

GAZE-CONTROLLABLE FACE GENERATION

By

HENGFEI WANG

A thesis submitted to
the University of Birmingham
for the degree of
DOCTOR OF PHILOSOPHY



Human-Centred Visual Learning Group
School of Computer Science
College of Engineering and Physical Sciences
University of Birmingham
December 2024

UNIVERSITY OF
BIRMINGHAM

University of Birmingham Research Archive

e-theses repository

This unpublished thesis/dissertation is copyright of the author and/or third parties. The intellectual property rights of the author or third parties in respect of this work are as defined by The Copyright Designs and Patents Act 1988 or as modified by any successor legislation.

Any use made of information contained in this thesis/dissertation must be in accordance with that legislation and must be properly acknowledged. Further distribution or reproduction in any format is prohibited without the permission of the copyright holder.

ABSTRACT

Gaze-controllable face generation systems produce facial images with adjustable gazes, enabling a wide range of applications such as virtual reality, digital humans, and computer-generated filmmaking. Traditional methods achieve gaze control through image replacement or graphical modeling techniques. However, to make gaze-controllable face generation more accessible and practical, researchers are increasingly focusing on leveraging deep neural networks, particularly generative models, in both 2D and 3D frameworks. My work addresses three outstanding problems regarding quality, speed, and usability in this field: 1) It is hard to model both rigid eyeball rotation and non-rigid eyelid deformation in a unified 3D representation; 2) The inference of 3D gaze-controllable generation models is still far from being real-time; 3) Using explicit gaze labels to control gaze is inconvenient for general use.

In this dissertation, I propose three approaches to tackle these problems separately: modeling a hybrid transformation from observation space to canonical space with one unified 3D model for high-fidelity generation, distilling face shape prior into an efficient 3D representation for real-time generation, controlling gaze image generation with the guidance of natural language for convenient use.

I present three key contributions in this dissertation. First, I introduce an eye-controllable face NeRF model handling rigid eyeball rotation and non-rigid eyelid deformation. The model can generate high-fidelity images under novel head poses and gaze directions. Second, I propose a real-time 3D-aware gaze redirection module generating multi-view gaze images from

a single face image. It takes 61ms to process the input and generate results, outperforming other methods by a large margin. Third, I present the first text-to-gaze dataset and a novel framework enabling language-guided gaze-controllable face generation. It provides a more natural way to control gaze image generation.

DEDICATION

Dedicated to the person I was, the person I am, and the person I aspire to be.

ACKNOWLEDGMENTS

I would like to express my deepest gratitude to my supervisor, Dr. Hyung Jin Chang, for his invaluable guidance and inspiration throughout my PhD journey. His mentorship has been instrumental in shaping my growth from a novice researcher to someone capable of tackling challenges independently. During this brief yet transformative doctoral journey, he has shared with me not only profound professional knowledge but also invaluable life lessons.

I am also deeply grateful to my collaborators, Yihua Cheng and Zhongqun Zhang. They have played pivotal roles during my PhD study. I am fortunate to have them around whenever challenges arise, and their suggestions have always been insightful and worth exploring.

I extend my heartfelt thanks to my labmates, who have brought energy and vibrancy to the lab. Their efforts in organizing various activities have made my time in the UK enjoyable and memorable.

Lastly, I owe my sincerest thanks to my family. Their unwavering support and encouragement have been the cornerstone of my ability to complete this PhD journey successfully.

PUBLICATIONS

This dissertation includes work published and under-reviewed in the following:

- Wang, H., Zhang, Z., Cheng, Y., & Chang, H. J. (2023, November). **High-Fidelity Eye Animatable Neural Radiance Fields for Human Face**. In 34th British Machine Vision Conference 2023 (BMVC) (Oral).
- Wang, H., Zhang, Z., Cheng, Y., & Chang, H. J. (2024, October). **Textgaze: Gaze-controllable face generation with natural language**. In Proceedings of the 32nd ACM International Conference on Multimedia (MM) (pp. 7143-7151).
- Wang, H., Zhang, Z., Cheng, Y., & Chang, H. J. (2024, November). **LiveGaze: Real-Time 3D-Aware Gaze Redirection from a Single Image**. Under review (CVPR2025).

Other publications from work not included in this dissertation:

- Wang, H., Oh, J. O., Chang, H. J., Na, J. H., Tae, M., Zhang, Z., & Choi, S. I. (2023). **Gazecaps: Gaze estimation with self-attention-routed capsules**. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 2669-2677).
- Zhang, Z., Wang, H., Yu, Z., Cheng, Y., Yao, A., & Chang, H. J. (2025). **NL2Contact: Natural Language Guided 3D Hand-Object Contact Modeling with Diffusion Model**. In European Conference on Computer Vision (ECCV) (pp. 284-300) (Oral). Springer.

Contents

	Page
1 Introduction	1
1.1 Motivation	1
1.2 Outstanding Problems	3
1.3 My Approaches	5
1.4 Contributions	7
1.5 Structure	7
2 Literature Review	9
2.1 Human Visual System	11
2.1.1 Eyeball Structure	11
2.1.2 Eye-Region Deformation	12
2.2 Gaze Estimation	14
2.2.1 Model-Based Approaches	15
2.2.2 Feature-Based Approaches	16
2.2.3 Appearance-Based Approaches	16
2.2.4 Challenges	19
2.3 Gaze-Controllable Face Generation	21
2.3.1 Traditional Approaches	22
2.3.2 2D Learning-Based Approaches	24
2.3.3 3D Learning-Based Approaches	25

2.3.4	Modern Generation Models	28
2.4	Datasets	30
2.4.1	Gaze Datasets	31
2.4.2	Dataset Selection	32
2.5	Conclusion	34
3	High-Fidelity Gaze-Controllable 3D Face Generation with Multi-View Images	35
3.1	Related Work	38
3.1.1	Neural Radiance Field	38
3.1.2	Eye Image Synthesis	38
3.2	Methodology	39
3.2.1	Multi-view Face Tracking	39
3.2.2	Dynamic Eye-aware Neural Radiance Field	41
3.2.3	Training	43
3.3	Experiments	44
3.3.1	Setup	44
3.3.2	Face Rendering under Novel Pose and Gaze	44
3.3.3	Comparison with Face Rendering Methods	45
3.3.4	Improving Gaze Estimation Performance	47
3.3.5	Multi-view Face Tracking Visualization	48
3.4	Conclusion	49
4	Real-Time Gaze-Controllable 3D Face Generation from a Single Image	51
4.1	Related Work	54
4.1.1	Gaze Redirection	54
4.1.2	3D-Aware One-Shot Portrait Generation	55
4.2	Methodology	56

4.2.1	Preliminary	56
4.2.2	Network Architecture	57
4.2.3	Distillation of 3D Face Prior	58
4.2.4	Training Objectives	60
4.3	Experiments	61
4.3.1	Experimental Setup	62
4.3.2	Efficient 3D-Aware Gaze Redirection	63
4.3.3	Face Rendering under Novel Views and Gazes	67
4.3.4	Ablation Study	67
4.3.5	Limitations	70
4.3.6	Additional Qualitative Results	70
4.4	Conclusion	73
5	Gaze-Controllable Face Generation with Nature Language	74
5.1	Related Work	77
5.1.1	Eye-Control Face Generation	77
5.1.2	Text-to-Image Generation	78
5.2	ToG: Text of Gaze Dataset	78
5.2.1	Overview	78
5.2.2	Text Generation using LLMs	79
5.2.3	ToG Dataset	81
5.2.4	Budget Efficiency of the Generated Data	84
5.3	Methodology	85
5.3.1	Text-to-Gaze Generation	85
5.3.2	Gaze-Controllable Face Generation	87
5.3.3	Implementation Details	88
5.4	Experiment	89

5.4.1	Datasets	89
5.4.2	Comparison on Text-Guided Generation	90
5.4.3	Ablation Study	93
5.4.4	Controlling Shape and Expression	95
5.4.5	Generation with Different Roll Angles	96
5.4.6	Identity Control	96
5.4.7	Comparison with Other SOTA Face Generation Models	97
5.4.8	Additional Qualitative Results	97
5.4.9	Discussion	99
5.5	Future Work	100
5.6	Conclusion	101
6	Discussion and Conclusion	102
6.1	Contributions	103
6.2	Limitations	104
6.3	Future Work	105
6.3.1	Compositional Animation of Eye Region	105
6.3.2	Generation without Finetuning	106
6.4	Final Remarks	107
	References	108

List of Figures

2.1	Eyeball structure	12
2.2	Extraocular muscles	13
2.3	Data collection setup and examples of ETH-XGaze	32
2.4	Examples of FFHQ dataset	34
3.1	Pipeline of DeNeRF and its application	36
3.2	Overview of DeNeRF model	41
3.3	Face rendering under novel gaze and novel head poses from DeNeRF	45
3.4	Qualitative comparison between DeNeRF and other SOTA methods	47
3.5	Fitting results of multi-view face tracking	49
4.1	3D-aware gaze redirection results from LiveGaze	52
4.2	Overview of the LiveGaze pipeline	57
4.3	Qualitative comparison between LiveGaze and other SOTA methods	64
4.4	Visualization of generated results under novel views and gazes	66
4.5	Visualization of ablation on fusion feature choices	68
4.6	Ablation on 3D face prior distillation	69
4.7	Failure cases under challenging conditions	69
4.8	Additional visualization of LiveGaze qualitative comparison	71
4.9	Additional visualization of generated images under novel gazes and novel views	72
5.1	Overview of our method and motivation	75
5.2	Visualization of head and gaze scores of different subjects	83

5.3	The pipeline of TextGaze	86
5.4	Qualitative comparison between TextGaze and other SOTA methods	91
5.5	Visualization of results from Label-Guided Model and Sketch-Guided Model	94
5.6	Interpolation of shape and expression	96
5.7	Generated results under different roll angles	96
5.8	Generated results with different identities	97
5.9	Additional qualitative results of TextGaze	98

List of Tables

3.1	Quantitative comparison between DeNeRF and other SOTA methods	46
3.2	Improving gaze estimation performance	48
4.1	Quantitative comparison between LiveGaze and other SOTA methods	65
4.2	Ablation on feature choices	67
5.1	Examples from ToG dataset	81
5.2	Text-image datasets	82
5.3	Examples of extent synonym in ToG	84
5.4	Comparison on costs	85
5.5	Quantitative comparison between TextGaze and other SOTA methods	93
5.6	Comparison of LGM and SGM in terms of head error and gaze error	94
5.7	Ablation on TAM module	95

Chapter One

Introduction

Gaze-controllable face generation is a task to control gaze direction during the process of generating a face image. The eyes are one of the most prominent features of the human face and play a crucial role in how we perceive the world. We rely heavily on vision to gather information in our daily lives, and people tend to trust what they see rather than what they hear. Moreover, gaze can reveal a great deal about personal traits, including desires, emotions, cognitive states, and interpersonal dynamics. For example, by observing the gaze of an audience, one can gauge their interest in a presentation—if they are engaged, their gaze is likely focused and attentive. Gaze-controllable face generation holds significant potential for applications in virtual reality (Zhang et al., 2022; Zheng et al., 2023), digital humans (Jack et al., 2015; Tse et al., 2022; Wang et al., 2023b), and CG (Computer Graphics) film-making (Yang et al., 2022; Blanz et al., 2004).

1.1 Motivation

To make gaze-controllable face generation widely applicable, it should meet requirements in three key aspects: quality, speed, and usability.

Quality. Image quality is always important to generation tasks (Gauthier, 2014; Goodfellow

et al., 2014). High-quality image generation is a prerequisite for the applicability of generation methods. Facial generation methods focus mainly on generation in the face region, regardless of changes in the eye region (Hong et al., 2022; Athar et al., 2022). The eye patch gaze redirection methods focus only on generation in the eye region under different gaze directions (He et al., 2019; Yu et al., 2019a). Regarding gaze-controllable face generation, it not only requires high-quality generation of facial features but also needs to produce photo-realistic and consistent eye regions under different gaze directions. Compared to other facial features, even minor imperfections in the eye region are highly noticeable. This makes gaze-controllable face generation even more challenging.

Speed. In terms of gaze-controllable face generation, it is not enough to generate natural and photorealistic images. Slow inference can also significantly hinder its practical deployment. In general, gaze-controllable face generation methods can be categorized into two groups: 2D-based methods (Ganin et al., 2016; Zheng et al., 2020) and 3D-based methods (Ruzzi et al., 2023; Li et al., 2022a). 2D-based methods do not perform well in 3D consistency. 3D-based methods suffer from low inference speed. There are many potential applications requiring both good 3D consistency and real-time performance, especially in immersive games and metaverse.

Usability. Usability is indispensable for the widespread adoption of gaze-controllable face generation. Gaze redirection methods tend to use angles of gaze direction to guide the generation process (Zheng et al., 2020; Ruzzi et al., 2023). This is because the angles are accurate and the generated images can be used for some downstream tasks such as gaze estimation (Zheng et al., 2020). However, it is much more natural for people to control generation with natural language in most cases (Feng et al., 2024). When generating a face image that appears to gaze in a specific direction, we do not aim for a precise orientation, but rather consider a certain acceptable range. If the initial generation does not meet expectations, we can further refine the results using language-based adjustments. For users,

language-guided methods are more efficient and convenient.

Therefore, my work focuses on improving quality, speed, and usability in gaze-controllable face generation and addresses related problems.

1.2 Outstanding Problems

This work tackles three critical challenges in the field of gaze estimation.

It is hard to model both rigid eyeball rotation and non-rigid eyelid deformation in a unified 3D representation. Combining a 3D face representation with gaze control is beneficial for ensuring 3D consistency in gaze-controllable face generation. However, modeling both rigid eyeball rotation and non-rigid eyelid deformation within a single framework presents significant challenges. Eyeball rotation involves rigid transformations, where the eyeball moves from one direction to another, maintaining its shape and structure. In contrast, eyelid deformation is non-rigid, as it involves intricate muscle motions and the flexible deformation of facial skin. Typically, non-rigid facial deformations are learned through neural networks conditioned on expression parameters (Hong et al., 2022) or latent deformation codes (Park et al., 2021). Hong et al. (2022) propose to rotate the eye latent features using an eye rotation matrix to model the eyeball rotation. However, performing the rotation in feature space often results in low-quality rendered images. Li et al. (2022a) separate the 3D space into eyeball space and outside space with an explicit eyeball surface model. Then they transfer the points in eyeball space and outside space into a common canonical space through rigid transformations and a learned warp field for volume rendering. However, the introduction of the explicit eyeball surface model increases the training difficulty and the simplified eyeball model also limits the generation quality. This raises an important question: is it possible to construct a unified 3D model that effectively handles both rigid eyeball rotation and non-rigid eyelid deformation?

The inference of 3D gaze-controllable generation models is still far from being real-time. Although 3D models excel at maintaining 3D consistency during deformation, they often suffer from slow inference speeds. Unlike 2D methods, 3D approaches must process and store more extensive information across three dimensions, requiring more computational resources. As a result, 3D methods typically rely on complex and resource-intensive models, significantly hindering their efficiency during inference. For instance, Li et al. (2022a) propose to perform dense sampling across a 3D space to render high-resolution images using Neural Radiance Field (NeRF) (Mildenhall et al., 2021). While this approach achieves exceptional visual fidelity, it often requires around half a minute per image, making it unsuitable for real-time applications. Additionally, this model requires training on a large set of personalized multi-view images, which makes it impractical for everyday applications. (Hong et al., 2022) build a generic 3D gaze-controllable face generation model with an efficient NeRF representation and a super-resolution module, achieving real-time rendering. Unfortunately, this method requires an inversion process that takes up to one minute to obtain the input facial parameters before rendering, when provided with a face image. This raises a crucial question: can we develop a 3D gaze-controllable face generation model capable of delivering real-time performance without compromising the benefits of 3D consistency?

Using explicit gaze labels to control gaze is inconvenient for general use. Existing gaze redirection methods typically rely on explicit gaze labels, such as pitch and yaw angles, to guide the generation process (Zheng et al., 2020; Ruzzi et al., 2023). These approaches are originally intended to produce gaze datasets, including gaze images and corresponding labels, for downstream tasks like gaze estimation. However, gaze control is also crucial for animatable digital avatars, which have extensive applications in the metaverse and 3D gaming. In such scenarios, requiring users to specify a precise target direction and provide it to the model is neither intuitive nor user-friendly (Delmas et al., 2022; Feng et al., 2024). Explicit gaze labels are inconvenient for general use, especially in interactive and creative applications.

This raises a compelling question: can gaze-controllable face generation be guided in a more natural and user-friendly way, such as through language descriptions?

1.3 My Approaches

To tackle these three challenges, I propose three distinct approaches: modeling a hybrid transformation from observation space to canonical space with one unified 3D model, distilling face shape prior into an efficient 3D representation, controlling gaze with the guidance of natural language.

Modeling a hybrid transformation from observation space to canonical space with one unified 3D model. Face rendering using NeRF (Mildenhall et al., 2021) 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 (Hong et al., 2022; Athar et al., 2022), which holds importance for various downstream tasks. I aim to learn a face NeRF model that is sensitive to eye movements from multi-view images. I 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, I first fit FLAME (Li et al., 2017), a well-established parametric face model, to the multi-view images considering multi-view consistency. Subsequently, I 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. I design an eye deformation field for the transformation, including rigid transformation, e.g., eyeball rotation, and non-rigid deformation in the eye region. The developed model is capable of generating high-fidelity images with accurate eyeball rotation and non-rigid periocular deformation, even under novel viewing angles. I show the details of this approach in Chapter 3.

Distilling face shape prior into an efficient 3D representation. Recent methods usually struggle with good 3D consistency (Zheng et al., 2020) or have limited efficiency and quality (Ruzzi et al., 2023), thereby limiting their applications. To address the problem, I present a real-time and high-quality gaze direction method. My method leverages recent advancements in real-time radiance fields (Chan et al., 2022) and fuses gaze with the extracted high-frequency features for redirection through a cross-attention mechanism. Consequently, I distill a lightweight module from a 3D portrait generator (Trevithick et al., 2023), which provides prior knowledge of face geometry. The final redirected image is attained via differentiable volume rendering (Mildenhall et al., 2021). The developed system achieves real-time 3D-aware gaze redirection with a feedforward network (i.e., ≈ 0.06 sec/image), $1000\times$ faster than other 3D-aware methods (Ruzzi et al., 2023; Hong et al., 2022). I show the details of this approach in Chapter 4.

Controlling gaze with the guidance of natural language. Generating face images with specific gaze information has attracted considerable attention in recent years. Existing approaches typically input gaze values directly for face generation, which is unnatural and requires annotated gaze datasets for training, thereby limiting its application (Zheng et al., 2020; Ruzzi et al., 2023). To address the problem, I present a novel gaze-controllable face generation model guided by natural language. The approach inputs textual descriptions that describe the human gaze and head behavior and generates corresponding face images. I first introduce a text-of-gaze dataset containing over 90k text descriptions spanning a dense distribution of gaze and head poses. I further propose a gaze-controllable text-to-face method. My model contains a sketch-conditioned face diffusion module and a model-based sketch diffusion module. I define a face sketch based on facial landmarks and an eye segmentation map. It provides a structured and detailed foundation for generating facial images. The face diffusion module generates face images from the face sketch, and the sketch diffusion module employs a 3D face model to generate a face sketch from the text description. I show the

details of this approach in Chapter 5.

1.4 Contributions

My work makes three key contributions.

First, I introduce an eye-controllable face NeRF model trained on multi-view images. This model enables high-fidelity face rendering under novel eyeball and head poses. Extensive experiments demonstrate that the rendered images significantly improve the performance of downstream gaze estimation tasks.

Second, I present a lightweight, real-time 3D-aware gaze redirection module that employs a cross-attention-based gaze fusion mechanism, ensuring robust 3D consistency. Evaluated on the ETH-XGaze dataset, our method outperforms state-of-the-art approaches in both inference speed and image quality.

Third, I propose a novel task: gaze-controllable face generation guided by textual descriptions of human gaze behavior. To support this task, I present the first text-to-gaze dataset, comprising over 90k descriptions. Additionally, I propose a two-stage text-to-face framework that enables language-guided, gaze-controllable face generation.

1.5 Structure

This dissertation is organized as follows:

Chapter 1 serves as an introduction, detailing the research challenges in the field, summarizing the approaches adopted to address them, and clearly stating the specific contributions of my work.

Chapter 2 provides an overview of gaze-controllable face generation and gaze estimation. It

reviews the mainstream approaches for gaze-controllable face generation and gaze estimation, highlighting their relationship and key connections. It then describes the gaze datasets and the dataset used in my work.

Chapter 3 proposes *DeNeRF*, a high-fidelity dynamic face NeRF model from multi-view images. It learns a unified canonical space to construct a rotation-aware manifold for eyeball rotation and the deformation of the periocular region. This model achieves high-fidelity face rendering under novel eyeball poses and head poses.

Chapter 4 presents *LiveGaze*, a real-time 3D-aware gaze controllable face generation model. The model simplifies the input to a single image and achieves real-time performance by distilling 3D face priors into a lightweight module.

Chapter 5 introduces *TextGaze*, a language-guided gaze-controllable face generation model. I propose the first text-to-gaze dataset to link text and gaze. This model enables users to control gaze image generation more naturally through natural language.

Chapter 6 concludes this dissertation by summarizing its key contributions, discussing its limitations, and proposing potential directions for future exploration.

Chapter Two

Literature Review

Gaze, as a crucial means for humans to observe the world, has always been an important subject of research for centuries. Gaze estimation and gaze-controllable face generation are two main tasks in gaze research. Gaze estimation is to estimate gaze directions from face images while gaze-controllable face generation is to generate face images under different gaze directions. Even though the goal of gaze-controllable face generation is different from gaze estimation, it is helpful to improve gaze estimation in different aspects. For instance, the generated gaze images can be used to augment gaze datasets to improve user-specific (Yu et al., 2019b; Wang et al., 2023b) or generic gaze estimation (Zheng et al., 2020; He et al., 2019) and gaze-controllable face generation can be used to learn a robust gaze representation (Yu et al., 2020; Park et al., 2019a) which is helpful to improve the generalization performance of gaze estimation models. Besides, a trained gaze estimator can also be used to improve the accuracy of gaze control (Zheng et al., 2020; Ruzzi et al., 2023) in gaze-controllable face generation. Regarding datasets, gaze-controllable face generation generally uses the same gaze datasets as gaze estimation. It is because gaze datasets have a large amount of gaze images with different gaze directions which is beneficial to train models generating gaze images under various gazes.

There are some tasks that are close to gaze-controllable face generation, like gaze correction and gaze redirection. Gaze correction is to adjust or align a person’s gaze (in an image, video, or real-time scenario) to make it appear as if they are making eye contact, typically for the purpose of improving visual alignment in video conferencing. Gaze redirection refers to the process of redirecting a person’s gaze to a specific target direction without altering the identity or other environmental elements. In general, gaze redirection is guided by a specific target gaze label which contains a yaw angle and a pitch angle of the gaze direction in the camera coordinate. Different from gaze correction, gaze-controllable face generation approaches are able to redirect gaze continuously in a certain range while gaze correction approaches generally only allow the person to look at the camera for making eye contact. As for gaze redirection, gaze-controllable face generation has the same goal with it regarding redirecting gazes, but gaze-controllable face generation does not strictly enforce the invariance of identity and the guidance from gaze labels. For example, gaze-controllable face generation can only take gaze guidance as input and generate multiple results with different identities. And the gaze guidance can be gaze labels or others like gaze text descriptions or face sketches including details in the eye region.

Gaze-controllable face generation can be realized through traditional approaches or learning-based approaches. Traditional approaches do not involve learning algorithms and they generally use image replacement or build a graphic model to realize gaze control. With the development of deep neural networks, learning-based methods have become more and more popular. Many researchers have shifted their focus to learning-based methods, leading to numerous significant research achievements. Based on whether explicit 3D facial representations are utilized, they can be further categorized into 2D-based methods and 3D-based methods. 2D-based methods operate on 2D images by warping image pixels or generating gaze images at the 2D level. In contrast, 3D-based methods construct gaze-controllable 3D face representations for each input using techniques like the neural radiance field (NeRF).

3D-based methods have better 3D consistency in face generation as they refer to the same 3D face representation when changing head poses.

In this chapter, I begin by briefly describing the eyeball structure and deformation in the eye region to help understand how the human visual system works. Following that, I introduce the mainstream methods of gaze estimation and show how gaze-controllable face generation can be used to improve gaze estimation. I then discuss existing gaze-controllable face generation methods, including traditional approaches, 2D learning-based approaches, and 3D learning-based approaches. I also discuss the modern generation models that can be applied to gaze-controllable face generation. I provide an overview of the latest advancements in the field and emphasize the specific research challenges that my work seeks to address. Finally, I detail the datasets utilized in gaze research. I select some datasets in my work and provide a rationale for their selection.

2.1 Human Visual System

Before delving into how gaze estimators and gaze-controllable face generation models work, it is essential to understand the structure of the human eye. In this section, I will introduce the anatomy of the eyeball and explain how the eye region deforms when we move our eyes.

2.1.1 Eyeball Structure

The eye is an image-forming organ that functions similarly to a camera. It captures light from the external environment and focuses it onto a plane to form an image. As shown in Fig. 2.1, when light enters the eye, it first passes through the cornea, which acts as the primary refractive medium, providing most of the eye’s focusing power. The light then travels through the pupil, whose size is regulated by the iris to control the amount of light entering the eye. After passing through the pupil, the light is further focused by the lens, which

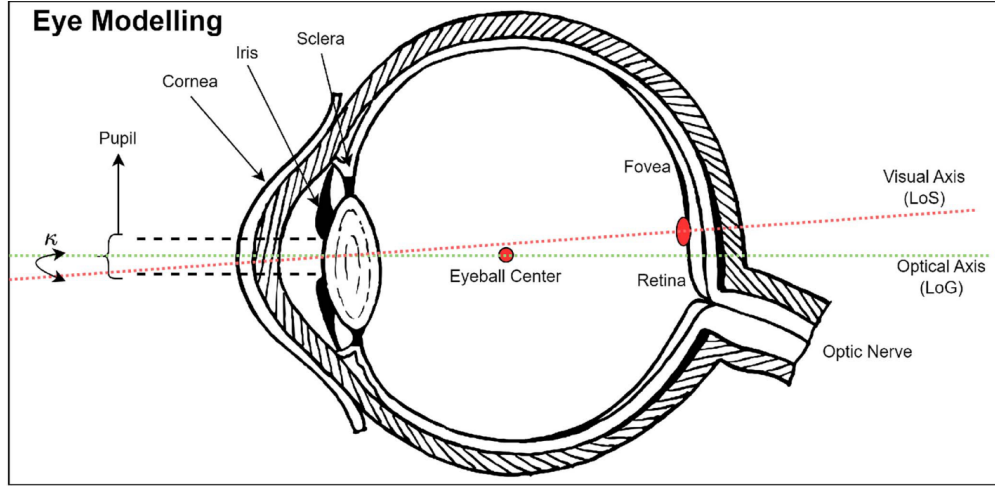


Figure 2.1: Eyeball structure (Ghosh et al., 2023).

adjusts its curvature to direct the light onto the retina. Within the eye, the optical axis (LoG) represents the geometric centerline of the eyeball, while the visual axis (LoS) is the path of light that passes through the center of the pupil and reaches the fovea, the region of the retina responsible for high-resolution vision. Due to the asymmetry of the eye's structure, there exists a small angular offset, denoted as κ , between the optical axis and the visual axis, which is critical for precise gaze estimation and eye modeling. According to previous studies (Carpenter, 1988; Guestrin et al., 2006), the fovea is positioned approximately $4\text{-}5^\circ$ horizontally and 1.5° vertically below the optical axis. However, this alignment can vary by up to 3° across different individuals (Guestrin et al., 2006). Finally, the photoreceptor cells in the retina convert the incident light into neural signals, which are transmitted via the optic nerve to the visual cortex in the brain. The brain processes and reconstructs this information, resulting in the perception of the visual image.

2.1.2 Eye-Region Deformation

Eye-region deformation mainly contains eyeball rotation and eyelid deformation. The muscles around the eyes, known as the extraocular muscles, control the rotation of the eyeball by

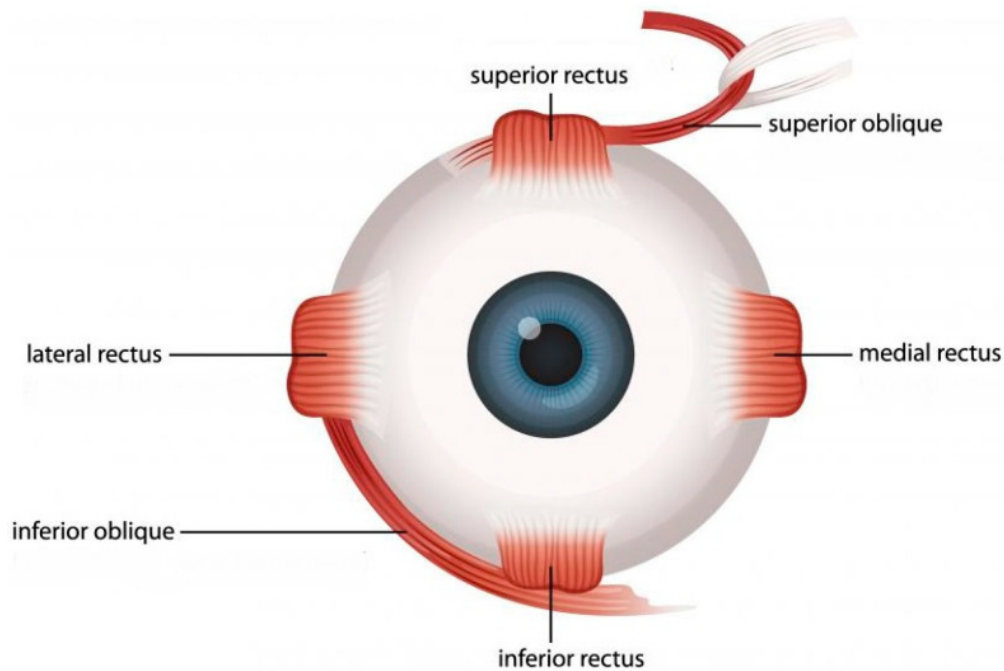


Figure 2.2: Extraocular muscles (Titova et al., 2024).

working in coordinated pairs to produce specific movements. As shown in Fig. 2.2, there are six primary extraocular muscles (Titova et al., 2024): Medial Rectus moves the eye inward (towards the nose) and is responsible for adduction; Lateral Rectus moves the eye outward (away from the nose) and is responsible for abduction; Superior Rectus moves the eye upward and is responsible for elevation; Inferior Rectus moves the eye downward and is responsible for depression; Superior Oblique rotates the eye downward and outward (laterally) and is responsible for intorsion (rotation of the top of the eye towards the nose); Inferior Oblique rotates the eye upward and outward and is responsible for extorsion (rotation of the top of the eye away from the nose).

These muscles work in a precise manner to control eye movements in different directions. When one muscle contracts, it pulls the eye in a specific direction, and the antagonist muscle on the opposite side may relax or contract in the opposite direction to balance the movement. This coordination allows for smooth and accurate eye rotations, which are essential for proper

vision, focus, and gaze direction. Eyeball rotation can be regarded as a rigid transformation as the six muscles act on the outer surface of the eyeball and rotate the eyeball as a whole object.

When the eyeball moves, the eyelid deforms with the movement. Different from eyeballs, eyelids are made of soft skins instead of rigid objects. Therefore, eyelid deformation can not be modeled simply by rigid transformation. Malbouisson et al. (2005) propose to describe upper and lower eyelid saccades as a harmonic oscillator function, which provides a mathematical model for the dynamics of eyelid movements. Based on the model, Wood et al. (2016a) develop a 3D morphable eye region model that enables natural eyelid deformation following eyeball rotation. However, individuals have varying eyelid shapes and different movement speeds for their upper and lower eyelids. With the development of deep neural networks, more researchers use neural networks to learn eyelid deformation from data (He et al., 2019; Zheng et al., 2020; Ruzzi et al., 2023), which my work also focuses on.

2.2 Gaze Estimation

Over the past few decades, researchers have proposed numerous gaze estimation methods. The earliest approaches involve sensors placed around the eyes to detect eye movement patterns such as fixation, saccades, and smooth pursuit (Young et al., 1975; Mowrer et al., 1935; Schott, 1922). However, these methods are often intrusive. With advancements in computer vision, more convenient non-intrusive methods have emerged. The devices used in gaze estimation have evolved from skin electrodes and infrared (IR) cameras to web cameras, which are more accessible in everyday life. These methods typically fall into three categories: model-based, feature-based, and appearance-based approaches (Cheng et al., 2021). Model-based approaches geometrically model the common physical structures of the human eye to compute a 3D gaze direction vector. Feature-based methods usually regress a gaze direction

or gaze point with the detected geometric features like pupil center. Appearance-based approaches directly regress gaze directions from the whole gaze or face appearance.

In this section, I will introduce model-based, feature-point-based, and appearance-based gaze estimation methods, highlighting the existing challenges that gaze-controllable face generation can help address.

2.2.1 Model-Based Approaches

The parameters used to reconstruct a geometric model of the eye can be categorized into extrinsic, fixed eye intrinsic, and variable parameters (Beymer et al., 2003). Extrinsic parameters include the 3D eye position (center of the eyeball) and the optical axis. Fixed eye intrinsic parameters encompass corneal radii, the angles between the visual and optical axes, refraction parameters, iris radius, and the distance between the pupil center and the cornea center. Variable parameters, which modify the shape of the eye model, include the pupil radius. These parameters are often estimated using infrared (IR) cameras and light sources to capture corneal glints (Guestrin et al., 2006; Zhu et al., 2007). By analyzing the positions of the detected glints along with the positions of the cameras and light sources, the eye’s intrinsic parameters can be estimated. Typically, this approach involves constructing a subject-specific geometric eye model that can fit the detected geometric features, such as infrared corneal reflections, pupil centers (Valenti et al., 2011), and iris contours (Alberto Funes Mora et al., 2014). The advantage of the model-based method lies in its accuracy and stability across different domains, as it relies on specialized equipment to extract precise geometric features from the human eye and model the relationships between different eye components. This allows for high precision in gaze estimation. However, there are some notable disadvantages. First, this method requires specialized equipment, such as IR cameras and light sources. While less intrusive than sensor-based methods, it still necessitates a controlled setup with specific hardware. Second, because human eye structures vary (e.g.,

corneal curvature, angular offset between visual and optical axes), personal calibration must be performed for each individual, which can be complex and time-consuming.

2.2.2 Feature-Based Approaches

Feature-based methods regress 2D eye features to human gaze by learning a mapping function which can be the polynomials (Morimoto et al., 2005; Stampe, 1993), and the neural networks (Ji et al., 2002). Hansen et al. (2002) combine a fast mean-shift color tracker for eye tracking with a specialized model to locate eye features, which are then mapped to gaze directions using Gaussian processes. To improve the reliability of eye trackers under varying gaze directions and head poses, researchers have also explored using additional features extracted from key facial points. For instance, Skodras et al. (2015) enhance the accuracy of gaze estimation along the vertical direction by incorporating features that provide information about eyelid positioning. However, it remains unclear which eye features are most relevant and should be prioritized. While geometric features can aid in gaze direction estimation, humans rely on the overall appearance of the eye region to infer gaze. Changes in gaze are not only reflected in eyeball rotation but also in the deformation of facial features. As a result, appearance-based information plays a critical role in gaze estimation. This has led to the development of appearance-based approaches to improve gaze estimation accuracy.

2.2.3 Appearance-Based Approaches

Appearance-based methods for gaze estimation do not explicitly extract geometric features but instead directly use the image content to estimate gaze. These methods typically take image features, such as pixel values (Lu et al., 2014) or deep features (Zhang et al., 2015a), as input to a regression model. Various regression models have been proposed for this purpose, including neural networks (Baluja et al., 1994), Gaussian process regression models (Williams et al., 2006), and adaptive linear regression models (Lu et al., 2014). Since appearance-based

methods do not require geometric features, they can work with images captured from standard web cameras, making them more accessible in everyday settings.

The general workflow of appearance-based methods involves four steps: 1) Collecting training samples with accurate gaze labels, which typically requires special equipment and a significant amount of time; 2) Extracting gaze-related features from raw image data using an effective feature extractor; 3) Learning the mapping function from appearance to gaze direction; 4) Fine-tuning the regression model using samples from specific domains or individuals to improve accuracy.

Earlier appearance-based methods often employed artificial features, such as histograms of oriented gradients (HOG) (Martinez et al., 2012), combined with traditional regression techniques. However, more recent approaches leverage deep learning models to automatically extract relevant features from raw images, improving performance and reducing the need for manual feature engineering.

In recent years, deep learning methods have become dominant in the field of gaze estimation. Unlike conventional appearance-based methods, deep learning approaches typically use deep features, which are automatically extracted through convolutional neural networks (CNNs), as input to a regression model (usually a neural network) for gaze estimation.

Zhang et al. (2015a) propose a method that concatenates head pose information with eye features extracted from eye images to estimate the gaze vector in the camera coordinate system. A significant contribution of this work is the collection of the MPIIGaze dataset, one of the most widely used datasets in gaze estimation, which was made publicly available. In 2017, Zhang et al. (2017d) further improve the model by replacing the LeNet network with VGG16, resulting in a significant increase in model accuracy. This transition to deeper architectures allowed the model to learn more complex features from the images, improving its ability to estimate gaze more accurately across different conditions.

Both of the works mentioned above only use images of a single eye as input to the model. However, Cheng et al. (2018) argue that incorporating gaze information from both eyes can improve gaze estimation accuracy. They propose a method that takes images of both eyes as input and adjusts the weights of the loss functions according to the quality of each eye’s image. Specifically, the eye with higher image quality is assigned a larger weight. Moreover, not all information from eye images is equally useful for gaze estimation. Park et al. (2018a) address this by simplifying the gaze representation to focus only on the locations of the eyeball and iris, which they call the “gazemap.” In their model, they first extract the gazemap from a single eye image using a stacked hourglass network. Then, they estimate the gaze vector using a regression network. This simplified representation helps the model focus on the most informative features for gaze estimation, improving both efficiency and accuracy.

The methods mentioned above require eye images as input, but they typically rely on separate models to locate the eyes and estimate head poses. This adds complexity and reduces the convenience of the system. It would be more efficient if a model could directly estimate gaze from a full-face image, enabling end-to-end training. However, this is challenging because the eyes are relatively small compared to the full face, making it difficult for the model to focus on the relevant features. Zhang et al. (2017b) address this challenge by incorporating an attention mechanism into their model. Their approach learns an average weight map, which increases the emphasis on the eye region. By combining the image features with this weight map, the model can effectively focus on the eyes and estimate the gaze vector more accurately. This method helps overcome the difficulty of distinguishing the small eye region within the larger context of the face, improving gaze estimation from full-face images.

In addition to estimating gaze from static images, several works have been proposed to incorporate temporal information to enhance gaze estimation. Gaze continuity across video frames is a key insight, as the gaze direction in a given frame is often closely related to the gaze in preceding and subsequent frames. This continuity provides valuable context

that can improve the accuracy of gaze estimation. For example, Kellnhofer et al. (2019a) propose using several bidirectional LSTM models to integrate information from multiple frames. Each LSTM model processes one frame and outputs gaze information, with the bidirectional nature allowing for the utilization of both past and future frames. Similarly, Zhou et al. (2019) use LSTM-based networks to develop a 3D gaze estimator that benefits from temporal dependencies. Other approaches, such as Wang et al. (2019), have also explored the use of optical flow to capture movement between frames and improve gaze estimation performance. These methods leverage temporal dynamics to provide more robust and accurate gaze predictions by considering the broader context of gaze behavior over time.

2.2.4 Challenges

Despite the availability of several models for gaze estimation, their performance still falls short of meeting the demands of everyday use. This can be attributed to three main challenges: dataset collection, personalization, and generalization.

Dataset Collection. Accurate gaze labeling is a significant challenge in gaze estimation. Unlike other tasks, gaze labels are invisible, which makes them hard to obtain. As a result, additional equipment and complex procedures are often needed to accurately label gaze data, which makes data collection both time-consuming and costly. For example, in (Funes Mora et al., 2014b), participants were instructed to focus on a ping pong ball while a depth camera recorded the process. After data collection, algorithms or manual methods were employed to label the eye center and ping-pong ball positions in the RGB video. These positions were then mapped to the 3D point captured by the depth camera, allowing the calculation of gaze direction by subtracting the two 3D positions. In (Zhang et al., 2015a), gaze was estimated by transforming the eye position (obtained using a 3D six-key point model) and gaze target into camera coordinates, after which the gaze direction was calculated using public camera parameters. There are also synthetic-based methods (Wood et al., 2015a; Wood et al., 2016b;

Wang et al., 2018; Shrivastava et al., 2017b) designed to overcome the challenge of gaze labeling. However, the realism of synthetic images is limited, which can introduce a large bias when training neural networks, making synthetic-based methods less effective for real-world applications.

Additionally, existing gaze estimation datasets often lack sufficient data continuity, limiting their ability to cover a wide range of head poses and gaze angles. Many datasets also fail to account for the diversity in eye shapes and gaze patterns across different age groups, genders, and races, as well as the effects of changing environmental factors such as lighting. As a result, models trained on these datasets often perform well only within the narrow scope of those datasets, but struggle to generalize to new users or different conditions.

Gaze-controllable face generation offers an alternative approach to gaze data collection. It allows for the generation of numerous labeled gaze images, eliminating the need for expensive data collection systems and complex calibration (Zheng et al., 2020; He et al., 2019). Moreover, it enables the creation of gaze data across a large and continuous range by interpolating in the latent space, as opposed to the discrete sampling used in traditional methods. This approach also provides greater flexibility in enriching datasets by incorporating additional controls over factors such as identity, illumination, and background.

Personalization. There are variations in eye structures across individuals, which lead to personalization issues. As mentioned in Section 2.1.1, the true gaze direction differs from the optical axis, with a small angular offset between them. These offsets vary from person to person and are not visible in images, making it impossible to infer this information directly from visual data. This discrepancy results in errors when attempting to estimate the gaze of different individuals.

Researchers have proposed various methods to address this issue. Liu et al. (2019) use a differential approach to learn gaze variations from images of the same person. Other

approaches model person-specific parameters using random effects, where fixed effects, shared across individuals, may still introduce errors (Lindén et al., 2019; Xiong et al., 2019). Additionally, some studies fine-tune a generic gaze estimator using multiple images from the same person to improve personalization (Yu et al., 2019b; Park et al., 2019b).

In practice, the personalization problem arises due to the scarcity of gaze data for individual users. This challenge can be addressed by constructing a personalized generative model to generate gaze images specifically for fine-tuning, which is also the focus of gaze-controllable face generation (Yu et al., 2019b; Wang et al., 2023b).

Generalization. Generalization is crucial for the widespread adoption of gaze estimation. Most existing models are trained on specific datasets and only perform well within those datasets. The key to improving generalization in gaze estimation is learning a robust gaze representation that disentangles gaze from other variations. Gaze-controllable face generation shares this goal, making the knowledge learned from these models transferable to gaze estimation tasks. For instance, Park et al. (2019a) propose a Disentangling Transforming Encoder-Decoder (DT-ED) to learn a robust gaze extractor by separating gaze from other variations. Additionally, Yu et al. (2020) introduce an unsupervised framework that learns a gaze representation via a gaze redirection module. Both approaches enable rapid adaptation for gaze estimation with a limited number of gaze images.

2.3 Gaze-Controllable Face Generation

In this section, I present the literature review on gaze controllable face generation by discussing on traditional approaches which are not learning-based as well as 2D and 3D learning-based approaches.

2.3.1 Traditional Approaches

Eye Replacement. Replacing eye images directly is straightforward and feasible to realize gaze control. To ensure a natural fusion, the eye image used for replacement should come from the same individual. Wolf et al. (2010) propose a gaze-replacement system which detects the eye region and replaces it with the eye image under a different gaze direction. The developed system is for correcting gaze directions in a videoconferencing setting. Specifically, the system contains three stages: 1) First, an approximate eye pose estimation is learned from a dataset of annotated images offline. 2) In the second stage, the user is prompted to look directly at the camera at the start of the video conferencing session. This enables the system to create a dataset of images capturing the user’s direct gaze. 3) Finally, during runtime, the system continuously tracks the eye’s position in the current frame, identifies the most similar image from the database by extracting SIFT (Lowe, 2004) features, and replaces it. Qin et al. (2015) further propose a zone of correction to formulate the range of head poses and eye openness when the gaze correction should be applied. Shu et al. (2016) use a 3D morphable model to warp the reference image first to align with the target image regarding poses and expressions. Then they use a newly designed Background-Foreground Guided Poisson (BFGP) compositing to blend the eye patch and other facial features in a more robust way. Even though the replaced eye image patch is realistic, replacement methods do not consider the movements of the eyelid in fusion which may cause unnatural artifacts in the synthesized image. Besides, this method requires pre-recording a large number of diverse images with the direct gaze and its performance will significantly degrade if the pre-prepared data is insufficient. In short, eye replacement is selecting the best choice from limited data while my work focuses on exploring continuous latent space of gaze, therefore not suffering from the above limitations.

3D Eye Morphable Model. A 3D Morphable Model (3DMM) is a statistical model used to represent and generate 3D shapes and textures, primarily for faces or other objects. It

is constructed by analyzing a set of 3D scans or samples using dimensionality reduction techniques such as Principal Component Analysis (PCA). The resulting model allows the generation of new 3D instances by varying a small set of parameters, enabling flexible and compact representation of complex 3D data. Researchers have also made numerous attempts to model the eyeballs using 3DMM models and control gaze with related parameters. Banf et al. (2009) introduce a model for example-based, photorealistic rendering of eye movements in 3D facial animation. Utilizing 3D scans of a face in various gaze directions, the model accurately captures the motion of the eyeball along with the corresponding deformations of the eyelids and surrounding skin. To avoid the artifacts caused by the gap and the sharp transition between face and eyeball, the visible portion of the eyeball is integrated into a continuous face mesh, with iris displacements and lid occlusions represented using a texture mapping technique. However, a unified mesh does not perform well under large gaze directions and the model represents eyelashes as one part of the texture which makes the model appear correct only in near-frontal poses and gaze directions. To solve the problems, Wood et al. (2016a) propose a multi-part eye model and build separate eyeball models. This design allows controlling eyeball rotation separately. They also use a carefully designed generic eye region topology (Wood et al., 2016c) and project eyelid vertices onto the eyeball mesh to avoid the unnatural transition between face region and eyeball. Their model is adopted by a well-known 3D face morphable model FLAME (Li et al., 2017) to control gaze in whole faces. Please note that both above models are able to animate eyelids when changing gaze directions. Even though 3D eye morphable models can control gaze through image fitting and parameter interpolation, there still is a big gap in realism between the synthesized images and real images. Based on the work from Wood et al. (2016a), Wood et al. (2018) propose to get an optical flow field from the fitted model and the redirected model. The flow field is used to warp the eyelid in the origin image to the target shape. This method enhances the realism of the facial part while the eyeballs are still synthesized by the morphable model. Therefore the realism issue still remains. In addition, building a 3D eye morphable model

requires a great number of high-accuracy 3D scans which are expensive. My work focuses on realizing gaze-controllable face generation with only 2D images and keeping realism the same as real images.

2.3.2 2D Learning-Based Approaches

Warping-Based Methods. Warping-based methods refer to the models that utilize geometric transformations to modify input data (e.g., images, features) by mapping it from a source space to a target space. The models rely on a warping function that governs the transformation. Since gaze changing involves not only rigid eyeball rotation but also complex non-rigid motion of eye muscles and eyelids, obtaining the warping function is non-trivial. Learning-based methods are usually used to complete complex tasks by learning from data. Therefore, such a method is a good choice for getting the warping function. Kononenko et al. (2015) propose to learn eye flow vectors which are the offsets of pixel positions between input images and gaze-redirected images with an ensemble of randomized decision trees. However, the proposed model is limited to training for a specific gaze redirection angle, lacking the capability to continuously adjust gaze within a defined range. To address this problem, Ganin et al. (2016) propose Deepwarp which predicts a warping field in a coarse-to-fine manner by taking an eye image, feature points as well as a correction angle as inputs. Deepwarp allows redirecting gaze continuously in a certain range by specifying different input correction angles and further improves the performance by introducing a lightness correction module. Unfortunately, both of the models above require extensive annotated data for training. Yu et al. (2019a) propose to train an image-warping neural network with synthesized data. They further introduce a cycle consistency loss and a gaze redirection loss from a generic gaze estimator trained with real data to bridge the domain gap between synthetic data and real data. This method is further refined by Yu et al. (2020) with an unsupervised gaze representation learning network.

Generator-Based Methods. Even though warping-based methods can realize gaze control, such methods are unable to handle scenarios where part of the eye is occluded, as they rely solely on replacing pixels with existing ones from the original image without generating new pixels. With the rapid progress in deep generative models, advanced tools leveraging these technologies have become widely adopted for gaze redirection tasks. He et al. (2019) propose to generate gaze-redirected images with a conditional generative adversarial network (GAN) (Goodfellow et al., 2020) architecture. They incorporate perceptual and cycle consistency losses to enhance the generated image quality. Park et al. (2019a) presents an encoder-decoder framework that transforms eye images into latent vectors, which are subsequently modified using rotation matrices to generate synthetic images with redirected gaze. Building on the approach introduced by Park et al. (2019a), Zheng et al. (2020) propose an innovative self-transforming encoder-decoder architecture (ST-ED) that effectively disentangles and controls both explicit and extraneous factors. This architecture enables ST-ED to perform both head and gaze redirection in full-face images, delivering precise and highly accurate results. ReDirTrans (Jin et al., 2023) builds upon ST-ED by introducing a method to project modified embeddings back into the original latent space. This approach facilitates attribute replacement with minimal interference in other features while maintaining the integrity of the latent distribution. Combining with a predefined feature space provided by StyleGAN (Karras et al., 2020), ReDirTrans is able to generate high-quality full-face images with 1024x1024 resolution.

2.3.3 3D Learning-Based Approaches

Although effective, 2D-based methods frequently face challenges in maintaining 3D consistency due to the absence of an explicit 3D facial representation. As a result, researchers continue to explore ways to incorporate 3D facial representations into gaze-controllable face generation to enhance realism and consistency. In the last few years, Neural Radiance

Fields (NeRF), a neural network-based method for representing 3D scenes as continuous volumetric representations in an implicit manner, have shown great capability in high-fidelity 3D reconstruction and animation.

Li et al. (2022a) introduce **EyeNeRF**, a method that integrates explicit surface modeling of the eyeball with implicit volumetric representations of the surrounding eye structures. To be specific, they employ a lightweight capture setup consisting of a few static cameras and lights, complemented by a single handheld camera with a co-located light source. The eyeball surface is represented as an explicit mesh, while the canonical shape of the periocular skin region and the interior eye volume is modeled using an implicit volumetric representation. It employs ray tracing to compute reflection and refraction rays by determining their intersection points with the explicit eyeball surface. These rays are subsequently utilized for ray marching through the implicit volumetric representation. Surrounding skin and hair deformations are modeled using a learned warp field applied to the canonical volume. To enable relightability, they represent the underlying reflectance using spherical harmonics coefficients, allowing the outgoing radiance to be computed as the product of reflectance and environmental illumination in the 3D frequency domain. They jointly optimize the eyeball mesh’s shape and pose, along with the density and reflectance within the implicit volume. This process is guided by minimizing the photometric loss between the modeled outgoing radiance and the pixel values in the captured video. This approach allows for high-fidelity gaze redirection with photorealistic results, achieved using a minimal setup of lights and cameras. However, EyeNeRF suffers from prolonged training and evaluation times, requiring 4 days on 8 NVIDIA V100 GPUs for training and approximately 30 seconds to process a single 800x800 resolution image during evaluation.

GazeNeRF (Ruzzi et al., 2023) introduces a two-stream multilayer perceptron (MLP) architecture to separately represent the face and eye regions using neural radiance fields. Specifically, they utilize the two-stream MLP structure to independently predict feature maps

for the eyeballs (eyes) and the rest of the face region (face-only). The features corresponding to the eyes are transformed using a desired 3D rotation matrix before compositing the eye and face regions through differentiable volume rendering. By explicitly separating the eyeballs, GazeNeRF applies rigid rotations to the 3D features, which they demonstrate as advantageous for improving gaze redirection accuracy. To enable effective training, they propose a feature composition mechanism after the two-stream MLPs, along with additional training losses specifically designed to enhance gaze redirection performance.

Although these methods exhibit strong 3D consistency, they often rely on complex and resource-intensive models, which constrain their feasibility for real-time applications. EyeNeRF (Li et al., 2022a) is a personalized model that aims to model the eye region with high fidelity under varying lighting conditions while controlling gaze. It requires retraining for each subject on multiview data and takes around 4 days to train on 8 NVIDIA V100 GPUs for a new subject. It takes around 30 seconds to generate a single image with a resolution of 800x800 only taking rendering process into account. On the other hand, GazeNeRF (Ruzzi et al., 2023) is a generalized model that aims to realize gaze-controllable face generation for any subject with only a single image. Once trained, it only requires a fitting process to generate a new subject’s model, which takes around 1 minute on a single NVIDIA 3090 GPU. The rendering process takes around 0.06 seconds for a 512x512 image with an efficient render module. EyeNeRF excels in photorealism when trained on multiview data, whereas GazeNeRF is more efficient in training/fitting and rendering times for gaze-controllable face generation on new subjects. However, neither method is suitable for real-time applications due to the extensive time required for training/fitting and rendering. My work seeks to achieve real-time performance regarding gaze-controllable face generation on new subjects from a single image while preserving high image quality.

2.3.4 Modern Generation Models

In this section, I will discuss some advanced generation models which can be used for gaze-controllable face generation. These techniques are based on recent advances in deep learning and computer vision and have been shown to produce high-quality generation results.

Neural Radiance Fields (NeRF) (Mildenhall et al., 2021) is a method for synthesizing novel views of complex scenes. NeRF represents a 3D scene as a continuous 5D function that maps 3D coordinates and viewing directions to radiance. The function is trained using a large set of posed images of the scene. NeRF can render novel views of the scene by querying the function at different 3D coordinates and viewing directions. NeRF has been shown to produce high-quality novel views of complex scenes, but it is computationally expensive and requires a large amount of training data. NeRF has been applied to gaze-controllable face generation because of its good 3D consistency and high-quality rendering (Li et al., 2022a; Ruzzi et al., 2023). As mentioned in Section 2.3.3, even though these methods have shown promising results, they are computationally expensive and require an extensive test-time optimization process to generate gaze-controllable faces.

3D Gaussian Splatting (3DGS) (Kerbl et al., 2023) is a novel 3D scene representation and rendering technique that efficiently models complex scenes by representing them as a set of oriented, anisotropic Gaussian functions (“splats”). Unlike traditional methods, such as NeRF, which rely on dense volumetric sampling, Gaussian Splatting leverages lightweight, differentiable Gaussians that adapt their position, size, orientation, and color parameters during training. This results in rapid, high-quality novel view synthesis, achieving real-time rendering performance even for detailed, photorealistic scenes. Its compact representation significantly reduces computational costs while maintaining high visual fidelity, making it suitable for applications in virtual reality, augmented reality, and real-time interactive graphics. Wei et al. (2024) proposed GazeGaussian, a novel gaze-controllable face generation method

that leverages Gaussian Splatting to synthesize high-quality, gaze-controllable face images. Same to GazeNeRF, they employ a dual-stream 3DGS model to distinctly represent facial and ocular areas. Utilizing the unstructured characteristics of 3DGS, they create an innovative eye representation for achieving rigid eye rotation aligned with the desired gaze direction. To improve the synthesis generalization across different individuals, they incorporate an expression-conditional module to guide the neural renderer. However, GazeGaussian still requires a fitting process (around 30 seconds) to generate gaze-controllable faces, which is not ideal for real-time applications.

Diffusion Models are a powerful class of generative models designed to create high-quality data such as images (Rombach et al., 2022; Du et al., 2023; Zhang et al., 2023), videos (Ho et al., 2022; Blattmann et al., 2023; Ceylan et al., 2023), or audio (Kong et al., 2020; Huang et al., 2023; Schneider, 2023) by iteratively refining random noise into structured outputs. Inspired by thermodynamic diffusion processes, these models first learn to gradually introduce noise into data (forward diffusion), and then reverse this process (reverse diffusion) to reconstruct realistic and diverse samples. Notable for their exceptional image synthesis capabilities, diffusion models, such as Stable Diffusion (Rombach et al., 2022) and Denoising Diffusion Probabilistic Models (DDPM) (Ho et al., 2020), have outperformed previous generative techniques like GANs (Goodfellow et al., 2020) in terms of stability, diversity, and photorealism. Widely adopted in generative art, image editing, and content creation, diffusion models have significantly reshaped the landscape of deep generative modeling. Diffusion models also can be extended to 3D data generation by incorporating 3D spatial information into the diffusion process (Liu et al., 2023; Liu et al., 2024). However, these methods do not perform well in generating 3D-consistent results with pure diffusion models, which is crucial for gaze-controllable face generation.

Large Language Models (LLMs) are advanced artificial intelligence systems built upon deep neural networks, primarily Transformers (Vaswani et al., 2017), trained on extensive

text datasets. These models can understand, generate, and reason about natural language at unprecedented levels of complexity and coherence. Popular examples include OpenAI’s GPT series (Achiam et al., 2023), Google’s Gemini (Team et al., 2023), and Meta’s LLaMA (Grattafiori et al., 2024). Due to their ability to perform tasks such as text generation, question answering, summarization, and language translation, LLMs have transformed various applications, ranging from conversational AI assistants to code generation and creative writing. Their remarkable capacity to generalize and adapt to diverse tasks has positioned LLMs at the forefront of artificial intelligence research and innovation. Recently, LLMs have been used for text data generation because of their ability to generate high-quality and diverse text data (Tang et al., 2023; Meng et al., 2022b). In the context of gaze-controllable face generation, controlling the generation process using natural language instructions can be much more convenient and intuitive for users compared to other methods redirecting gaze with numerical values. However, current gaze-related datasets do not have textual descriptions of gaze directions. Generating high-quality and diverse gaze-related textual descriptions and building a language-guided gaze-controllable face generation system can be a promising research direction.

2.4 Datasets

Datasets for gaze-controllable face generation should include facial image data with diverse gaze directions. Explicit gaze labels are not essential, as gaze can be guided through other modalities, such as text descriptions or sketches. Furthermore, datasets with a wide range of identities and high-resolution images are particularly advantageous.

In this section, I first briefly introduce gaze datasets which are originally for gaze estimation, and then describe the datasets I choose to use in my work as well as the reason I choose them.

2.4.1 Gaze Datasets

Zhang et al. (2015b) introduced the MPIIGaze dataset, widely used for appearance-based gaze estimation. It contains 213,659 images from 15 subjects collected over several months, with no constraints on head pose. The dataset includes both 2D and 3D gaze annotations and a standard evaluation set with 3,000 images per subject (1,500 left-eye and 1,500 right-eye). The dataset was further extended in (Zhang et al., 2019; Zhang et al., 2017c), providing additional face images and manual landmark annotations.

The EyeDiap (Funes Mora et al., 2014a) dataset includes 94 video clips from 16 participants in a laboratory setting, featuring three visual target sessions: continuous moving targets, discrete moving targets, and a floating ball. Each subject participated in six sessions, with static and free head movements. Data was captured using an RGBD and an HD camera. However, the dataset lacks illumination variation, limiting its use in dynamic lighting conditions.

The GazeCapture dataset (Krafka et al., 2016), collected via crowdsourcing, contains 2,445,504 images from 1,474 participants using mobile phones or tablets. Participants were asked to gaze at a circle on the device, with no restrictions on head movement, resulting in a diverse dataset that includes varying lighting conditions and head motions.

Fischer et al. (2018) introduce the RT-Gene dataset, which collects gaze data using a dedicated eye-tracking device, and later removes the device from the images through post-processing. Kellnhofer et al. (2019b) propose the Gaze360 dataset, consisting of 238 subjects in both indoor and outdoor environments, capturing 3D gaze across a wide range of head poses and distances. Zhang et al. (2020) present the ETH-XGaze dataset, offering high-resolution images that cover extreme head poses and include 16 different illumination conditions to explore the effects of lighting variations.

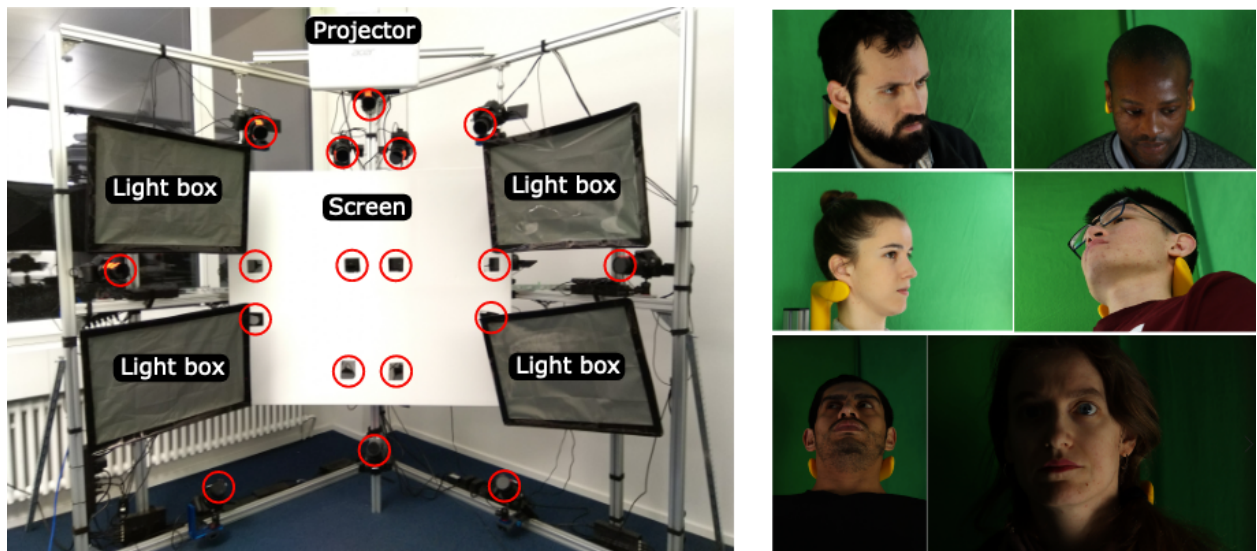


Figure 2.3: Data collection setup and examples of ETH-XGaze (Zhang et al., 2020).

2.4.2 Dataset Selection

ETH-XGaze Dataset. ETH-XGaze (Zhang et al., 2020) is a comprehensive, large-scale gaze dataset consisting of high-resolution images that exhibit extreme variations in head pose and gaze direction. The dataset was captured using a multi-view camera system under diverse lighting conditions. It includes 756k frames of 80 subjects in the training set, with each frame comprising 18 images taken from different camera angles. Additionally, a person-specific test set features 15 subjects, each contributing 200 images accompanied by ground truth gaze labels. The dataset thoroughly samples a wide range of variations in head poses, up to the maximum where both eyes remain visible ($\pm 70^\circ$ from directly facing the camera), and encompasses a comprehensive set of gaze directions ($\pm 50^\circ$ in the head coordinate system). Data was collected from 110 participants of diverse ethnicity, age, and gender—some wearing glasses and some without—ensuring a rich and varied dataset.

Each participant provided over 500 gaze directions under full-on illumination, along with an additional 90 samples collected under 15 different lighting conditions. The setup used 18 Canon 250D digital SLR cameras positioned at different viewpoints to capture a broad

spectrum of head poses. Four Walimex Daylight 250 light boxes surrounded a large screen (120×100 cm), which displayed stimuli controlled by a Raspberry Pi and projected by a projector. Participants were seated approximately one meter from the screen, with their heads stabilized using a headrest to minimize unintentional motion. During data collection, participants focused on a shrinking circle on the screen and clicked a mouse when the circle turned into a dot, indicating the gaze point. Gaze points were randomly distributed across the screen to ensure comprehensive coverage. For most of the data collection, all four light boxes were fully illuminated to maximize brightness. To simulate diverse lighting conditions, the dataset also included data captured with various combinations of the light boxes switched on and off, simulating 15 different illumination conditions.

I choose to use the ETH-XGaze dataset as it covers a wide range of head poses as well as gaze directions and contains high-resolution multi-view images. I show its data collection setup and some examples in Fig. 2.3. I do not use the pre-processed versions of ETH-XGaze since they only preserve partial face regions in a low resolution. In my work, I process the raw images from ETH-XGaze to get full-face images in a higher resolution.

FFHQ Dataset. The dataset (Karras et al., 2019) contains 70k high-quality PNG images at a resolution of 1024×1024 pixels, featuring significant variation in age, ethnicity, and image backgrounds. It also includes a diverse range of accessories, such as eyeglasses, sunglasses, and hats. The images were sourced from Flickr and were automatically aligned and cropped.

I choose the FFHQ dataset because it covers a large number of subject identities which is beneficial to the diversity of generation. I show some examples of the FFHQ dataset in Fig. 2.4 Even though it does not cover a wide range of head poses and gaze directions as the ETH-XGaze dataset, the FFHQ dataset still presents various poses of head and gaze which are enough to learn a gaze-controllable face generation model.



Figure 2.4: Examples of FFHQ dataset (Karras et al., 2019).

2.5 Conclusion

In this chapter, I discuss gaze-controllable face generation, providing an overview of the evolution of this task. I review existing methodologies, including both non-learning-based and learning-based approaches, highlighting their limitations and identifying the specific challenges addressed in my work. Lastly, I introduce two benchmark datasets utilized in this dissertation for evaluation, detailing their features and relevance to this research.

Chapter Three

High-Fidelity Gaze-Controllable 3D Face Generation with Multi-View Images

This chapter presents work that has been published at the 34th British Machine Vision Conference 2023 in Aberdeen, UK (Wang et al., 2023b).

Face rendering is an important task in computer vision and computer graphics. Conventional methods fit a parametric face model based on a face template mesh (Li et al., 2017; Feng et al., 2021). Generative adversarial networks (GAN) directly render photo-realistic images with deep neural networks (Goodfellow et al., 2020; Antipov et al., 2017; Gauthier, 2014). Recent research has incorporated the Neural Radiance Fields (NeRF) (Mildenhall et al., 2021) for face rendering. NeRF models encode 3D geometry and exhibit great multi-view consistency that enables face rendering under novel viewpoints. Additionally, various previous works explore parametric face NeRF models (Hong et al., 2022), 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

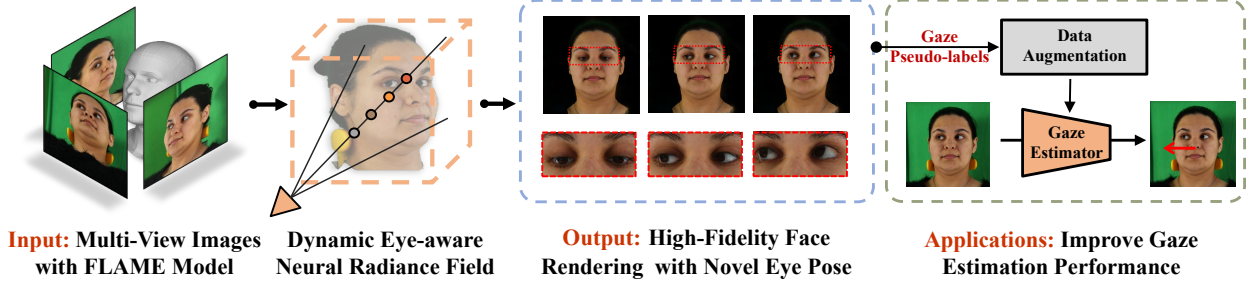


Figure 3.1: Pipeline of DeNeRF and its application. The Dynamic Eye-aware Neural Radiance Field (DeNeRF) is designed to render high-fidelity faces with animatable eyes using a set of multi-view images. It enables face rendering under novel view and eye pose. By leveraging DeNeRF, we are able to obtain pseudo gaze labels from the model, which can be utilized to enhance gaze estimation methods.

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 (Ruzzi et al., 2023; Zheng et al., 2020). 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 chapter, I 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. I 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, I begin by fitting a well-established parametric face model, FLAME (Li et al., 2017), to the multi-view face images. FLAME is originally designed for face tracking from a single face image, and I 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 I then use for DeNeRF learning.

I further define a unified canonical space in DeNeRF. Given a pixel in the observed images, I sample 3D positions based on view directions in the observation space. I transform the 3D positions from the observation space into the canonical space with given poses. I design an eye deformation field for the transformation, including rigid transformation (*e.g.*, eyeball rotation and head rotation) and non-rigid transformation. I input the 3D positions in the canonical space into a NeRF model and render the pixel for alignment. To reduce the computational costs, I adopt a patch-based sampling approach (Schwarz et al., 2020). I sample patches in images for alignment in each iteration. To enforce realistic eye region, I generate an eye mask for each image and enlarge the sampling ratio in the eye region.

Overall, our contributions are three-fold:

- I propose DeNeRF which learns a dynamic face NeRF model from multi-view images. To capture the eyeball pose accurately, I design a new fitting process for the FLAME model, ensuring consistency across multiple views.
- I define a unified canonical space to construct a rotation-aware manifold. I 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.

3.1 Related Work

3.1.1 Neural Radiance Field

NeRF (Mildenhall et al., 2021) 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

$$H_{\theta} : (\gamma(\mathbf{x}), \gamma(\mathbf{d})) \rightarrow (\mathbf{c}, \sigma), \quad (3.1)$$

where θ represents the parameters of the network and γ refers to a positional encoding function (Mildenhall et al., 2021; Vaswani et al., 2017) 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 (Park et al., 2021; Pumarola et al., 2021; Gafni et al., 2021; Hong et al., 2022; Peng et al., 2021; Freer et al., 2022) adapt NeRF to represent dynamic scenes. Hong et al. (2022) 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 (Guo et al., 2021; Ye et al., 2023b; Du et al., 2023). They usually overlook the controllable eye movement in face rendering.

3.1.2 Eye Image Synthesis

The precise manipulation of realistic eye imagery has proven essential across multiple domains of application. Qin et al. (2022) 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. (2016c) build a virtual 3D morphable eye model with computer graphic algorithms. They can synthesize eye images with arbitrary gaze directions (Wood et al., 2016c; Wood et al., 2016a; Wood et al., 2015b). 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 (Shrivastava et al., 2017a; He et al., 2019; Park et al., 2019a; Yu et al., 2019a). He et al. (2019) utilize GAN for gaze redirection task. They can synthesize large-scale gaze data by performing gaze redirection task on one eye image.

However, above generative models are not 3D-aware models and cannot maintain good 3D consistency during generation. Ruzzi et al. (2023) use 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. Li et al. (2022a) propose EyeNeRF, a hybrid model that combines explicit surface modeling of the eyeball with implicit volumetric representations of the surrounding eye structures for gaze-controllable face generation. They employ a simple parametric model to represent eyeball in 3D space. However, their eyeball model is not flexible enough to represent various eye shapes and sizes. In contrast, our method combines a more flexible implicit neural representation for eyeball modeling and the FLAME model for rotation awareness. Our model provides more flexibility in eye shape and size representation while maintaining 3D rotation awareness.

3.2 Methodology

3.2.1 Multi-view Face Tracking

I first perform multi-view face tracking to acquire eyeball poses for training. I use the well-known parametric face model FLAME (Li et al., 2017) for face tracking. FLAME model

fits facial parameters (such as pose and expression) from a single face image. I 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, I input multi-view images and corresponding camera poses to the face tracker. I detect face landmarks, pupil centers, and face masks in each image (Bulat et al., 2017; Yu et al., 2018), where face masks are used to remove the background in the images. I initialize the textured FLAME face model with zero parameters and project the face model to all views. I 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. I can also obtain face landmarks and pupil centers from the FLAME model. I align them with the detected results for structure consistency. I use L_1 loss for all alignments where I denote them as face appearance loss $\mathcal{L}_{appear}$, facial landmark loss $\mathcal{L}_{face}$ and pupil center loss $\mathcal{L}_{pupil}$. Overall, I optimize the multi-view face tracker by minimizing

$$\mathcal{L}_{tracking} = \left(\sum_{i=1}^N (\alpha \mathcal{L}_{pupil}^i + \beta \mathcal{L}_{face}^i + \gamma \mathcal{L}_{appear}^i) \right) / N, \quad (3.2)$$

where α, β, γ are hyper-parameters, and I empirically set them as 10, 5, and 30. N is the number of views which is 13 in our experiment.

All three losses are mean absolute errors. The losses are defined as follows:

$$\mathcal{L}_{appear} = \left(\sum_{i=1}^N |\hat{C}_i - C_i| \right) / N, \quad (3.3)$$

where N is the number of multi-view images from one frame and N is 13 in our setting, C_i is the colors of all pixels from true images in one frame. $\hat{C}_i$ is the colors of all pixels from rendered images in the same frame.

$$\mathcal{L}_{face} = \left(\sum_{i=1}^N |\hat{X}_i - X_i| \right) / N, \quad (3.4)$$

where $X_i \in \mathbb{R}^{68 \times 2}$ is the 2D positions of 68 facial landmarks of true faces from one frame,

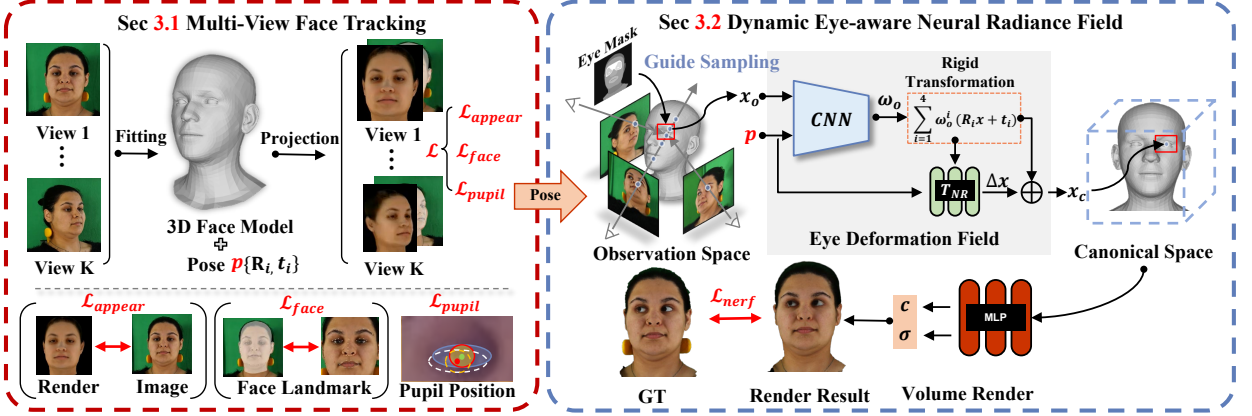


Figure 3.2: Overview of DeNeRF model. DeNeRF fits a set of multi-view images to a face model including poses for the eyeballs, jaw, and neck. It defines a canonical space and introduces an eye deformation field that can transform points from observation space into the canonical space. This field is composed of both rigid and non-rigid transformation modules, enabling it to handle the complexity of eye movement and deformation.

$\hat{X}_i \in \mathbb{R}^{68 \times 2}$ is the 2D positions of 68 facial landmarks of rendered faces from the same frame.

$$\mathcal{L}_{pupil} = \left(\sum_{i=1}^N \left| \hat{P}_i - P_i \right| \right) / N, \quad (3.5)$$

where $P_i \in \mathbb{R}^{2 \times 2}$ is the two pupil center positions of true faces from one frame. $\hat{P}_i \in \mathbb{R}^{2 \times 2}$ is the two pupil center positions of rendered faces from the same frame.

After fitting the FLAME model, I obtain facial parameters as the outcome. For DeNeRF Learning, I 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, I refer to all four poses as $\mathbf{p}$.

3.2.2 Dynamic Eye-aware Neural Radiance Field

The input of DeNeRF contains multi-view images and poses $\mathbf{p}$ 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 I can learn a unified eyeball model in the canonical space. I perform the similar operation in DeNeRF. I newly design an *eye deformation field* which can handle both rigid and non-rigid transformations. *The eye deformation field* allows us to transform a point from the observation space into the canonical space. I obtain the color and density of the point in the canonical space based on Eq. (3.1) and learn a unified NeRF model in the canonical space.

Eye Deformation Field. I aim to learn a deformation field which transforms a point $\mathbf{x}_o$ in the observation space into the canonical space. Intuitively, I 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, I propose a solution that involves applying various transformations, including head rotation and eyeball rotation, and combining them using learnable weights (Lewis et al., 2000):

$$T_{\text{rigid}}(\mathbf{x}_o, p) = \sum_{i=1}^4 w_o^i(\mathbf{x}_o) (R_i \mathbf{x}_o + t_i). \quad (3.6)$$

Note that I 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 , I learn w_c^i as described in (Weng et al., 2022), from which w_o^i can be computed:

$$w_o^i(\mathbf{x}_o) = \frac{w_c^i(R_i \mathbf{x}_o + t_i)}{\sum_{k=1}^4 w_c^k(R_k \mathbf{x}_o + t_k)}. \quad (3.7)$$

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, I 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, I incorporate pose information $\mathbf{p}$ as supplementary inputs to the MLP. Consequently, the complete eye deformation field can be expressed as follows:

$$\mathbf{x}_c = \hat{\mathbf{x}} + T_{NR}(\hat{\mathbf{x}}, \mathbf{p}), \quad (3.8)$$

where $\hat{\mathbf{x}} = T_{rigid}(\mathbf{x}_o, \mathbf{p})$ and $\mathbf{x}_c$ represents the position in the canonical space. Finally, I input the position $\mathbf{x}_c$ into a vanilla NeRF to obtain the color and density as Eq. (3.1).

3.2.3 Training

I 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, I adopt a patch-based sampling approach (Schwarz et al., 2020). I randomly sample six patches with size 32×32 pixels from input images and 128 points for each ray. I use LPIPS loss (Zhang et al., 2018) ($\mathcal{L}_{LPIPS}$) and MSE loss ($\mathcal{L}_{MSE}$) for training as

$$\mathcal{L}_{DeNeRF} = \mathcal{L}_{LPIPS}(\mathbf{P}, \hat{\mathbf{P}}) + \lambda \mathcal{L}_{MSE}(\mathbf{P}, \hat{\mathbf{P}}), \quad (3.9)$$

where $\mathbf{P}$ is the image patch and $\hat{\mathbf{P}}$ is the ground truth. I 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, I introduce an eye mask (Yu et al., 2018) to guide the ray sampling in DeNeRF. Specifically, I assign a larger sampling ratio to the eye region than to other facial parts, thus directing more attention to the eye region.

For training, I 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.

3.3 Experiments

3.3.1 Setup

Data Preparation. I select ETH-XGaze (Zhang et al., 2020) 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. I 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. I select 9 frames which roughly cover the gaze in nine different directions for each subject. I remove the view under extreme pose and each frame finally provides 13-view images, *i.e.*, I use 9×13 images for training on one subject.

Evaluation Metrics. I conduct qualitative and quantitative comparison to demonstrate image quality, where SSIM (Wang et al., 2004), PSNR, and LPIPS (Zhang et al., 2018) are reported for quantitative comparison. I also show the advantage of our method in a downstream task. I render images for data augmentation in gaze estimation task. I use angular degree error for gaze estimation metric (Cheng et al., 2021; Wang et al., 2023a).

3.3.2 Face Rendering under Novel Pose and Gaze

I present the qualitative results for image generation under novel gazes and head poses in Fig. 3.3. 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

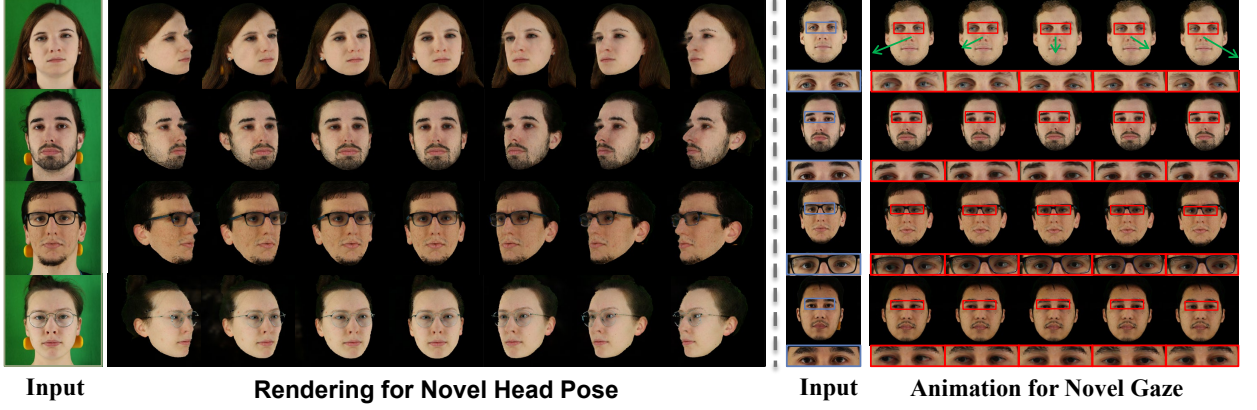


Figure 3.3: Face rendering under novel gaze and novel head poses from DeNeRF. The left images show the rendering under novel head pose. DeNeRF preserves great multi-view consistency. We also use the DeNeRF to render images under novel gaze in the right image. DeNeRF generates high-fidelity face images in a large range of head pose and gaze directions.

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.

3.3.3 Comparison with Face Rendering Methods

I conduct a comprehensive comparison of our method with the SOTA methods in 2D and 3D face rendering with eye animation, STED (Zheng et al., 2020) and GazeNeRF (Ruzzi et al., 2023), and a NeRF-based face rendering approach called HeadNeRF (Hong et al., 2022), 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

Table 3.1: The quantitative comparison between DeNeRF and other SOTA methods in terms of rendered image quality.

Methods	SSIM $\uparrow$	PSNR $\uparrow$	LPIPS $\downarrow$
STED (Zheng et al., 2020)	0.726	17.279	0.306
HeadNeRF (Hong et al., 2022)	0.718	15.262	0.300
GazeNeRF (Ruzzi et al., 2023)	0.728	15.322	0.297
Ours	0.732	19.144	0.265

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. 3.4 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.

In addition, I show the quantitative results in Table 3.1. 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

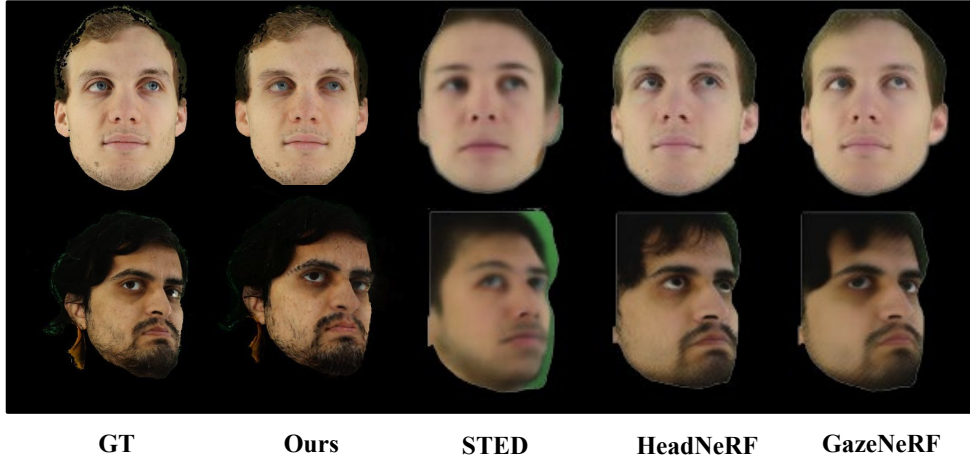


Figure 3.4: The qualitative comparison between DeNeRF and other SOTA methods.

space, which makes training much easier.

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.

3.3.4 Improving Gaze Estimation Performance

My method is capable of rendering face images from novel views and eyeball poses. Furthermore, I 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, I can accurately calculate the final gaze direction.

To demonstrate the capacity of our model in enhancing gaze estimation, I conducted experiments on four renowned gaze datasets including EyeDiap (Funes Mora et al., 2014a), MPIIFace (Zhang et al., 2017c), RT-Gene (Fischer et al., 2018), and Gaze360 (Kellnhofer et al., 2019b). Due to the scarcity of annotated images per subject in the test person specific

Table 3.2: 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.

	EyeDiap	MPIIFace	RT-Gene	Gaze360
×	41.829	36.441	42.172	17.709
✓	34.074	28.567	34.632	16.351

set of ETH-XGaze, I randomly select seven subjects from the train set of ETH-XGaze for the experiment. I 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, I evaluated the performance of the augmented versions on the annotated data of each subject. I trained the state-of-the-art gaze estimator, GazeTR (Cheng et al., 2022), separately on each of the four original datasets and their augmented versions for comparison.

I present the gaze error in Table 3.2. 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.

3.3.5 Multi-view Face Tracking Visualization

I show the visualization of face tracking in Fig. 3.5. The detected landmarks used for supervision are estimated using the approach proposed by Bulat et al. (2017). Our multi-view

face tracker takes multi-view images and corresponding camera poses as inputs. I initialize the textured FLAME face model with zero parameters and project the face model to all views. I render face images and obtain face landmarks and pupil centers from the fitted FLAME model. Then I perform pixel-wise alignment in the face appearance and position alignment for facial landmarks and pupil centers. The result shows that our multi-view face tracking can get accurate facial parameters while keeping multi-view consistency.



Figure 3.5: Fitting results of multi-view face tracking.

3.4 Conclusion

In this chapter, I 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, I perform face tracking on the multi-view images to capture the eye pose. Then, I fit a parametric 3D face model, FLAME, considering the multi-view

consistency. Next, I construct a rotation-aware manifold to model the rotation of the eyeball. I define a canonical space for DeNeRF and transform 3D points from different observation spaces into this space. Finally, I 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. I evaluate our method on the ETH-XGaze dataset and show that it can render high-fidelity face images from novel viewpoints and eye poses. I also mix our rendered images with the original training set for data augmentation, which further improves performance. In future work, I aim to reduce the requirement for multi-view images and lower computational costs.

Chapter Four

Real-Time Gaze-Controllable 3D Face Generation from a Single Image

This chapter presents work that is under review by The IEEE/CVF Conference on Computer Vision and Pattern Recognition 2025.

Gaze is one of the most important facial features and it conveys human attention and intention in interaction. Gaze redirection involves redirecting the gaze of a face image to a given target direction without changing the identity. It has diverse applications, including virtual reality (Pai et al., 2016; Mania et al., 2021; Henrysson et al., 2005; Zhang et al., 2022; Zheng et al., 2023), digital human modeling (Demirel et al., 2007; Jack et al., 2015; Leyvand et al., 2006; Tse et al., 2022), and computer-generated (CG) filmmaking (Yang et al., 2022; Vanezis et al., 1989; Blanz et al., 2004). Besides, it is also used to generate training data for downstream tasks such as gaze estimation as gaze data collection is complex and time-consuming (Wang et al., 2023c).

Existing gaze redirection methods can be broadly divided into two categories: 2D-based and 3D-based, depending on whether they incorporate 3D representations. 2D-based methods achieve gaze redirection either by warping pixels in the input image (Ganin et al., 2016) or by

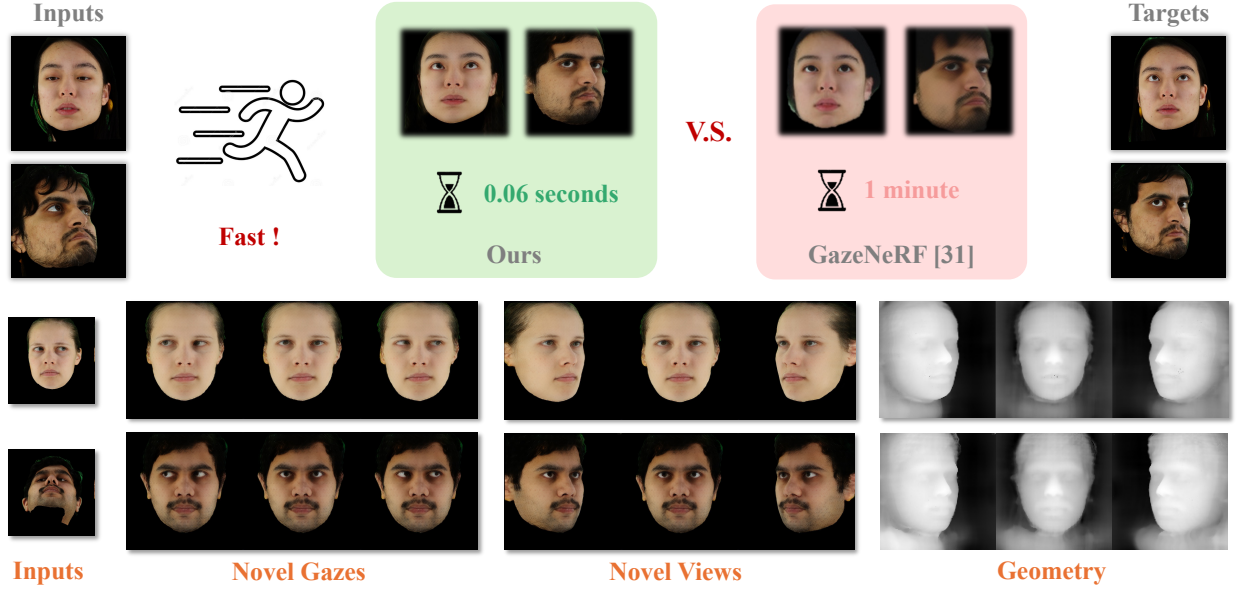


Figure 4.1: 3D-aware gaze redirection results from our proposed LiveGaze, which generates photo-realistic face images under novel gazes and views with good 3D consistency in real time. Compared to the state-of-the-art 3D-aware gaze redirection method GazeNeRF (Ruzzi et al., 2023), which requires approximately one minute during inference, our approach achieves real-time performance at **60ms** while maintaining superior image quality.

generating new gaze images through deep generative models such as Generative Adversarial Networks (GANs) (He et al., 2019; Jindal et al., 2023), encoder-decoder networks (Park et al., 2019a), and Variational Autoencoders (VAEs) (Zheng et al., 2020). While effective to some extent, these methods do not capture the inherently 3D nature of gaze redirection, resulting in suboptimal performance under larger head poses.

3D-based methods, on the other hand, construct a 3D representation of each input face image using techniques like the neural radiance field (NeRF) (Mildenhall et al., 2020). Once trained, these models can generate a full 3D face and, by adjusting camera poses, produce images with varied head orientations, ensuring strong 3D consistency across a wide range of poses. Among these, GazeNeRF (Ruzzi et al., 2023) is the state-of-the-art, employing

two separate multilayer perceptrons (MLPs) to model the radiance fields for the face and eyes independently. GazeNeRF generates novel views using latent codes and gaze labels, but during inference, it requires fine-tuning and updating learnable latent codes before rendering (Hong et al., 2022; Ruzzi et al., 2023), a process that is time-consuming and delays gaze redirection. Balancing 3D consistency with real-time performance, therefore, remains an open challenge in gaze redirection.

In this chapter, I tackle the challenge of real-time 3D-aware gaze redirection by distilling a 3D portrait generator into a lightweight module (Trevithick et al., 2023; Chan et al., 2022) that requires only a single image as input. I introduce a novel method, *LiveGaze*, for real-time gaze redirection with 3D awareness.

As illustrated in Fig. 4.2, our streamlined module takes an image and gaze label as inputs. Drawing inspiration from Stable Diffusion (Rombach et al., 2022), I incorporate the gaze label by merging it with image features through cross-attention mechanism. The fused features are then encoded into a triplane representation (Chan et al., 2022) for volumetric rendering, enabling the model to be trained with a reconstruction loss. Direct optimization of appearance and shape in a compact model is challenging, particularly when deriving 3D geometry from a single image (Liu et al., 2023; Liu et al., 2024). To address this, I distill the prior knowledge of face geometry from a 3D portrait generator into our module, optimizing appearance and adjusting shape as needed.

Our system achieves *real-time* 3D-aware gaze redirection through a feedforward network, processing each frame in just **61ms** on a standard consumer GPU. Extensive quantitative and qualitative evaluations validate our approach, and a series of ablation studies confirm the effectiveness of our design choices. Compared with existing methods (Ruzzi et al., 2023; Hong et al., 2022), LiveGaze offers superior image quality with a significant boost in inference speed.

In summary, our contributions are as follows:

- I present a lightweight, real-time 3D-aware gaze redirection module that utilizes a cross-attention-based gaze fusion mechanism, maintaining strong 3D consistency.
- I introduce the distillation of 3D face priors from 3D GANs into a lightweight module, enhancing 3D face generation quality.
- Our method, tested on the ETH-XGaze dataset, surpasses state-of-the-art methods in both inference speed and image quality.

4.1 Related Work

4.1.1 Gaze Redirection

Gaze redirection methods can generally be divided into two categories: 2D-based methods and 3D-based methods.

Among 2D-based methods, Deepwarp (Ganin et al., 2016) employs warping maps learned from pairs of eye images with different gaze directions, which requires extensive annotated data. To reduce this reliance on annotated real data, Yu et al. (2019a) incorporate a pretrained gaze estimator with synthetic eye images, further refined by Yu et al. (2020) with an unsupervised gaze representation learning network. GAN-based approaches, such as the one proposed by He et al. (2019), enable gaze redirection by leveraging generative models. FAZE introduces an encoder-decoder framework that encodes eye images into latent vectors, which are then manipulated with rotation matrices to produce synthetic images featuring redirected gaze. ST-ED (Zheng et al., 2020) builds on this by disentangling latent representations to perform both head and gaze redirection for full-face images, achieving highly accurate results. Expanding on ST-ED, ReDirTrans (Jin et al., 2023) projects edited embeddings back into the original

latent space, allowing for attribute replacement with minimal impact on other features and preserving the latent distribution. While effective, 2D-based methods often struggle with 3D consistency, as they lack an explicit 3D facial representation.

In contrast, 3D-based methods offer improved 3D consistency. EyeNeRF (Li et al., 2022a) combines explicit surface modeling for the eyeball with implicit volumetric representations of surrounding eye structures, enabling high-fidelity gaze redirection with photorealistic effects using a minimal setup of lights and cameras. GazeNeRF (Ruzzi et al., 2023) employs a two-stream MLP architecture to separately model the face and eye regions via neural radiance fields, allowing for independent manipulation of the eyeball orientation. Additionally, HeadNeRF (Hong et al., 2022) integrates gaze labels as conditional inputs to support gaze redirection. Despite their robust 3D consistency, these methods often require complex, resource-intensive models, limiting their real-time applicability. To address these limitations, our proposed method distills 3D facial geometry from 3D GANs into a streamlined, 3D-aware gaze redirection module, achieving real-time performance.

4.1.2 3D-Aware One-Shot Portrait Generation

Keeping good 3D consistency is challenging when generating a 3D-aware portrait from one single image. Efficient 3D representations, like neural radiance fields (Mildenhall et al., 2020) or 3D meshes (Li et al., 2017), are commonly used to impose geometric constraints on neural rendering to enhance view consistency. ROME (Meng et al., 2022a) is a mesh-based method that estimates the head mesh as well as the neural texture. Although it achieves good view consistency, the limited resolution of polygonal meshes restricts the neural renderer’s ability to produce high-fidelity geometric and appearance details. HeadNeRF (Hong et al., 2022) and MofaNeRF (Zhuang et al., 2022) support direct control of the head pose of the generated images with NeRF-based parametric model. However, these methods require intensive test-time optimization, limiting their usability in real-time applications, and often struggle to

preserve the source identity due to the use of compact latent vectors. EG3D (Chan et al., 2022) utilizes an advanced triplane-based neural field representation that efficiently encodes the 3D structure and appearance of an avatar’s head, enabling detailed and structured 3D modeling. However, it still needs to update the latent codes for one specific head via GAN inversion (Abdal et al., 2019; Abdal et al., 2020; Bai et al., 2022). Live3D (Trevithick et al., 2023) distills a lightweight neural network from EG3D to realize 3D portrait generation from one single image in real time. Our work relies on the structure of Live3D and inherits its real-time performance. While working efficiently, it lacks the disentanglement between the appearance and gaze direction and is unable to impose various driving gazes onto the input. To address the limitation, I propose a gaze fusion module to inject gaze as a condition for gaze controlling and train the model with pairs of images with different gazes. In addition, I further propose a distillation model of 3D face priors from the pre-trained Live3D model (Trevithick et al., 2023) to supplement shape information.

4.2 Methodology

4.2.1 Preliminary

3D-aware GANs (Chan et al., 2022; Trevithick et al., 2023) have proven effective in generating highly realistic 3D images by leveraging collections of single-view images. To improve computational efficiency and image quality of 3D GANs, EG3D (Chan et al., 2022) proposes an efficient triplane-based 3D representation for portrait reconstruction. However, EG3D must do GAN inversion before 3D reconstruction given an image. The GAN inversion makes the whole inference process slow. To address this limitation, I rely on Live3D (Trevithick et al., 2023), a state-of-the-art single-image 3D portrait reconstruction model that distills a lightweight module from the pre-trained EG3D model.

Live3D takes an image $\mathbf{I}$ as input instead of latent codes. It encodes the input image with two

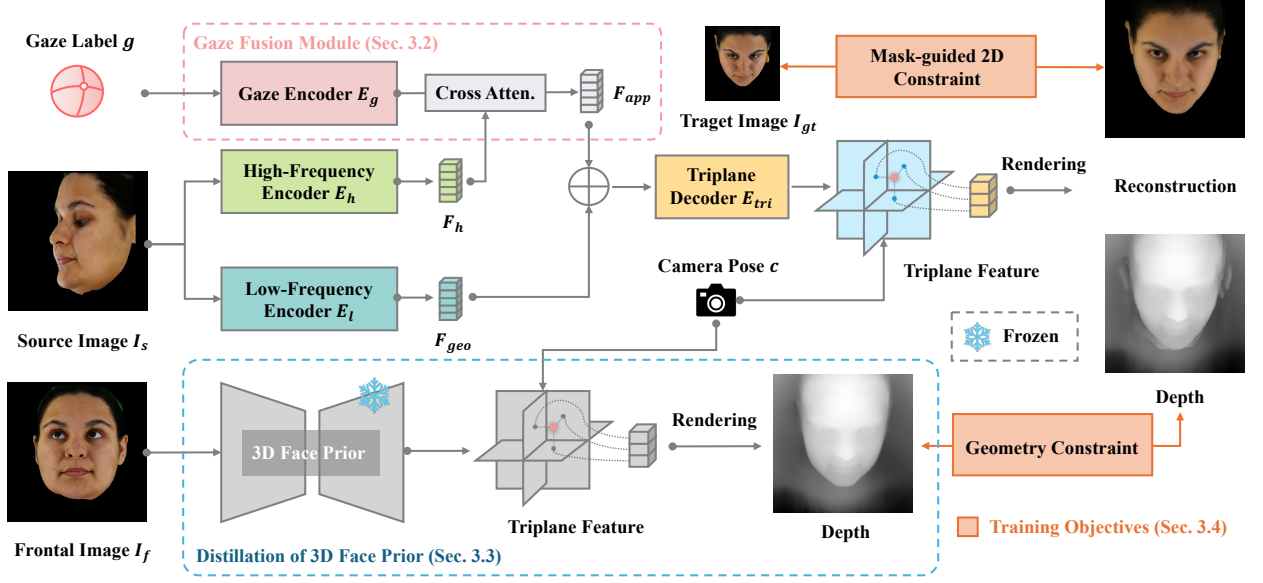


Figure 4.2: Overview of the LiveGaze pipeline. During training, the model extracts high-frequency and low-frequency features from the source image and fuses them with gaze information to generate a 3D face representation. This is used to render gaze-redirected images with depth maps, guided by 2D and 3D geometry constraints. During inference, the model enables real-time gaze redirection using only a single image and gaze labels.

encoders, E_h and E_l , to extract the features with high frequency and low frequency. Then, a ViT-based decoder E_{tri} is used to transfer the concatenated features into the triplane $\mathbf{T}$:

$$\mathbf{T} = E_{tri}(E_h(\mathbf{I}), E_l(\mathbf{I})). \quad (4.1)$$

The triplane $\mathbf{T}$ is followed with a volume renderer to render low-resolution images under the given camera pose c . High-resolution images are obtained via a super-resolution module which is the same as EG3D (Chan et al., 2022).

4.2.2 Network Architecture

The two encoders, E_h and E_l , are designed to extract high-frequency features and low-frequency features from the input image in Live3D. 3DPE (Bai et al., 2025) analyzes the two kinds of features by separately disabling the features and visualizing them. They find the

features from E_h keep appearance information while losing the shape and the features from E_l retain the geometry while fail to capture the appearance. Therefore, the high-frequency features are also referred to as appearance features and the low-frequency features are referred to as geometry features. The human eyeball is close to a sphere and the geometric changes in eye region are tiny. Compared with geometric changes, redirecting gaze usually causes bigger changes in appearance. To this end, I choose to fuse the gaze condition with the high-frequency features from E_h to modify the face appearance and leave face geometry unchanged.

Injecting Gaze as Condition. Given a source image $\mathbf{I}_s$ and a gaze label $\mathbf{g}$, I first use the gaze label encoder E_g to create gaze embeddings whose dimensions are aligned with the high-frequency features $\mathbf{F}_h$. Inspired by Stable Diffusion (Rombach et al., 2022), I fuse the gaze embedding and $\mathbf{F}_h$ via cross-attention mechanism. Specifically, I add a cross-attention layer after E_h to obtain the modified appearance features $\mathbf{F}_{app}$:

$$\mathbf{F}_{app} = \text{CrossAttention} (E_h(\mathbf{I}_s), \mathbf{F}_h), \quad (4.2)$$

In the cross attention, $\mathbf{F}_h$ functions as the query, while $E_g(\mathbf{g})$ is used as both the key and value. The modified appearance features $\mathbf{F}_{app}$ and geometry feature $\mathbf{F}_{geo}$ are subsequently fed into the decoder E_{tri} to infer the triplane $\mathbf{T}_g$:

$$\mathbf{T}_g = E_{tri} (\mathbf{F}_{app}, \mathbf{F}_{geo}), \quad (4.3)$$

The gaze-redirected image $\mathbf{I}_g$ and the depth map $\mathbf{D}_g$ under the target camera pose $\mathbf{c}$ are generated by a triplane renderer G same as EG3D (Chan et al., 2022):

$$\mathbf{I}_g, \mathbf{D}_g = G (\mathbf{T}_g, \mathbf{c}). \quad (4.4)$$

4.2.3 Distillation of 3D Face Prior

Directly optimizing appearance and shape from a single image is challenging, as it lacks sufficient constraints to accurately represent a 3D scene. In this work, I distill face geometric

priors from a pre-trained model to enhance the quality of the generated images. Notably, the pre-trained model is typically not designed for gaze redirection. I found that leveraging face geometric priors also improves the overall quality of the generated face images.

In detail, I distill face geometric prior from a pre-trained 3D portrait generation model (Trevithick et al., 2023). I find the model has the best performance when the face in the input image is oriented frontally. Therefore, I input a frontal image whose identity and gaze are the same as the target image to the pre-trained model and obtain a NeRF representation like the triplane feature $\mathbf{T}$. The triplane feature enables the rendering of multi-view images and dense depth maps by predicting the color and density of 3D points along camera rays. In this work, I focus solely on depth map generation and apply a distillation loss on the predicted depth, as the synthesized images do not always maintain appearance consistency with the target images. Specifically, I sample 3D points along the camera rays $\mathbf{r}$ under the target camera pose $\mathbf{c}$. For each ray $\mathbf{r}$, the depth value $D(\mathbf{r})$ is computed as:

$$D(\mathbf{r}) = \sum_{i=1}^N W_i (1 - \exp(-\sigma_i \delta_i)) \mathbf{d}_i, \quad (4.5)$$

and

$$W_i = \exp\left(-\sum_{j=1}^{i-1} \sigma_j \delta_j\right). \quad (4.6)$$

N is the number of samples along the ray, $\mathbf{d}_i$ is the distance between the i -th sample and the camera, σ_i is the density of the i -th sample, and δ_i is the distance between the i -th and $(i + 1)$ -th samples. The depth map $\mathbf{D}$ under the camera pose $\mathbf{c}$ is obtained by integrating all depth values. I compute the ground truth depth map $\mathbf{D}_{gt}$ from the pre-trained model and the depth map $\mathbf{D}_g$ from our model, applying an L1 loss to enforce depth consistency between them.

$$\mathcal{L}_{\mathcal{D}} = \|\mathbf{D}_g - \mathbf{D}_{gt}\|_1. \quad (4.7)$$

4.2.4 Training Objectives

I train the model using a pair of images ($\mathbf{I}_s$ and $\mathbf{I}_{gt}$) with the same identity and different gazes. The 3D face prior is obtained from an additional frontal image $\mathbf{I}_f$ with the same identity and gaze as $\mathbf{I}_{gt}$. I optimize our model using the following objective function:

$$\mathcal{L} = \lambda_{\mathcal{R}}\mathcal{L}_{\mathcal{R}} + \lambda_{\mathcal{D}}\mathcal{L}_{\mathcal{D}} + \lambda_{\mathcal{P}}\mathcal{L}_{\mathcal{P}}, \quad (4.8)$$

where $\mathcal{L}_{\mathcal{R}}$, $\mathcal{L}_{\mathcal{D}}$, $\mathcal{L}_{\mathcal{P}}$ represent the reconstruction loss, depth loss, and perceptual loss, respectively. As illustrated in Fig. 4.2, the reconstruction loss defines our mask-guided 2D constraint and the depth loss builds our geometry constraint.

Mask-Guided 2D Reconstruction Loss. To achieve realistic gaze redirection, I introduce a 2D reconstruction loss that minimizes the differences between the generated image $\mathbf{I}_g$ and the target image $\mathbf{I}_{gt}$ in pixel level. To improve the generation quality of eyes, I apply a face region mask and an eye region mask to the reconstruction loss and enhance eye region reconstruction by setting the loss in eye region with a larger weight:

$$\mathcal{L}_{\mathcal{R}} = \lambda_{\mathcal{R}_f}\mathcal{L}_{\mathcal{R}_f} + \lambda_{\mathcal{R}_e}\mathcal{L}_{\mathcal{R}_e}, \quad (4.9)$$

where $\mathcal{L}_{\mathcal{R}_f}$ and $\mathcal{L}_{\mathcal{R}_e}$ stand for face region reconstruction loss and eye region reconstruction loss respectively. I set $\lambda_{\mathcal{R}_e} > \lambda_{\mathcal{R}_f}$ in our case.

The face region reconstruction loss $\mathcal{L}_{\mathcal{R}_f}$ is formulated as:

$$\mathcal{L}_{\mathcal{R}_f} = \frac{1}{|M_f \odot \mathbf{I}_{gt}|} \|M_f \odot (\mathbf{I}_g - \mathbf{I}_{gt})\|_1, \quad (4.10)$$

where M_f is the face region mask and $\odot$ denotes the pixel-wise Hadamard product operator.

And the eye region reconstruction loss $\mathcal{L}_{\mathcal{R}_e}$ is formulated as:

$$\mathcal{L}_{\mathcal{R}_e} = \frac{1}{|M_e \odot \mathbf{I}_{gt}|} \|M_e \odot (\mathbf{I}_g - \mathbf{I}_{gt})\|_1, \quad (4.11)$$

where M_e is the eye region mask.

Perceptual Loss. Perceptual loss (Johnson et al., 2016) is to assess the quality of generated images by comparing high-level feature representations rather than pixel-level differences. It leverages a pre-trained network (such as VGG (Simonyan et al., 2014)) to capture the semantic and structural similarity between generated and target images. The models trained with perceptual loss tend to produce more visually realistic and detailed outputs by focusing on perceptual similarity, which better aligns with human visual perception. It has been widely used in image synthesis (Hong et al., 2022; Jindal et al., 2023) and approved as effective. I utilize a perceptual loss function to ensure perceptual alignment between the gaze-redirectioned image $\mathbf{I}_g$ and the target image $\mathbf{I}_{gt}$:

$$\mathcal{L}_{\mathcal{P}} = \sum_i \frac{1}{|\phi_i(\mathbf{I}_{gt})|} \|\phi_i(\mathbf{I}_g) - \phi_i(\mathbf{I}_{gt})\|_1, \quad (4.12)$$

where ϕ_i denotes the i -th layer of a VGG16 (Simonyan et al., 2014) network pre-trained on ImageNet (Krizhevsky et al., 2012).

During inference, our model only takes a single 2D portrait image and a specified gaze direction as input and produces a gaze-redirectioned, triplane-based 3D face NeRF. Our model enables photorealistic view synthesis, allowing for highly realistic visualizations from multiple perspectives.

4.3 Experiments

To demonstrate the effectiveness of LiveGaze, I train our model on the ETH-XGaze (Zhang et al., 2020) dataset and compare it to the state-of-the-art methods regarding efficiency and image quality both qualitatively and quantitatively. Then I show the results of novel view synthesis and gaze redirection. Finally, I validate our design choices via ablation studies.

4.3.1 Experimental Setup

Dataset. I train our model on the ETH-XGaze dataset which is widely used in gaze redirection task (Ruzzi et al., 2023; Wang et al., 2023c) and gaze estimation task (Cheng et al., 2023; Wang et al., 2023a). ETH-XGaze is a large-scale gaze dataset comprising high-resolution images that capture a wide range of head poses and gaze directions. It was collected using a multi-camera setup under various lighting conditions to enhance diversity. The training set includes 756K frames across 80 subjects, with each frame featuring images from 18 distinct camera angles. Additionally, a personalized test set contains 15 subjects, with each providing 200 images along with accurate ground truth gaze labels for evaluation. Following GazeNeRF (Ruzzi et al., 2023), I train LiveGaze with 14.4K images from 10 frames per subject, 18 images with different views per frame, and 80 subjects on the ETH-XGaze training set. I test all the models on the personalized test set.

Data Preparation. The resolution of raw images in ETH-XGaze (Zhang et al., 2020) is 6Kx4K. I first normalize the raw images using the method proposed by Zhang et al. (2020) and get the normalized head poses and gaze directions. The normalized distance between the camera and the center of the face is fixed at 950mm and the focal length in the normalized camera projection matrice is set to 1600. To align the data format with Live3D (Trevithick et al., 2023) and EG3D (Chan et al., 2022), I resize the normalized images to 512x512 and estimate camera poses (Deng et al., 2019). To apply our mask-guided 2D constraint, I use the face parsing model (Yu et al., 2018) to segment the whole and the eye region. I use the detected landmarks (Bulat et al., 2017) to do the segmentation when the face parsing model does not work when processing some challenging images. I convert the provided gaze labels into pitch-yaw labels in the head coordinate system for convenience of gaze controlling in 3D space.

Implementation Details. Our model is trained in an end-to-end manner. I employ Adamw

(Loshchilov et al., 2017) as our optimizer for whole model. The learning rates are set to $1e^{-5}$ and $1e^{-5}$ for the encoding part and the rendering part respectively. I train our model with a batch size of 4 for 50 epochs. I empirically set the loss coefficients ($\mathcal{L}_{\mathcal{R}}$, $\mathcal{L}_{\mathcal{D}}$, $\mathcal{L}_{\mathcal{P}}$) in equation (4.8) to 1, 1, 0.8 respectively. The coefficients ($\mathcal{L}_{\mathcal{R}_f}$ and $\mathcal{L}_{\mathcal{R}_e}$) of in equation (4.9) are assigned with 1 and 2 separately. It takes around 18 hours to train the whole model on two NVIDIA A100 GPUs with 40GB memory.

The personalized test set in ETH-XGaze includes 200 labeled images for each subject. I split the personalized test set into an input group and a target group following GazeNeRF (Ruzzi et al., 2023). The input group contains 100 images for each subject and the target group includes the other 100 images. I train our model with 80 subjects in the train set of ETH-XGaze first and then finetune the model with images from the input group for 10 epochs. I generate the images in the target group during evaluation.

Evaluation Metrics. I evaluate our model with various metrics regarding model efficiency and generated image quality. To evaluate the efficiency of models, I report the encoding time and rendering time measured on a single NVIDIA 3090 GPU in the inference stage with an average of 100 samples. To evaluate the quality of generated image, I report four widely used metrics including Structure Similarity Index(SSIM) (Wang et al., 2004), Peak Signal-to-Noise Ratio(PSNR), Learned Perceptual Image Patch Similarity(LPIPS) (Zhang et al., 2018), and Fréchet Inception Distance(FID).

4.3.2 Efficient 3D-Aware Gaze Redirection

Baseline Methods. I compare our model against the state-of-the-art gaze redirection methods including 2D-based method ST-ED (Zheng et al., 2020) and 3D-based method GazeNeRF (Ruzzi et al., 2023). ST-ED realizes gaze redirection on full-face images by disentangling latent vectors with a novel self-transforming encoder-decoder architecture.

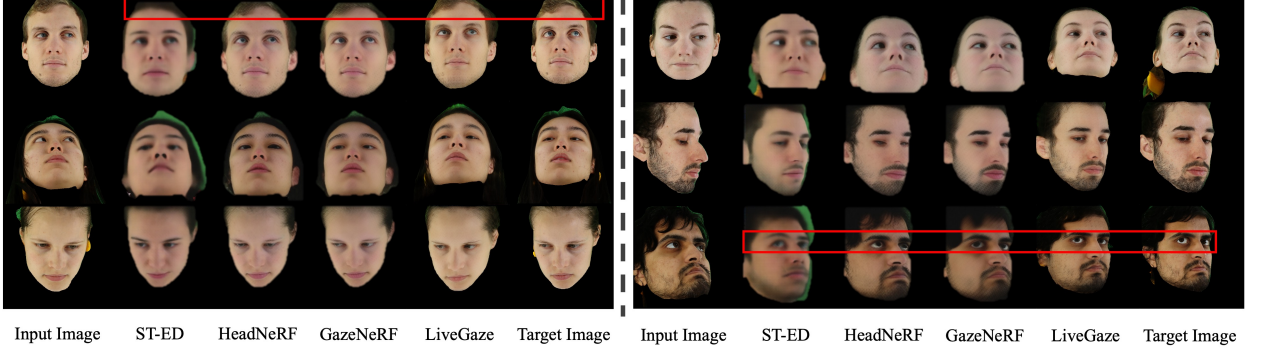


Figure 4.3: Qualitative comparisons. I conduct the comparison on ETH-XGaze dataset. The background is removed by applying face masks. Please note the hair and eye generations in the two red boxes. Our model generates results with higher quality.

GazeNeRF disentangles eye and face with two-stream MLPs and achieves 3D-aware gaze redirection based on NeRF representation. I also compare our model with HeadNeRF, a state-of-the-art NeRF-based 3D portrait generation model. It is adapted to gaze redirection task by simply adding two-dimension gaze labels as additional input.

Qualitative Results. I show the qualitative results of the comparison with SOTA methods in Fig. 4.3. Following GazeNeRF (Ruzzi et al., 2023), I pair the images with the different gazes from the personalized test set of ETH-XGaze to get the input images and target images. Our model takes one single image as well as a target gaze label as input and generates a photorealistic gaze-redirectioned image. As shown in Fig. 4.3, ST-ED (Zheng et al., 2020) suffers from preserving identity information tending to generate similar faces with different inputs. Besides, the results from ST-ED preserve the unmasked green background by mistake which is barely found in 3D-based methods (HeadNeRF, GazeNeRF, and our LiveGaze). 2D-based methods only learn a mapping from the input image and gaze label to the target image, while 3D-based methods are trained to build 3D face representations by integrating extensive multi-view information. It explains the robustness of 3D-based methods in handling defective inputs, which verifies our choice of NeRF-based architecture. Even though HeadNeRF (Hong

Table 4.1: Quantitative comparisons regarding image quality (SSIM, PSNR, LPIPS, FID) and inference speed (Encode Time, Render Time, Total Time). For fairness, I report inference speed metrics only for 3D methods. Image quality is evaluated on the personalized test set from ETH-XGaze, while inference speed is averaged over 100 samples on a single NVIDIA 3090 GPU.

Method	3D-based	SSIM $\uparrow$	PSNR $\uparrow$	LPIPS $\downarrow$	FID $\downarrow$	Enc Time $\downarrow$	Render Time $\downarrow$	Total Time $\downarrow$
ST-ED		0.726	17.530	0.300	115.020	-	-	-
HeadNeRF		0.720	15.298	0.294	69.487	60s	0.058s	60.058s
GazeNeRF		0.733	15.453	0.291	81.816	60s	0.060s	60.060s
Ours		0.715	19.007	0.272	38.346	0.026s	0.035s	0.061s

et al., 2022) generates face images with the correct identity, it fails to redirect the gaze accurately and loses details in eye regions. GazeNeRF (Ruzzi et al., 2023) generates gaze-redirected images whose gazes are aligned with target images, while it struggles to preserve more facial details including facial texture and fine-grained hair as shown in the red box on the left. In contrast, our model can generate photorealistic face images with extensive details while maintaining the ability to redirect gaze accurately. Notably, our model works in real time which outperforms all the existing 3D-aware methods.

Quantitative Results. I evaluate our model against other state-of-the-art methods in terms of generated image quality, using widely used metrics including SSIM, PSNR, LPIPS, and FID. To ensure fairness, I compare inference speed with other 3D-based methods, examining encoding time, rendering time, and total time. Image quality is assessed on the personalized test set from the ETH-XGaze dataset, and inference speed is measured by averaging results from 100 samples on a single NVIDIA 3090 GPU. I show the quantitative results in Table 4.1. For inference speed, LiveGaze achieves real-time performance in both encoding and rendering stages, with a total processing time of 61ms per image. This is attributed to the efficient triplane-based lightweight module distilled from a pre-trained 3D GAN (Trevithick et al.,



Figure 4.4: Visualization of generated results under novel views and gazes.

2023), as well as the avoidance of the inversion process by requiring only a single image as input. In contrast, both HeadNeRF and GazeNeRF are based on the same parametric head model with NeRF representation. Their inputs are parameters of a specific head instead of images. Therefore, they have to conduct an inversion process to update the parameters with the input image, which takes a great amount of time like one minute. They suffer from slower encoding times due to the involved inversion process. Regarding image quality, LiveGaze beats the other SOTA methods on most metrics (PSNR, LPIPS, FID) and achieves a comparable result on SSIM. Notably, our model outperforms other methods on FID by a large margin.

Table 4.2: Ablation on feature choices. The results indicate that injecting gaze label into low frequency feature cannot achieve competitive performance.

Injecting into	FID ↓	ID ↑	Gaze ↓	Head ↓
Low-Frequency Feature	67.298	38.517	18.973	5.409
High-Frequency Feature	38.346	60.708	9.047	3.631

4.3.3 Face Rendering under Novel Views and Gazes

To showcase the effectiveness of our model in generating 3D-consistent results and achieving consistent gaze redirection, I show the visualization of face rendering under novel views and gazes in Fig. 4.4. I set the gaze as looking forward during the generation under novel views and interpolate the gaze from left to right under a frontal view in the generation under novel gazes. The results demonstrate that our model can generate face images with strong 3D consistency and enables smooth and coherent gaze interpolation. The good performance relies on our expressive triplane-based 3D face representation and the simple but effective gaze fusion module. It is important to note that our model allows the generation of photorealistic face images across a large range of head poses and gaze directions.

4.3.4 Ablation Study

Ablation on Feature Choices. LiveGaze separates high-frequency features (related to appearance) and low-frequency features (related to geometry) before mapping them to the triplane representation. I perform an ablation study on these features by reporting FID for image quality evaluation, identity similarity (ID) for identity preservation evaluation, and gaze/head redirection errors for redirection accuracy evaluation. Identity similarity (ID) is assessed using the face recognition model from FaceX-Zoo (Wang et al., 2021). This model evaluates the discrepancies in identity between the redirected images and the corresponding ground truth images.

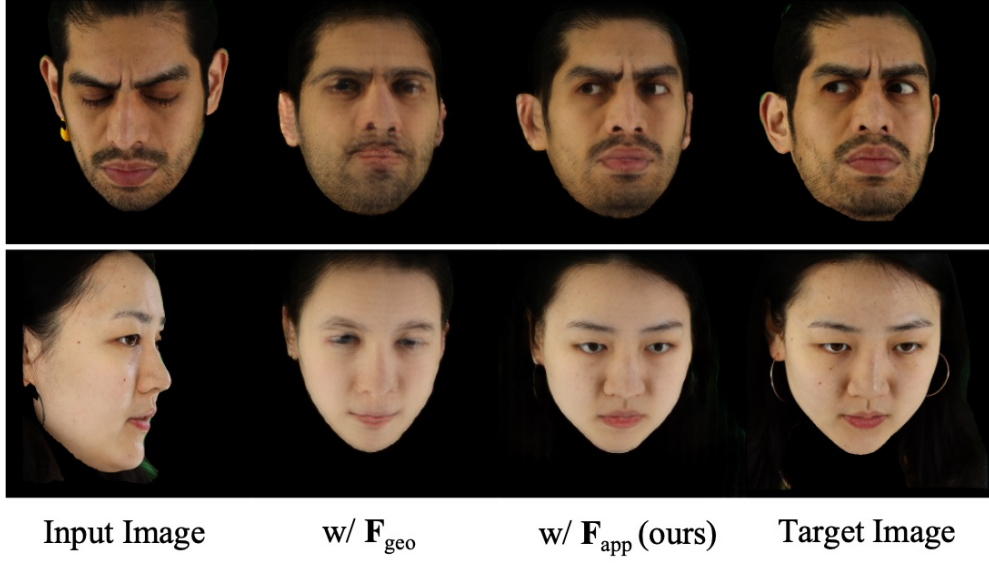


Figure 4.5: Visualization of ablation on fusion feature choices. The results highlight that using only low-frequency geometric features results in blurry images and inaccurate gaze redirection.

The quantitative results are shown in Table 4.2. The results indicate that injecting gaze label into low frequency feature cannot achieve competitive performance. I also try to inject gaze prompt into both low-frequency and high-frequency features, but the training fails to converge due to the difficulty of modifying appearance and geometry simultaneously. We present the generated images with the fusion of low-frequency features in Fig. 4.5. Images generated using only low-frequency features appear blurry across the entire image, with gaze redirection failing to achieve accurate results. When the gaze embedding is fused with low-frequency features, the geometry of the 3D face becomes entangled with the input gaze labels. Ideally, the model should modify only the geometry around the eye region; however, without specific geometric constraints, the model struggles to focus on the eye region alone. Instead, it tends to alter the entire face, leading to noticeable instability in the generated results. In our approach, I fuse the high-frequency features with the gaze embedding while keeping the geometric features unchanged. This enables the model to perform gaze redirection

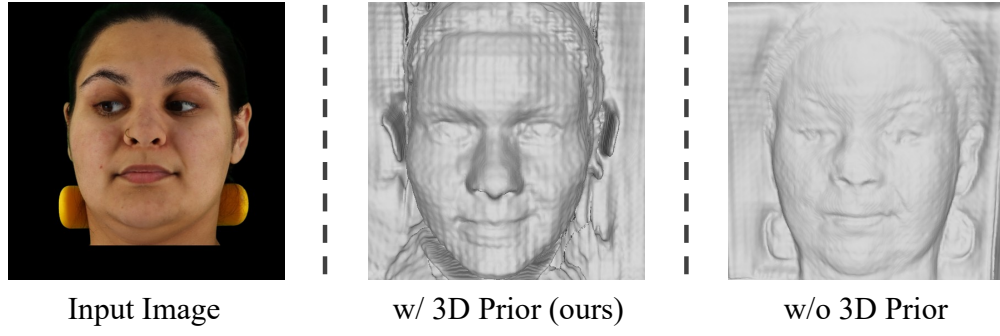


Figure 4.6: Ablation on 3D face prior distillation.

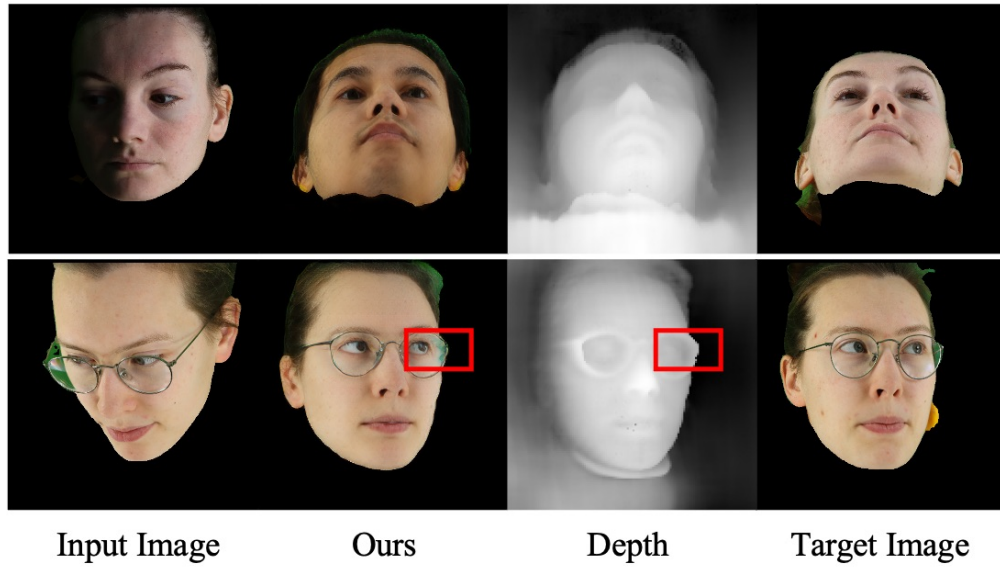


Figure 4.7: Failure cases under challenging conditions.

by adjusting only the appearance of the eye region, ensuring a stable face shape throughout the process.

Ablation on 3D Face Prior Distillation. I conduct an ablation study on 3D face prior distillation and show the reconstructed 3D meshes in Fig. 4.6. Our model, leveraging 3D priors, clearly reconstructs 3D face shapes with high fidelity. In contrast, the mesh generated without 3D priors appears flat across the face, lacking significant depth information.

4.3.5 Limitations

While LiveGaze is capable of generating photorealistic gaze-redirected face images in real time, it faces challenges under certain conditions, such as varying illumination and the presence of glasses. Failure cases are illustrated in Fig. 4.7. In low-light environments, our model may mistakenly perceive faces as belonging to different subjects with darker facial tones. Additionally, its performance declines when processing faces with glasses. As shown in the second row of Fig. 4.7, the model struggles to reconstruct glasses correctly, and their presence interferes with the generation of the eye region. Despite these limitations, it is worth noting that our model successfully redirects gaze correctly in both scenarios, underscoring its robust gaze redirection capability.

4.3.6 Additional Qualitative Results

In Fig. 4.8, I show additional qualitative results comparing our model to the SOTA methods. All the models are evaluated on the personalized test set of the ETH-XGaze dataset. Our model generates images with superior quality while ST-ED (Zheng et al., 2020) encounters difficulties in preserving identity information and both HeadNeRF (Hong et al., 2022) and GazeNeRF (Ruzzi et al., 2023) struggle to retain facial details.

In Fig. 4.9, I show additional qualitative results of generation under novel gazes and novel views. Our model can generate 3D faces with controllable gazes from a single input image. It produces photorealistic face images across a wide range of head poses and gaze directions. The results under novel viewpoints demonstrate the model’s strong 3D consistency throughout the generation process. Additionally, its capability to produce consistent gaze images is validated by the results under novel gaze directions.

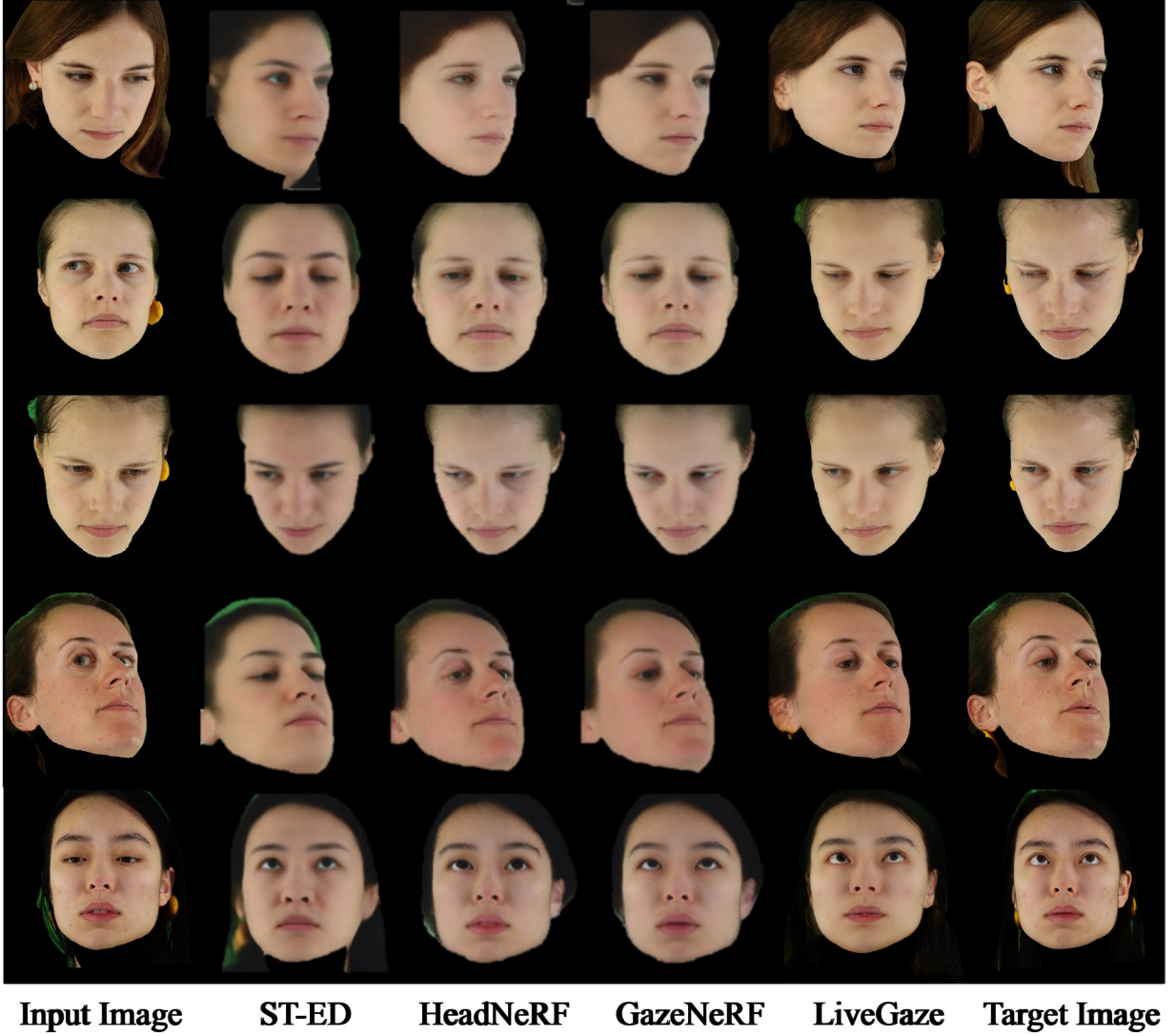


Figure 4.8: Additional visualization of generated images from ETH-XGaze with our LiveGaze, ST-ED, HeadNeRF, and GazeNeRF. The background is eliminated using face masks. Our model can generate photo-realistic images with extensive details. In contrast, ST-ED struggles to preserve identity information. HeadNeRF and GazeNeRF face challenges in maintaining facial details.



Figure 4.9: Additional visualization of generated images under novel gazes and novel views. The upper part is the generated images under novel gazes which are from left to right. The lower part is the generated results under novel views. Our model generates 3D faces with controllable gazes from a single image, producing photorealistic results across diverse head poses and gaze directions while maintaining 3D and gaze consistency.

4.4 Conclusion

In this chapter, I propose LiveGaze, a real-time 3D-aware gaze redirection method from a single image. Our model achieves real-time inference by employing an efficient triplane-based NeRF architecture and distilling the prior knowledge from 3D GANs into a lightweight module. Benefiting from the simple but effective gaze fusion module and the dedicated choice of fused features, our model realizes accurate gaze redirection while maintaining superior image quality. With its exceptional real-time performance and high-quality generation, our model holds great potential for numerous downstream applications, particularly in scenarios with high real-time requirements. While LiveGaze offers significant advantages, its performance declines under challenging conditions like varying lighting and the presence of glasses. Addressing these challenges is reserved for future work.

Chapter Five

Gaze-Controllable Face Generation with Nature Language

This chapter presents work that has been published at the 32nd ACM International Conference on Multimedia 2024 in Melbourne, Australia (Wang et al., 2024).

Face image generation has made significant progress in recent years, owing to the remarkable capabilities of adversarial networks (Goodfellow et al., 2014) and diffusion models (Ramesh et al., 2022; Rombach et al., 2022). It has various applications including virtual reality (Zhang et al., 2022; Zheng et al., 2023), digital human (Jack et al., 2015; Tse et al., 2022; Wang et al., 2023c) and CG film-making (Yang et al., 2022; Blanz et al., 2004). The text-to-face task, which aims to control the face generation using natural language descriptions, has also gained considerable attention (Hou et al., 2023; Sun et al., 2022). This task generates face images corresponding to input text, attracting interest due to its natural interaction and vast practical potential.

Recent text-to-face methods typically utilize natural language to manipulate semantic feature generation, such as “black hair” and “young boy”. These semantic features are manually annotated for each image and provide paired data for the generation model training. In

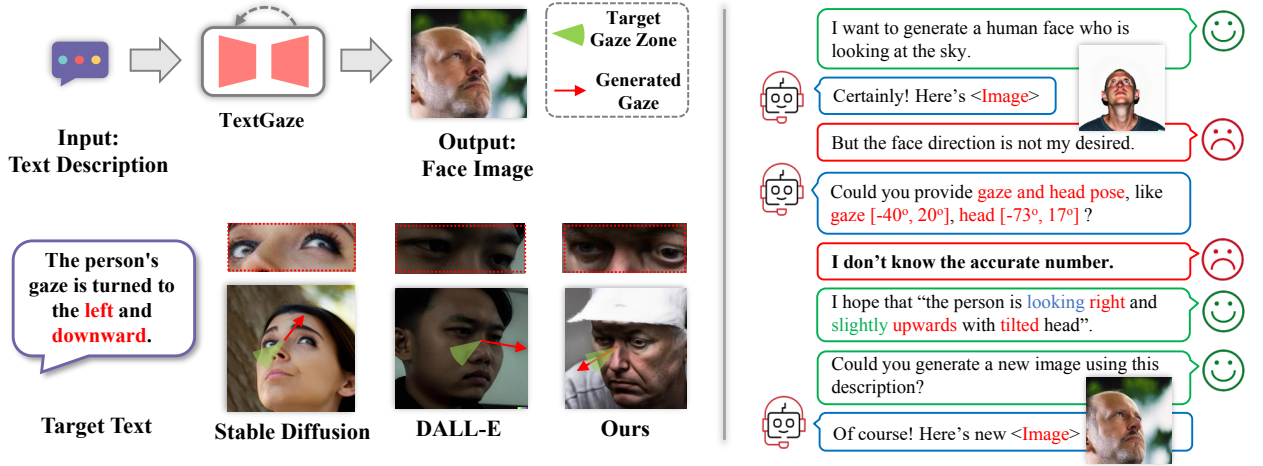


Figure 5.1: **Left:** Overview of our method. Generating gaze images with natural language. Our method significantly outperforms Stable Diffusion and DALL-E. **Right:** Our motivation. People tend to describe head pose and gaze direction using direction and extent words instead of labels.

In addition to these semantic features, human behaviors are also crucial components evident in human faces. Human gaze reflects human intention and is highly demanded in numerous applications (Cheng et al., 2023; Cheng et al., 2024; Bao et al., 2020). However, to the best of our knowledge, there is no work for the text-driven gaze-controllable face generation.

Gaze-controllable face generation methods aim to produce face images that correspond to a specified gaze input. Recent approaches typically involve inputting a specific numerical gaze value (Cheng et al., 2021) to generate the corresponding face images (He et al., 2019; Ruzzi et al., 2023). However, I argue that using gaze values as input may not be user-friendly, particularly for non-expert users as shown in Fig. 5.1. Representing gaze through numerical values could pose challenges for individuals unfamiliar with the technical aspects of gaze estimation. This raises the need for a more user-friendly and intuitive approach to specifying gaze in the generation process.

In this work, I propose a novel gaze-controllable text-to-face generation task, where the

input is the text descriptions of human gaze behavior, such as “the person looks forward”. Unlike using gaze values, employing natural language is more aligned with common expressive habits, making it more accessible and acceptable to a broader range of users. However, a key challenge arises due to the scarcity of text descriptions of gaze behavior. The development of an accurate and diverse text description dataset is highly demanded.

I first introduce a text of gaze (ToG) dataset containing over 90k gaze descriptions. Our approach decomposes human gaze into eye and head rotations, employing uniform sampling to obtain a dense gaze distribution. Manually annotating text descriptions for each gaze sample is time-consuming, and the expertise of annotators greatly influences the annotation outcomes. Inspired by the success of Large Language Models (LLMs) in natural language, I utilize LLMs to annotate gaze behaviors. I construct prompts based on gaze samples and employ LLM to generate diverse text descriptions.

I further introduce a gaze-controllable text-to-face generation method, consisting of a sketch-conditioned face diffusion module and a model-based sketch diffusion model. Our approach utilizes CLIP to obtain feature embeddings from text and introduces a text attention module to extract gaze-related and head-related features from these embeddings. By learning gaze patterns from the extracted features, I convert gaze patterns into face sketches based on a 3D face model. The face diffusion model then generates face images from these sketches. Unlike conventional methods that rely on annotated gaze datasets for training, our method generates diverse face images without the need for such datasets, using face sketches instead.

Overall, our contributions are three-fold.

- I introduce a novel gaze controllable face generation task where the input is a text description of human gaze behavior. I provide the first text-to-gaze dataset containing more than 90k descriptions. LLMs are used to produce accurate and diverse annotations. To the best of our knowledge, our work is the first to leverage LLMs for gaze image

generation.

- I propose a two-stage text-to-face method. Our method leverages the prior information from 3D face models, eliminating the need for gaze dataset. I also propose a text attention module. The module aggregates gaze and head information from text feature embedding, and improve the performance of gaze generation.
- I conduct experiments on FFHQ datasets. The experiment demonstrates our method achieves better gaze-controllable face generation performance than compared methods.

5.1 Related Work

5.1.1 Eye-Control Face Generation

Face generation methods often prioritize controlling semantic information (Sun et al., 2022; Liu et al., 2022), such as appearance and hair color, but frequently overlook gaze control. One closely related task is gaze redirection, where the goal is to generate face images aligned with a specified gaze direction. Deepwarp Kononenko et al. (2017) employs a deep network to learn warping maps between pairs of eye images with different gaze directions. Yu et al. (2019b) utilizes a pretrained gaze estimator and trains a network to generate eye images. He et al. (2019) utilize GAN for gaze redirection and synthesize large-scale gaze data by performing gaze redirection tasks on one eye image. Ruzzi et al. (2023) train a NeRF model based on given gaze directions. They generate 3D face models but produce low-quality rendered images. However, these methods share a common limitation in that they all require images with gaze annotations for training, significantly restricting their applicability. This requirement poses challenges for most image datasets where obtaining an annotated gaze may be impractical.

5.1.2 Text-to-Image Generation

Text-to-image generation shows significant achievement within the context of generative adversarial (Goodfellow et al., 2014). They leverage text description as a condition and train a conditional GAN (Mirza et al., 2014) based on a pair of image-text samples (Reed et al., 2016). Later, stacked structure (Zhang et al., 2017a) and attention mechanism (Xu et al., 2018) are proposed to improve the quality of image generation. Crowson et al. (2022) leverage CLIP (Radford et al., 2021) embeddings to guide image editing use VQ-GAN (Esser et al., 2021). They train the model over large-scale data for diverse generations. Recently, there are many large text-to-image models such as Imagen (Saharia et al., 2022), DALL-E2 (Ramesh et al., 2022), and Stable Diffusion (Rombach et al., 2022). These models leverage the diffusion model and produce unprecedented generation.

5.2 ToG: Text of Gaze Dataset

I utilize large language models (LLMs) to generate text-of-gaze (ToG) dataset. Our dataset contains over 90k text descriptions.

5.2.1 Overview

Text-to-image tasks typically rely on paired text-image data $\{\mathbf{I}_i, \mathbf{T}_i\}$ for training, where $\mathbf{I}_i$ denotes images and $\mathbf{T}_i$ represents the corresponding text description. However, the collection of such data is both time-consuming and costly, as annotators are required to manually generate text descriptions for each image.

Recently, LLMs attract significant attention and demonstrate remarkable capabilities in natural language. This inspires us to explore the potential of leveraging LLMs for data generation. In this section, I propose a cost-effective method for generating text data using LLMs. I consider following two questions:

1) *How can LLMs be utilized for the data generation?* While recent LLM models exhibit proficiency in understanding images, their functionality may not be entirely convincing. In our approach, I disentangle gaze information from images. Rather than generating paired data $\{\mathbf{I}_i, \mathbf{T}_i\}$, I generate paired data $\{\mathbf{g}_i, \mathbf{T}_i\}$, where $\mathbf{g}_i \in \mathbb{R}^2$ represents the gaze in the form of (Yaw, Pitch). Compared to high-dimensional image data, gaze directions offer greater accuracy and ease of interpretation. Additionally, I can further estimate gaze from images to establish the connection between text and images.

2) *How can the validity of the generated data be confirmed?* To assess the quality of the generated data, I sample text descriptions from our dataset and conduct a user study for evaluation.

Next, I introduce the LLM-based data generation method in Sec. 5.2.2. I present an overview of the ToG dataset statistics and discuss the user study evaluating our dataset in Sec. 5.2.3.

5.2.2 Text Generation using LLMs

I aim to utilize LLMs to generate text descriptions from gaze. However, relying solely on gaze values may not accurately capture human facial features. Human gaze is influenced by both head pose and eye rotation, introducing ambiguity that can result in unclear text descriptions. Therefore, I further decompose the gaze into head pose and eye pose to address this challenge.

Dense Gaze Sample Generation: I respectively define the ranges for head pose and eye pose, and calculate gaze based on these values (Deng et al., 2017). Specifically, I define head pose in both the yaw and pitch axes, sampling from the range of -70° to 70° at intervals of 10° . This results in 15×15 head rotation samples. I also sample eyeball rotation from the range of -50° to 50° at intervals of 10° , generating 11×11 eye rotation samples. By combining these values, I generate a total of 27,225 gaze samples. To ensure visibility of the

eyes, I exclude gaze samples that fall outside the range of -70° to 70° in the pitch axis and -120° to 120° in the yaw axis (Zhang et al., 2020). Consequently, I obtain a total of 23,887 samples.

Prompt Design: I further employ LLMs to generate text description from gaze. I design the following prompt:

“Imagine describing someone’s head and gaze movement. Given four numbers representing yaw and pitch for both head and gaze (positive for left/up, negative for right/down), craft a detailed description. Your narrative should:

- 1. Avoid numerical values for angles;*
- 2. Offer rough direction descriptions regardless of extent;*
- 3. Be one sentence and within 10 words;*
- 4. Begin with ‘The person’ and describe ‘the head’ and ‘the gaze’ impersonally.”*

The first requirement ensures that the LLM does not produce annotations based on simple templates such as *“The person’s gaze is 60 degrees upward”*. The second and the third requirements set limitations to the precision and the length of the annotations. I usually describe gaze behavior roughly or accurately in different scenarios. The two requirements make LLM to generate only low-precision descriptions, such as *“The person tilts the head right, gazing up and left”*. Besides, I design another two requirements which require LLM to generate high-precision annotations:

- “2. Offer precise direction descriptions with a clear extent;*
- 3. Be concise, between 20 and 30 words;”*

It allows LLM to generate detailed pose annotation like *“The person gently rotated the head to the right and deeply dipped it, whereas the gaze darted left before sharply descending, aligning slightly off-center”*. In addition, human faces have many attributes such as gender and age. To make the descriptions focused on describing pose, I set limitations to the narrative format

Table 5.1: I present text descriptions from the ToG dataset. I define two levels of precision and leverage LLMs to generate different descriptions for each precision level. The last two columns present the corresponding gaze and head labels.

Precision	Descriptions	Head label	Gaze label
Low	The person’s head turns left, gaze shifts left and slightly up.	$(60^\circ, 0^\circ)$	$(109^\circ, 10^\circ)$
	The person kept the head and the gaze straight ahead.	$(0^\circ, 0^\circ)$	$(0^\circ, 0^\circ)$
	The person tilted the head right, directing the gaze sharply downwards.	$(-70^\circ, -70^\circ)$	$(-115^\circ, -20^\circ)$
High	The person’s head turns significantly left, remaining level, while the gaze shifts sharply left and slightly upwards, indicating keen interest or examination.	$(60^\circ, 0^\circ)$	$(109^\circ, 10^\circ)$
	The person maintained a direct, steady posture, with the head and the gaze fixed straightforward, indicating intense focus.	$(0^\circ, 0^\circ)$	$(0^\circ, 0^\circ)$
	The person directed the head significantly to the right and downward, while the gaze extended far right, with a minimal decline.	$(-70^\circ, -70^\circ)$	$(-115^\circ, -20^\circ)$

with the fourth requirement. Another advantage is that I can easily replace the pronouns in sentences like replacing "*The person*" with "*The boy*", "*The girl*", or "*The farmer*". It expands the capability of our method for diversification.

To further expand the diversity of the dataset, I ask LLM to generate three different descriptions in high precision for each sample simply by adding "*Please give me three different descriptions.*" to the prompt. I show some examples of different precisions in Table 5.1. The prompt encourages LLM to replace words, especially the ones describing the extent.

5.2.3 ToG Dataset

I utilize ChatGPT-4 Turbo for data generation. In total, I sample 23,887 gaze values, considering variations in both eye rotation and head rotation. For each gaze value, I

Table 5.2: I list some text-image datasets in various tasks. Our work provides the first text-gaze dataset ToG, containing large and diverse text descriptions.

Datasets	Task	#Texts
BABEL (Punnakkal et al., 2021)	Text-Body	63,000
HumanML3D (Guo et al., 2022)	Text-Body	44,970
PoseScript (Delmas et al., 2022)	Text-Body	300,000
Text2FaceGAN (Nasir et al., 2019)	Text-Face	60,000
CelebAText-HQ (Sun et al., 2021)	Text-Face	150,100
ToG (Ours)	Text-Gaze	95,548

generated four descriptions, one with low precision and the other three with high precision given that low-precision descriptions exhibit limited variability. Consequently, I obtained four descriptions for each sample, resulting in a total of 95,548 annotations. Each sample in our ToG dataset is denoted as $\{\mathbf{g}_i, \mathbf{h}_i, \mathbf{T}_i\}$, where $\mathbf{g}_i$ represents the gaze value, $\mathbf{h}_i$ denotes the head pose, and $\mathbf{T}_i$ is the text description.

I compare our ToG dataset with existing human-centered text-to-video/image datasets in Table 5.2. BABEL (Punnakkal et al., 2021), HumanML3D (Guo et al., 2022), PoseScript (Delmas et al., 2022) focus on text pose generation. Text2FaceGAN (Nasir et al., 2019) and CelebAText-HQ (Sun et al., 2021) are for text face generation. Our ToG dataset includes over 90k text descriptions, making it comparable in size to existing datasets. However, unlike existing datasets, ToG is specifically designed for text-guided gaze generation, addressing a unique need that other datasets cannot fulfill.

To evaluate the generated dataset, I randomly select 20 samples from our ToG dataset. I get the famous gaze dataset ETH-XGaze dataset (Zhang et al., 2020), which provides corresponding gaze values and image pairs. I match each description with an image in

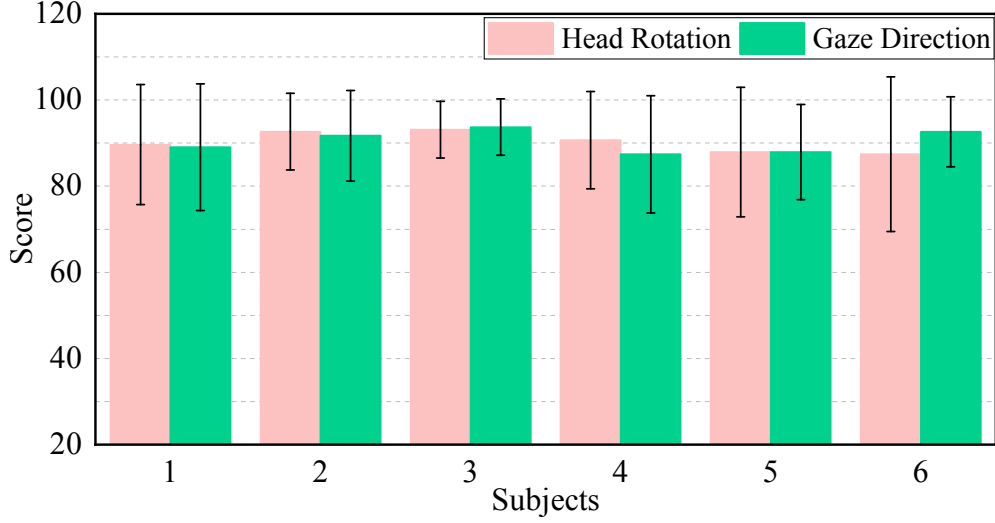


Figure 5.2: Visualization of head and gaze scores of different subjects.

ETH-XGaze dataset by finding the closet pose labels, *i.e.*, gaze value and head pose. As the result, I acquire image-text pairs.

I invited six users to rate the correspondence between text descriptions and images on a hundred-point scale, with lower scores indicating poorer correspondence. Specifically, I asked them to score the correspondence between the image and the text description of the head, as well as between the image and the text description of the gaze. I show the head and gaze scores from different users in Fig. 5.2. The results show that both scores for head and gaze are around 90 with small variances, which indicates that our text descriptions match both head and gaze values well. The good results validate the quality and accuracy of the generated data.

LLMs are famous for the great diversity of generated languages. Our dataset generated by LLMs inherits this advantage. Our ToG dataset has **475** different adverbs for extent description. In contrast, the well-known text pose datasets BABEL (Punnakkal et al., 2021) and PoseScript (Delmas et al., 2022) only have **250** unique action categories and **87** unique posecodes separately. I show some examples of extent description words in Table 5.3.

Table 5.3: Examples of extent synonym in ToG.

Extent	Synonym Examples in ToG
Slightly	mildly, faintly, subtly, marginally, a little, a bit, gently, minimally, ...
Moderately	reasonably, medium, fairly, modestly, somewhat, gradually, plainly, gracefully, ...
Significantly	considerably, substantially, notably, sharply, seriously, majorly, vastly, heavily, ...

5.2.4 Budget Efficiency of the Generated Data

I compare the budget efficiency against the manual annotation on Amazon Mechanical Turk. According to the academic report (Hara et al., 2018), Amazon Mechanical Turk pays between 1 dollar and 6 dollars per hour for a typical Human Intelligence Task (HIT) task. In the comparison, I use the lowest price which is 1 dollar per hour. I assume each worker can annotate one pair of head pose and gaze directions with two text descriptions in one minute. For LLMs annotation, I use ChatGPT4 Turbo. According to (OpenAI, 2024), 1 token is approximately 0.75 words. The price is 0.01 dollars per 1K tokens for input and 0.03 dollars per 1K tokens for output.

Based on the above price, I show the estimated cost of our annotation task using Amazon Mechanical Turk and ChatGPT4 Turbo in Table 5.4. It indicates that using LLMs like ChatGPT has better budget efficiency than the manual annotation on Amazon Mechanical Turk.

Table 5.4: Comparison on costs.

Annotation Method	Costs (\$) ↓
Amazon Mechanical Turk	796.23
ChatGPT4 Turbo	114.02

5.3 Methodology

Given a text description $\mathbf{T}$ describing human gaze behavior, our objective is to generate face images $\mathbf{I}$ that align with the description. In the previous section, I present the ToG dataset, which provides paired data $\{\mathbf{g}_i, \mathbf{h}_i, \mathbf{T}_i\}$. Additionally, I have an image dataset $\{\mathbf{I}_i\}$ containing diverse face images. Matching face images with text descriptions is a challenge in this task. While some gaze estimation datasets offer paired data $\{\mathbf{g}_i, \mathbf{h}_i, \mathbf{I}_i\}$, where the gaze labels are obtained through calibration, these datasets typically lack diverse face appearances.

To tackle this challenge, I propose a two-stage text-to-face image generation method. Our method contains a sketch-conditioned face diffusion module and a model-based sketch diffusion model. I define a face sketch based on facial landmarks and eye segmentation, and the face diffusion model generates face images from face sketch. Importantly, the model is trained without gaze annotation. The sketch diffusion model, on the other hand, learns gaze patterns from text descriptions. To convert the gaze pattern into a face sketch, I utilize a 3D face model and compute the face region information based on eye and head rotation.

5.3.1 Text-to-Gaze Generation

I first build a diffusion model to learn gaze pattern from text description. In general, I learn a diffusion model $\mathcal{G}$, where $\{\mathbf{g}, \mathbf{h}\} = \mathcal{G}(\mathbf{T})$. I obtain feature embeddings from the text description and propose a text attention module (TAM) to capture head and gaze information in this section.

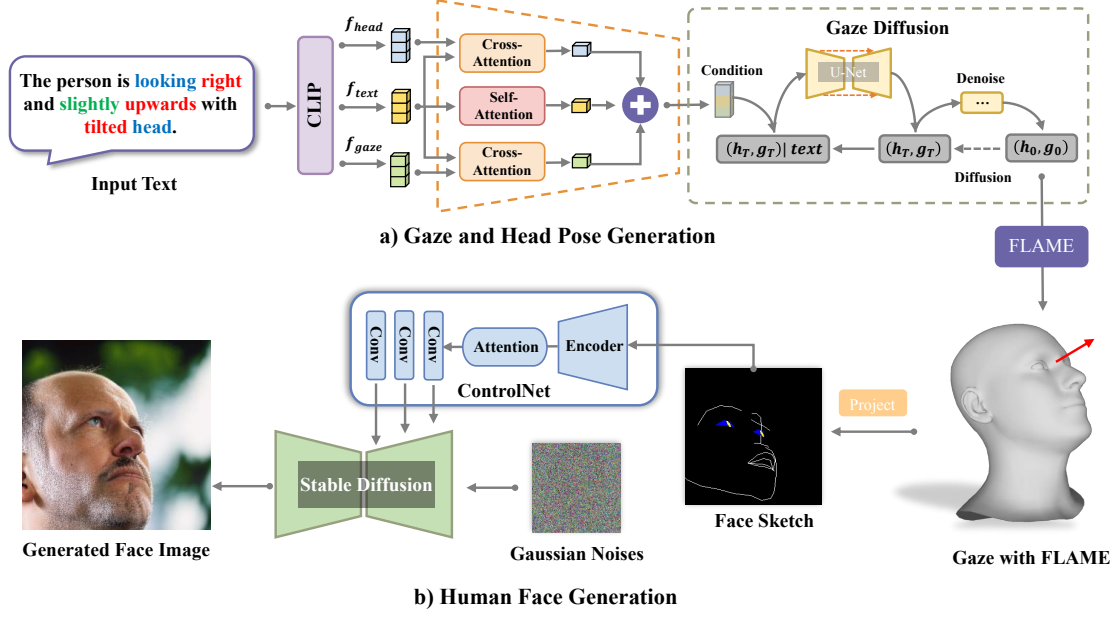


Figure 5.3: The pipeline of TextGaze. The model operates in two stages: pose generation and face generation.

Specifically, given a text $\mathbf{T}$, I first convert it into feature embedding $\{f_{text}^i, i = 1 \dots L\}$, where L is the length of $\mathbf{T}$. I utilize CLIP (Radford et al., 2021) to obtain the embedding feature for each word. To effectively extract pose information from the text, I propose the TAM. Our main idea is to extract gaze-related and head-related text information from $\{f_{text}^i\}$. In particular, CLIP preserves feature similarity for two similar words due to contrastive learning. I acquire additional word embedding for the words ‘gaze’ and ‘head’ via CLIP, denoted as f_{gaze} and f_{head} respectively. These embeddings are used to aggregate gaze-related and head-related information from f_{word}^i . TAM is consisted of two cross-attention modules and one self-attention module. The inputs of the two cross-attention modules are $\{f_{gaze}, f_{text}^i\}$ and $\{f_{head}, f_{text}^i\}$ respectively, allowing them to aggregate text features related to gaze and head. Additionally, I input $\{f_{text}^i\}$ into a self-attention module to extract sentence feature. TAM combines the three extracted features as the output, which are then used as conditions to generate gaze and head via diffusion models.

The diffusion model $\mathcal{G}$ generates $\{\mathbf{g}, \mathbf{h}\}$ from $\mathbf{T}$. I then convert the gaze information into an intermediate representation, face sketch, for face image generation. In detail, I leverage the 3D face model FLAME (Li et al., 2017) in our work. I rotate the 3D face model based on $\{\mathbf{g}, \mathbf{h}\}$ and project it into 2D space. I acquire 2D landmarks and connect these landmarks of different region to generate face sketch. Considering the importance of eye region, I highlight the eye region and pupil using different values. I will explain the advantage of the face sketch in the next section.

5.3.2 Gaze-Controllable Face Generation

I aim to generate face images using gaze information $\{\mathbf{g}, \mathbf{h}\}$. Gaze redirection methods directly input gaze information as a condition and generate satisfied images. However, it means I need paired data $\{\mathbf{g}, \mathbf{h}, \mathbf{I}\}$ for training. In this section, I first convert gaze information into face sketches using a 3D face model and then perform face-sketched condition face generation. This pipeline enables us to perform gaze-controllable face generation without a gaze label.

As shown in Fig. 5.3, the face sketch is made of facial layout landmarks and iris segmentation, which give information on head pose and gaze direction. I use the segmentation map to highlight the eye region since the eye region is important for the gaze-controllable generation. More concretely, given an image $\mathbf{I}$, I perform facial landmark detection and eye segmentation on $\mathbf{I}$ (Bulat et al., 2017; Park et al., 2018b). I connect landmarks and combine it with eye segmentation to produce a face sketch. As the result, I obtain the paired data $\{\mathbf{I}, \mathbf{I}_s\}$, where $\mathbf{I}_s$ represents the face sketch. I use the data to train a conditional diffusion model for gaze-controllable face generation.

I use ControlNet (Zhang et al., 2023) to generate face images from facial sketches. I learn the marginal distribution of face image $\mathbf{I}$ conditioned on the sketch $\mathbf{I}_{sketch}$.

$$\mathcal{L} = \mathbb{E}_{\mathbf{z}_0, \mathbf{t}, \mathbf{c}_t, \mathbf{c}_f, \epsilon \sim \mathcal{N}(0,1)} [\|\epsilon - \epsilon_\theta(\mathbf{z}_t, \mathbf{t}, \mathbf{c}_t, \mathbf{c}_f)\|_2^2] \quad (5.1)$$

I use Eq. (5.1) as the objective function while I learn added noise ϵ rather than image $\mathbf{x}_0$. I also fix the Stable Diffusion as the backbone and only train the control module in training (Zhang et al., 2023).

In the inference stage, given gaze information $\{\mathbf{g}, \mathbf{h}\}$, I leverage a 3D face model to generate the face sketch. I rotate the 3D face model based on $\{\mathbf{g}, \mathbf{h}\}$ and project it into 2D space. I acquire 2D face and eye landmarks, and then connect them for face sketching. I also highlight the eye region as an eye segmentation map based on eye landmarks.

5.3.3 Implementation Details

Our proposed baseline is implemented by Pytorch. I generate 256×256 image from the text description. I use ControlNet (Zhang et al., 2023) as the diffusion model for face image generation. I use a transformer with **6** layers **8** heads for all three branches in the conditioning part of the Text-to-Gaze diffusion model. The Text-to-Gaze diffusion model is built based on Latent Diffusion Model (Rombach et al., 2022). I adopt the Adam optimizer and the learning rate of 1.0e-06 for training the two models. I use a batch size of 8 with 20 epochs for the face diffusion model and a batch size of 16 with 10 epochs for the text-to-gaze diffusion model. I use Stable Diffusion v2.1 as the backbone in the face diffusion model. I train the face diffusion model and the text-to-gaze diffusion model for 24 hours and 7 hours separately on one RTX 3090 GPU.

5.4 Experiment

5.4.1 Datasets

ToG Dataset. I divide our ToG dataset into a training set and a test set. The training set ratio is 0.9. I use the training set to train the text-to-gaze generation model.

FFHQ Dataset. The face generation model undergoes training utilizing the FFHQ dataset (Karras et al., 2019), a collection of 70,000 superior-quality facial images sourced from the internet, frequently employed in generative model training (Karras et al., 2019; Rombach et al., 2022). FFHQ’s native resolution stands at 1024×1024 , while our experiments employ its images resized to 256×256 . I get the style prompts of FFHQ dataset with BLIP (Li et al., 2022b). To evaluate our model, I split the FFHQ dataset into a training part and a test part. The training part ratio is 0.9.

Sketch Generation. The generated sketch includes face layout landmarks and iris segmentation. I get face landmarks using face-alignment (Bulat et al., 2017). I get iris landmarks and eyelid landmarks with the eye landmark detector (Park et al., 2018b). I connect the landmarks of different parts and make it a sketch. I further mask out the invisible iris region with eyelid landmarks to generate a realistic sketch. I use the sketches generated from a 3D face model in the reference stage. To mitigate the gap between the sketches in the training stage and the inference stage, I align the sketches by setting the camera parameters in the 3D face model same as FFHQ dataset. Specifically, I first get the camera parameters used in FFHQ dataset by fitting the images to the face model and then fix the parameters when generating new sketches.

5.4.2 Comparison on Text-Guided Generation

Baselines. The existing text-to-image generation methods (Esser et al., 2021; Ramesh et al., 2022) are not optimized specifically for gaze image generation. As far as I know, we are the first work to do gaze-controllable face generation with natural language. To make a fair comparison, I first match the images in the FFHQ training set with the text descriptions in our ToG dataset and then train a latent diffusion model (LDM) (Esser et al., 2021) and a ControlNet model (Zhang et al., 2023) using the matched data. Specifically, I apply a pre-trained pose estimator to gauge head poses and gaze directions on FFHQ’s training set (Zheng et al., 2020). Then I find the closest pose labels and the corresponding descriptions to the estimated ones. Finally, I match the selected descriptions with the images to get the training data. For the LDM model, I combine the pose descriptions with the style prompts as the input. The LDM model is trained from scratch. For the ControlNet model, I input the style prompt as (Zhang et al., 2023) and adopt the control network to accept the pose descriptions as the additional input condition. The ControlNet model also uses Stable Diffusion v2.1 as the backbone. Both models are trained with a batch size of 8 in an end-to-end way for 30 epochs.

Metrics. I report the standard evaluation metrics Inception Score (IS) (Salimans et al., 2016), Fréchet Inception Distance (FID) and Kernel Inception Distance (KID) (Heusel et al., 2017) on the FFHQ test set to evaluate the quality of generated images. I report KID* which is $1000 \times \text{KID}$ to make a more precise comparison. I also report CLIP scores (Radford et al., 2021) on the same data to evaluate the correlation between prompts and generated images. I use DDIM as the sampler and use 50 steps to sample each image. I calculate FID and KID metrics on FFHQ test set. I combine style prompt and pose description as one prompt when calculating CLIP scores. I further perform a user study to evaluate our model’s effectiveness. I randomly chose 20 generated images with the same pose prompts for our model, ControlNet, and LDM. I get 20 sets with a total of 60 images. Twelve independent blinded users were



Figure 5.4: Visualization of generated images from our model and two baselines (ControlNet and LDM). I show 12 sets of comparisons. In each set, pose description is listed on the left and the images generated by our model, ControlNet, and LDM are listed side by side on the right. "head", "gaze" and directional words are highlighted for better visualization.

invited to join the study and they were shown 20 side-by-side images each generated by our model, ControlNet, and LDM. The users were given the task of selecting the image which matches the given descriptions best in terms of gaze direction and head pose separately. If none of the images matches the description, the user could select a "None" option. I report user preferences for head and gaze according to the user study. Note that CLIP model is not optimized specifically for precise pose descriptions. Therefore, CLIP score does not reveal how well the generated images match the pose descriptions. I deem that the user preferences for gaze and head alignment from the user study are more suitable metrics for our work.

Qualitative Results. The qualitative results are shown in Fig. 5.4. I give 12 sets of images

and their pose descriptions. Thanks to the capability of ControlNet, our model can also control the style of the generated images. I split the samples into three style collections for better visualization. The pose descriptions and the images are listed side by side. All three models can generate meaningful face images. The images generated by our model and ControlNet are of good quality while the results of LDM have some artifacts. It is because our model and ControlNet only trained an additional network to add additional conditions leaving the Stable Diffusion backbone unchanged. Therefore, they always generate images with a similar quality to Stable Diffusion. However, the whole LDM model is optimized in the training, which might make the model easily overfit the training data and unstable. Even though the generated images from ControlNet and LDM are meaningful, they do not follow the head pose, gaze direction, or both depicted in the input text. Both baseline models can generate various head poses and gaze while the contents are not related to the text descriptions. Their disability to render correct gaze images under text instructions shows that capturing geometric information from text descriptions directly is hard. To solve the problem, I design two stages including pose generation and face generation. The pose generation model estimates the head pose and gaze direction from text and transfers the pose information to face sketches explicitly through a 3D face model. The face generation model takes the face sketches as input and renders face images. The two-stage design alleviates the difficulty of direct matching between text and images. As shown in the results, our generated images effectively capture the head and gaze information and maintain good quality simultaneously.

Quantitative Results. I show the quantitative results in Table 5.5. Our model achieves the best performance on Inception Score and KID and comparable performance on FID compared with the other two baselines. It indicates that our model can generate high-quality face images with the additional eye-controllable function. Our model achieves a comparable result on CLIP score with ControlNet while outperforming LDM by a large margin. I show the user preferences for head and gaze in the last two columns. The results demonstrate

Table 5.5: Comparison to other SOTA methods on the FFHQ dataset in terms of image quality (IS, FID, KID), correspondence of text and image (CLIP-score, User Preference for Head and Gaze). KID* and UP stands for KIDx1000 and User Preference. **Bold** indicates the best number.

Method	IS $\uparrow$	FID $\downarrow$	KID* $\downarrow$	CLIP-score $\uparrow$	UP (Head) $\uparrow$	UP (Gaze) $\uparrow$
LDM	3.89	55.45	45.34	27.36	0.07 ± 0.07	0.05 ± 0.05
ControlNet	7.11	61.67	35.94	30.84	0.11 ± 0.07	0.11 ± 0.10
Ours	7.22	59.91	35.73	30.63	0.76 ± 0.16	0.75 ± 0.20

that users predominantly prefer our generated images, which consistently align well with the provided pose descriptions.

The good performance is attributed to our newly designed two-step image generation. Our high-performance two-step image generation process starts with obtaining predicted head pose and gaze directions using our text-to-pose model, trained on the ToG dataset. This model effectively extracts gaze and head pose information from text, generating them within a reasonable range. I then use the FLAME model to simulate head and eyeball rotations, producing a rotated 3D face model for sketch generation. In the second stage, I use a diffusion model trained with ControlNet to generate images that accurately incorporate the predicted head pose and gaze data. This ensures that the head and gaze details are consistently preserved from the initial text description through to the final images.

5.4.3 Ablation Study

Verification of Face Sketch. I advocate for the use of face sketches in facial generation due to the scarcity of gaze labels. To verify the effectiveness of our face sketch, I applied a pre-trained VGG-based gaze estimator to determine head poses and gaze directions on the FFHQ training subset. I adapted the condition encoding segment of a ControlNet model to

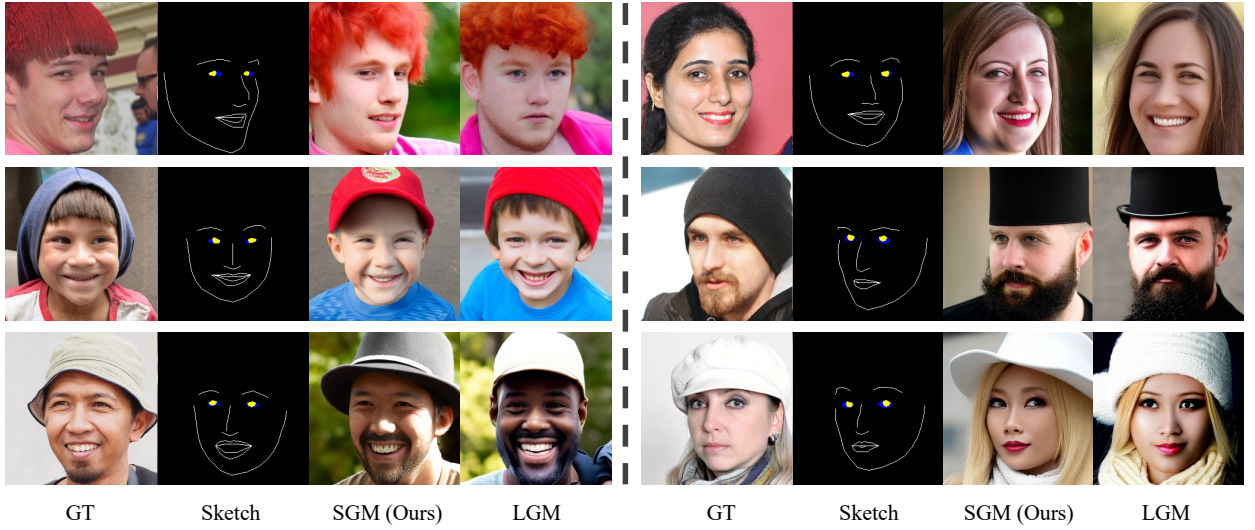


Figure 5.5: Visualization of results from Label-Guided Model and Sketch-Guided Model.

Table 5.6: Comparison of LGM and SGM in terms of head error and gaze error in degree on the test set of FFHQ dataset.

Method	Head ↓	Gaze ↓
LGM	4.37	4.79
SGM (ours)	3.62	4.35

accept pose inputs, keeping other components intact. Subsequently, I developed a label-guided model. I refer to the label-guided model as LGM and our sketch-guided model as SGM. LGM is trained on the FFHQ training set with estimated pose labels. Similar to our face generation model, it is trained with a batch size of 8 for 20 epochs. I evaluate the models on the FFHQ test set.

I show the qualitative comparison with our sketch-guided model in Fig. 5.5. The label-guided model shows limited control of head pose and gaze direction in the generation. I estimate the head poses and gaze directions of the generated images from the two models using another pre-trained ResNet50 (He et al., 2016)-based gaze estimator from (Zheng et al., 2020). I calculate the head error and gaze error in degree and report them in Table 5.6. The results

Table 5.7: Ablation on TAM module in terms of generated head error and generated gaze error in degree on the test set of ToG dataset.

Method	Generated Head ↓	Generated Gaze ↓
w/o TAM	16.94	21.62
w/ TAM (concat)	15.45	20.28
w/ TAM (add)	15.23	20.00

show that our sketch-guided model generates images with higher accuracy in head pose and gaze direction. These results underscore the efficacy of our approach, confirming the validity of our choice to rely on sketch-guided modeling.

Verification of Text Attention Module. To show the effect of our designed text attention module, I train a pose generation model with only a self-attention module under the same training settings for comparison. I further test two different ways (adding and concatenation) to fuse the features from different branches. I report the generated angle errors on the test set of our ToG dataset. The angle errors are calculated via cosine similarity between the estimated poses and ground truth poses. I show the results in Table 5.7 The two models with TAM get comparable results and achieve better performance compared with the pure self-attention model. The results indicate that our TAM is more effective in capturing pose information from text descriptions.

5.4.4 Controlling Shape and Expression

Thanks to our sketch-guided model and the 3D face model, I can control the face shape and expression by changing the input sketches by interpolating the shape and expression code of the 3D face model. I show the visualization results in Fig. 5.6. This demonstrates that our model can generate faces with varying shapes and expressions while preserving consistent head poses and gaze directions. Such versatility significantly enhances the potential

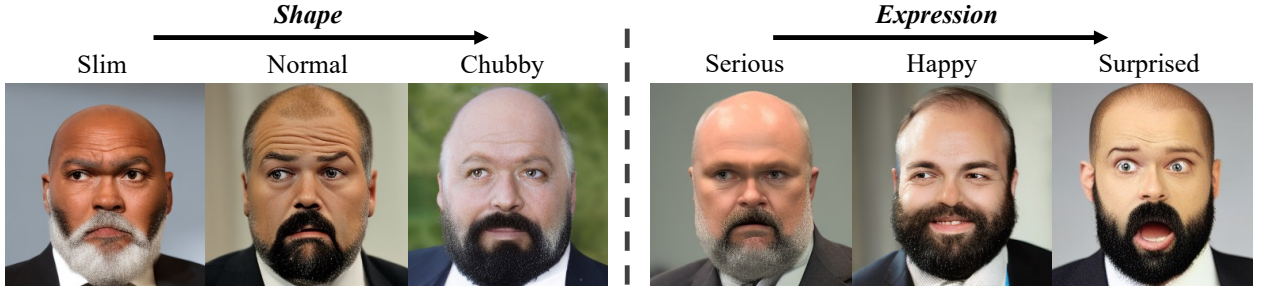


Figure 5.6: Interpolation of shape and expression.

applications of our model, making it suitable for diverse fields.

5.4.5 Generation with Different Roll Angles

Human eye movement is typically represented by (Yaw, Pitch) since the eyeball has no freedom in the roll axis, and I maintain consistency with head representation. However, our model is capable of generating face images with different roll angles of the head. I can perform roll rotation on the face model and generate the corresponding face sketch. I do not need to re-train the face diffusion model, as the network is trained using detected face sketches and the face image dataset includes roll rotations. An example is shown in Fig. 5.7.



Figure 5.7: Generated results under different roll angles.

5.4.6 Identity Control

Our existing method enables identity-controllable generation using two text inputs without re-training: one for gaze control and one for identity control. This capability arises from our use of the face sketch as an intermediate representation. Consequently, the face diffusion

network generates the face image conditioned on face sketches. I inject this condition into an existing model rather than retraining a new one. The original network uses one text input for generating different face appearances, while our additional face sketch condition maintains gaze control. I use different identity texts with the same gaze text and show the results in Fig. 5.8.



Figure 5.8: Generated results with different identities.

5.4.7 Comparison with Other SOTA Face Generation Models

IP-Adapter (Ye et al., 2023a) and PhotoMaker (Li et al., 2024) are two SOTA face generation models. To make our results more convincing, I conduct an additional comparison with these two methods. The head/gaze user preferences for **our model**, **PhotoMaker** and **IP-Adapter** are **0.79/0.83**, **0.20/0.13**, **0.01/0.03** separately. It is clear that they cannot achieve the gaze control. Besides, the FIDs ($\downarrow$) for the three models are **59.91**, **65.05**, and **68.90**. Our model achieves the best performance on generated image quality. In fact, conventional text-to-image methods, even in the face generation domain, have not addressed the problem of gaze control. Our work is the first to explore it and naturally achieves the best performance.

5.4.8 Additional Qualitative Results

The additional results of our model are presented in Fig. 5.9. It contains 12 sets of results including text description and generated image. They are in three different styles. For better visualization, I annotate the possible gaze directions using the green sector and the head



Figure 5.9: Additional qualitative results. It includes 12 result sets, each with a text description and a generated image in three styles. Gaze directions are annotated with a green sector, and head direction is marked with a blue arrow indicating left/right. Please refer to the text for head pose in the vertical direction. The alignment between descriptions and images demonstrates the model’s effectiveness. The blue arrow only shows left/right direction.

direction with a blue arrow same as the main chapter. Please note the blue arrow only shows left/right direction. Please refer to the text for head pose in the vertical direction. The close alignment of the text descriptions with the generated images demonstrates the effectiveness of our model.

5.4.9 Discussion

Diffusion Model vs. Physical Model-Based System. Instead of directly employing a physical model-based system, such as FLAME (Li et al., 2017), I developed a sketch-guided diffusion model to generate the final result. Hair generation using physical model-based systems encounters significant challenges, including high computational costs arising from complex numerical calculations, sensitivity to various physical parameters requiring meticulous tuning, and computationally intensive collision detection that frequently leads to visual artifacts (Ward et al., 2007; Müller et al., 2005). Furthermore, there remains a notable gap in realism between images synthesized by physical model-based systems and actual photographs. In contrast, the sketch-guided diffusion model I employed leverages extensive training on large-scale image datasets containing abundant real-world examples, thus enabling the generation of high-quality and realistic hair.

Face Sketch vs. Face Landmarks. The face sketch utilized in this work consists of facial layout lines and a segmentation map of the iris, which is highlighted with distinct colors. The primary difference between a face sketch and face landmarks is that the former is composed of continuous lines, while the latter is represented by discrete landmark points. As a guidance signal, lines provide richer spatial information regarding the facial structure, whereas landmarks offer only sparse and localized guidance. Consequently, face sketches yield superior performance in generating the facial region due to their clearer and more comprehensive guidance compared to the sparse nature of face landmarks.

5.5 Future Work

Analysing and Mitigating the Bias of Synthetic Datasets. While large language models (LLMs) serve as powerful tools for generating extensive synthetic textual datasets, facilitating research by mitigating issues of data scarcity and privacy concerns, it is crucial to acknowledge that these generated datasets may inadvertently inherit, amplify, or introduce biases stemming from their training corpus or intrinsic generative mechanisms (Sheng et al., 2019; Yeh et al., 2023; Lin et al., 2024). Specifically, biases related to demographic attributes, linguistic representations, or cultural stereotypes may manifest subtly within the synthetic text produced by LLMs. Consequently, future research should systematically analyze such biases through rigorous statistical assessments (Sun et al., 2024), intersectional analyses across multiple demographic and contextual dimensions (Ma et al., 2023), and the deployment of interpretability methods (e.g., attention visualization, causal inference, or feature attribution) (Linardatos et al., 2020) to precisely identify their origins. Additionally, dedicated mitigation strategies must be explored, including bias-aware fine-tuning of LLMs, strategically balanced prompt design, targeted data augmentation techniques, algorithmic post-processing to correct biases, and the incorporation of human-in-the-loop evaluations for nuanced quality assurance. Proactively addressing these considerations will significantly enhance the fairness, reliability, ethical validity, and societal acceptance of textual datasets synthesized by generative language models.

Identity Preservation and 3D Lifting. *TextGaze* is a prototype for language-guided eye-controllable face generation. Even though it realizes the basic function of controlling gaze with natural language, it can not control the identity of generated faces and the generated results are in 2D not 3D. For now, *TextGaze* allows style controlling through additional prompts like Stable Diffusion, but it can not fix the identities for generations. Controlling identities is important in many personalized applications. It has been shown that the intermediate features of diffusion models have exceptional communication capabilities, allowing for pixel-

level fine control over images (Xie et al., 2024; Chang et al., 2023). Therefore, incorporating a diffusion-based appearance reference module could be an effective solution for controlling identity. On the other hand, 3D consistency is crucial for many immersive applications, such as the metaverse and online conferences. Generating 2D images separately can diminish the realism of the scene. Combining 3D face generation with natural language guidance is a promising direction, as it not only maintains strong 3D consistency but also enhances usage convenience through language control. Recent advancements in text-to-3D generation offer potential solutions for this task. For instance, Bai et al. (2025) integrate text prompts as conditions to modify generated 3D portraits, using a cross-attention module inspired by Stable Diffusion. 3D-aware gaze-controllable face generation with language guidance could be similarly achieved by adapting such techniques.

5.6 Conclusion

In this chapter, I introduce an innovative task of gaze-controllable face generation, driven by textual descriptions of human gaze behaviors. I present the pioneering text-to-gaze dataset, featuring over 90k descriptions that align with various gaze behaviors. This dataset encompasses a wide range of gaze directions, enriched by diverse natural language annotations derived from Large Language Models (LLMs). Building upon this dataset, I develop a novel two-stage face diffusion model. Initially, I create detailed face sketches, which are then transformed into face images using conditional diffusion models. Additionally, our approach infers face sketches from text descriptions, incorporating a 3D face model prior. Significantly, our method operates without the need for explicit gaze annotations and can generate gaze-controllable face images effectively. The efficacy of the proposed approach is further demonstrated through experiments conducted on the FFHQ dataset.

Chapter Six

Discussion and Conclusion

To make gaze-controllable face generation widely applicable, I address the challenges faced in quality, speed, and usability. 1) **It is hard to model both rigid eyeball rotation and non-rigid eyelid deformation in a unified 3D representation.** A unified 3D face representation ensures 3D consistency in gaze-controllable face generation but poses challenges in modeling both rigid eyeball rotation and non-rigid eyelid deformation. Eyeball rotation involves simple rigid transformations, while eyelid deformation is complex and non-rigid, relying on muscle motion and skin flexibility. I have shown that a unified rotation-aware face NeRF model on the canonical space can render high-fidelity face images from novel viewpoints and eye poses. It achieves superior 3D consistency and natural deformation in eye regions. 2) **The inference of 3D gaze-controllable generation models is still far from being real-time.** For gaze-controllable face generation, creating natural and realistic images is not enough and slow processing also limits its practical use. 2D-based methods lack 3D consistency, while 3D-based methods are too slow. Many applications, like immersive games and the metaverse, need both real-time speed and good 3D consistency. I have shown that 3D-aware gaze-controllable face generation can achieve real-time performance by employing an efficient triplane-based NeRF architecture and distilling the prior knowledge from 3D GANs into a lightweight module. 3) **Using explicit gaze labels to control gaze is**

inconvenient for general use. Convenience is crucial for the adoption of gaze-controllable face generation. While gaze angles are precise and useful for tasks like gaze estimation, natural language is more intuitive for most users. Exact gaze orientation is often unnecessary; an acceptable range suffices, and language-based input allows for easy refinement of results if needed. I have developed a gaze-controllable face-generation system guided by natural language and shown the effectiveness of language on this task.

6.1 Contributions

The primary contributions of my work are outlined as follows:

High-Fidelity Generation Quality. In Chapter 3, I propose **DeNeRF**, a high-fidelity eye-controllable face NeRF model that learns a unified representation for both rigid and non-rigid transformations in the eye region. DeNeRF generates high-quality images with precise eyeball rotation and realistic periocular deformations, even from novel viewing angles, achieving a significant improvement over state-of-the-art methods.

Real-Time Inference Speed. In Chapter 4, I present **LiveGaze**, a real-time 3D-aware gaze redirection module capable of generating 3D faces with controllable eyes from a single image. It is the first method to achieve real-time performance in 3D-aware gaze redirection.

Language-Guided Usability. In Chapter 5, I introduce **TextGaze**, a language-guided eye-controllable face generation model. It is the first approach to enable gaze control using natural language. Additionally, I propose the first text-to-gaze dataset to incorporate textual descriptions into gaze-controllable face generation.

6.2 Limitations

Even though my work achieves great improvement in quality, speed, and usability, limitations remain in this area.

First, while *LiveGaze* excels at generating photorealistic gaze-redirected face images in real time, it encounters challenges under certain conditions, such as varying lighting and the presence of glasses. In low-light environments, the model may mistakenly interpret faces with darker tones as belonging to different individuals. This is because the environmental elements are not well disentangled from gaze redirection. Additionally, its performance is affected when handling faces with glasses, as it struggles to reconstruct the glasses accurately, leading to interference in generating the eye region.

Second, I show the ToG dataset in Chapter 5. Our dataset boasts an extensive collection of gaze descriptions, yet it falls short in diversity when compared to human annotations. Despite implementing diverse criteria for precision, length, and sequence in the generation of descriptions, there is a noticeable uniformity in style across the generated content. Beyond using directional words for pose description, there are numerous alternative methods like reference objects and metaphors. Investigating the potential of Large Language Models (LLMs) to produce a variety of descriptions, including metaphors and diverse reference points, presents a promising avenue for future research. For example, people often use environmental references like the sky, floor, or ceiling to indicate directions, and professionals from different fields tend to have unique descriptive preferences. However, our dataset encompasses only a limited range of these reference objects. Additionally, metaphorical language is a prevalent tool in describing poses, yet our dataset scarcely incorporates such metaphors.

6.3 Future Work

Despite the limitations mentioned above, I believe my work has achieved promising results and has paved the way for future research opportunities.

6.3.1 Compositional Animation of Eye Region

In Chapter 3, I address both the rigid rotation of the eyeball and the non-rigid deformation of the periocular region, particularly the deformation of the eyelids. To this end, I propose a unified NeRF-based representation capable of modeling the animation of the entire eye region. Similarly, EyeNeRF (Li et al., 2022a) and GazeNeRF (Ruzzi et al., 2023) decompose eye-region deformation into rigid eyeball rotation and non-rigid periocular deformation. EyeNeRF employs an explicit 3D eyeball model to isolate the eyeballs from other facial components, applying rigid transformations to the eyeballs while learning a warping field via neural networks to capture non-rigid deformation. GazeNeRF introduces a two-stream MLP architecture to process the eyeballs and the surrounding facial regions separately, and explicitly applies gaze rotation matrices to the learned eyeball features.

However, as one of the most intricate facial regions, the eye area comprises not only the eyeball and eyelids but also the eyebrows and eyelashes. The deformation of the eyebrows and eyelashes differs fundamentally from that of the eyeball and eyelids, as they are composed of hair and are simultaneously influenced by the deformation of the underlying skin. Given the distinct attributes of the eyeballs, eyelids, eyebrows, and eyelashes, employing specifically designed representations in a compositional framework allows for more accurate modeling of their respective animations, better capturing their unique characteristics and enhancing the realism of gaze-controllable facial generation.

Li et al. (2023) introduce the first learning-based framework for single-view 3D eyebrow reconstruction, representing the eyebrow as a set of fiber curves and formulating the reconstruction

task as a fiber growth problem. This approach enables the modeling of diverse eyebrow styles and lengths, ranging from short and sparse to long and bushy. Moreover, this framework holds the potential for extending to the modeling of eyelashes.

6.3.2 Generation without Finetuning

Finetuning is crucial for existing gaze-controllable face generation models to maintain accurate identity preservation. Personalized models, such as EyeNeRF (Li et al., 2022a) and my own work DeNeRF, require multi-view images and extensive finetuning to adapt to new subjects. However, this adaptation typically takes hours, and acquiring multi-view data for individual subjects is challenging in practice.

Generalized models, including GazeNeRF (Ruzzi et al., 2023) and my recent work LiveGaze, are trained on datasets encompassing diverse subjects. Yet, these models also necessitate finetuning to perform well on new individuals. Specifically, GazeNeRF involves a time-consuming inversion step to obtain personalized latent codes as input. LiveGaze partially addresses this issue by employing a lightweight model for feed-forward processing of a single image, though a brief finetuning phase is still required to fully retain identity details. This ongoing dependency on finetuning hinders the widespread adoption of gaze-controllable face generation methods.

Eliminating the need for finetuning calls for models with strong generalization capabilities. Recent studies have demonstrated that training 3D reconstruction models on large-scale datasets significantly enhances their generalization (Hong et al., 2023; Yang et al., 2025; Kirschstein et al., 2025; Wang et al., 2025). These methods leverage highly scalable transformer-based architectures combined with extensive training data, resulting in powerful, generalized models. Likewise, training gaze-controllable face generation models on large-scale datasets containing numerous identities may effectively eliminate the necessity for finetuning, promoting practical

and widespread usability.

6.4 Final Remarks

The field of gaze-controllable face generation has seen rapid advancements in recent years, driven by breakthroughs in deep generative models. These techniques have provided many possibilities for gaze-controllable face generation, enabling improvements in generation quality, inference speed, and user experience. My work contributes to this progress by addressing key challenges in the field.

I develop a high-fidelity face NeRF model with controllable eyes, significantly enhancing image quality. Additionally, I introduce a real-time gaze redirection model that achieves real-time performance as well as high-quality image generation. To improve usability, I propose a language-guided model specifically designed for gaze-controllable face generation.

I believe this dissertation advances the field by presenting novel methodologies and addressing existing limitations. The findings not only enhance our understanding of gaze behavior in human-computer interaction but also open pathways for future research in areas like virtual reality, accessibility technologies, and personalized user experiences.

References

- Abdal, Rameen, Yipeng Qin, and Peter Wonka (2019). “Image2stylegan: How to embed images into the stylegan latent space?” In: *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4432–4441.
- (2020). “Image2stylegan++: How to edit the embedded images?” In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8296–8305.
- Achiam, Josh et al. (2023). “Gpt-4 technical report”. In: *arXiv preprint arXiv:2303.08774*.
- Alberto Funes Mora, Kenneth and Jean-Marc Odobez (2014). “Geometric generative gaze estimation (g3e) for remote rgb-d cameras”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1773–1780.
- Antipov, Grigory, Moez Baccouche, and Jean-Luc Dugelay (2017). “Face aging with conditional generative adversarial networks”. In: *2017 IEEE international conference on image processing (ICIP)*. IEEE, pp. 2089–2093.
- Athar, ShahRukh et al. (2022). “RigNeRF: Fully Controllable Neural 3D Portraits”. In: *Computer Vision and Pattern Recognition (CVPR)*.
- Bai, Qingyan et al. (2022). “High-fidelity gan inversion with padding space”. In: *European Conference on Computer Vision*. Springer, pp. 36–53.
- Bai, Qingyan et al. (2025). “Real-time 3d-aware portrait editing from a single image”. In: *European Conference on Computer Vision*. Springer, pp. 344–362.

-
- Baluja, Shumeet and Dean Pomerleau (1994). *Non-intrusive gaze tracking using artificial neural networks*. Tech. rep. CARNEGIE-MELLON UNIV PITTSBURGH PA DEPT OF COMPUTER SCIENCE.
- Banf, Michael and Volker Blanz (2009). “Example-based rendering of eye movements”. In: *Computer Graphics Forum*. Vol. 28. 2. Wiley Online Library, pp. 659–666.
- Bao, Yiwei et al. (2020). “Adaptive Feature Fusion Network for Gaze Tracking in Mobile Tablets”. In: *International Conference on Pattern Recognition (ICPR)*.
- Beymer, David and Myron Flickner (2003). “Eye gaze tracking using an active stereo head”. In: *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings*. Vol. 2. IEEE, pp. II–451.
- Blanz, Volker et al. (2004). “Exchanging faces in images”. In: *Computer Graphics Forum*. Vol. 23. 3. Wiley Online Library, pp. 669–676.
- Blattmann, Andreas et al. (2023). “Align your latents: High-resolution video synthesis with latent diffusion models”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22563–22575.
- Bulat, Adrian and Georgios Tzimiropoulos (2017). “How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks)”. In: *Proceedings of the IEEE international conference on computer vision*, pp. 1021–1030.
- Carpenter, Roger HS (1988). *Movements of the Eyes, 2nd Rev.* Pion Limited.
- Ceylan, Duygu, Chun-Hao P Huang, and Niloy J Mitra (2023). “Pix2video: Video editing using image diffusion”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 23206–23217.
- Chan, Eric R et al. (2022). “Efficient geometry-aware 3d generative adversarial networks”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16123–16133.

-
- Chang, Di et al. (2023). “MagicPose: Realistic Human Poses and Facial Expressions Retargeting with Identity-aware Diffusion”. In: *Forty-first International Conference on Machine Learning*.
- Cheng, Yihua and Feng Lu (2022). “Gaze Estimation using Transformer”. In: *ICPR*.
- (Oct. 2023). “DVGaze: Dual-View Gaze Estimation”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20632–20641.
- Cheng, Yihua, Feng Lu, and Xucong Zhang (2018). “Appearance-Based Gaze Estimation via Evaluation-Guided Asymmetric Regression”. en. In: *Computer Vision – ECCV 2018*. Ed. by Vittorio Ferrari et al. Vol. 11218. Cham: Springer International Publishing, pp. 105–121.
- Cheng, Yihua et al. (2021). “Appearance-based Gaze Estimation With Deep Learning: A Review and Benchmark”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- Cheng, Yihua et al. (2024). “What Do You See in Vehicle? Comprehensive Vision Solution for In-Vehicle Gaze Estimation”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1556–1565.
- Crowson, Katherine et al. (2022). “Vqgan-clip: Open domain image generation and editing with natural language guidance”. In: *European Conference on Computer Vision*. Springer, pp. 88–105.
- Delmas, Ginger et al. (2022). “Posescript: 3d human poses from natural language”. In: *European Conference on Computer Vision*. Springer, pp. 346–362.
- Demirel, H Onan and Vincent G Duffy (2007). “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*. Springer, pp. 824–832.

- Deng, Haoping and Wangjiang Zhu (2017). “Monocular Free-Head 3D Gaze Tracking With Deep Learning and Geometry Constraints”. In: *The IEEE International Conference on Computer Vision*.
- Deng, Yu et al. (2019). “Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pp. 0–0.
- Du, Chenpng et al. (2023). “DAE-Talker: High Fidelity Speech-Driven Talking Face Generation with Diffusion Autoencoder”. In: *arXiv preprint arXiv:2303.17550*.
- Esser, Patrick, Robin Rombach, and Bjorn Ommer (2021). “Taming transformers for high-resolution image synthesis”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12873–12883.
- Feng, Yao et al. (2021). “Learning an animatable detailed 3d face model from in-the-wild images”. In: *ACM Transactions on Graphics (ToG)* 40.4, pp. 1–13.
- Feng, Yao et al. (2024). “Chatpose: Chatting about 3d human pose”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2093–2103.
- Fischer, Tobias, Hyung Jin Chang, and Yiannis Demiris (2018). “RT-GENE: Real-Time Eye Gaze Estimation in Natural Environments”. In: *The European Conference on Computer Vision*.
- Freer, Jonathan et al. (2022). “Novel-View Synthesis of Human Tourist Photos”. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3069–3076.
- Funes Mora, Kenneth Alberto, Florent Monay, and Jean-Marc Odobez (2014a). “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*. DOI: [10.1145/2578153.2578190](https://doi.org/10.1145/2578153.2578190).
- Funes Mora, Kenneth Alberto, Florent Monay, and Jean-Marc Odobez (2014b). “Eyediap: A database for the development and evaluation of gaze estimation algorithms from rgb

- and rgb-d cameras”. In: *Proceedings of the Symposium on Eye Tracking Research and Applications*, pp. 255–258.
- Gafni, Guy et al. (2021). “Dynamic neural radiance fields for monocular 4d facial avatar reconstruction”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8649–8658.
- Ganin, Yaroslav et al. (2016). “DeepWarp: Photorealistic Image Resynthesis for Gaze Manipulation”. In: *The European Conference on Computer Vision*.
- Gauthier, Jon (2014). “Conditional generative adversarial nets for convolutional face generation”. In: *Class project for Stanford CS231N: convolutional neural networks for visual recognition, Winter semester 2014.5*, p. 2.
- Ghosh, Shreya et al. (2023). “Automatic gaze analysis: A survey of deep learning based approaches”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 46.1, pp. 61–84.
- Goodfellow, Ian et al. (2014). “Generative adversarial nets”. In: *Advances in neural information processing systems* 27.
- (2020). “Generative adversarial networks”. In: *Communications of the ACM* 63.11, pp. 139–144.
- Grattafiori, Aaron et al. (2024). “The llama 3 herd of models”. In: *arXiv preprint arXiv:2407.21783*.
- Guestrin, Elias Daniel and Moshe Eizenman (2006). “General theory of remote gaze estimation using the pupil center and corneal reflections”. In: *IEEE Transactions on biomedical engineering* 53.6, pp. 1124–1133.
- Guo, Chuan et al. (2022). “Generating diverse and natural 3d human motions from text”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5152–5161.
- Guo, Yudong et al. (2021). “Ad-nerf: Audio driven neural radiance fields for talking head synthesis”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5784–5794.

- Hansen, Dan Witzner et al. (2002). “Eye typing using Markov and active appearance models”. In: *Sixth IEEE Workshop on Applications of Computer Vision, 2002.(WACV 2002). Proceedings*. IEEE, pp. 132–136.
- Hara, Kotaro et al. (2018). “A data-driven analysis of workers’ earnings on Amazon Mechanical Turk”. In: *Proceedings of the 2018 CHI conference on human factors in computing systems*, pp. 1–14.
- He, Kaiming et al. (2016). “Deep residual learning for image recognition”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778.
- He, Zhe et al. (2019). “Photo-Realistic Monocular Gaze Redirection Using Generative Adversarial Networks”. In: *The IEEE International Conference on Computer Vision*.
- Henrysson, Anders, Mark Billinghurst, and Mark Ollila (2005). “Face to face collaborative AR on mobile phones”. In: *Fourth ieee and acm international symposium on mixed and augmented reality (ismar’05)*. IEEE, pp. 80–89.
- Heusel, Martin et al. (2017). “Gans trained by a two time-scale update rule converge to a local nash equilibrium”. In: *Advances in neural information processing systems* 30.
- Ho, Jonathan, Ajay Jain, and Pieter Abbeel (2020). “Denoising diffusion probabilistic models”. In: *Advances in neural information processing systems* 33, pp. 6840–6851.
- Ho, Jonathan et al. (2022). “Video diffusion models”. In: *Advances in Neural Information Processing Systems* 35, pp. 8633–8646.
- Hong, Yang et al. (2022). “Headnerf: A real-time nerf-based parametric head model”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20374–20384.
- Hong, Yicong et al. (2023). “Lrm: Large reconstruction model for single image to 3d”. In: *arXiv preprint arXiv:2311.04400*.
- Hou, Xianxu et al. (2023). “TextFace: Text-to-Style Mapping Based Face Generation and Manipulation”. In: *IEEE Transactions on Multimedia* 25, pp. 3409–3419.

- Huang, Rongjie et al. (2023). “Make-an-audio: Text-to-audio generation with prompt-enhanced diffusion models”. In: *International Conference on Machine Learning*. PMLR, pp. 13916–13932.
- Jack, Rachael E and Philippe G Schyns (2015). “The human face as a dynamic tool for social communication”. In: *Current Biology* 25.14, R621–R634.
- Ji, Qiang and Xiaojie Yang (2002). “Real-time eye, gaze, and face pose tracking for monitoring driver vigilance”. In: *Real-time imaging* 8.5, pp. 357–377.
- Jin, Shiwei et al. (2023). “Redirtrans: Latent-to-latent translation for gaze and head redirection”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5547–5556.
- Jindal, Swati and Xin Eric Wang (2023). “Cuda-ghr: Controllable unsupervised domain adaptation for gaze and head redirection”. In: *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 467–477.
- Johnson, Justin, Alexandre Alahi, and Li Fei-Fei (2016). “Perceptual losses for real-time style transfer and super-resolution”. In: *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part II 14*. Springer, pp. 694–711.
- Karras, Tero, Samuli Laine, and Timo Aila (2019). “A style-based generator architecture for generative adversarial networks”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4401–4410.
- Karras, Tero et al. (2020). “Analyzing and improving the image quality of stylegan”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8110–8119.
- Kellnhofer, Petr et al. (Oct. 2019a). “Gaze360: Physically Unconstrained Gaze Estimation in the Wild”. en. In: *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. Seoul, Korea (South): IEEE, pp. 6911–6920.

-
- (2019b). “Gaze360: Physically Unconstrained Gaze Estimation in the Wild”. In: *The IEEE International Conference on Computer Vision*.
- Kerbl, Bernhard et al. (July 2023). “3D Gaussian Splatting for Real-Time Radiance Field Rendering”. In: *ACM Transactions on Graphics* 42.4. URL: <https://repo-sam.inria.fr/fungraph/3d-gaussian-splatting/>.
- Kirschstein, Tobias et al. (2025). “Avat3r: Large Animatable Gaussian Reconstruction Model for High-fidelity 3D Head Avatars”. In: *arXiv preprint arXiv:2502.20220*.
- Kong, Zhifeng et al. (2020). “Diffwave: A versatile diffusion model for audio synthesis”. In: *arXiv preprint arXiv:2009.09761*.
- Kononenko, Daniil and Victor Lempitsky (2015). “Learning To Look Up: Realtime Monocular Gaze Correction Using Machine Learning”. In: *The IEEE Conference on Computer Vision and Pattern Recognition*.
- Kononenko, Daniil et al. (2017). “Photorealistic monocular gaze redirection using machine learning”. In: *IEEE transactions on pattern analysis and machine intelligence* 40.11, pp. 2696–2710.
- Krafka, Kyle et al. (2016). “Eye Tracking for Everyone”. In: *The IEEE Conference on Computer Vision and Pattern Recognition*.
- Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E Hinton (2012). “Imagenet classification with deep convolutional neural networks”. In: *Advances in neural information processing systems* 25.
- Lewis, John P, Matt Cordner, and Nickson Fong (2000). “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*, pp. 165–172.
- Leyvand, Tommer et al. (2006). “Digital face beautification”. In: *ACM Siggraph 2006 Sketches*, 169–es.

-
- Li, Chenghong et al. (2023). “EMS: 3D Eyebrow Modeling from Single-view Images”. In: *ACM Transactions on Graphics (TOG)* 42.6, pp. 1–19.
- Li, Gengyan et al. (2022a). “EyeNeRF: a hybrid representation for photorealistic synthesis, animation and relighting of human eyes”. In: *ACM Transactions on Graphics (TOG)* 41.4, pp. 1–16.
- Li, Junnan et al. (2022b). “BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation”. In: *ICML*.
- Li, Tianye et al. (2017). “Learning a model of facial shape and expression from 4D scans.” In: *ACM Trans. Graph.* 36.6, pp. 194–1.
- Li, Zhen et al. (2024). “Photomaker: Customizing realistic human photos via stacked id embedding”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8640–8650.
- Lin, Luyang et al. (2024). “Investigating bias in llm-based bias detection: Disparities between llms and human perception”. In: *arXiv preprint arXiv:2403.14896*.
- Linardatos, Pantelis, Vasilis Papastefanopoulos, and Sotiris Kotsiantis (2020). “Explainable ai: A review of machine learning interpretability methods”. In: *Entropy* 23.1, p. 18.
- Lindén, Erik, Jonas Sjostrand, and Alexandre Proutiere (2019). “Learning to personalize in appearance-based gaze tracking”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pp. 0–0.
- Liu, Gang et al. (2019). “A differential approach for gaze estimation”. In: *IEEE transactions on pattern analysis and machine intelligence*.
- Liu, Minghua et al. (2024). “One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization”. In: *Advances in Neural Information Processing Systems* 36.
- Liu, Ruoshi et al. (2023). “Zero-1-to-3: Zero-shot one image to 3d object”. In: *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9298–9309.
- Liu, Yuchen et al. (2022). “3d-fm gan: Towards 3d-controllable face manipulation”. In: *European Conference on Computer Vision*. Springer, pp. 107–125.

- Loshchilov, Ilya and Frank Hutter (2017). “Decoupled weight decay regularization”. In: *arXiv preprint arXiv:1711.05101*.
- Lowe, David G (2004). “Distinctive image features from scale-invariant keypoints”. In: *International journal of computer vision* 60, pp. 91–110.
- Lu, Feng et al. (2014). “Adaptive linear regression for appearance-based gaze estimation”. In: *IEEE transactions on pattern analysis and machine intelligence* 36.10, pp. 2033–2046.
- Ma, Weicheng et al. (2023). “Intersectional stereotypes in large language models: Dataset and analysis”. In: *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 8589–8597.
- Malbouisson, Jorge MC et al. (2005). “Upper and lower eyelid saccades describe a harmonic oscillator function”. In: *Investigative ophthalmology & visual science* 46.3, pp. 857–862.
- Mania, Katerina, Ann McNamara, and Andreas Polychronakis (2021). “Gaze-aware displays and interaction”. In: *ACM SIGGRAPH 2021 Courses*, pp. 1–67.
- Martinez, Francis, Andrea Carbone, and Edwige Pissaloux (2012). “Gaze estimation using local features and non-linear regression”. In: *2012 19th IEEE International Conference on Image Processing*. IEEE, pp. 1961–1964.
- Meng, Kevin et al. (2022a). “Locating and editing factual associations in GPT”. In: *Advances in Neural Information Processing Systems* 35, pp. 17359–17372.
- Meng, Yu et al. (2022b). “Generating training data with language models: Towards zero-shot language understanding”. In: *Advances in Neural Information Processing Systems* 35, pp. 462–477.
- Mildenhall, Ben et al. (2020). “NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis”. In: *ECCV*.
- (2021). “Nerf: Representing scenes as neural radiance fields for view synthesis”. In: *Communications of the ACM* 65.1, pp. 99–106.
- Mirza, Mehdi and Simon Osindero (2014). “Conditional generative adversarial nets”. In: *arXiv preprint arXiv:1411.1784*.

- Morimoto, Carlos H and Marcio RM Mimica (2005). “Eye gaze tracking techniques for interactive applications”. In: *Computer vision and image understanding* 98.1, pp. 4–24.
- Mowrer, OH, Theodore C Ruch, and NE Miller (1935). “The corneo-retinal potential difference as the basis of the galvanometric method of recording eye movements”. In: *American Journal of Physiology-Legacy Content* 114.2, pp. 423–428.
- Müller, Matthias et al. (2005). “Meshless deformations based on shape matching”. In: *ACM transactions on graphics (TOG)* 24.3, pp. 471–478.
- Nasir, Osaïd Rehman et al. (2019). “Text2facegan: Face generation from fine grained textual descriptions”. In: *2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM)*. IEEE, pp. 58–67.
- OpenAI (2024). *Pricing*. <https://openai.com/pricing>.
- Pai, Yun Suen et al. (2016). “GazeSim: simulating foveated rendering using depth in eye gaze for VR”. In: *ACM SIGGRAPH 2016 Posters*, pp. 1–2.
- Park, Keunhong et al. (2021). “Nerfies: Deformable neural radiance fields”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5865–5874.
- Park, Seonwook, Adrian Spurr, and Otmar Hilliges (2018a). “Deep Pictorial Gaze Estimation”. In: *arXiv:1807.10002 [cs]* 11217, pp. 741–757. arXiv: [1807.10002 \[cs\]](https://arxiv.org/abs/1807.10002).
- Park, Seonwook et al. (2018b). “Learning to find eye region landmarks for remote gaze estimation in unconstrained settings”. In: *Proceedings of the 2018 ACM symposium on eye tracking research & applications*, pp. 1–10.
- Park, Seonwook et al. (2019a). “Few-Shot Adaptive Gaze Estimation”. In: *The IEEE International Conference on Computer Vision*.
- (2019b). “Few-shot adaptive gaze estimation”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9368–9377.
- Peng, Sida et al. (2021). “Animatable neural radiance fields for modeling dynamic human bodies”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14314–14323.

-
- Pumarola, Albert et al. (2021). “D-nerf: Neural radiance fields for dynamic scenes”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10318–10327.
- Punnakkal, Abhinanda R et al. (2021). “BABEL: Bodies, action and behavior with english labels”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 722–731.
- Qin, Jiawei, Takuru Shimoyama, and Yusuke Sugano (2022). “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*, pp. 4981–4991.
- Qin, Yalun et al. (2015). “Eye gaze correction with a single webcam based on eye-replacement”. In: *Advances in Visual Computing: 11th International Symposium, ISVC 2015, Las Vegas, NV, USA, December 14-16, 2015, Proceedings, Part I 11*. Springer, pp. 599–609.
- Radford, Alec et al. (2021). “Learning transferable visual models from natural language supervision”. In: *International conference on machine learning*. PMLR, pp. 8748–8763.
- Ramesh, Aditya et al. (2022). “Hierarchical text-conditional image generation with clip latents”. In: *arXiv preprint arXiv:2204.06125* 1.2, p. 3.
- Reed, Scott et al. (2016). “Generative adversarial text to image synthesis”. In: *International conference on machine learning*. PMLR, pp. 1060–1069.
- Rombach, Robin et al. (June 2022). “High-Resolution Image Synthesis With Latent Diffusion Models”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695.
- Ruzzi, Alessandro et al. (2023). “GazeNeRF: 3D-Aware Gaze Redirection with Neural Radiance Fields”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*.
- Saharia, Chitwan et al. (2022). “Photorealistic text-to-image diffusion models with deep language understanding”. In: *Advances in Neural Information Processing Systems* 35, pp. 36479–36494.

-
- Salimans, Tim et al. (2016). “Improved techniques for training gans”. In: *Advances in neural information processing systems* 29.
- Schneider, Flavio (2023). “Archisound: Audio generation with diffusion”. In: *arXiv preprint arXiv:2301.13267*.
- Schott, Eduard (1922). “Über die Registrierung des Nystagmus und anderer Augenbewegungen verm itteles des Saitengalvanometers”. In: *Deut Arch fur klin Med* 140, pp. 79–90.
- Schwarz, Katja et al. (2020). “Graf: Generative radiance fields for 3d-aware image synthesis”. In: *Advances in Neural Information Processing Systems* 33, pp. 20154–20166.
- Sheng, Emily et al. (2019). “The woman worked as a babysitter: On biases in language generation”. In: *arXiv preprint arXiv:1909.01326*.
- Shrivastava, Ashish et al. (2017a). “Learning From Simulated and Unsupervised Images Through Adversarial Training”. In: *The IEEE Conference on Computer Vision and Pattern Recognition*.
- (2017b). “Learning from simulated and unsupervised images through adversarial training”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2107–2116.
- Shu, Zhixin et al. (2016). “Eyeopener: Editing eyes in the wild”. In: *ACM Transactions on Graphics (TOG)* 36.1, pp. 1–13.
- Simonyan, Karen and Andrew Zisserman (2014). “Very deep convolutional networks for large-scale image recognition”. In: *arXiv preprint arXiv:1409.1556*.
- Skodras, Evangelos, Vasileios G Kanas, and Nikolaos Fakotakis (2015). “On visual gaze tracking based on a single low cost camera”. In: *Signal Processing: Image Communication* 36, pp. 29–42.
- Stampe, Dave M (1993). “Heuristic filtering and reliable calibration methods for video-based pupil-tracking systems”. In: *Behavior Research Methods, Instruments, & Computers* 25.2, pp. 137–142.

-
- Sun, Jianxin et al. (2021). “Multi-caption text-to-face synthesis: Dataset and algorithm”. In: *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 2290–2298.
- Sun, Jianxin et al. (2022). “Anyface: Free-style text-to-face synthesis and manipulation”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18687–18696.
- Sun, Kun, Rong Wang, and Anders Søgaard (2024). “Comprehensive Reassessment of Large-Scale Evaluation Outcomes in LLMs: A Multifaceted Statistical Approach”. In: *arXiv preprint arXiv:2403.15250*.
- Tang, Chao et al. (2023). “Graspgpt: Leveraging semantic knowledge from a large language model for task-oriented grasping”. In: *IEEE Robotics and Automation Letters* 8.11, pp. 7551–7558.
- Team, Gemini et al. (2023). “Gemini: a family of highly capable multimodal models”. In: *arXiv preprint arXiv:2312.11805*.
- Titova, Angelina et al. (2024). “Extreme Tolerance of Extraocular Muscles to Diseases and Aging: Why and How?” In: *International Journal of Molecular Sciences* 25.9, p. 4985.
- Trevithick, Alex et al. (2023). “Real-time radiance fields for single-image portrait view synthesis”. In.
- Tse, Tze Ho Elden et al. (2022). “S 2 Contact: Graph-Based Network for 3D Hand-Object Contact Estimation with Semi-supervised Learning”. In: *European Conference on Computer Vision*. Springer, pp. 568–584.
- Valenti, Roberto, Nicu Sebe, and Theo Gevers (2011). “Combining head pose and eye location information for gaze estimation”. In: *IEEE Transactions on Image Processing* 21.2, pp. 802–815.
- Vanezis, P et al. (1989). “Application of 3-D computer graphics for facial reconstruction and comparison with sculpting techniques”. In: *Forensic science international* 42.1-2, pp. 69–84.

- Vaswani, Ashish et al. (2017). “Attention is all you need”. In: *Advances in Neural Information Processing Systems*, pp. 5998–6008.
- Wang, Hengfei et al. (2023a). “GazeCaps: Gaze Estimation With Self-Attention-Routed Capsules”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2668–2676.
- Wang, Hengfei et al. (2023b). “High-Fidelity Eye Animatable Neural Radiance Fields for Human Face”. In: *34th British Machine Vision Conference 2023, BMVC 2023, Aberdeen, UK, November 20-24, 2023*. BMVA. URL: <https://papers.bmvc2023.org/0105.pdf>.
- (2023c). “High-fidelity eye animatable neural radiance fields for human face”. In: *BMVC*.
- (2024). “Textgaze: Gaze-controllable face generation with natural language”. In: *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 7143–7151.
- Wang, Jianyuan et al. (2025). “VGGT: Visual Geometry Grounded Transformer”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- Wang, Jun et al. (2021). “Facex-zoo: A pytorch toolbox for face recognition”. In: *Proceedings of the 29th ACM international conference on multimedia*, pp. 3779–3782.
- Wang, Kang, Rui Zhao, and Qiang Ji (2018). “A hierarchical generative model for eye image synthesis and eye gaze estimation”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 440–448.
- Wang, Zhiyong, Jinxiang Chai, and Shihong Xia (2019). “Realtime and accurate 3D eye gaze capture with DCNN-based iris and pupil segmentation”. In: *IEEE transactions on visualization and computer graphics* 27.1, pp. 190–203.
- Wang, Zhou et al. (2004). “Image quality assessment: from error visibility to structural similarity”. In: *IEEE transactions on image processing* 13.4, pp. 600–612.
- Ward, Kelly et al. (2007). “A survey on hair modeling: Styling, simulation, and rendering”. In: *IEEE transactions on visualization and computer graphics* 13.2, pp. 213–234.

- Wei, Xiaobao et al. (2024). “GazeGaussian: High-Fidelity Gaze Redirection with 3D Gaussian Splatting”. In: *arXiv preprint arXiv:2411.12981*.
- Weng, Chung-Yi et al. (2022). “Humannerf: Free-viewpoint rendering of moving people from monocular video”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16210–16220.
- Williams, Oliver, Andrew Blake, and Roberto Cipolla (2006). “Sparse and Semi-supervised Visual Mapping with the S⁺ 3GP”. In: *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’06)*. Vol. 1. IEEE, pp. 230–237.
- Wolf, Lior, Ziv Freund, and Shai Avidan (2010). “An eye for an eye: A single camera gaze-replacement method”. In: *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, pp. 817–824.
- Wood, E. et al. (Dec. 2015a). “Rendering of Eyes for Eye-Shape Registration and Gaze Estimation”. In: *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 3756–3764.
- Wood, Erroll et al. (2015b). “Rendering of Eyes for Eye-Shape Registration and Gaze Estimation”. In: *The IEEE International Conference on Computer Vision*.
- Wood, Erroll et al. (2016a). “A 3D Morphable Eye Region Model for Gaze Estimation”. In: *The European Conference on Computer Vision*.
- Wood, Erroll et al. (Mar. 2016b). “Learning an Appearance-Based Gaze Estimator from One Million Synthesised Images”. en. In: *Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications*. Charleston South Carolina: ACM, pp. 131–138.
- (2016c). “Learning an Appearance-Based Gaze Estimator from One Million Synthesised Images”. In: *Proceedings of the ACM Symposium on Eye Tracking Research & Applications*, pp. 131–138. ISBN: 9781450341257. DOI: [10.1145/2857491.2857492](https://doi.org/10.1145/2857491.2857492).
- Wood, Erroll et al. (2018). “Gazedirector: Fully articulated eye gaze redirection in video”. In: *Computer Graphics Forum*. Vol. 37. 2. Wiley Online Library, pp. 217–225.

- Xie, You et al. (2024). “X-portrait: Expressive portrait animation with hierarchical motion attention”. In: *ACM SIGGRAPH 2024 Conference Papers*, pp. 1–11.
- Xiong, Yunyang, Hyunwoo J Kim, and Vikas Singh (2019). “Mixed effects neural networks (menets) with applications to gaze estimation”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7743–7752.
- Xu, Tao et al. (2018). “Attngan: Fine-grained text to image generation with attentional generative adversarial networks”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1316–1324.
- Yang, Guo-Wei et al. (2022). “Recursive-NeRF: An efficient and dynamically growing NeRF”. In: *IEEE Transactions on Visualization and Computer Graphics*.
- Yang, Jianing et al. (2025). “Fast3R: Towards 3D Reconstruction of 1000+ Images in One Forward Pass”. In: *arXiv preprint arXiv:2501.13928*.
- Ye, Hu et al. (2023a). “Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models”. In: *arXiv preprint arXiv:2308.06721*.
- Ye, Zhenhui et al. (2023b). “Geneface: Generalized and high-fidelity audio-driven 3d talking face synthesis”. In.
- Yeh, Kai-Ching et al. (2023). “Evaluating interfaced llm bias”. In: *Proceedings of the 35th Conference on Computational Linguistics and Speech Processing (ROCLING 2023)*, pp. 292–299.
- Young, Laurence R and David Sheena (1975). “Survey of eye movement recording methods”. In: *Behavior research methods & instrumentation* 7.5, pp. 397–429.
- Yu, Changqian et al. (2018). “Bisenet: Bilateral segmentation network for real-time semantic segmentation”. In: *Proceedings of the European conference on computer vision (ECCV)*, pp. 325–341.
- Yu, Yu, Gang Liu, and Jean-Marc Odobez (2019a). “Improving Few-Shot User-Specific Gaze Adaptation via Gaze Redirection Synthesis”. In: *The IEEE Conference on Computer Vision and Pattern Recognition*.

-
- (2019b). “Improving few-shot user-specific gaze adaptation via gaze redirection synthesis”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11937–11946.
- Yu, Yu and Jean-Marc Odobez (2020). “Unsupervised Representation Learning for Gaze Estimation”. In: *The IEEE Conference on Computer Vision and Pattern Recognition*.
- Zhang, Han et al. (2017a). “Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks”. In: *Proceedings of the IEEE international conference on computer vision*, pp. 5907–5915.
- Zhang, Lvmin, Anyi Rao, and Maneesh Agrawala (2023). “Adding conditional control to text-to-image diffusion models”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3836–3847.
- Zhang, Richard et al. (2018). “The unreasonable effectiveness of deep features as a perceptual metric”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595.
- Zhang, Xucong et al. (June 2015a). “Appearance-Based Gaze Estimation in the Wild”. In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4511–4520. arXiv: [1504.02863](https://arxiv.org/abs/1504.02863).
- (2015b). “Appearance-Based Gaze Estimation in the Wild”. In: *The IEEE Conference on Computer Vision and Pattern Recognition*.
- (2017b). “It’s written all over your face: Full-face appearance-based gaze estimation”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 51–60.
- Zhang, Xucong et al. (2017c). “It’s written all over your face: Full-face appearance-based gaze estimation”. In: *The IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 2299–2308.
- (2017d). “Mpiigaze: Real-world dataset and deep appearance-based gaze estimation”. In: *IEEE transactions on pattern analysis and machine intelligence* 41.1, pp. 162–175.

-
- (Jan. 2019). “MPIIGaze: Real-World Dataset and Deep Appearance-Based Gaze Estimation”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41.1, pp. 162–175. ISSN: 1939-3539. DOI: [10.1109/TPAMI.2017.2778103](https://doi.org/10.1109/TPAMI.2017.2778103).
- Zhang, Xucong et al. (2020). “ETH-XGaze: A Large Scale Dataset for Gaze Estimation under Extreme Head Pose and Gaze Variation”. In: *The European Conference on Computer Vision*.
- Zhang, Zhongqun et al. (2022). “Trans6D: Transformer-Based 6D Object Pose Estimation and Refinement”. In: *European Conference on Computer Vision*. Springer, pp. 112–128.
- Zheng, Linfang et al. (2023). “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*, pp. 17163–17173.
- Zheng, Yufeng et al. (2020). “Self-Learning Transformations for Improving Gaze and Head Redirection”. In: *Advances in Neural Information Processing Systems*.
- Zhou, Xiaolong et al. (2019). “Learning a 3D gaze estimator with improved Itracker combined with bidirectional LSTM”. In: *2019 IEEE international conference on Multimedia and expo (ICME)*. IEEE, pp. 850–855.
- Zhu, Zhiwei and Qiang Ji (2007). “Novel eye gaze tracking techniques under natural head movement”. In: *IEEE TRANSACTIONS on biomedical engineering* 54.12, pp. 2246–2260.
- Zhuang, Yiyu et al. (2022). “Mofanerf: Morphable facial neural radiance field”. In: *European conference on computer vision*. Springer, pp. 268–285.