Unsupervised Gaze Representation Learning from Multi-view Face Images

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

Abstract

Annotating gaze is an expensive and time-consuming endeavor, requiring costly eye-trackers or complex geometric calibration procedures. Although some eye-based unsupervised gaze representation learning methods have been proposed, the quality of gaze representation extracted by these methods degrades severely when the head pose is large. In this paper, we present the Multi-View Dual-Encoder (MV-DE), a framework designed to learn gaze representations from unlabeled multi-view face images. Through the proposed Dual-Encoder architecture and the multi-view gaze representation swapping strategy, the MV-DE successfully disentangles gaze from general facial information and derives gaze representations closely tied to the subject's eyeball rotation without gaze label. Experimental results illustrate that the gaze representations learned by the MV-DE can be used in downstream tasks, including gaze estimation and redirection. Gaze estimation results indicates that the proposed MV-DE displays notably higher robustness to uncontrolled head movements when compared to state-of-the-art (SOTA) unsupervised learning methods.

1. Introduction

Vision is one of the most important sense for humans. Human gaze reveals the direction of visual attention, which is an important cue for understanding how humans perceive the surrounding world. Thus, gaze estimation techniques have become an vital tool in numerous applications, such as Virtual Realty and Augmented Reality[2, 30, 34], automotive safety [13, 26, 27] and healthcare [3, 20, 23]. In recent years, appearance-based gaze estimation methods have drawn a lot of attention, since these methods only require simple web cameras, eliminating the need for expensive eye trackers with dedicated devices such as infrared cameras. Among these methods, Convolutional Neural Networks (CNN) based approaches exhibit exceptional performance in unconstrained environments.

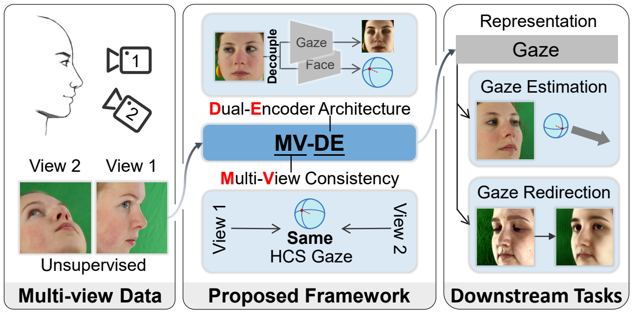


Figure 1. We propose the Multi-View Dual-Encoder, a face-based unsupervised gaze representation learning framework that is robust to large head poses.

CNN-based gaze estimation methods are usually trained in an end-to-end manner with substantial amount of labeled data. The performance of CNN-based gaze estimation methods highly relies on the quantity and diversity of the training dataset. Unfortunately, annotating gaze directions is difficult. Unlike common Computer Vision tasks like Object Detection, gaze cannot be reliably annotated without specialized hardware. Gaze direction is either measured by costly eye trackers under controlled environment [29], or obtained by the line connecting the 3D face center to the 3D position of the gaze target [39]. It takes great effort and time to collect diverse labeled training data, since the participants need to stare at large amount of gaze targets as instructed.

To address the challenge of annotation, a number of unsupervised representation learning approaches have been proposed [4, 14, 19]. These approaches extract common visual representations from the input image without annotation. However, such methods do not perform well in gaze estimation, as gaze estimation differs significantly from common visual tasks. Common visual tasks like object classification and detection require representations of overall appearance of the subject. On the other hand, gaze is an direction vector that manifested as rotations of the subtle eye structure in the image. Efforts have been made to design unsupervised gaze representation learning methods

[*] Corresponding Author. This work was supported by the National Natural Science Foundation of China (NSFC) under Grant 62372019.

[33, 35], which incorporate cropped eye images as input to eliminate undesired visual contents. However, existing unsupervised gaze representation learning methods could only handle samples within limited head pose range, since the eye appearance changes dramatically when the head rotation approaches 90°.

Recent supervised gaze estimation methods usually employ face images as their input [1, 6, 8, 21, 37] to handle samples with large head poses. Unfortunately, the rich information in the face images becomes a double-edged sword in unsupervised settings. These pieces of information become interference due to the absence of gaze label as constraint. For example, head pose and gaze are strongly coupled. They behave exactly the same in common data augmentation methods such as image flipping and rotation, makes it difficult to separate them even for latest contrastive learning approaches. Thus, unsupervised gaze representation learning method with face images is a challenging task that remain to be solved.

A potential solution to tackle this challenge is to harness the extra constraint introduced by multi-view settings. Multi-view gaze estimation is a recent hot research topic. A number of supervised multi-view gaze estimation methods [7, 24] have been proposed. Multi-view image pairs provide a set of images with the same eyeball rotation, *i.e*. gaze direction within the Head Coordinate System (HCS), as they are captured simultaneously. Our aim is to design a face-based multi-view unsupervised gaze representation learning method that works well on samples with large head poses.

In this paper, we present the Multi-View Dual-Encoder (MV-DE), an unsupervised gaze representation learning framework that employs face images as input. The proposed MV-DE framework extracts gaze representation by separating eyeball rotation from general facial information. The MV-DE framework consists of two encoders: the Gaze Encoder and the Face Encoder. First, we train the Face Encoder to extract face representations including head pose and appearance, while excluding gaze information. Then, we freeze the Face Encoder and introduce the gaze feature extracted by the Gaze Encoder to compensate for the missing gaze information. Based on the consistent eyeball rotation across different views within the same frame, we integrate the face representations with gaze representations from different views to reconstruct the original image. At the inference time, the MV-DE extracts gaze representations from single-view face images.

Experimental results illustrate that the MV-DE framework successfully derives gaze representations across a wide range of head poses. The learned gaze representation can be used for varies downstream tasks such as gaze estimation and gaze redirection. The contribution of this paper are summarized in three folds:

- We propose the Multi-View Dual-Encoder , a face-based unsupervised gaze representation learning framework under uncalibrated multi-view settings. The MV-DE is robust to free head movements.
- We uncouple gaze from other interference such as head pose by the proposed Dual-Encoder architecture and the multi-view gaze representation swapping strategy.
- Extensive experiments demonstrate the effectiveness of the MV-DE framework in unconstrained environments. Qualitative analyses prove that the extracted gaze representation is disentangled from head pose and appearance.

2. Related Work

2.1. Supervised Gaze Estimation

Appearance-based gaze estimation approaches aim to estimate gaze from eye or face appearance directly. Early methods estimate gaze directions from eye images [9, 28, 36]. Zhang *et al*. first propose to utilize full face images and outperform eye-image-based methods [37]. Since then, most CNN-based gaze estimation methods employ face images as input. Chen *et al*. propose to utilize dilated convolution for gaze estimation [5]. Cheng *et al*. propose to employ the Transformer architecture for gaze estimation [6]. Some recent methods utilize both face and eye images as input for better estimation performance [1, 8, 17, 22]. Above methods are trained in an end-to-end manner with gaze annotations. A number of gaze estimation datasets are published. These datasets are collected under different scenario with varies devices, including web cameras [11, 37], 360° cameras [21], high-resolution cameras [39], eye trackers [29] and mobile devices [22]. With sufficient labeled data, gaze feature disentanglement is achieved by using GAN [10] and Nerf [32].

2.2. Multi-view Gaze Estimation

Before the development of Deep Learning methods, most conventional model-based methods utilize multiple cameras for gaze estimation [16]. Model-based methods reconstruct 3D eyeball models and obtain the gaze direction based on the optical geometry of human eye structures [15]. Model-based methods achieve remarkable estimation accuracy. However, these methods have high requirements for the shooting angle of cameras. Thus, these methods are usually employed in the Head Mounted Devices, such as Meta Quest Pro and Microsoft Hololens, where the cameras are approximately stationary relative to the eye of users.

Multi-view settings have also drawn a lot of attention in CNN-based gaze estimation methods. A few multi-view gaze estimation datasets have been proposed. Zhang *et al*. propose the ETH-XGaze dataset, which employs 18 high-resolution cameras to capture face images with large head pose range [39]. Park *et al*. propose the EVE dataset [29]. They use 3 web cameras and a industrial camera to capture user face images. Qin *et al*. propose to utilize 3D face alignment approaches to generate multiple rotated images from a given sample [31]. These rotated images can be regarded as samples from different virtual cameras. Based on these datasets, CNN-based multi-view gaze estimation approaches have been proposed. These methods integrate features from different views by concatenating [24] or self attention [7] to improve estimation accuracy

2.3. Unsupervised Gaze Representation Learning

Unsupervised representation learning has always been a hot topic in Computer Vision community. Recently, Contrastive Learning approaches have achieves satisfying performance in common Computer Vision tasks. Contrastive Learning methods generate multiple views of a given sample by different data augmentation methods and constrain the model to extract similar representation from these views [4, 14]. He *et al*. propose the Masked Autoencoders, a self-supervised learner trained by simply reconstructing masked images. However, these methods are designed to extract common visual representations, which do not perform well in the gaze estimation task.

To learn gaze representations in an unsupervised manner, Yu *et al*. propose to utilize the gaze redirection task to extract two-dimensional representations that relate to gaze pitch and yaw angles [35]. Sun *et al*. propose the Cross-Encoder, which learn gaze representations by a latent-code-swapping mechanism on eye-consistent image pairs and gaze-similar pairs [33]. Gideon *et al*. further propose to adapt the Cross-Encoder architecture to the multi-view setting [12]. Above methods all employ eye images as their input, as there are less gaze-irrelevant visual contents. However, appearance of eyes changes dramatically as head pose increases, makes it difficult for these methods to handle samples with large head pose distribution. Our aim is to introduce an unsupervised gaze representation learning approach that employs face images as input, which is robust to unconstrained head movements.

3. Method

We propose the Multi-View Dual-Encoder, a face-based multi-view unsupervised gaze representation learning method. Our aim is to extract gaze representations from the input face images without using the gaze label. The main challenge of this task is to separate gaze from the other facial information, especially head pose. Since gaze and head pose are both physical directions, common data augmentations like rotation and flipping are not able to separate them.

To achieve our goal, the proposed MV-DE framework introduces two encoders: a Face Encoder and a Gaze Encoder. We first train the Face Encoder to extract general facial representation including head pose and appearance, while excluding gaze information. Then, we freeze the Face Encoder and train the Gaze Encoder to extract gaze representation which compensates for the missing gaze information. We constrain the Gaze Encoder to derive similar representation from different views within a given frame since the eyeball rotation is the same. In this way, we isolate gaze representation from other facial features without using the gaze label. In the following sections, we introduce the training pipeline of two encoders in detail.

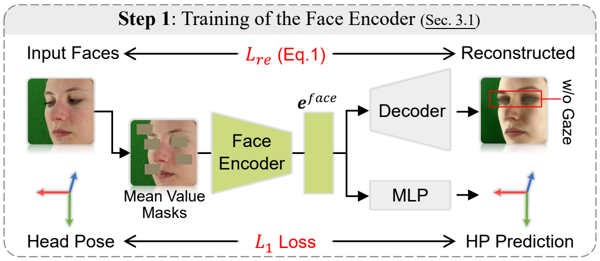


Figure 2. Training process of the Face Encoder. The proposed Face Encoder extracts face representations that contain general facial information while excluding gaze information. It is trained by two tasks: the head pose estimation task and the image reconstruction task.

3.1. Training of the Face Encoder

The target of the Face Encoder is to extract representation of general facial information except gaze. We design two training strategies to achieve this goal: eye masking and multi-task learning including head pose estimation and image reconstruction, as shown in Fig. 2.

Given a training face image $\boldsymbol{x}_{i,j}$ where i is the frame index and j is the camera index, we mask the two eye areas by the average pixel value. We further add three random masks to prevent the Face Encoder from directly estimating head pose based on the position of eye masks. Given that input face images are normalized to 224 × 224 pixels, we set the size of masks to 55 × 33 according to the average size of eyes. Then, we input the masked face images to the Face Encoder for the proposed multi-task learning.

First, the Face Encoder extracts the face representation $\boldsymbol{e}^{f}_{i,j}$ from the masked face image. Then, we add a regression MLP and a decoder for the head pose estimation task and the image reconstruction task, respectively. In the head pose estimation task, the target of the Face Encoder is to estimate the 3D Euler Angles of the subject's head pose. We use $\mathcal{L}_1$ Loss function for the head pose estimation task: $\mathcal{L}_{hp} = \mathcal{L}_1(\hat{\boldsymbol{y}}_{i,j}, \boldsymbol{y}_{i,j})$, where $\hat{\boldsymbol{y}}_{i,j}$ is the estimated head pose and $\boldsymbol{y}_{i,j}$ is the ground truth head pose. The purpose of the head pose estimation task is to ensure that the face representation $\boldsymbol{e}^{f}_{i,j}$ encodes head pose information.

In the image reconstruction task, the Face Encoder and the decoder are trained in an adversarial way. The target

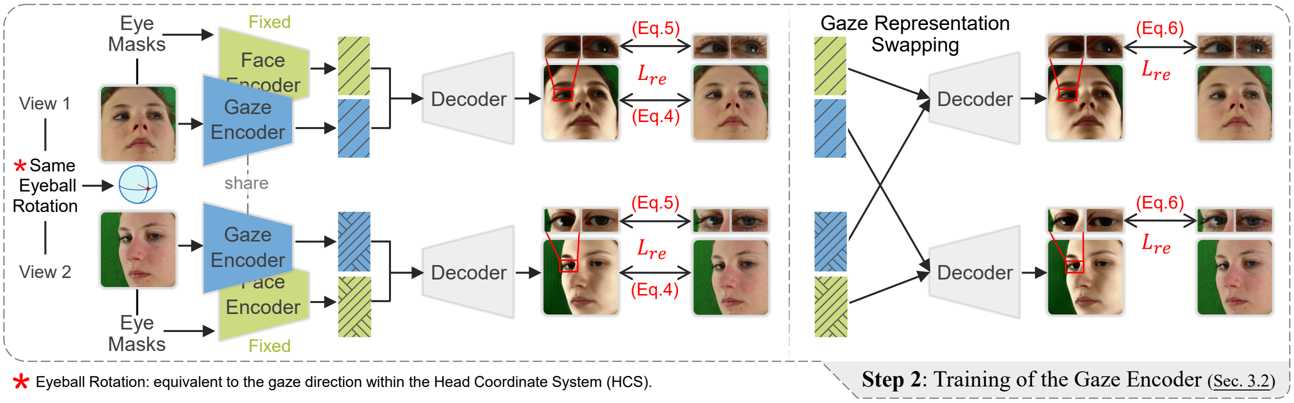


Figure 3. Training process of the Gaze Encoder. The Gaze Encoder extracts gaze representations that represent the eyeball rotation, *i.e.* gaze direction within HCS. The Gaze Encoder is trained by the proposed multi-view gaze representation swapping strategy.

of the Face Encoder is to reconstruct the image with eye-masks. The target of the decoder is to reconstruct the original image from the face representation $\boldsymbol{e}^f_{i,j}$, including the eye area. Both the Face Encoder and the decoder are constrained by the Image Reconstruction Loss $\mathcal{L}_{re}$:

$$\mathcal{L}_{re}(\hat{\boldsymbol{x}}, \boldsymbol{x}) = \alpha_1 \mathcal{L}_2(\hat{\boldsymbol{x}}, \boldsymbol{x}) + \alpha_2 \mathcal{L}_{pcpt}(\hat{\boldsymbol{x}}, \boldsymbol{x}), \tag{1}$$

where $\hat{\boldsymbol{x}}$ is the reconstructed image, $\boldsymbol{x}$ is the target image, $\mathcal{L}_2$ is the Mean Squared Error (MSE) Loss function and $\mathcal{L}_{pcpt}$ is the Perception Loss function. α_1, α_2 are the coefficients of loss functions. The purpose of the image reconstruction tasks is to encode general facial information in the face representation while excluding gaze information. Images constructed from the $\boldsymbol{e}^f_{i,j}$ are basically the original face images without iris and pupil.

Overall, we train the Face Encoder by minimizing $\mathcal{L}_{FE} = \beta_1 \mathcal{L}_{hp} + \mathcal{L}_{re}$, where β_1 is coefficient to balance two tasks. Once the training of the Face Encoder is completed, we freeze the parameters of the Face Encoder and proceed to train the Gaze Encoder via the multi-view constraint.

3.2. Training of the Gaze Encoder

The target of the Gaze Encoder is to extract gaze representation $\boldsymbol{e}^g_{i,j}$ that represents the direction of gaze in the HCS without any gaze label. Since the face representation $\boldsymbol{e}^f_{i,j}$ encodes general face information except gaze, we combine the gaze representation $\boldsymbol{e}^g_{i,j}$ with $\boldsymbol{e}^f_{i,j}$ to fully reconstruct the original image, so that $\boldsymbol{e}^g_{i,j}$ compensates for the missing gaze information. The training strategy of the Gaze Encoder is shown in Fig. 3.

Given a training sample $\boldsymbol{x}_{i,j}$, we extract the gaze representation from the original image and the face representation from the eye-masked image. Then, we input both representations to the decoder to reconstruct the original image. To reconstruct the original image, the Gaze Encoder has to capture gaze information which is missing in the face representation. Considering that the region occupied by two eyes represents only a small portion of the face image, we additionally compute the Image Reconstruction Loss for both eye regions:

$$\begin{aligned} \hat{\boldsymbol{x}}_{i,j} &= G(\boldsymbol{e}^f_{i,j}, \boldsymbol{e}^g_{i,j}), \\ \mathcal{L}_{GE} &= \mathcal{L}_{re}(\hat{\boldsymbol{x}}_{i,j}, \boldsymbol{x}_{i,j}) + \mathcal{L}_{re}(\hat{\boldsymbol{x}}^e_{i,j}, \boldsymbol{x}^e_{i,j}), \end{aligned} \tag{2}$$

where $G(\cdot)$ is the decoder, $\hat{\boldsymbol{x}}^e_{i,j}$ and $\boldsymbol{x}^e_{i,j}$ are the two eye regions of the reconstructed image and the original image, respectively.

However, the gaze representation can also encode other facial information such as head pose during the training process. In an extreme case, the decoder might theoretically reconstruct the entire image solely based on the gaze representation, since the Gaze Encoder utilizes the full face image as input. We utilize the consistency of eyeball rotation in the multi-view settings to exclude gaze-irrelevant information. First, we randomly sample an image $\boldsymbol{x}_{i,j'}$ from another view within the same frame where $j' \neq j$. Since the eyeball rotations from both views are consistent, we swap the gaze representations of two views and reconstruct the eye regions of the original views. For a training image pair $\{\boldsymbol{x}_{i,j}, \boldsymbol{x}_{i,j'}\}$, the final loss function of the Gaze Encoder $\mathcal{L}_{GE}$ is:

$$\mathcal{L}_{GE} = \beta_2 \mathcal{L}^{face}_{re} + \beta_3 \mathcal{L}^{eyes}_{re} + \beta_4 \mathcal{L}^{swap}_{re}, \tag{3}$$

where $\beta_2, \beta_3, \beta_4$ are the coefficients, $\mathcal{L}^{face}_{re}$ is the Image Reconstruction Loss of full face images from both view:

$$\mathcal{L}^{face}_{re} = \mathcal{L}_{re}(\hat{\boldsymbol{x}}_{i,j}, \boldsymbol{x}_{i,j}) + \mathcal{L}_{re}(\hat{\boldsymbol{x}}_{i,j'}, \boldsymbol{x}_{i,j'}). \tag{4}$$

$\mathcal{L}^{eyes}_{re}$ is the Image Reconstruction Loss of eye regions from both view:

$$\mathcal{L}^{eys}_{re} = \mathcal{L}_{re}(\hat{\boldsymbol{x}}^e_{i,j}, \boldsymbol{x}^e_{i,j}) + \mathcal{L}_{re}(\hat{\boldsymbol{x}}^e_{i,j'}, \boldsymbol{x}^e_{i,j'}). \tag{5}$$

$\mathcal{L}_{re}^{swap}$ is the Image Reconstruction Loss of the generated eye regions after gaze representation swapping:

$$\begin{aligned}\mathcal{L}_{re}^{swap} =& \mathcal{L}_{re}(G(\boldsymbol{e}^f_{i,j}, \boldsymbol{e}^g_{i,j'})^e, \boldsymbol{x}^e_{i,j}) + \\ & \mathcal{L}_{re}(G(\boldsymbol{e}^f_{i,j'}, \boldsymbol{e}^g_{i,j})^e, \boldsymbol{x}^e_{i,j'}).\end{aligned} \tag{6}$$

At test time, the Gaze Encoder extracts gaze representations from the given single-view face images, similar to common supervised single-view gaze estimation methods. We rotate the estimated HCS gaze to CCS by the head pose label for evaluation.

3.3. Implementation Details

Training Details: The proposed method is implemented by PyTorch using two RTX 3090 GPU. We employ the Adam optimizer with a leaning rate of 10^{-3} for all the encoders and decoders. The Face Encoder is trained for 13 epochs with a batch size of 100. Learning rate is decayed by 0.2 every 4 epochs. The Gaze Encoder is trained for 5 epochs with a batch size of 50. Learning rate is decayed by 0.2 every 2 epochs. We use the data rectification method from [37] and histogram equalization to normalize the input face images. (α_1, α_2) are set to $(10, 1)$ and $(10, 0.5)$ during the training of the Face Encoder and Gaze Encoder, respectively. β_1 is set to 0.1 and $(\beta_2, \beta_3, \beta_4)$ is set to $(0.3, 0.35, 0.35)$. We use the first 6 convolutional layers of a ImageNet pretrained VGG-16 to calculate the Perceptual Loss.

Network Architecture: Both the Face Encoder and the Gaze Encoder use the ResNet-18 [18] as backbone. For the Face Encoder, we take the $(512 * 7 * 7)$ feature map after the last residual block as the Face Representation. The output channels of the last Linear Layer is set to 3 for head pose estimation. For the Gaze Encoder, the Linear layers is replaced by a 1×1 Convolutionallayer to compress the $(512 * 7 * 7)$ feature map to $(1 * 7 * 7)$ gaze representation. In the decoder, the channel of the gaze representation is expanded to 10 by a 1×1 Convolutionallayer and then concatenated with the face representation. The decoder also has four residual blocks with target channel sizes of $[256, 128, 63, 32]$. The feature maps are up-scaled to twice their size before each residual block. We upscale the feature map again and use two Convolutional layers to generate the final $(3 \times 224 \times 224)$ image. During the training of the Face Encoder, we set the gaze representation to zeros. Note that we train the decoder from scratch during the training of the Gaze Encoder, as the decoder has learned to ignore the 0 replaced gaze representation during the training of the Face Encoder.

4. Experiments

4.1. Data Preparation

We conduct experiments on four different gaze estimation datasets: ETH-XGaze [39], MPII-NV [31], EVE [29] and EyeDiap [11]. We visualize the head pose and gaze distribution of each dataset in Fig. 4. The distribution of these datasets varies significantly. below we introduce each dataset in detail.

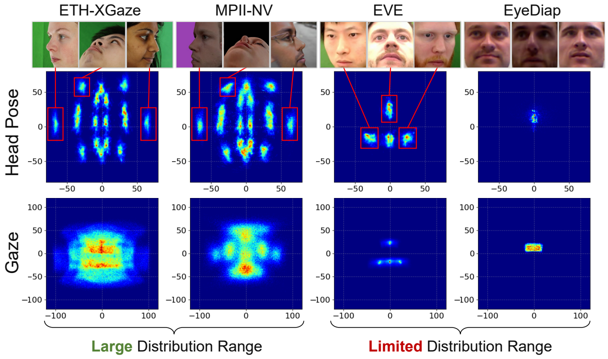


Figure 4. Head pose (top row) and gaze direction distributions (bottom row) of 4 different datasets. The EVE and the EyeDiap datasets only provide very limited head pose and gaze range with frontal faces.

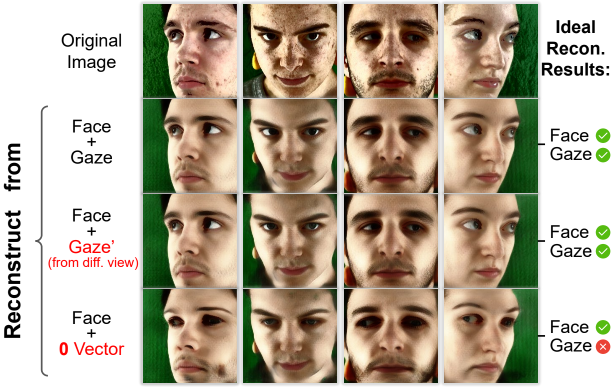


Figure 5. Image reconstruction results of combining face representations with different gaze representations in the ETH-XGaze dataset.

ETH-XGaze ($\mathcal{D}_E$): A multi-view dataset collected by 18 high-resolution cameras. It provides very large head pose and gaze distributions. ETH-XGaze dataset contains 80 subjects, results in over 750,000 images in total. We divide the last 5 subjects as the labled test set and the first 75 subjects as the multi-view unsupervised training set. We follow [25] to optimize the gaze and head pose label to ensure multi-view consistency.

MPII-NV ($\mathcal{D}_M$): A synthesized multi-view dataset. We follow [31] to reconstruct 3D faces from MPIIFaceGaze [37] dataset and rotate every 3D face to generate 18 different views, referring to the setting of ETH-XGaze. MPII-NV dataset contains 15 subjects and 359,984 images in total.

Table 1. Gaze estimation error of few-shot experiments in degrees. **Bold** numbers are the best results and underline numbers are the second best results.

Dataset	ETH-XGaze				MPII-NV				EVE				Average
Calibration Num	50	100	200	Avg.	50	100	200	Avg.	50	100	200	Avg.	
BYOL[14]	25.67	26.04	25.29	25.67	10.19	12.65	11.26	11.37	10.36	9.74	9.74	9.95	15.66
SimCLR[4]	26.81	24.98	23.86	25.22	12.55	12.13	12.14	12.27	12.53	11.36	11.48	11.79	16.43
Cross-Encoder[33]	16.23	14.72	14.54	15.16	11.47	11.36	11.75	11.53	**8.07**	**7.81**	**7.48**	**7.79**	11.49
DE (ours w/o MV)	23.32	19.93	17.76	20.34	11.64	10.73	10.74	11.04	10.79	9.73	9.44	9.99	13.79
MV-DE (ours)	**8.66**	**8.08**	**7.77**	**8.17**	**7.28**	**6.52**	**6.39**	**6.73**	9.18	8.6	8.68	8.82	**7.91**
Gain	▼ 7.57	▼ 6.64	▼ 6.77	▼ 6.99	▼ 2.91	▼ 4.21	▼ 4.35	▼ 4.31	▲ 1.11	▲ 0.79	▲ 1.20	▲ 1.03	▼ 3.58

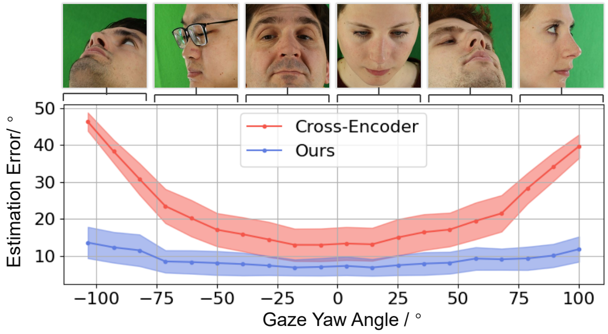


Figure 6. Gaze estimation error relative to the gaze yaw angle in the ETH-XGaze dataset. The proposed MV-DE performs significantly better when the gaze yaw angle is large.

We use the last subject as the labeled test set.

EVE ($\mathcal{D}_V$)**:** A multi-view dataset with 4 camera views. As shown in the Fig. 4, the EVE dataset only provides frontal faces with limited head pose and gaze distribution. We use the training and testing set split as defined in the original dataset and sample 3 images per second from the original videos, result in 527,896 images for unsupervised training and 64,464 images for testing.

EyeDiap ($\mathcal{D}_D$)**:** A single-view dataset with limited head pose and gaze distribution with 16,674 images in total. We only use the EyeDiap in the cross-domain experiments. Thus, we use the whole dataset for training or testing.

We follow [38] to normalize the face images. We employ Histogram Equalization and normalize the pixel values to $[-1, 1]$.

4.2. Evaluation of the Learned Representations

In this section, we validate the effectiveness of the learned face and gaze representations through image reconstruction. The propose MV-DE is supposed to decouple representations of the eyeball rotation from general facial information. In Fig. 5, we combine the face representation with different gaze representations and reconstruct the face image for validation. As shown in row 2, the decoder effectively reconstructs the input face image from the original face and gaze representation. When replacing the original gaze representation with gaze representation from another view, the reconstruction results are similar with the original one. It proves that the MV-DE successfully learned the multi-view consistency of the eyeball rotation. In the last row, the gaze representation is replaced by zero vectors. The head pose and appearance of the reconstructed images are almost identical with the original image, while the eyeball rotations are rather random. These results confirm that the MV-DE successfully separate eyeball rotation from general facial information.

4.3. Application of the MV-DE: Gaze Estimation

The learned gaze representation can be used for gaze estimation by adding a MLP head calibrated under the few-shot learning setting. The estimation error is also an evaluation of the learned gaze representation. In the MV-DE , the MLP takes the gaze representation and the head pose estimation as input. The estimated head pose is first encoded to a 30 dimensional embedding and then concatenated with the gaze representation through 3 linear layers with target dimensions of $(64, 64, 2)$. The MLP head is trained to estimate the HCS gaze. Then, we rotate the HCS gaze by head rotation to obtain CCS gaze. Sigmoid function is employed in the activation layer.

In Tab. 1, we compare the MV-DE with 3 SOTA unsupervised learning methods. BYOL [14] and SimCLR [4] are two SOTA contrastive learning methods. We modify them to adapt the multi-view gaze representation learning task. We employ face images from different camera views in the same frame instead of data augmentation methods to generate the positive pairs. Cross-Encoder [33] is a SOTA **eye-based** unsupervised gaze representation learning method. Gideon *et al.* also propose an eye-based method [12], but their method requires sample pairs with different eyeball rotation while keeping the head stable within a short video clip. Since such sample pairs are not available in $\mathcal{D}_E$ and $\mathcal{D}_M$, we exclude their moethod in comparison. We also remove the multi-view constraint in the MV-DE as an ablation study (named DE). Without the multi-view constraint, the Gaze Encoder is trained by the image reconstruc-

Table 2. Estimation error of different designs of the MLP head with 200 calibration samples.

Dimension	MLP inputs & outputs	$\mathcal{D}_E$	$\mathcal{D}_M$	$\mathcal{D}_V$
$d = 32$	$\boldsymbol{e}_g \rightarrow$ CCS Gaze	24.34	42.82	9.83
	$\boldsymbol{e}_g + HP \rightarrow$ CCS Gaze	20.59	42.29	9.57
	$\boldsymbol{e}_g \rightarrow$ HCS Gaze	8.89	10.49	8.66
	$\boldsymbol{e}_g + HP \rightarrow$ HCS Gaze	**8.63**	**7.26**	**8.60**
$d = 64$	$\boldsymbol{e}_g \rightarrow$ CCS Gaze	22.27	41.70	9.98
	$\boldsymbol{e}_g + HP \rightarrow$ CCS Gaze	18.11	34.56	9.55
	$\boldsymbol{e}_g \rightarrow$ HCS Gaze	8.43	7.44	8.68
	$\boldsymbol{e}_g + HP \rightarrow$ HCS Gaze	**7.84**	**6.39**	**8.67**
$d = 256$	$\boldsymbol{e}_g \rightarrow$ CCS Gaze	20.97	40.99	10.32
	$\boldsymbol{e}_g + HP \rightarrow$ CCS Gaze	18.88	34.05	10.39
	$\boldsymbol{e}_g \rightarrow$ HCS Gaze	8.07	6.50	**8.72**
	$\boldsymbol{e}_g + HP \rightarrow$ HCS Gaze	**7.59**	**5.71**	8.87

tion task described in Eq. (2).

Overall, the average accuracy of the MV-DE outperforms other SOTA methods significantly. BYOL and SimCLR perform worse than the gaze-specialized Cross-Encoder and the MV-DE, proves that common visual representations are not suitable for the gaze estimation task.

Most importantly, the MV-DE outperforms the Cross-Encoder significantly in the ETH-XGaze and MPII-NV datasets, where the head pose distributes across a wide range. On the other hand, the MV-DE achieves comparable performance with the Cross-Encoder in the EVE dataset, because the EVE dataset only contains frontal face images with limited head pose range.

To further investigate the performance gap between the Cross-Encoder and the MV-DE, we visualize the gaze estimation error relative to the gaze yaw angle within the ETH-XGaze dataset in Fig. 6. Obviously, the estimation error of the Cross-Encoder increases severely when the gaze yaw angle approach $100°$. Samples with such extreme gaze yaw angle are generally associated with large head poses. Above results proves the point that the MV-DE are more robust to large head poses.

Results from the last 2 rows demonstrate the importance of the proposed multi-view gaze representation swapping strategy. The multi-view constraint brought a improvement as large as $5.88°$. But the Dual-Encoder (DE) still achieves better overall performance than SOTA constrastive learning methods, primary attributes to the proposed dual-encoder architecture.

4.3.1 Ablation Study: Design of the MLP Head

The MLP head is responsible for estimating gaze from the learned gaze representations under the few-shot learning setting. In Tab. 2, we verify the performances of different MLP designs with 200 calibration samples. For example, $[d = 64, \boldsymbol{e}_g + HP \rightarrow HCS\ Gaze]$ denotes that the MLP includes 64-dimensional hidden layers, takes gaze representations and head poses as the input vector, and predicts gaze directions within Head Coordinate System (HCS). Note that in the MV-DE, we rotate the estimated HCS gaze with head pose label physically.

Table 3. Cross-dataset gaze estimation errors of using the MV-DE as an unsupervised pretrain method. We first pretrain the model on the ETH-XGaze or the MPII-NV dataset unsupervisely. Then, we use the labeled training dataset to calibrate the MLP head.

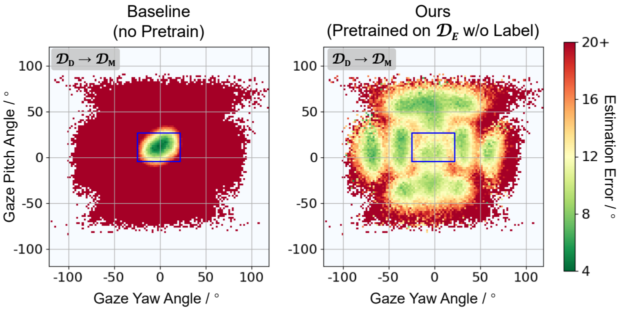
Method	Unsupervised	Train	Test Dataset		
	Pretrain	Dataset	ETH-Xgaze	MPII-NV	EyeDiap
Baseline	-	EVE	31.27	32.52	12.55
MV-DE	ETH-XGaze	EVE	19.67 ▼ 37.10%	10.45▼ 67.87%	11.36 ▼ 9.47%
MV-DE	MPII-NV	EVE	25.42 ▼ 18.71%	10.33 ▼ 68.23%	11.58 ▼ 7.73%
			ETH-Xgaze	MPII-NV	EVE
Baseline	-	EyeDiap	44.06	39.46	23.58
MV-DE	ETH-XGaze	EyeDiap	25.15 ▼ 42.92%	12.11 ▼ 69.31%	10.4 ▼ 55.89%
MV-DE	MPII-NV	EyeDiap	25.58 ▼ 41.94%	11.51 ▼ 70.83%	10.58 ▼ 55.13%

Results of Tab. 2 lead to 3 conclusions: (1) The MLP performs signicantly better on predicting HCS gaze, since the learned gaze representations represent eyeball rotation instead of the CCS Gaze. Theoretically, combination of $\boldsymbol{e}_g + HP$ should be enough to predict the CCS gaze, but the estimation error is still huge, probably due to the limited number of calibration samples and MLP parameters. (2) Head pose seems helpful for predicting the HCS gaze. This observation is consistent with previous research [37] that the additional information of the full face images helps in the gaze estimation task. (3) 256-dimensional hidden layers achieve the best performance in the 200-shot settings. But larger number of parameters require more calibration samples, the performance of 256-dimensional hidden layers degrades in 50 and 100 shot calibration settings. Considering the balance between the number of parameters and required calibration samples, we choose $d = 64$ for the MV-DE.

4.3.2 Additional Experiments: MV-DE as an Unsupervised Pretrain Method

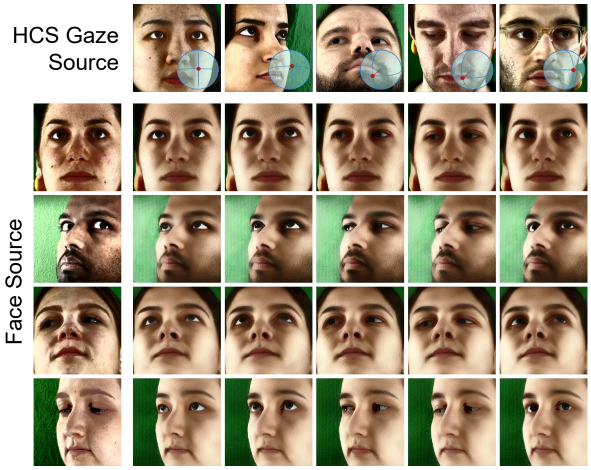
Another potential application of the MV-DE is to used as an unsupervised pretrain method, since it is more cost-effective to collect large number of unlabeled face images than labeled ones. In Tab. 3, we employ the MV-DE to pretrain the Gaze Encoder without gaze label in the ETH-XGaze or the MPII-NV datasets. Then, we freeze the Gaze Encoder and only train the MLP head in the EVE or the EyeDiap dataset with gaze label.

To establish a baseline, we directly train the Gaze Encoder and the regression MLP in an end-to-end manner with gaze label in the labeled training dataset, just like common supervised gaze estimation methods. The cross-dataset errors of the baseline method are extremely large on the ETH-



.

Figure 7. Dsitributions of cross-domain gaze estimation error with (right) and without (left) the proposed unsupervised pretrain on the ETH-XGaze. Blue rectangle indicates the gaze distribution range of the labeled training dataset (EyeDiap).



.

Figure 8. Results of using the MV-DE for gaze redirection. Spheres on the bottom-right corner indicates the gaze direction ground truth within the HCS, *i.e.* the eyeball rotation.

XGaze and MPII-NV dataset, since the gaze distributions of the training sets are significantly smaller than the test sets. When used as a unsupervised pretrain method, the MV-DE effectively improves the cross-dataset performance of the baseline model. The distributions of estimation errors in Fig. 7 demonstrate that with the pretrain of the MV-DE, the model generates reasonable estimations in a significantly larger range. Above results indicate that the MV-DE can be used as a pretrain method to improve cross-dataset performances when large amount of unlabeled face images are available.

4.4. Application of the MV-DE: Gaze Redirection

The learned representations of the MV-DE can also be used for the gaze redirection task. Results in Fig. 8 show that the MV-DE effectively redirects the gaze direction according to the HCS gaze of the reference images, while keeping the head pose and identity unchanged. Fig. 8 demonstrate that the learned gaze representation is subject-independent. The head pose and identity of the reconstructed images are controlled by the face representations, where the eyeabll rotation is controlled by the gaze representations as expected.

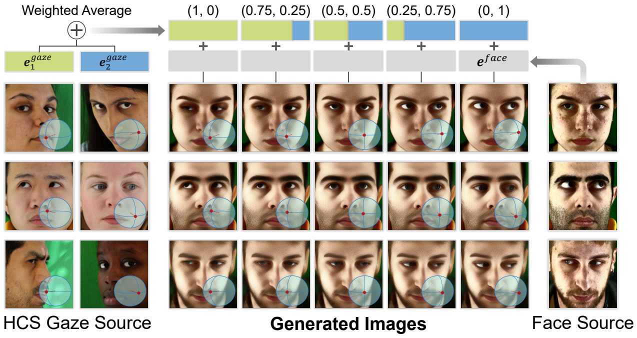


Figure 9. Gaze redirection results of combining different gaze representations linearly. Results demonstrate that the gaze representations exhibit good linearity invariance. Spheres on the bottom-right corner indicate the ground truth or the target eyeball rotation.

4.4.1 Linearity Invariance of the Gaze Representation

Gaze, as a physical direction vector, conforms to the principles of additivity. Ideally, well-learned gaze representations should possess similar characteristics. In Fig. 9, we combine gaze representations from different samples linearly with different ratios to reconstruct face images. The sphere in the bottom-left corner of reconstructed the images indicates the target eyeball rotation angle, which corresponds to the linear combination of eyeball rotations from the HCS gaze source images. Results demonstrate that the learned gaze representation satisfies Linearity Invariance, *i.e.* the eyeball rotation of reconstructed images undergo the same linear combination as the input gaze representations. Experiments in Fig. 9 prove that the MV-DE successfully captures and retains part of the physical properties of gaze during the unsupervised learning process.

5. Conclusion

In this paper, we present the Multi-View Dual-Encoder (MV-DE), an unsupervised gaze representation learning framework based on multi-view face images. We propose the Dual-Encoder architecture and the multi-view gaze representation swapping strategy to learn gaze representations that are separated from general facial information without gaze label. Experiments show that the learned gaze representations can be used for downstream tasks like gaze estimation and gaze redirection. Gaze estimation results show that the quality of learned gaze representations from the MV-DE are significantly better than other SOTA methods with unconstrained head movements.

References

[1] 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. 2

[2] Yiwei Bao, Jiaxi Wang, Zhimin Wang, and Feng Lu. Exploring 3d interaction with gaze guidance in augmented reality. In *2023 IEEE Conference Virtual Reality and 3D User Interfaces (VR)*, pages 22–32. IEEE, 2023. 1

[3] Ariella Fornachari Ribeiro Belan, Marcos Vasconcelos Pais, Marina von Zuben de Arruda Camargo, Livea Carla Fidalgo Garcêz Sant'Ana, Marcia Radanovic, and Orestes Vicente Forlenza. Diagnostic performance of an eye-tracking assisted visual inference language test in the assessment of cognitive decline due to alzheimer's disease. *Journal of Alzheimer's Disease*, (Preprint):1–15, 2023. 1

[4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 1, 3, 6

[5] 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] Yihua Cheng and Feng Lu. Gaze estimation using transformer. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 3341–3347. IEEE, 2022. 2

[7] Yihua Cheng and Feng Lu. Dvgaze: Dual-view gaze estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20632–20641, 2023. 2, 3

[8] 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. 2

[9] 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

[10] 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. 2

[11] Kenneth Alberto Funes Mora, Florent Monay, and Jean-Marc Odobez. Eyediap: A database for the development and evaluation of gaze estimation algorithms from rgb and rgb-d cameras. In *Proceedings of the Symposium on Eye Tracking Research and Applications*, pages 255–258, 2014. 2, 5

[12] John Gideon, Shan Su, and Simon Stent. Unsupervised multi-view gaze representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5001–5009, 2022. 3, 6

[13] Mohamed Waheed Gomaa, Rasha O Mahmoud, and Amany M Sarhan. A cnn-lstm-based deep learning approach for driver drowsiness prediction. *Journal of Engineering Research*, 6(3):59–70, 2022. 1

[14] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020. 1, 3, 6

[15] 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

[16] 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

[17] Junfeng He, Khoi Pham, Nachiappan Valliappan, Pingmei Xu, Chase Roberts, Dmitry Lagun, and Vidhya Navalpakkam. On-device few-shot personalization for real-time gaze estimation. In *Proceedings of the IEEE/CVF international conference on computer vision workshops*, pages 0–0, 2019. 2

[18] 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. 5

[19] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 1

[20] Daniel A Hofmaenner and Philipp K Buehler. The impact of eye-tracking on patient safety in critical care. *Journal of Clinical Monitoring and Computing*, 36(6):1577–1579, 2022. 1

[21] 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

[22] 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. 2

[23] Astar Lev, Yoram Braw, Tomer Elbaum, Michael Wagner, and Yuri Rassovsky. Eye tracking during a continuous performance test: Utility for assessing adhd patients. *Journal of Attention Disorders*, 26(2):245–255, 2022. 1

[24] 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. 2, 3

[25] Ruicong Liu and Feng Lu. Uvagaze: Unsupervised 1-to-2 views adaptation for gaze estimation. Unpublished manuscript. 5

[26] SI Lyapunov, II Shoshina, and IS Lyapunov. Tremor eye movements as an objective marker of driver's fatigue. *Human Physiology*, 48(1):71–77, 2022. 1

[27] Nada Mohammed Murad, Lilia Rejeb, and Lamjed Ben Said. Computing driver tiredness and fatigue in automobile via eye

tracking and body movements. *Periodicals of Engineering and Natural Sciences*, 10(1):573–586, 2022. 1

[28] 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

[29] Seonwook Park, Emre Aksan, Xucong Zhang, and Otmar Hilliges. Towards end-to-end video-based eye-tracking. In *European Conference on Computer Vision (ECCV)*, 2020. 1, 2, 5

[30] Thammathip Piumsomboon, Gun Lee, Robert W Lindeman, and Mark Billinghurst. Exploring natural eye-gaze-based interaction for immersive virtual reality. In *2017 IEEE symposium on 3D user interfaces (3DUI)*, pages 36–39. IEEE, 2017. 1

[31] Jiawei Qin, Takuru Shimoyama, and Yusuke Sugano. Learning-by-novel-view-synthesis for full-face appearance-based 3d gaze estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4981–4991, 2022. 3, 5

[32] Alessandro Ruzzi, Xiangwei Shi, Xi Wang, Gengyan Li, Shalini De Mello, Hyung Jin Chang, Xucong Zhang, and Otmar Hilliges. Gazenerf: 3d-aware gaze redirection with neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9676–9685, 2023. 2

[33] Yunjia Sun, Jiabei Zeng, Shiguang Shan, and Xilin Chen. Cross-encoder for unsupervised gaze representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3702–3711, 2021. 2, 3, 6

[34] Xin Yi, Yiqin Lu, Ziyin Cai, Zihan Wu, Yuntao Wang, and Yuanchun Shi. Gazedock: Gaze-only menu selection in virtual reality using auto-triggering peripheral menu. In *2022 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pages 832–842. IEEE, 2022. 1

[35] Yu Yu and Jean-Marc Odobez. Unsupervised representation learning for gaze estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7314–7324, 2020. 2, 3

[36] 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

[37] 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, 5, 7

[38] 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

[39] 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, 5