually overlook the controllable eye movement in face rendering.

Eye Image Synthesis. The precise manipulation of realistic eye imagery has proven essential across multiple domains of application. Qin *et al.* [31] reconstruct 3DMM face model from multi-view images. They can rotate a virtual camera to synthesize images in different camera poses but cannot rotate the eyeball. Wood *et al.* [42] build a virtual 3D morphable eye model with computer graphic algorithms. They can synthesize eye images with arbitrary gaze directions[41, 42, 43]. However, their synthetic images are usually not realistic enough. Recently, generative model shows great potential in image generation. Compared to 3D morphable model, generative model can generate more realistic images. Some methods use generative model to generate realistic eye images [16, 28, 34, 47]. He *et al.* [16] utilize GAN for gaze redirection task. They can synthesize large-scale gaze data by performing gaze redirection task on one eye image. Ruzzi *et al.* [32] uses NeRF for gaze redirection tasks. They train a NeRF model based on given gaze directions and perform rotation in feature space, which results in low-quality rendered images.

3 Methodology

3.1 Multi-view Face Tracking

We first perform multi-view face tracking to acquire eyeball poses for training. We use the well-known parametric face model FLAME [23] for face tracking. FLAME model fits facial

parameters (such as pose and expression) from a single face image. We modify the fitting process for the multi-view images by projecting the fitted face model into multiple views and adding consistency loss for all views.

In practice, we input multi-view images and corresponding camera poses to the face tracker. We detect face landmarks, pupil centers, and face masks in each image [3, 46], where face masks are used to remove the background in the images. We initialize the textured FLAME face model with zero parameters and project the face model to all views. We render face images and perform pixel-wise alignment in the face appearance. Note that, face masks are used to ensure the alignment is performed in the face region. We can also obtain face landmarks and pupil centers from the FLAME model. We align them with the detected results for structure consistency. We use L_1 loss for all alignments where we denote them as face appearance loss $\mathcal{L}_{appear}$, facial landmark loss $\mathcal{L}_{face}$ and pupil center loss $\mathcal{L}_{pupil}$. Overall, we optimize the multi-view face tracker by minimizing

$$\mathcal{L}_{tracking} = \left(\sum_{i=1}^{N} \left(\alpha \mathcal{L}^i_{pupil} + \beta \mathcal{L}^i_{face} + \gamma \mathcal{L}^i_{appear} \right) \right) / N, \tag{2}$$

where α, β, γ are hyper-parameters, and we empirically set them as 10, 5, and 30. N is the number of views which is 13 in our experiment. We show the details of the loss function and fitting result in the supplementary material.

After fitting the FLAME model, we obtain facial parameters as the outcome. For DeNeRF Learning, we specifically choose four poses, namely two for the eyeballs, one for the jaw, and one for the neck. These poses are denoted as $p_i = \{R_i, t_i\}$, where i ranges from 1 to 4. Collectively, we refer to all four poses as p.

3.2 Dynamic Eye-aware Neural Radiance Field

The input of DeNeRF contains multi-view images and poses p as well as camera poses. Conventional NeRF-based methods implicitly learn a static geometric model. Although they can render images under novel views, they cannot control the content in the 3D model.

Our key idea is to learn a unified face NeRF model in the canonical space. Given an eyeball model and its pose, it is easy to rotate the eyeball from the observation space into the canonical space, and we can learn a unified eyeball model in the canonical space. We perform the similar operation in DeNeRF. We newly design an *eye deformation field* on both rigid and non-rigid transformations. *The eye deformation field* allows us to transform a point from the observation space into the canonical space. We obtain the color and density of the point in the canonical space based on Eq. (1) and learn a unified NeRF model in the canonical space.

Eye Deformation Field. We aim to learn a deformation field which transforms a point $\mathbf{x}_o$ in the observation space into the canonical space. Intuitively, we would first rotate the 3D point based on the head pose, and then rotate it based on the eyeball pose if this point is in the eye region. Unfortunately, this approach is not feasible due to the absence of an explicit mesh model. To address this issue, we propose a solution that involves applying various transformations, including head rotation and eyeball rotation, and combining them using learnable weights [21]:

$$T_{\text{rigid}}(\mathbf{x}_o, p) = \sum_{i=1}^{4} w^i_o(\mathbf{x}_o) \left(R_i \mathbf{x}_o + t_i \right). \tag{3}$$

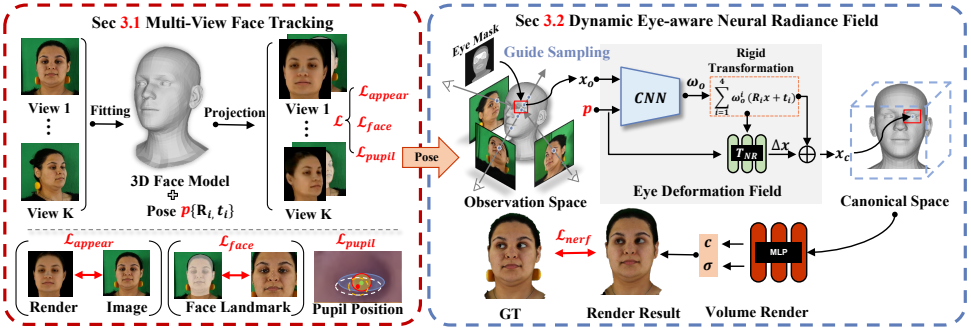

Figure 2: Overview of DeNeRF model which specializes in creating high-fidelity face images that are aware of the eyes. It achieves this by training on multi-view images, and leveraging parametric model-based face tracking to establish multi-view consistency. From this data, DeNeRF fits a face model that includes poses for the eyeballs, jaw, and neck. To facilitate the learning process, DeNeRF defines a canonical space and learns a unified model within it. Additionally, it introduces an eye deformation field that can transform points from observation space into the canonical space. This field is composed of both rigid and non-rigid transformation modules, enabling it to handle the complexity of eye movement and deformation.

Note that we first convert all poses R_i, t_i into a common coordinate system to facilitate the summation of transformed results. Instead of directly estimating w_o^i, we learn w_c^i as described in [40], from which w_o^i can be computed:

$$w_o^i(\mathbf{x}_o) = \frac{w_c^i(R_i\mathbf{x}_o + t_i)}{\sum_{k=1}^{4} w_c^k(R_k\mathbf{x}_o + t_k)}. \tag{4}$$

This process stabilizes the model and decreases the probability of collapse.

T_{rigid} exclusively deals with rigid transformations. However, facial appearance involves non-rigid transformations, such as periocular deformation. To address this, we introduce a non-rigid transformation field, denoted as T_{NR}, which is implemented as an MLP. This MLP generates an offset Δx to augment the output of T_{rigid}. Additionally, we incorporate pose information p as supplementary inputs to the MLP. Consequently, the complete eye deformation field can be expressed as follows:

$$\mathbf{x}_c = \hat{\mathbf{x}} + T_{NR}(\hat{\mathbf{x}}, p), \tag{5}$$

where $\hat{\mathbf{x}} = T_{rigid}(\mathbf{x}_o, p)$ and $\mathbf{x}_c$ represents the position in the canonical space. Finally, we input the position $\mathbf{x}_c$ into a vanilla NeRF to obtain the color and density as Eq. (1).

3.3 Training

We train DeNeRF to render 512×512 pixels images. DeNeRF is trained with a photometric reconstruction loss in an end-to-end manner. To handle the high computational costs, we adopt a patch-based sampling approach [33]. We randomly sample six patches with size 32×32 pixels from input images and 128 points for each ray. We use LPIPS loss [48] ($\mathcal{L}_{\text{LPIPS}}$) and MSE loss ($\mathcal{L}_{\text{MSE}}$) for training as

$$\mathcal{L}_{\text{DeNeRF}} = \mathcal{L}_{\text{LPIPS}}(\mathbf{P}, \hat{\mathbf{P}}) + \lambda \mathcal{L}_{\text{MSE}}(\mathbf{P}, \hat{\mathbf{P}}), \tag{6}$$

where $\mathbf{P}$ is the image patch and $\hat{\mathbf{P}}$ is the ground truth. We set $\lambda = 0.2$ in the training.

Eye-mask Guided Sampling. To address the issue of the eye region being too small compared to the rest of the face, we introduce an eye mask [46] to guide the ray sampling in DeNeRF. Specifically, we assign a larger sampling ratio to the eye region than to other facial parts, thus directing more attention to the eye region.

For training, we utilize the Adam optimizer with a learning rate of 5×10^{-4} and exponential learning decay. The sample ratio is set to 0.8, while the eye region sample ratio is set to 0.5 during training. Each subject is trained for 400K epochs using four A100 GPUs.

4 Experiments

4.1 Setup

Data Preparation. We select ETH-XGaze [51] for experiments since the dataset contains rich eye movement. ETH-XGaze is a large-scale gaze estimation dataset collected from 110 subjects with 18 cameras. The dataset provides over one million high-resolution images. We crop the face patch from raw images and resize it into 512×512 for training. Our model is trained based on a sequence of multi-view images. We select 9 frames which roughly cover the gaze in nine different directions for each subject. We remove the view under extreme pose and each frame finally provides 13-view images, *i.e.*, we use 9×13 images for training on one subject.

Evaluation Metrics. We conduct qualitative and quantitative comparison to demonstrate image quality, where SSIM [39], PSNR, and LPIPS [48] are reported for quantitative comparison. We also show the advantage of our method in a downstream task. We render images for data augmentation in gaze estimation task. We use angular degree error for gaze estimation metric [5, 38].

4.2 Face Rendering under Novel Pose and Gaze

We present the qualitative results for image generation under novel gazes and head poses in Fig. 3. The image on the left showcases the rendered images under novel head poses. Our DeNeRF model excels in generating high-fidelity face images while maintaining multi-view consistency. The skin and eye textures are clearly visible with vivid details, and the hair is accurately reconstructed. Surprisingly, the last two rows in the image demonstrate our model's remarkable generative ability in reconstructing subjects wearing glasses. It shows that our DeNeRF can effectively organize the multi-view information from sparse views for high-fidelity 3D face reconstruction. The image on the right showcases the eye animation generated under novel gaze directions. The animation displays a natural and continuous eye movement despite being trained only on nine sparse gaze directions. It demonstrates that our model has successfully learned to accurately represent the rotation of the eyeball. This is attributed to the precise eyeball pose achieved through multi-view face tracking using FLAME, as well as the deformation strategy based on canonical space. They enable us to integrate all information from multi-view images. Overall, the results clearly demonstrate the effectiveness of our model in both face reconstruction and eye animation.

4.3 Comparison with Face Rendering Methods

We conduct a comprehensive comparison of our method with the SOTA methods in 2D and 3D face rendering with eye animation (STED [54] and GazeNeRF [32]) and a NeRF-based

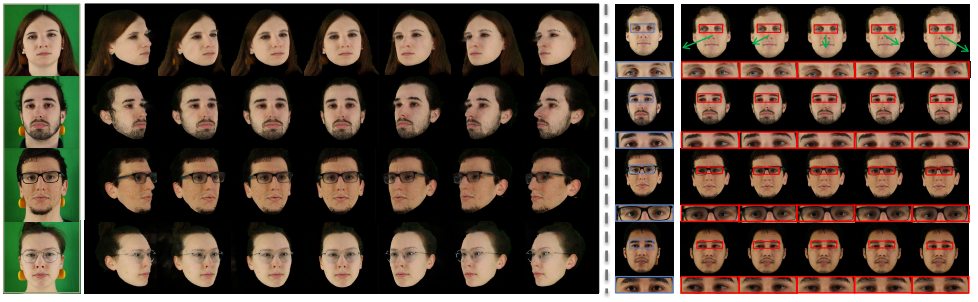

Figure 3: Face rendering under novel gaze and novel head poses. We train DeNeRF with multi-view images from only nine frames. The left images show the rendering under novel head pose. DeNeRF preserves great multi-view consistency. We also use the DeNeRF to render images under novel gaze in the right image. DeNeRF is a parametric NeRF model. We can directly change the eyeball pose to generate images under novel gaze. DeNeRF generates high-fidelity face images in a large range of head pose and gaze direction. This is the key advantage of DeNeRF.

face rendering approach called HeadNeRF [13], which can be adapted for eye animation. STED proposes an encoder-decoder structure to automate the disentanglement of gaze direction and head pose. GazeNeRF is a 3D-aware gaze redirection model that takes multi-view images and gaze labels as input and allows control over eye deformation via gaze input. HeadNeRF is a NeRF-based method for high-fidelity facial imaging and can be extended to include gaze direction as additional input for eye animation. Our comparison primarily focuses on the quality of rendered images, particularly in the eye region.

Fig. 4 shows the qualitative comparison with the SOTA methods. All of the methods are capable of face reconstruction. However, the faces generated by STED are blurry. Although the results from HeadNeRF and GazeNeRF are better than STED, they still struggle with achieving a natural skin and hair look. In contrast, our model not only produces a photo-realistic eye region but also accurately reconstructs other facial features and hair. In the region surrounding the eye, our method produces sharper images than other methods and the finer details of the periocular region, such as eyebrows and eyelids are clearly visible.

Methods	SSIM ↑	PSNR ↑	LPIPS ↓
STED [34]	0.726	17.279	0.306
HeadNeRF [13]	0.718	15.262	0.300
GazeNeRF [32]	0.728	15.322	0.297
Ours	**0.732**	**19.144**	**0.265**

Table 1: We show the quantitative comparison between DeNeRF and other SOTA methods in terms of rendered image quality. DeNeRF shows significant improvement in all metrics.

In addition, we show the quantitative results in Table 1. It shows that our method outperforms GazeNeRF by a large margin which supports the qualitative results above. GazeNeRF learns to map the latent code space onto various 3D faces. In contrast, DeNeRF defines a canonical space and warps each point in the observation space to a static canonical space where the deformation is handled by explicit face poses and learned linear blend skinning weights. Such a design provides the neural network with a clear objective of learning the static canonical space, which makes training much easier.

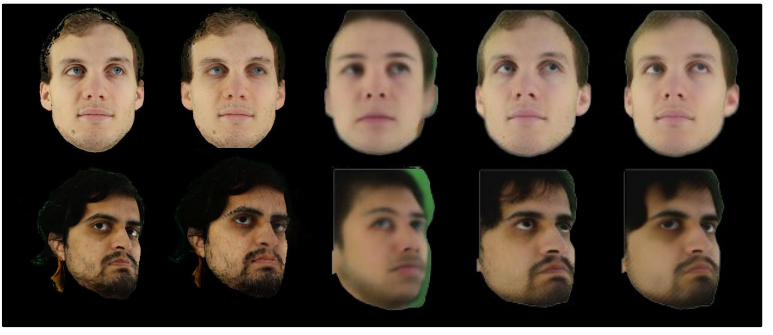

Figure 4: We show the comparison with face rendering methods. We report the result of compared methods from the SOTA face rendering method [32]. STED and GazeNeRF are designed for gaze redirection. They perform rotation in the feature space which degrades the quality of rendering images. HeadNeRF is adapted for eye animation using gaze direction as additional inputs. It is obvious that all compared methods have artifacts in the rendering images, while our method renders high-fidelity face images. This demonstrates the advantage of DeNeRF.

It is worth noting that our model is trained on only 13 views from 9 frames, whereas GazeNeRF is first pre-trained on all available 80 subjects in ETH-XGaze before being fine-tuned using all 18 available views from 100 frames. Despite the vast difference in the amount of training data, our model significantly outperforms GazeNeRF in rendered image quality.

4.4 Improving Gaze Estimation Performance

Our method is capable of rendering face images from novel views and eyeball poses. Furthermore, we can generate pseudo-labels for the rendered images based on eyeball rotation and head pose. The eyeball rotation is estimated through multi-view face tracking, while the head pose can be derived from camera pose. By combining the eyeball rotation and head pose, we can accurately calculate the final gaze direction.

To demonstrate the capacity of our model in enhancing gaze estimation, we conducted experiments on four renowned gaze datasets including EyeDiap [11], MPIIFace [49], RT-Gene [9], and Gaze360 [20]. Due to the scarcity of annotated images per subject in the test person specific set of ETH-XGaze, we randomly select seven subjects from the train set of ETH-XGaze for the experiment. We train our model on nine frames for each subject and then generate 324 images under random 36 head poses and nine gaze directions. These newly generated data were incorporated into the four gaze datasets to create their corresponding augmented version. Finally, we evaluated the performance of the augmented versions on the annotated data of each

	EyeDiap [11]	MPIIFace [50]	RT-Gene [9]	Gaze360 [20]
×	41.829	36.441	42.172	17.709
√	**34.074**	**28.567**	**34.632**	**16.351**

Table 2: We use rendering images to enhance gaze estimation performance. We train GazeTR in the four datasets and test it in the ETH-XGaze. The checkmark means we add rendering images into training set.

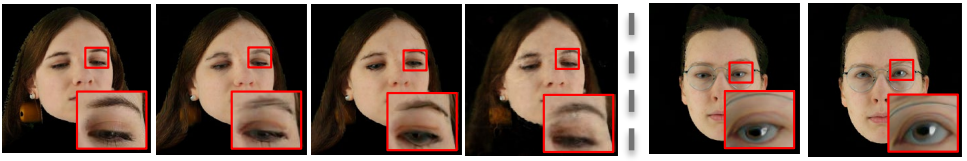

Figure 5: Ablation study on eye mask guided sampling (ES) and LPIPS loss. The result shows both ES and LPIPS loss improve the image quality in eye region. We design the non-rigid transformation part in the eye deformation field. The right figure demonstrates the effectiveness of the non-rigid transformation. It is obvious that our model accurately capture the periocular movement. This proves the advantage of the eye deformation field.

subject. We trained the state-of-the-art gaze estimator, GazeTR [4], separately on each of the four original datasets and their augmented versions for comparison.

We present the gaze error in Table 2. The rendering images bring significant improvements on the EyeDiap, MPIIFace, and RT-Gene datasets, with gains of 18.54%, 21.61%, and 17.88%, respectively. This improvement can be attributed to the ability of DeNeRF to generate a wider range of gaze and head poses than those present in the narrow dataset. This increased variation allows for more accurate gaze and head pose estimations. Despite Gaze360 already having a large gaze and head pose range, our augmented dataset still outperforms it with a 7.67% improvement. This result further highlights the potential of our method for the gaze estimation task.

4.5 Ablation Studies

We conduct ablation studies on eye mask guided sampling and LPIPS loss, as seen in Fig. 5. Our model without eye mask guided sampling generates an unnatural eyeball that is nearly completely black. The iris, which is the most crucial semantic information for gaze, is difficult to identify. In contrast, our full model produces a photo-realistic eyeball with a clear boundary between the iris and sclera. The effectiveness of our eye mask guided sampling can be attributed to its ability to address the deformation issue of small regions. Additionally, when our model is not trained with LPIPS loss, the rendered image appears blurry not only in the eye region but also in other face parts. In comparison, our full model produces sharper details in the rendered image, emphasizing the importance of the LPIPS loss for the image quality of our model. On the right image, we display the periocular movement as the subject look from bottom to top. It demonstrates that our model is capable of handling non-rigid periocular deformation.

5 Conclusion

In this paper, we present a novel dynamic eye-aware NeRF that allows for facial rendering from different perspectives and eye poses. DeNeRF utilizes multi-view images to train a NeRF model of the face. First, we perform face tracking on the multi-view images to capture the eye pose. Then, we fit a parametric 3D face model, FLAME, considering the multi-view consistency. Next, we construct a rotation-aware manifold to model the rotation of the eyeball. We define a canonical space for DeNeRF and transform 3D points from different

observation spaces into this space. Finally, we learn a unified face NeRF model on the canonical space while considering an eye deformation field for the transformation. The eye deformation field accounts for rigid transformation, including eyeball rotation, and non-rigid transformation, such as periocular deformation. We evaluate our method on the ETH-XGaze dataset and show that it can render high-fidelity face images from novel viewpoints and eye poses. We also mix our rendered images with the original training set for data augmentation, which further improves performance. In future work, we aim to reduce the requirement for multi-view images and lower computational costs.

Acknowledgment

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.2022-0-00608, Artificial intelligence research about multi-modal interactions for empathetic conversations with humans). The research utilized the Baskerville Tier 2 HPC service (https://www.baskerville.ac.uk/) funded by the Engineering and Physical Sciences Research Council (EPSRC) and UKRI through the World Class Labs scheme (EP/T022221/1) and the Digital Research Infrastructure programme (EP/W032244/1) operated by Advanced Research Computing at the University of Birmingham. Hengfei Wang and Zhongqun Zhang were supported by China Scholarship Council Grant No.202006210057 and No.202208060266, respectively.

References

[1] Grigory Antipov, Moez Baccouche, and Jean-Luc Dugelay. Face aging with conditional generative adversarial networks. In *2017 IEEE international conference on image processing (ICIP)*, pages 2089–2093. IEEE, 2017.

[2] Volker Blanz, Kristina Scherbaum, Thomas Vetter, and Hans-Peter Seidel. Exchanging faces in images. In *Computer Graphics Forum*, volume 23, pages 669–676. Wiley Online Library, 2004.

[3] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In *International Conference on Computer Vision*, 2017.

[4] Yihua Cheng and Feng Lu. Gaze estimation using transformer. *ICPR*, 2022.

[5] Yihua Cheng, Haofei Wang, Yiwei Bao, and Feng Lu. Appearance-based gaze estimation with deep learning: A review and benchmark. *arXiv preprint arXiv:2104.12668*, 2021.

[6] H Onan Demirel and Vincent G Duffy. Applications of digital human modeling in industry. In *Digital Human Modeling: First International Conference on Digital Human Modeling, ICDHM 2007, Held as Part of HCI International 2007, Beijing, China, July 22-27, 2007. Proceedings 1*, pages 824–832. Springer, 2007.

[7] Chenpng Du, Qi Chen, Tianyu He, Xu Tan, Xie Chen, Kai Yu, Sheng Zhao, and Jiang Bian. Dae-talker: High fidelity speech-driven talking face generation with diffusion autoencoder. *arXiv preprint arXiv:2303.17550*, 2023.

[8] Yao Feng, Haiwen Feng, Michael J Black, and Timo Bolkart. Learning an animatable detailed 3d face model from in-the-wild images. *ACM Transactions on Graphics (ToG)*, 40(4):1–13, 2021.

[9] Tobias Fischer, Hyung Jin Chang, and Yiannis Demiris. Rt-gene: Real-time eye gaze estimation in natural environments. In *The European Conference on Computer Vision*, 2018.

[10] Jonathan Freer, Kwang Moo Yi, Wei Jiang, Jongwon Choi, and Hyung Jin Chang. Novel-view synthesis of human tourist photos. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3069–3076, 2022.

[11] Kenneth Alberto Funes Mora, Florent Monay, and Jean-Marc Odobez. Eyediap: A database for the development and evaluation of gaze estimation algorithms from rgb and rgb-d cameras. In *Proceedings of the ACM Symposium on Eye Tracking Research & Applications*, 2014. doi: 10.1145/2578153.2578190.

[12] Guy Gafni, Justus Thies, Michael Zollhofer, and Matthias Nießner. Dynamic neural radiance fields for monocular 4d facial avatar reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8649–8658, 2021.

[13] Jon Gauthier. Conditional generative adversarial nets for convolutional face generation. *Class project for Stanford CS231N: convolutional neural networks for visual recognition, Winter semester*, 2014(5):2, 2014.

[14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[15] Yudong Guo, Keyu Chen, Sen Liang, Yong-Jin Liu, Hujun Bao, and Juyong Zhang. Ad-nerf: Audio driven neural radiance fields for talking head synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5784–5794, 2021.

[16] Zhe He, Adrian Spurr, Xucong Zhang, and Otmar Hilliges. Photo-realistic monocular gaze redirection using generative adversarial networks. In *The IEEE International Conference on Computer Vision*, 2019.

[17] Anders Henrysson, Mark Billinghurst, and Mark Ollila. Face to face collaborative ar on mobile phones. In *Fourth ieee and acm international symposium on mixed and augmented reality (ismar'05)*, pages 80–89. IEEE, 2005.

[18] Yang Hong, Bo Peng, Haiyao Xiao, Ligang Liu, and Juyong Zhang. Headnerf: A real-time nerf-based parametric head model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20374–20384, 2022.

[19] Rachael E Jack and Philippe G Schyns. The human face as a dynamic tool for social communication. *Current Biology*, 25(14):R621–R634, 2015.

[20] Petr Kellnhofer, Adria Recasens, Simon Stent, Wojciech Matusik, and Antonio Torralba. Gaze360: Physically unconstrained gaze estimation in the wild. In *The IEEE International Conference on Computer Vision*, 2019.

[21] John P Lewis, Matt Cordner, and Nickson Fong. Pose space deformation: a unified approach to shape interpolation and skeleton-driven deformation. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 165–172, 2000.

[22] Tommer Leyvand, Daniel Cohen-Or, Gideon Dror, and Dani Lischinski. Digital face beautification. In *ACM Siggraph 2006 Sketches*, pages 169–es. 2006.

[23] Tianye Li, Timo Bolkart, Michael J Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4d scans. *ACM Trans. Graph.*, 36(6):194–1, 2017.

[24] Katerina Mania, Ann McNamara, and Andreas Polychronakis. Gaze-aware displays and interaction. In *ACM SIGGRAPH 2021 Courses*, pages 1–67, 2021.

[25] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.

[26] Yun Suen Pai, Benjamin Tag, Benjamin Outram, Noriyasu Vontin, Kazunori Sugiura, and Kai Kunze. Gazesim: simulating foveated rendering using depth in eye gaze for vr. In *ACM SIGGRAPH 2016 Posters*, pages 1–2, 2016.

[27] Keunhong Park, Utkarsh Sinha, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Steven M Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5865–5874, 2021.

[28] Seonwook Park, Shalini De Mello, Pavlo Molchanov, Umar Iqbal, Otmar Hilliges, and Jan Kautz. Few-shot adaptive gaze estimation. In *The IEEE International Conference on Computer Vision*, 2019.

[29] Sida Peng, Junting Dong, Qianqian Wang, Shangzhan Zhang, Qing Shuai, Xiaowei Zhou, and Hujun Bao. Animatable neural radiance fields for modeling dynamic human bodies. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14314–14323, 2021.

[30] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10318–10327, 2021.

[31] Jiawei Qin, Takuru Shimoyama, and Yusuke Sugano. Learning-by-novel-view-synthesis for full-face appearance-based 3d gaze estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4981–4991, 2022.

[32] Alessandro Ruzzi, Xiangwei Shi, Xi Wang, Gengyan Li, Shalini De Mello, Hyung Jin Chang, Xucong Zhang, and Otmar Hilliges. Gazenerf: 3d-aware gaze redirection with neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

[33] Katja Schwarz, Yiyi Liao, Michael Niemeyer, and Andreas Geiger. Graf: Generative radiance fields for 3d-aware image synthesis. *Advances in Neural Information Processing Systems*, 33:20154–20166, 2020.

[34] Ashish Shrivastava, Tomas Pfister, Oncel Tuzel, Joshua Susskind, Wenda Wang, and Russell Webb. Learning from simulated and unsupervised images through adversarial training. In *The IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[35] Tze Ho Elden Tse, Zhongqun Zhang, Kwang In Kim, Ales Leonardis, Feng Zheng, and Hyung Jin Chang. S 2 contact: Graph-based network for 3d hand-object contact estimation with semi-supervised learning. In *European Conference on Computer Vision*, pages 568–584. Springer, 2022.

[36] P Vanezis, RW Blowes, AD Linney, AC Tan, R Richards, and R Neave. Application of 3-d computer graphics for facial reconstruction and comparison with sculpting techniques. *Forensic science international*, 42(1-2):69–84, 1989.

[37] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.

[38] Hengfei Wang, Jun O Oh, Hyung Jin Chang, Jin Hee Na, Minwoo Tae, Zhongqun Zhang, and Sang-Il Choi. Gazecaps: Gaze estimation with self-attention-routed capsules. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2668–2676, 2023.

[39] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[40] Chung-Yi Weng, Brian Curless, Pratul P Srinivasan, Jonathan T Barron, and Ira Kemelmacher-Shlizerman. Humannerf: Free-viewpoint rendering of moving people from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16210–16220, 2022.

[41] Erroll Wood, Tadas Baltrusaitis, Xucong Zhang, Yusuke Sugano, Peter Robinson, and Andreas Bulling. Rendering of eyes for eye-shape registration and gaze estimation. In *The IEEE International Conference on Computer Vision*, 2015.

[42] Erroll Wood, Tadas Baltrušaitis, Louis-Philippe Morency, Peter Robinson, and Andreas Bulling. Learning an appearance-based gaze estimator from one million synthesised images. In *Proceedings of the ACM Symposium on Eye Tracking Research & Applications*, page 131–138, 2016. ISBN 9781450341257. doi: 10.1145/2857491.2857492.

[43] Erroll Wood, Tadas Baltrušaitis, Louis Philippe Morency, Peter Robinson, and Andreas Bulling. A 3d morphable eye region model for gaze estimation. In *The European Conference on Computer Vision*, 2016.

[44] Guo-Wei Yang, Wen-Yang Zhou, Hao-Yang Peng, Dun Liang, Tai-Jiang Mu, and Shi-Min Hu. Recursive-nerf: An efficient and dynamically growing nerf. *IEEE Transactions on Visualization and Computer Graphics*, 2022.

[45] Zhenhui Ye, Ziyue Jiang, Yi Ren, Jinglin Liu, JinZheng He, and Zhou Zhao. Geneface: Generalized and high-fidelity audio-driven 3d talking face synthesis. 2023.

[46] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 325–341, 2018.

[47] Yu Yu, Gang Liu, and Jean-Marc Odobez. Improving few-shot user-specific gaze adaptation via gaze redirection synthesis. In *The IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

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

[49] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. It's written all over your face: Full-face appearance-based gaze estimation. In *The IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 2299–2308, 2017.

[50] 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, Jan 2019. ISSN 1939-3539. doi: 10.1109/TPAMI.2017.2778103.

[51] Xucong Zhang, Seonwook Park, Thabo Beeler, Derek Bradley, Siyu Tang, and Otmar Hilliges. Eth-xgaze: A large scale dataset for gaze estimation under extreme head pose and gaze variation. In *The European Conference on Computer Vision*, 2020.

[52] Zhongqun Zhang, Wei Chen, Linfang Zheng, Aleš Leonardis, and Hyung Jin Chang. Trans6d: Transformer-based 6d object pose estimation and refinement. In *European Conference on Computer Vision*, pages 112–128. Springer, 2022.

[53] Linfang Zheng, Chen Wang, Yinghan Sun, Esha Dasgupta, Hua Chen, Aleš Leonardis, Wei Zhang, and Hyung Jin Chang. Hs-pose: Hybrid scope feature extraction for category-level object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17163–17173, 2023.

[54] Yufeng Zheng, Seonwook Park, Xucong Zhang, Shalini De Mello, and Otmar Hilliges. Self-learning transformations for improving gaze and head redirection. *Advances in Neural Information Processing Systems*, 2020.