GazeGene: Large-scale Synthetic Gaze Dataset with 3D Eyeball Annotations

Yiwei Bao Zhiming Wang Feng Lu [*]
State Key Laboratory of VR Technology and Systems, School of CSE, Beihang University
{baoyiwei, zy2306418, lufeng}@buaa.edu.cn

Abstract

Thanks to the introduction of large-scale datasets, deep-learning has become the mainstream approach for appearance-based gaze estimation problems. However, current large-scale datasets contain annotation errors and provide only a single vector for gaze annotation, lacking key information such as 3D eyeball structures. Limitations in annotation accuracy and variety have constrained the progress in research and development of deep-learning methods for appearance-based gaze-related tasks. In this paper, we present GazeGene, a new large-scale synthetic gaze dataset with photo-realistic samples. More importantly, GazeGene not only provides accurate gaze annotations, but also offers 3D annotations of vital eye structures such as the pupil, iris, eyeball, optical and visual axes for the first time. Experiments show that GazeGene achieves comparable quality and generalization ability with real-world datasets, even outperforms most existing datasets on high-resolution images. Furthermore, its 3D eyeball annotations expand the application of deep-learning methods on various gaze-related tasks, offering new insights into this field. The dataset is available at: https://phi-ai.buaa.edu.cn/GazeGene/.

1. Introduction

Vision is one of the primary senses in humans. Thus, gaze direction provides valuable insights for understanding human interaction intentions and environmental perception processes. In recent years, gaze estimation technology has demonstrated significant potential for applications in human-computer interaction and cognitive analysis. For example, gaze is widely utilized for interaction in VR [6, 27, 38] and AR [29, 43] systems, and it also plays an important role in medical diagnosis [8, 24, 25] and intelligent cockpit systems [5, 34].

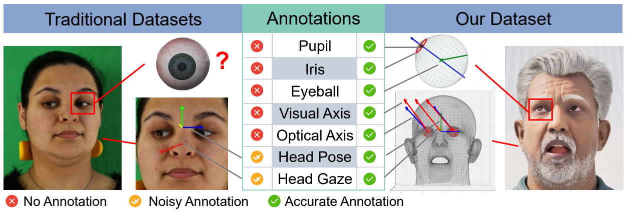

Figure 1. We propose a large-scale synthetic gaze datset GazeGene, which offers accurate and comprehensive 3D annotations of eyeball structures for the first time.

Gaze estimation techniques could be classified into two categories: model-based methods and appearance-based methods. Model-based methods usually estimate gaze by measuring eyeball structures through dedicated hardwares such as infrared cameras and lights. These methods are often deployed in specialized devices such as eye trackers for their high accuracy and cost. On the contrary, recent appearance-based methods directly estimate gaze from the eye or face images using deep-learning techniques. These methods require only a single web camera, giving it the potential to be applied in everyday devices such as smartphones.

Deep-learning gaze estimation methods have achieve tremendous progress in the last ten years, thanks to the advancements in computational hardware and the introduction of large datasets such as ETH-XGaze [52]. However, the annotation accuracy and variety of existing large gaze datasets are limited, hindering the further research and development of deep-learning based methods. Most large-scale appearance-based datasets annotate gaze direction by calibrated geometric settings. There is an inevitable error during calibration and 3D head pose estimation, result in inaccurate gaze annotations. Moreover, existing remote gaze datasets only provide annotation of a single gaze direction originated from the center of user faces, missing annotations for critical 3D eyeball structures such as eyeball center, radius and pupil. Consequently, deep-learning methods lack the capability to understand human eyeball structures or to perform more complex gaze-related tasks, such as optical and visual axes estimation.

The main challenge of providing accurate annotations

[*] Corresponding Author. This work was supported by Beijing Natural Science Foundation (L242019) and National Natural Science Foundation of China (NSFC, no. 62372019)

Table 1. Overview of popular gaze estimation datasets. Gaze range and head range are the approximated distribution range of yaw and pitch angles.

Datasets	**subjects**	**total**	**gaze range**	**head range**	**Annotation**	
			(yaw, pitch)	(yaw, pitch)	head gaze	3D eyeball
EyeDiap [17]	16	237 min	$\pm 25, 0 \sim 20$	$\pm 15, -10 \sim 30$	✓	✗
MPIIGaze [49]	15	213,659	$\pm 30, -30 \sim 10$	$\pm 30, \pm 40$	✓	✗
EVE [37]	54	4.2K videos	$\pm 50, \pm 30$	$\pm 40, -40 \sim 45$	✓	✗
Gaze360 [23]	238	172,000	$\pm 140, -60 \sim 20$	±90,unknown	✓	✗
ETH-XGaze [52]	110	1,083,492	$\pm 120, \pm 70$	$\pm 80, \pm 80$	✓	✗
GazeGene	56	1,008,000	$\pm 120, \pm 70$	$\pm 80, \pm 80$	✓	✓

for 3D eyeball structures is that eyeball structure is subtle and mostly hidden inside human face. Although 3D eyeball structures could be measured by dedicated hardwares similar to eye trackers, the distribution of gaze, head pose would be limited to a very small range. Attempts have been made to alleviate this problem. Wood *et al*. proposed UnityEyes, a synthetic gaze dataset [45]. But UnityEyes only synthesizes a small facial region around the left eye with simple gaze direction annotations. The significant domain gap between UnityEyes and real-world datasets limited its practical use.

To address this challenge, we propose the GazeGene, a synthesized large-scale full-face gaze estimation dataset with detailed annotations of 3D eyeball structures. To synthesize realistic and diverse samples, we employ a high-fidelity virtual character and design comprehensive data synthesis strategies to include wide distributions of gaze, head pose, expression and illumination. The proposed dataset contains over 1 million full-face images with comparable quality and diversity with the largest existing gaze dataset. More importantly, GazeGene provides thorough and accurate 3D annotations of eyeball structures, including eyeball center, eyeball radius, eyeball mesh, pupil, iris and so on. We hope that these data will expand the design possibilities for appearance-based methods and drive their further development.

Experiments show that **in traditional head gaze estimation tasks, GazeGene achieves comparable performance with existing gaze datasets, proves that the appearances of GazeGene is close to that of real-world datasets.** It even outperforms other real-world datasets significantly when tested on ETH-XGaze, the only large-scale, high-definition dataset. Moreover, **leveraging our comprehensive annotations, we conduct validation experiments to explore the potential of deep-learning methods in more complex tasks, *e.g*. 3D eyeball structure estimation, variousand visual axes estimation, *etc*.** The results of these verification experiments demonstrate that the unique annotations in our dataset have the potential to advance the application of deep-learning methods in various gaze-related tasks. The main contributions of this paper are threefold:

- We propose the GazeGene dataset, a large scale synthesized dataset (over 1 million image from 56 subjects) for gaze estimation covering wide gaze, head pose, expression and light condition distribution range with high gaze annotation accuracy and image resolution.
- Detailed and accurate annotations for 3D eye structures including pupil, iris, optical axis, visual axis, eyeball and face mesh. As far as we know, GazeGene is the first large scale remote gaze dataset providing such labels.
- Extensive experiments to demonstrate the variety and quality of the GazeGene dataset and explore several tasks regarding 3D eye structures, providing new insights for appearance-based gaze estimation methods.

2. Related Work

2.1. Gaze Estimation Techniques

Gaze estimation techniques are primarily categorized into two types, model-based methods and appearance-based methods. Model-based methods reconstruct 3D eyeball structures through anatomical prior and dedicated devices such as infrared cameras and lights [18, 19, 40]. Although model-based methods achieve satisfying accuracy, they are often deployed in specialized devices such as eye trackers and VR/AR headsets for their strict hardware requirements and high costs.

Appearance-based methods aim to directly estimate gaze from eye or face images[14]. In 2015, Zhang *et al*. proposed the first deep-learning gaze estimation method, which directly estimate gaze from eye images [48]. More eye-based deep-learning approaches have been proposed in the following years, with different techniques such as asymmetric network [12], pictorial feature [35] and few-shot learning [36]. Later, researchers found the extra information in the full-face images helpful in gaze estimation. Various face-based approaches have been proposed, including spatial attention [50], dilated-convolutions [9], coarse-to-

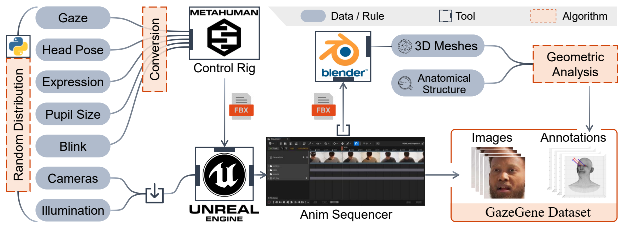

Figure 2. Overview of the GazeGene synthetic pipeline.

fine network [11], adaptive feature fusion [2] and transformer [10]. More recent works focus on unsupervised domain adaptation [3, 7, 33, 42] and domain generalization [1, 13, 32, 46] for gaze estimation. Ververas *et al*. align a 3D eyeball mesh according to the iris to estimate eyeball center in the normalized space [41]. Overall, appearance-based methods usually estimate a single gaze vector. Estimation of 3D eyeball structures in physical space is still challenging due to the lack of annotated data.

2.2. Gaze Estimation Datasets

The development of deep learning gaze estimation methods is inseparable from the introduction of various gaze datasets. There are two primary collection procedures for gaze datasets: eye-tracker-based and geometric-based. Eye-tracker-based datasets utilize eye trackers to collect user gaze annotation. RT-Gene dataset collected user gaze by eye-tracking glasses [16]. Park *et al*. recorded over 4K videos annotated by a desktop eye-tracker [37]. Eye-tracker based datasets often provide high accuracy annotations. But their gaze and head pose varieties are also limited by the eye-trackers.

Geometric-based gaze datasets annotate gaze direction by the vector from user head center to the gaze target. A number of gaze datasets are collected under desktop settings, such as EyeDiap [17], MPIIGaze [49], GazeCapture [28], ShanghaiTechGaze [30] and EVE [37]. To provide a wider gaze and head pose range, Kellnhofer *et al*. recorded subejcts and gaze targets by a 360 degree camera, presenting the Gaze360 dataset [23]. Zhang *et al*. collected the ETH-XGaze [52] dataset by 18 cameras under laboratory environment, which is currently the only high-resolution dataset with large gaze and head pose range. There are also some synthetic gaze datasets, *e.g*. UT-Multiview [39], UnityEye [45] and [31]. But they only synthesize eye region images with 2D annotations instead of full face images with 3D annotations.

In summary, existing datasets do not provide annotations of the 3D structures of the human eyeball due to the high difficulty of measurement. The lack of data hinders the further development of deep learning gaze estimation methods. Our GazeGene dataset fills this gap, offering high-resolution images with large gaze range and thorough annotation for 3D eyeball structures. A comprehensive comparison between the GazeGene and existing gaze estimation datasets is shown in Tab. 1

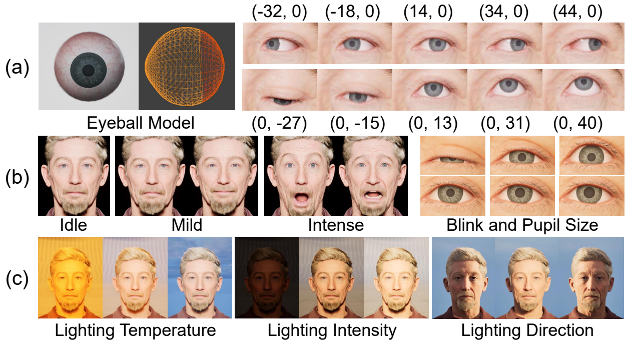

Figure 3. Examples of: (a) Eyeball model and eye appearance. Numbers are gaze yaw and pitch angles in degree. (b) Expression, blinking amplitude and pupil size. (c) Lighting temperature, intensity and direction.

3. GazeGene Dataset

Subtle anatomical eye structures such as pupil, iris, eyeball center and radius are extremely difficult to measure in real world, especially in the remote settings where the camera is positioned at a certain distance from the face. Thus, to provide accurate annotations, we propose to synthesize face images and corresponding annotations by virtual avatars. In the following sections, we introduce the data synthesis techniques, data synthesis strategy and data characteristics in detail.

3.1. Virtual Avatar and Platform

The virtual avatar and redering platform affect the appearance of the synthesized images significantly. We choose MetaHuman in the Unreal Engine (UE) as our virtual avatar.

MetaHuman is a complete framework that gives users the power to create and animate highly realistic digital human characters made by Epic. MetaHuman provides detailed eyeball models with over 800 vertices, which models the spherical shape of the cornea through meshes. The eyeball model and the eye appearance across different gaze directions of the MetaHuman is shown in Fig. 3, (a). Along with the detailed face mesh over 20K vertices, we are able to synthesize photo-realistic samples. However, as a tool for creating virtual avatars instead of annotating them, MetaHuman does not have any functionality for providing annotations. Incorporating anatomical priors, we calculate all annotations from the 3D meshes of the MetaHuman through a complex pipeline involving multiple platforms and tools.

The overall synthesis pipeline is shown in Fig. 2: first,

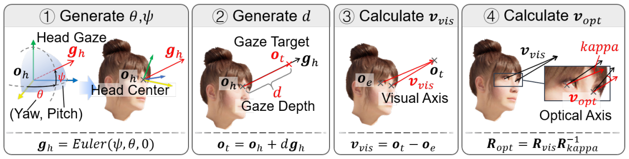

Figure 4. The process of determining eyeball rotation, *i.e*. the optic axis during data generation.

we random generate the desired face, eye and environmental parameters (introduced in Sec. 3.2). Then, we import these parameters into Unreal Engine to control the MetaHuman character and render face images. Lastly, we import the generated face animation sequences into Blender and calculate the 3D face and eyeball annotations from the original meshes. Note that according to the Epic Content License Agreement, using these data as a training input or prompt-based input into any Generative AI Programs is forbidden.

3.2. Data Synthesis Strategy

To ensure the diversity in the data distribution of our synthetic dataset, numerous factors must be considered, such as gaze direction, head pose, expression and illumination. In the following sections, we introduce the data synthesis strategy of the proposed GazeGene for each factor.

3.2.1. Gaze and Head Pose

Gaze and head pose are two of the most basic elements of gaze datasets. In this section, we introduce the synthesis strategy of head pose and head gaze. Specifically, we define head gaze $\mathbf{g}_h$ as the unit direction vector from the subject's head center $\mathbf{o}_h$ to the gaze target $\mathbf{o}_t$. This definition is commonly used in the previous datasets [17, 23, 50, 52]. We randomly generate head gaze yaw and pitch angles from range $[-36.8°, 36.84°]$ and $[-26.54°, 28.53°]$ within the Head Coordinate System (HCS), respectively. For head pose, we randomly generate yaw and pitch angles from range $[-40°, 40°]$. This distribution range is comparable with the ETH-XGaze dataset, the largest existing dataset in terms of distribution range.

3.2.2. Eyeball Parameters

To precisely control the rotation of two eyeballs of the virtual character, we need to take more complex factors into consideration, including kappa angle, gaze vergence depth, optical and visual axes. Fundamentally speaking, the rotation of each eye is determined by the visual axis, *i.e*. the line connecting the fovea, eyeball center $\mathbf{o}_e$, and the gaze target $\mathbf{o}_t$ [18]. We first determine the position of the gaze target by $\mathbf{o}_t = \mathbf{o}_h + d\mathbf{g}_\mathbf{h}$, where d is the gaze vergence depth. The direction of the visual axis $\mathbf{v}_{vis}$ are acquired by:

$$\hat{\mathbf{v}}_{vis} = \frac{\mathbf{o}_t - \mathbf{o}_e}{\|\mathbf{o}_t - \mathbf{o}_e\|}. \tag{1}$$

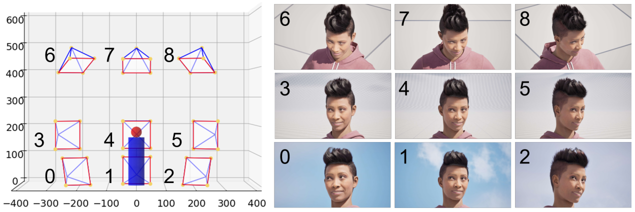

Figure 5. Visualization of the 9 cameras in GazeGene and their captured image samples.

However, fovea is not modeled in MetaHuman characters. We have to calculate the rotation of the eyeball from the optical axis. Given the visual axis $\hat{\mathbf{v}}_{vis}$, its corresponding rotation matrix $\mathbf{R}_{vis}$ with no roll, and the rotation matrix of kappa under the Eyeball Coordinate System $\mathbf{R}_{kappa}$, we calculate the rotation matrix of the optical axis $\mathbf{R}_{opt}$ by:

$$\mathbf{R}_{opt} = \mathbf{R}_{visl} * \mathbf{R}_{kappa}^{-1}, \tag{2}$$

the last column of $\mathbf{R}_{opt}$ is the direction of the optical axis $\hat{\mathbf{v}}_{opt}$. $\hat{\mathbf{v}}_{opt}$ is consistent with the direction from the eyeball center $\mathbf{o}_e$ to pupil center $\mathbf{p}$, which determines the rotation of each eyeball. The whole process is shown in Fig. 4

For the gaze vergence depth d, we randomly generate values within the range of $[0.2, 5]$ meters for 75% of the data to simulate the indoor scenario. We set the depth of the remaining 25% of the data to 1000 meters to simulate the scenario of distant gaze. The average kappa angle various across different literature and different measurement methods [4, 18, 20]. We set the average horizontal and vertical kappa angles to around 4 and 1 degree, with a standard deviation around 0.5 degree across different subjects.

3.2.3. Expressions and Illumination

Expressions and illumination are also important characteristics in gaze datasets since they significantly impact facial appearance. In our dataset, we control the overall facial expressions and eye appearances separately. MetaHuman provides an expression library containing six distinct expression types, each available at four different intensity levels. We set 25% of the data to contain subtle expressions, generated by randomly combining the two lowest intensity expressions. Another 50% of the data is set to contain intense expressions, generated by a linear combination of four expressions at random intensities. The last 25% of the data remains default expressions. Independently of facial expressions, we randomly set the degree of blinking and pupil size to introduce more variety, as shown in Fig. 3, (b).

The illumination in our dataset mainly comes from a directional light actor in Unreal Engine. The light source is positioned above the virtual character, as in most real-world scenarios. We randomly rotate the light source around the character and change the light intensity from 0.25 lux to 18

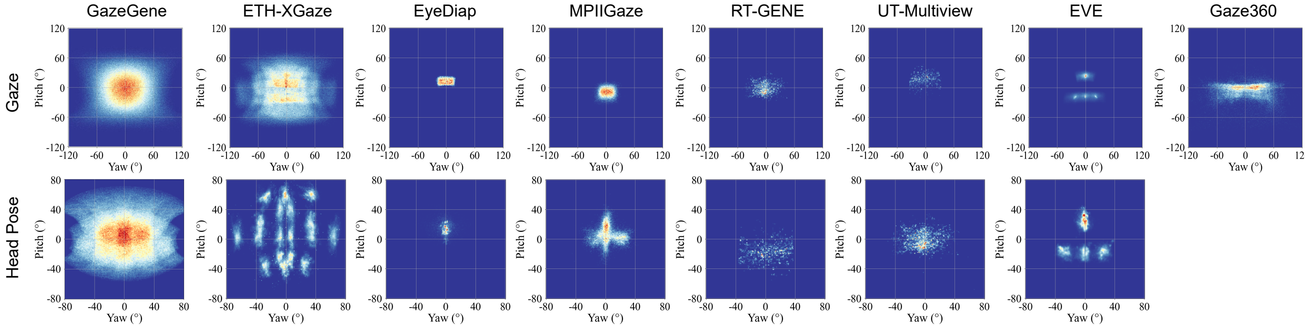

Figure 6. Gaze (top) and head pose (bottom) distribution of different datasets.

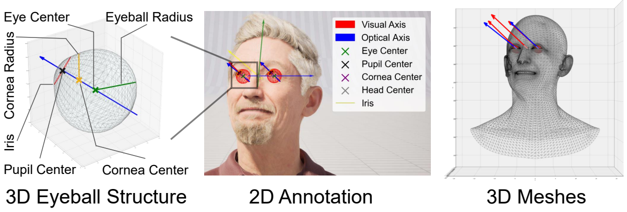

Figure 7. Visualization of annotations including gaze, 3D meshes and 3D eyeball structures in GazeGene.

lux. In addition, we also randomly change the temperature of the light source from 1700 to 12000, simulating different light sources in real-world such as LED and the sunlight, as shown in Fig. 3, (c).

3.2.4. Camera Settings

We capture the virtual character with 9 cameras simultaneously to support multi-view algorithms. All virtual cameras use standard perspective projection, with a sensor size of 23.76×13.365 mm, a resolution of 2560×1440, and a focal length of 200 mm. All cameras are approximately positioned on a sphere centered on the character, with a radius of about 5 meters, as shown in Fig. 5. We provide precise camera intrinsics and extrinsics with no errors along with the dataset.

3.3. Data Characteristics

In total, we synthesize data from 56 virtual characters, with different gender, ethnicity, age and appearance. Subjects of GazeGene can be easily scaled, since any digital human generated by the MetaHuman Creator can be incorporated into our synthesis pipeline without any modification. We synthesize 2000 frames for each subject, result in $1,080,000$ images from all 9 cameras. As shown in Fig. 6, the GazeGene dataset provide comparable gaze and head pose range with the largest existing dataset ETH-XGaze [52], with even larger gaze extremum angles as large as $\pm 170°$ and $\pm 90°$ for yaw and pitch respectively. Considering that ETH-XGaze only public annotations of about $800,000$ frames, the GazeGene dataset is the largest appearance-based gaze dataset in terms of data volume and distribution range as far as we know.

3.3.1. Comprehensive and Precise Annotations

More importantly, the GazeGene dataset provides comprehensive and precise annotations 3D eyeball structures, which are not available in any existing appearance-based datasets. Based on the shape of the 3D eyeball mesh and prior knowledge of eyeball anatomy, we use geometric analysis such as spherical fitting to compute the 3D coordinates of structures including the eyeball center, pupil, and iris. Then, we calculate the directions of the optical and visual axes correspondingly. Moreover,we also provide detailed eyeball parameters including eyeball radius, cornea radius and kappa angles for each subject. Visualization of these annotations is shown in Fig. 7.

Our dataset provides massive amount of data with accurate and comprehensive 3D labels for eyeball structures, which is the underlying cause of eye gaze. Thus, these annotations could potentially be helpful for appearance-based gaze estimation algorithms. Besides, the proposed dataset allows further gaze estimation researches to explore the potential of deep-learning techniques on various gaze-related tasks under remote settings, *e.g.* optical and visual axes estimation, 3D pupil, iris and eyeball estimation, *etc*.

4. Experiments

4.1. GazeGene as a Traditional Gaze Dataset

In this section, we validate the overall quality and variety of the proposed GazeGene dataset as a traditional gaze dataset, *i.e.* using only the head gaze annotations in GazeGene, similar to existing datasets. We validate our dataset in both within-domain and cross-domain experiments.

4.1.1. Datasets and Implementation Details

Other than the proposed dataset, we conduct experiments on 5 commonly used gaze datasets: ETH-XGaze (ETH) [52],

Table 2. Within domain gaze estimation error in degrees. Results of SOTA methods on datasets except GazeGene comes from [14].

Method	GazeGene	ETH	MPII	G360	ED
FullFace[50]	3.54	7.38	4.93	14.99	6.53
Dilated-Net[9]	3.08	N/A	4.42	13.73	6.19
GazeTR-H [10]	3.37	N/A	4.00	10.62	5.17
ResNet-18	3.76	5.73	4.70	12.23	6.13
ResNet-50	3.36	5.32	4.20	11.61	5.12

MPIIFaceGaze (MPII) [50], Gaze360 (G360) [23], EyeDiap (ED) [17] and EVE [37]. **ETH-XGaze:** We divide the last 10 subjects as the test set. **Gaze360:** We only use samples with frontal faces in our experiments and follow the official training and testing splits. **MPIIFaceGaze:** We use the $45k$ images from the official test set in our experiments. **EyeDiap:** $16k$ frames captured from the original videos. Datasets except for Gaze360 is normalized following [51], including GazeGene. For the GazeGene dataset, we divide the last 10 subjects as the test set.

We use ResNet-18 and ResNet-50 [21] as backbone models, both implemented in PyTorch. The Adam optimizer [26] is applied with beta values of $(0.5, 0.95)$ and a learning rate of 10^{-4}. The models are trained for 10 epochs with batch sizes of 512 and 256 for ResNet-18 and ResNet-50, respectively. The loss function is the L1 loss between the predicted 3D unit head gaze vector and the ground truth. Face images are resized to 224×224.

4.1.2. Within-Domain Evaluation

Within-domain evaluation is a commonly used measurement for evaluating gaze estimation methods. We test 3 SOTA gaze estimation methods FullFace [50], Dilated-Net [9] and GazeTR [10] along with the two backbone models in 5 different datasets. As shown in Tab. 2, all 5 models achieve the best performance on the GazeGene dataset, which proves that the high annotation accuracy of our dataset is beneficial for gaze estimation model training.

Given the similarity between our dataset and ETH-XGaze in terms of gaze range, image number and resolution, we further analyze the experimental results on both datasets. Compared to the results on ETH-XGaze, the accuracies of two ResNet models are around 2 degree lower on GazeGene. One possible reason is that ETH-XGaze suffers from the inaccurate annotation caused by the inevitable errors in the camera calibration and 3D head pose detection. To verify such hypothesis, we convert gaze and head pose annotations from different cameras into camera 0 and calculate the angular difference between the the converted results and the annotations from camera 0 in ETH-XGaze. As shown in Fig. 8, there is a $2.91°$ and $4.96°$ multi-view inconsistency between different camera views for gaze and head pose, respectively, which is an approximation of the annotation errors in the ETH-XGaze. The magnitude of the gaze annotation inconsistency in ETH-XGaze is similar to the accuracy difference between ETH-XGaze and GazeGene, which supports our hypothesis.

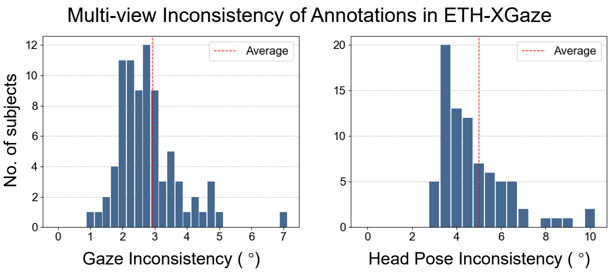

Figure 8. The inconsistency of gaze and head pose annotation between different camera views in the ETH-XGaze dataset. This inconsistency demonstrates the errors in annotations of real-world datasets.

4.1.3. Cross-Domain Evaluation

One of the main concerns for synthesized datasets is the domain gap between synthesized images and real images. In this section, we conduct extensive cross-domain evaluations to verify the domain gap between GazeGene and other real-world datasets. Random translation and color jitter is applied for all datasets. As shown in Tab. 3, our dataset achieves the highest accuracy on ETH-XGaze and MPIIFaceGaze datasets and comparable performances on other datasets. GazeGene outperforms other datasets significantly on the ETH-XGaze dataset, demonstrating the advantage of GazeGene in terms of gaze range and image quality. Overall, results in Tab. 3 proves that our dataset generalizes well to other real-world datasets.

In Fig. 9, we further visualize the distribution of gaze estimation error with respect to image brightness, head pose deviation, gaze yaw and pitch angles on ETH-XGaze. It is clear that the performance of the GazeGene is consistently better than other datasets across different factors. These results demonstrate that GazeGene provides not only high-quality samples and annotations but also sufficient diversity.

In summary, both within-domain and cross-domain evaluations prove that the overall quality and diversity of our dataset are better than most existing datasets, even without the extra 3D annotations.

4.2. New Possibilities enabled by GazeGene

The proposed GazeGene dataset is the first remote gaze dataset that offers precise 3D annotations of eyeball structures, enabling new deep-learning research and methods for various gaze-related tasks. In this section, we use optical and visual axis estimation, 3D eyeball structure estimation and gaze vergence depth estimation as examples to illus-

Table 3. Cross domain evaluation of different datasets on ResNet-18 and ResNet-50.

Train \ Test	ResNet-18						ResNet-50					
	ETH	MPII	G360	ED	EVE	GazeGene	ETH	MPII	G360	ED	EVE	GazeGene
ETH	—	7.2	**19.47**	**7.88**	**9.44**	15.42	—	7.33	**18.72**	**8.9**	**9.78**	16.32
MPIIGaze	37.57	—	27.16	13.57	15.77	33.21	37.19	—	26.45	13.38	14.64	32.75
Gaze360	18.51	7.89	—	9.06	10.26	18.8	16.7	6.94	—	7.86	11.53	17.14
EyeDiap	44.12	16.52	36.26	—	25.07	39.54	43.86	18.52	36.29	—	25.66	40.61
EVE	30.18	10.38	27.15	9.44	—	28.26	30.65	11.59	29.23	9.37	—	28.81
GazeGene	**12.87**	**6.97**	24.2	11.33	10.25	—	**12.37**	**6.86**	26.81	11.62	11.58	—

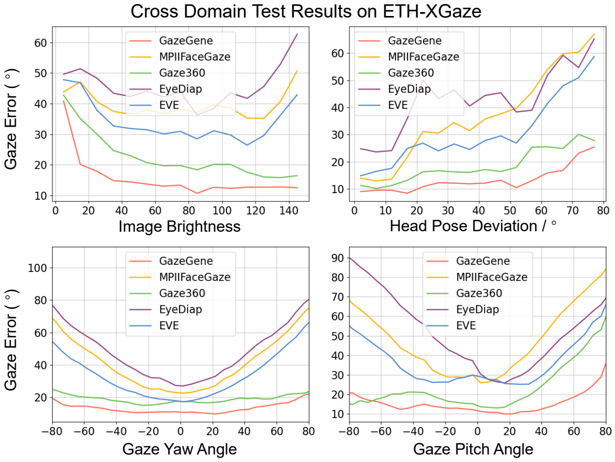

Figure 9. The distribution of gaze estimation errors with respect to image brightness, head pose deviation , gaze yaw and pitch angle. Results are from cross-domain evaluations on the ETH-XGaze dataset.

Table 4. "Visual Axis" and "Optical+Kappa": Angular error of gaze estimation by 1) visual axis regression and 2) applying ground truth kappa to optical axis regression results, respectively. "Optical axis": Angular error of optical axis estimation. Small gray numbers are STDs.

Model	Visual Axis		Optical+Kappa		Optical axis	
	Left	Right	Left	Right	Left	Right
ResNet-18	4.93 .30	4.82 .32	4.89 .20	4.77 .26	4.87 .19	4.75 .25
ResNet-50	4.22 .21	4.14 .16	4.20 .29	4.06 .11	4.19 .29	4.05 .12

trate the applications of GazeGene, offering new insights for deep-learning based methods in gaze-related tasks.

Note that different from Sec. 4.1, experiments in Sec. 4.2 do not implement the normalization technique from [51]. The reason is that the normalization introduces a certain distortion along the z-axis, which could potentially compromise the validity of the conclusion of these 3D experiments.

Table 5. Evaluation of eyeball structure estimation on the GazeGene dataset. The error units for 2D and 3D estimations are pixels and centimeters, respectively.

Method	Eyeball		Pupil		Iris	Optic
	2D	3D	2D	3D	2D	Axis
3DGazeNet	2.29	N/A	N/A	N/A	1.26	17.63
ResNet-18	0.87	0.11	1.39	0.15	1.84	4.98
ResNet-50	1.09	0.13	1.45	0.16	1.73	4.72

4.2.1. Optic, Visual Axes and Kappa

Fundamentally speaking, the eye gaze of human is the visual axis of each eye. However, visual axis is determined by eyeball structures that are not visible in the image, such as fovea. Although some of the structures that determine the optical axis are visible, the kappa angle still various from person to person. Whether deep learning methods can learn patterns from large amount of data to estimate the kappa angle and the visual axis remains an open question.

To explore this question, we train baseline models on GazeGene to estimate optical axis and visual axis respectively. Specifically, the outputs of each networks are two 3D vectors, correspond to optical axis and visual axis of both eyes. L1 loss is used as loss function. The average angular error of the last 5 epochs are shown in Tab. 4. The angular errors of visual axis estimation are about 0.1° higher than the error of optical axis across different eyes and networks. It is reasonable to hypothesis that this small but stable performance gap comes from the unpredictable differences of kappa angle across subjects. To verify it, we derive the visual axis by applying the ground truth kappa angles on the optical axis prediction and result in slightly better accuracy than the visual axis estimation. From above results, we can conclude that **even for data-driven methods, the visual axis is still more challenging to estimate than the optical axis. The approach of estimating the optical axis and calibrating the kappa angle has the potential to achieve higher accuracy.**

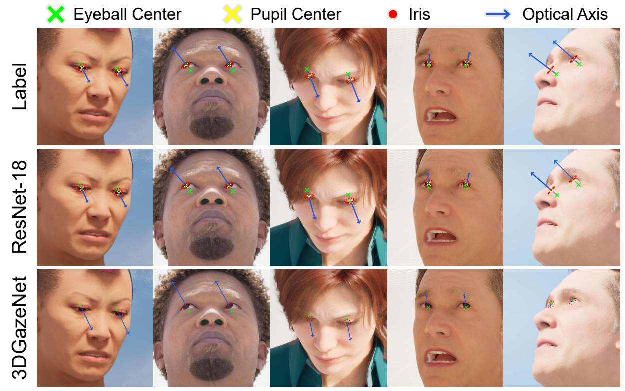

Figure 10. Random examples of the eyeball structure estimation and corresponding gaze estimation results on the GazeGene dataset. Results of 3DGazeNet come from the model released by the original author.

4.2.2. 3D Eyeball Structure Estimation

Eye gaze is closely related to eyeball structures such as eyeball center, pupil and iris. In this section, we investigate whether deep-learning methods can directly estimate the 3D position of these structures and calculate the optical axis.

Specifically, the target of the network is to estimate 3D eyeball center, 3D pupil center and 2D iris contour consists of 10 points. We apply 4 losses: L1 loss for 1) 3D eyeball center coordinates, 2) 3D pupil center coordinates 3) iris contour pixel coordinates, and 4) angular error between the optical axis ground truth and the optical axis calculated by the estimated eyeball center and pupil center. Results in Tab. 5 show that estimating gaze direction by directly reconstructing 3D eyeball structures is feasible. We also test the pretrained 3DGazeNet [41] released by the authors on GazeGene for reference. Visualization in Fig. 10 shows that the main gap between the pretrained 3DGazeNet and the ResNet-18 trained on GazeGene is the accuracy of eyeball center estimation. Since the eyeball center is not visible, 3DGazeNet learns from the pseudo label generated by eyeball mesh alignment [41], which inherently contain some error. Both quantitative and qualitative results prove that **estimating gaze from 3D eyeball structures remotely is feasible, and the precise annotations of GazeGene greatly aid in estimating hidden structures such as eyeball center and gaze direction.**

4.2.3. Gaze Vergence Depth

Gaze vergence depth has been proven to be helpful for interaction in Augmented Reality [44, 47]. While former methods were implemented in near-eye settings, we explore whether deep-learning methods could estimate gaze vergence depth from remote face images using GazeGene.

We select samples with vergence depth within $[0.2m, 1.5m]$ for this task to simulate the classical desktop interaction range. To capture the subtle changes of the eye, we stack two eye patches and input them into a ResNet-50 model to directly estimate vergence depth. L1 is used as loss function. The ResNet-50 model achieves an average error of 27cm. Results in Fig. 11 indicate that **the model can learn a certain level of depth discrimination from the data. However, this discriminative ability diminishes as depth increases.**

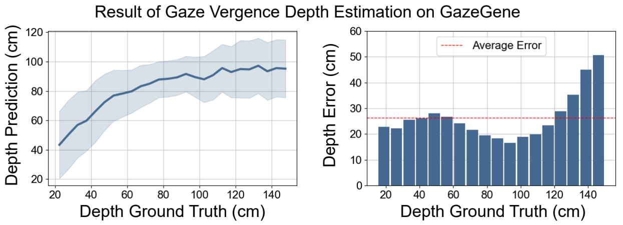

Figure 11. Distribution of gaze vergence depth prediction and error (cm) with respect to vergence depth ground truth on GazeGene.

5. Discussions and Limitations

Applications of the GazeGene: Except for the examples in Sec. 4.2, there are many other gaze-related tasks could also benefit from GazeGene such as eye contact detection [15, 22]. In this task, samples with gaze deviation within 5 degrees are typically considered positive samples [22], requiring high accuracy in gaze annotations. In addition, there is a systematic error in real-world datasets due to the deviation between the camera and the actual eye contacting target. Our dataset offers a new solution to these problems. Moreover, results in Sec. 4.2 did not reach the full potential of GazeGene, since we only use the most basic models and straight-forward designs for verification purpose.

Limitations: Although experiments in Sec. 4.1 already proves the satisfying generalization ability and variety of our dataset, there is still rooms for improvement. Currently, our dataset does not include random backgrounds. Introducing random backgrounds could further increase dataset variety, which we have reserved for future work.

6. Conclusion

We present a new large-scale synthetic gaze estimation dataset GazeGene, offering over 1M samples with wide data distribution, high-resolution, and most importantly, accurate and comprehensive 3D eyeball structure annotations that are not available in existing datasets. In basic gaze estimation experiments, GazeGene demonstrates comparable quality and diversity with existing datasets. Verification experiments in Sec. 4.2 show that GazeGene enables deep-learning applications and research across various gaze-related tasks, providing a valuable resource for future work in appearance-based gaze estimation.

References

[1] Yiwei Bao and Feng Lu. From feature to gaze: A generalizable replacement of linear layer for gaze estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1409–1418, 2024. 3

[2] Yiwei Bao, Yihua Cheng, Yunfei Liu, and Feng Lu. Adaptive feature fusion network for gaze tracking in mobile tablets. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 9936–9943. IEEE, 2021. 3

[3] Yiwei Bao, Yunfei Liu, Haofei Wang, and Feng Lu. Generalizing gaze estimation with rotation consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4207–4216, 2022. 3

[4] Hikmet Basmak, Afsun Sahin, Nilgun Yildirim, Thanos D Papakostas, and A John Kanellopoulos. Measurement of angle kappa with synoptophore and orbscan ii in a normal population, 2007. 4

[5] Olga Vl Bitkina, Jaehyun Park, and Hyun K Kim. The ability of eye-tracking metrics to classify and predict the perceived driving workload. *International Journal of Industrial Ergonomics*, 86:103193, 2021. 1

[6] Alisa Burova, John Mäkelä, Jaakko Hakulinen, Tuuli Keskinen, Hanna Heinonen, Sanni Siltanen, and Markku Turunen. Utilizing vr and gaze tracking to develop ar solutions for industrial maintenance. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2020. 1

[7] Xin Cai, Jiabei Zeng, Shiguang Shan, and Xilin Chen. Source-free adaptive gaze estimation by uncertainty reduction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22035–22045, 2023. 3

[8] Nora Castner, Thomas C Kuebler, Katharina Scheiter, Juliane Richter, Thérése Eder, Fabian Hüttig, Constanze Keutel, and Enkelejda Kasneci. Deep semantic gaze embedding and scanpath comparison for expertise classification during opt viewing. In *ACM Symposium on Eye Tracking Research and Applications*, pages 1–10, 2020. 1

[9] Zhaokang Chen and Bertram E Shi. Appearance-based gaze estimation using dilated-convolutions. In *Asian Conference on Computer Vision*, pages 309–324. Springer, 2018. 2, 6

[10] Yihua Cheng and Feng Lu. Gaze estimation using transformer. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 3341–3347. IEEE, 2022. 3, 6

[11] Yihua Cheng, Shiyao Huang, Fei Wang, Chen Qian, and Feng Lu. A coarse-to-fine adaptive network for appearance-based gaze estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10623–10630, 2020. 3

[12] Yihua Cheng, Xucong Zhang, Feng Lu, and Yoichi Sato. Gaze estimation by exploring two-eye asymmetry. *IEEE Transactions on Image Processing*, 29:5259–5272, 2020. 2

[13] Yihua Cheng, Yiwei Bao, and Feng Lu. Puregaze: Purifying gaze feature for generalizable gaze estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 436–443, 2022. 3

[14] Yihua Cheng, Haofei Wang, Yiwei Bao, and Feng Lu. Appearance-based gaze estimation with deep learning: A review and benchmark. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 2, 6

[15] Eunji Chong, Elysha Clark-Whitney, Audrey Southerland, Elizabeth Stubbs, Chanel Miller, Eliana L Ajodan, Melanie R Silverman, Catherine Lord, Agata Rozga, Rebecca M Jones, et al. Detection of eye contact with deep neural networks is as accurate as human experts. *Nature communications*, 11(1):6386, 2020. 8

[16] Tobias Fischer, Hyung Jin Chang, and Yiannis Demiris. Rt-gene: Real-time eye gaze estimation in natural environments. In *Proceedings of the European conference on computer vision (ECCV)*, pages 334–352, 2018. 3

[17] Kenneth Alberto Funes Mora, Florent Monay, and Jean-Marc Odobez. Eyediap: A database for the development and evaluation of gaze estimation algorithms from rgb and rgb-d cameras. In *Proceedings of the Symposium on Eye Tracking Research and Applications*, pages 255–258, 2014. 2, 3, 4, 6

[18] Elias Daniel Guestrin and Moshe Eizenman. General theory of remote gaze estimation using the pupil center and corneal reflections. *IEEE Transactions on biomedical engineering*, 53(6):1124–1133, 2006. 2, 4

[19] Dan Witzner Hansen and Qiang Ji. In the eye of the beholder: A survey of models for eyes and gaze. *IEEE transactions on pattern analysis and machine intelligence*, 32(3):478–500, 2009. 2

[20] Hassan Hashemi, Mehdi KhabazKhoob, Kamran Yazdani, Shiva Mehravaran, Ebrahim Jafarzadehpur, and Akbar Fotouhi. Distribution of angle kappa measurements with orbscan ii in a population-based survey. *Journal of Refractive Surgery*, 26(12):966–971, 2010. 4

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6

[22] Thorsten Hempel, Magnus Jung, Ahmed A Abdelrahman, and Ayoub Al-Hamadi. Nitec: Versatile hand-annotated eye contact dataset for ego-vision interaction. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4437–4446, 2024. 8

[23] Petr Kellnhofer, Adria Recasens, Simon Stent, Wojciech Matusik, and Antonio Torralba. Gaze360: Physically unconstrained gaze estimation in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6912–6921, 2019. 2, 3, 4, 6

[24] Jess Kerr-Gaffney, Amy Harrison, and Kate Tchanturia. Eye-tracking research in eating disorders: A systematic review. *International Journal of Eating Disorders*, 52(1):3–27, 2019. 1

[25] Andrew J King, Gregory F Cooper, Gilles Clermont, Harry Hochheiser, Milos Hauskrecht, Dean F Sittig, and Shyam Visweswaran. Leveraging eye tracking to prioritize relevant medical record data: comparative machine learning study. *Journal of medical Internet research*, 22(4):e15876, 2020. 1

[26] Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[27] Robert Konrad, Anastasios Angelopoulos, and Gordon Wetzstein. Gaze-contingent ocular parallax rendering for virtual

reality. *ACM Transactions on Graphics (TOG)*, 39(2):1–12, 2020. 1

[28] Kyle Krafka, Aditya Khosla, Petr Kellnhofer, Harini Kannan, Suchendra Bhandarkar, Wojciech Matusik, and Antonio Torralba. Eye tracking for everyone. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2176–2184, 2016. 3

[29] Mikko Kytö, Barrett Ens, Thammathip Piumsomboon, Gun A Lee, and Mark Billinghurst. Pinpointing: Precise head-and eye-based target selection for augmented reality. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2018. 1

[30] Dongze Lian, Lina Hu, Weixin Luo, Yanyu Xu, Lixin Duan, Jingyi Yu, and Shenghua Gao. Multiview multitask gaze estimation with deep convolutional neural networks. *IEEE transactions on neural networks and learning systems*, 30 (10):3010–3023, 2018. 3

[31] Shijun Liang, Haofei Wang, and Feng Lu. Eyeir: Single eye image inverse rendering in the wild. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024. 3

[32] Ziyang Liang, Yiwei Bao, and Feng Lu. De-confounded gaze estimation. In *Computer Vision–ECCV 2024: 18th European Conference*. Springer, 2024. 3

[33] Yunfei Liu, Ruicong Liu, Haofei Wang, and Feng Lu. Generalizing gaze estimation with outlier-guided collaborative adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3835–3844, 2021. 3

[34] José M Morales, Carolina Díaz-Piedra, Héctor Rieiro, Joaquín Roca-González, Samuel Romero, Andrés Catena, Luis J Fuentes, and Leandro L Di Stasi. Monitoring driver fatigue using a single-channel electroencephalographic device: A validation study by gaze-based, driving performance, and subjective data. *Accident Analysis & Prevention*, 109:62–69, 2017. 1

[35] Seonwook Park, Adrian Spurr, and Otmar Hilliges. Deep pictorial gaze estimation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 721–738, 2018. 2

[36] Seonwook Park, Shalini De Mello, Pavlo Molchanov, Umar Iqbal, Otmar Hilliges, and Jan Kautz. Few-shot adaptive gaze estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9368–9377, 2019. 2

[37] Seonwook Park, Emre Aksan, Xucong Zhang, and Otmar Hilliges. Towards end-to-end video-based eye-tracking. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*, pages 747–763. Springer, 2020. 2, 3, 6

[38] Vincent Sitzmann, Ana Serrano, Amy Pavel, Maneesh Agrawala, Diego Gutierrez, Belen Masia, and Gordon Wetzstein. Saliency in vr: How do people explore virtual environments? *IEEE transactions on visualization and computer graphics*, 24(4):1633–1642, 2018. 1

[39] Yusuke Sugano, Yasuyuki Matsushita, and Yoichi Sato. Learning-by-synthesis for appearance-based 3d gaze estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1821–1828, 2014. 3

[40] Kentaro Takemura and Kenta Yamagishi. A hybrid eye-tracking method using a multispectral camera. In *2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 1529–1534. IEEE, 2017. 2

[41] Evangelos Ververas, Polydefkis Gkagkos, Jiankang Deng, Michail Christos Doukas, Jia Guo, and Stefanos Zafeiriou. 3dgazenet: Generalizing 3d gaze estimation with weak-supervision from synthetic views. In *European Conference on Computer Vision*, pages 387–404. Springer, 2025. 3, 8

[42] Yaoming Wang, Yangzhou Jiang, Jin Li, Bingbing Ni, Wenrui Dai, Chenglin Li, Hongkai Xiong, and Teng Li. Contrastive regression for domain adaptation on gaze estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19376–19385, 2022. 3

[43] Zhimin Wang, Huangyue Yu, Haofei Wang, Zongji Wang, and Feng Lu. Comparing single-modal and multimodal interaction in an augmented reality system. In *2020 IEEE International Symposium on Mixed and Augmented Reality Adjunct, ISMAR 2020 Adjunct, Recife, Brazil, November 9-13, 2020*, pages 165–166. IEEE, 2020. 1

[44] Zhimin Wang, Yuxin Zhao, and Feng Lu. Gaze-vergence-controlled see-through vision in augmented reality. *IEEE Transactions on Visualization and Computer Graphics*, 28 (11):3843–3853, 2022. 8

[45] Erroll Wood, Tadas Baltrušaitis, Louis-Philippe Morency, Peter Robinson, and Andreas Bulling. Learning an appearance-based gaze estimator from one million synthesised images. In *Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications*, pages 131–138, 2016. 2, 3

[46] Mingjie Xu, Haofei Wang, and Feng Lu. Learning a generalized gaze estimator from gaze-consistent feature. In *Proceedings of the AAAI conference on artificial intelligence*, pages 3027–3035, 2023. 3

[47] Chenyang Zhang, Tiansu Chen, Eric Shaffer, and Elahe Soltanaghai. Focusflow: 3d gaze-depth interaction in virtual reality leveraging active visual depth manipulation. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–18, 2024. 8

[48] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. Appearance-based gaze estimation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4511–4520, 2015. 2

[49] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. Mpiigaze: Real-world dataset and deep appearance-based gaze estimation. *IEEE transactions on pattern analysis and machine intelligence*, 41(1):162–175, 2017. 2, 3

[50] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. 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*, pages 51–60, 2017. 2, 4, 6

[51] Xucong Zhang, Yusuke Sugano, and Andreas Bulling. Revisiting data normalization for appearance-based gaze estimation. In *Proceedings of the 2018 ACM symposium on eye tracking research & applications*, pages 1–9, 2018. 6, 7

[52] Xucong Zhang, Seonwook Park, Thabo Beeler, Derek Bradley, Siyu Tang, and Otmar Hilliges. Eth-xgaze: A large scale dataset for gaze estimation under extreme head pose and gaze variation. In *European Conference on Computer Vision*, pages 365–381. Springer, 2020. 1, 2, 3, 4, 5