Cross-Encoder for Unsupervised Gaze Representation Learning

Yunjia Sun[1,2], Jiabei Zeng[1], Shiguang Shan[1,2] , Xilin Chen[1,2]
[1]Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences (CAS),
Institute of Computing Technology, CAS, Beijing 100190, China
[2]University of Chinese Academy of Sciences, Beijing 100049, China
{sunyunjia18z, jiabei.zeng, sgshan, xlchen}@ict.ac.cn

Abstract

In order to train 3D gaze estimators without too many annotations, we propose an unsupervised learning framework, Cross-Encoder, to leverage the unlabeled data to learn suitable representation for gaze estimation. To address the issue that the feature of gaze is always intertwined with the appearance of the eye, Cross-Encoder disentangles the features using a latent-code-swapping mechanism on eye-consistent image pairs and gaze-similar ones. Specifically, each image is encoded as a gaze feature and an eye feature. Cross-Encoder is trained to reconstruct each image in the eye-consistent pair according to its gaze feature and the other's eye feature, but to reconstruct each image in the gaze-similar pair according to its eye feature and the other's gaze feature. Experimental results show the validity of our work. First, using the Cross-Encoder-learned gaze representation, the gaze estimator trained with very few samples outperforms the ones using other unsupervised learning methods, under both within-dataset and cross-dataset protocol. Second, ResNet18 pretrained by Cross-Encoder is competitive with state-of-the-art gaze estimation methods. Third, ablation study shows that Cross-Encoder disentangles the gaze feature and eye feature.

1. Introduction

Gaze indicates where someone is looking toward. It serves as one of the cues in understanding human desires, intents and states of mind. 3D gaze estimation retrieves the direction of the line from an observer's eye to a sight. Automatically estimating gaze direction show potential applications in psychological research [22], human-computer interaction [26], driver distraction detection [1], and other areas. Recently, significant efforts have been devoted to developing non-intrusive gaze direction estimator based on facial or eyes images. Specially, the growing strength of Convolutional Neural Network (CNN) [11] makes it relatively easy to deal with some practical problems in gaze estimation like head pose variations, eye occlusions, and variable eye shapes [7, 8, 31, 38, 40].

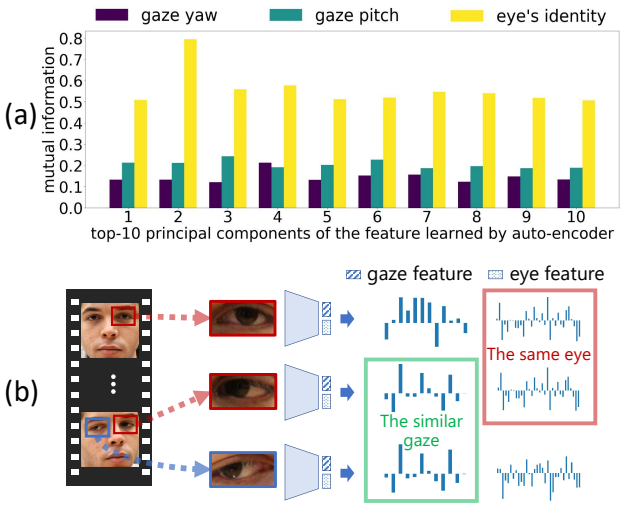


Figure 1. (a) The mutual information between the auto-encoder learned representation and the eye's identity or the gaze. (b) Disentangled gaze features and eye features learned by Cross-Encoder. The images of the same eye from different frames have consistent eye features. The images of the two eyes from one frame have similar gaze features.

Despite its representational power, a well-performed CNN-based method is usually trained on sufficiently large and diverse labeled data. However, acquiring precise gaze labels is difficult. The gaze direction could not be measured directly, but be measured by complicated setups and computations according to the geometry [20, 41]. The limited access to labeled data hinders the development of gaze estimation methods. When trained on a small amount of monotonous annotated samples, supervised learning methods are easy to overfit the training data and preserve features that do not represent the gaze. The redundant and unrelated features lead methods to perform poorly on data beyond the training ones.

Acknowledgement: This work is partially supported by National Key R&D Program of China (No. 2017YFA0700800) and National Natural Science Foundation of China (No. 61976203, 61702481).

Various unsupervised or self-supervised learning strategies are proposed and show potentials in addressing the issue of scarce annotations [4, 6, 29]. Most of them focus on learning a representation for relatively general purposes, *e.g.,* image classification [4], object detection [9], segmentation [28]. However, these representations and methods are not the best for gaze estimation. It is also controversial that a universal representation exists for all the tasks.

Learning a good representation for gaze is non-trivial because the features of gaze direction are always intertwined with those of what the eye looks like. Fig. 1(a) illustrates the mutual information between the top 10 principal components of the unsupervised features learned by auto-encoder and the eye-identity or the gaze. The eye-identity is defined as the distinguishing shape and appearance of an eye. The left and the right eye of a same person have *different* eye-identities. As can be seen, if we learn representations from eye images in an unsupervised way without any priors, the learned features are more related to the eye-identity than to the gaze. To our knowledge, Yu and Odobez [37] are the first to learn gaze-specific representation without annotations. Unlike their work, which leverages the gaze redirection task and ignores the intertwined factors, we explicitly disentangle the features of gaze and those of what makes the eye look like the eye, in an unsupervised manner.

To this end, we propose an unsupervised learning framework to learn disentangled gaze feature and eye feature from eye images. The key component is an auto-encoder-like architecture, dubbed as Cross-Encoder. It is trained with two types of paired images simultaneously: Paired images of the same eye or with similar gaze directions. Fig. 1(b) shows the intuition of our method. As can be seen, we select the same right (or left) eyes of a subject from different frames to constitute the eye-consistent pairs. For the gaze-similar pair, we use the right and left eyes of the subject in one frame, because when someone is looking at a distant object, the gaze direction of the two eyes are nearly parallel [27]. Cross-Encoder aims at encoding eye's identity information in the eye feature and gaze information in the gaze feature. Our contributions are summarized in three folds. 1) We propose a simple and effective unsupervised representation learning method called Cross-Encoder. It disentangles the representation by reconstructing the images according to switched features. 2) We learn unsupervised gaze-specific representation using Cross-Encoder by introducing two strategies to select the training pairs. 3) Extensive experiments demonstrate the advantage of the learned gaze representation and validate the effectiveness of each component in Cross-Encoder.

2. Related Work

Gaze estimation methods: Gaze estimation includes 2D gaze estimation and 3D gaze estimation. 2D gaze estimation attempts to infer a fixation point on a 2D plane, *e.g.,* a screen. However, different devices result in different relative position of the 2D plane to the camera. Therefore, 2D gaze estimation is hard to generalize to new devices. 3D gaze estimation aims at predicting a line of sight in the 3D world regardless of devices. To estimate the gaze, geometric based methods and appearance based methods were proposed. As detailed in Kar et al. [19], the former requires special hardware like NIR LEDs and IR LEDs, whereas the latter only needs images taken by ordinary RGB cameras. Thus, the appearance-based methods attracts more and more researchers. Our work focuses on learning gaze representation for appearances-based 3D gaze estimation.

Early works on appearances-based gaze estimation focus on designing gaze features from eye images(*e.g.,* the eye landmark [2] and the iris shape [25]), and then use the extracted features to train gaze estimators by off-the-shelf methods (*e.g.,* Principal Component Analysis [36], Support Vector Machine [3]). Recent deep learning methods train the representation and gaze estimator in an end-to-end manner and achieve promising performance [7, 10, 30, 31, 33, 35, 38, 40]. Gaze estimation also benefits from varied sources beyond the eyes' images. For example, Zhang et al. [40] used an attention block to extract the most gaze-related information from the whole face. Cheng et al. [8] correct the asymmetric performance of the left and the right eye using an evaluation network. Note that the "asymmetric" here refers to the performance rather than the physical difference, thus not conflicting our assumption that the left and the right eye share similar gaze. Cheng et al. [7] predicted a coarse gaze direction from face image and corrected it by two eye images. At the same time learning from side information, recent methods learned redundant information that was mixed in the features as well.

Some works make efforts to eliminate the influence of factors other than gaze. Zhang et al. [38] reduced the influence of head pose by normalizing the eyes images according to the rotating a virtual cameras [33]. Deng et al. [10] explicitly learned head poses, gaze directions in different coordinate systems, and the transformation between them. Park et al. [30] disentangled the features of appearance, head pose, and gaze by using an auto-encoder to reconstruct an image of a same person under different head pose with different gaze. Although promising performance is achieved by reducing irrelevant information, the methods require supervisions other than gaze.

Unsupervised representation learning: Unsupervised representation learning aims to learn a representation without access to labeled data. Many novel unsupervised representation learning methods were proposed, *e.g.,* deep clustering [4] and contrastive learning [6] [12]. However, these methods focus on learning general visual features from images and videos. Most of them show good performance on

image classification [4], object detection [9], semantic segmentation [28].

To learn a task-related representation, efforts were made on separating the unsupervised representations into target parts. Locatello et al. [24] stated that unsupervised disentangling is impossible if there is no inductive bias, *e.g.*, special model architecture or prior knowledge for the data. Many works used paired training data and exploited the prior relations between the pairs to learn disentangled representations [5, 17, 18]. Jha et al. [18] disentangled the features into two complementary parts, by alternatively applying cycle-consistency on one part and reconstructing the inputs from the switched features. Chen et al. [5] disentangled features of information which is provided with similarity labels between two images in a pair, by modeling probability of the similarity. Jakab et al. [17] transferred landmark features into explicit heatmaps of one image and concatenated them with appearance feature of another image for reconstruction. Li et al. [23] disentangled the representation of facial actions and head motions by learning how to warp the source image to the target ones that only one factor is changed. Yu et al. [37] is the first, to our best knowledge, to learn gaze-specific representation in a self-supervised learning manner. They took advantages of gaze redirection task and trained the method on paired eye's images of the same subject. Although Yu et al. [37] have acquired simple and effective gaze representation, their method has some limitations. It requires the input pairs to be with similar head poses and to be strictly aligned. Rather than aligning the inputs by extra components, the proposed Cross-Encoder randomly choose two images of the eyes from a subject to get a usable gaze feature.

3. Method

We propose an unsupervised learning framework to extract gaze-specific representation, which excludes the intertwined irrelevant information that has negative effects on gaze estimation. Below, we first introduce a novel Cross-Encoder architecture as the main component with a latent-code-swapping mechanism. Then, we use Cross-Encoder to learn disentangled gaze feature and eye feature by introducing two strategies to choose the training pairs, *i.e.,* the eye-consistent pair and the gaze similar pair.

3.1. Cross-Encoder

A conventional auto-encoder encodes the input into a vector-like embedding, according to which the decoder reconstructs the input. To disentangle the embedding, the Cross-Encoder modifies the conventional auto-encoder by separating the embedding into two parts and taking two paired images as the input. Then, the Cross-Encoder encodes each image into two features called the shared feature and the specific feature. Each image is reconstructed according to its specific feature and the other's shared feature.

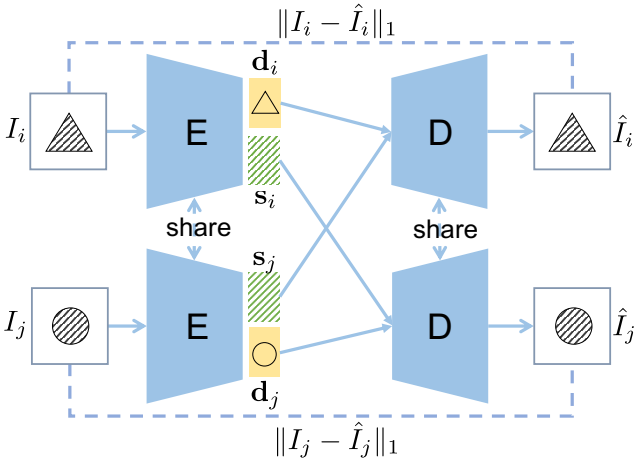


Figure 2. Architecture of the Cross-Encoder. The input images I_i and I_j are encoded as $[\mathbf{d}_i, \mathbf{s}_i]$ and $[\mathbf{d}_j, \mathbf{s}_j]$, respectively. The Cross-Encoder is forced to reconstruct I_i and I_j when $\mathbf{s}_i$ and $\mathbf{s}_j$ are switched. Therefore, $\mathbf{d}_i$ and $\mathbf{d}_j$ are expected to encode the differences between I_i and I_j, and $\mathbf{s}_i$ and $\mathbf{s}_j$ are expected encode the shared feature.

Fig. 2 illustrates the proposed Cross-Encoder architecture. As can be seen, a pair of training images I_i and I_j are fed into the weight-shared encoder E, and are encoded as features $[\mathbf{d}_i, \mathbf{s}_i]$ and $[\mathbf{d}_j, \mathbf{s}_j]$, respectively. We suppose that $\mathbf{d}_i$ and $\mathbf{d}_j$ are the specific features that encode the differences between the input I_i and I_j, *i.e.,* the shape. $\mathbf{s}_i$ and $\mathbf{s}_j$ encode the shared features of the inputs, *i.e.,* the texture. In the conventional auto-encoder, the input I_i is reconstructed from $[\mathbf{d}_i, \mathbf{s}_i]$. In Cross-Encoder, since $\mathbf{s}_i$ and $\mathbf{s}_j$ are consistent, I_i is reconstructed according to $\mathbf{d}_i$ and $\mathbf{s}_j$. Similarly, I_j is reconstructed according to $\mathbf{d}_j$ and $\mathbf{s}_i$.

If the two features were disentangled, the shared features $\mathbf{s}_i$ and $\mathbf{s}_j$ should convey nothing about the specific features $\mathbf{d}_i$ and $\mathbf{d}_j$, and vice versa. To ensure the former, we train the Cross-Encoder by minimizing the loss function

$$\mathcal{L} = \sum_{(i,j)} \|I_i - \hat{I}_i\|_1 + \|I_j - \hat{I}_j\|_1 + \alpha\mathcal{R}, \tag{1}$$

where (I_i, I_j) are the selected training pairs. $\hat{I}_i$ and $\hat{I}_j$ denote the reconstruction of I_i and I_j. The first two items are the *reconstruction loss* that force the reconstructed images to be pixel-wise consistent to the original ones. If $\mathbf{s}_i$ and $\mathbf{s}_j$ were different, the reconstructed images would be inconsistent to the original ones. The item $\mathcal{R} = \|(I_i - \hat{I}_i) - (I_j - \hat{I}_j)\|_1$ is the *residual loss* that regularizes the two residuals between the reconstructed and the original images to be similar. It further prevents the Cross-Encoder to encode the difference between the input pair into the shared features $\mathbf{s}_i$ and $\mathbf{s}_j$. α is a coefficient that balances the importance of the reconstruction loss and residual loss.

Although little information about the differences is en-

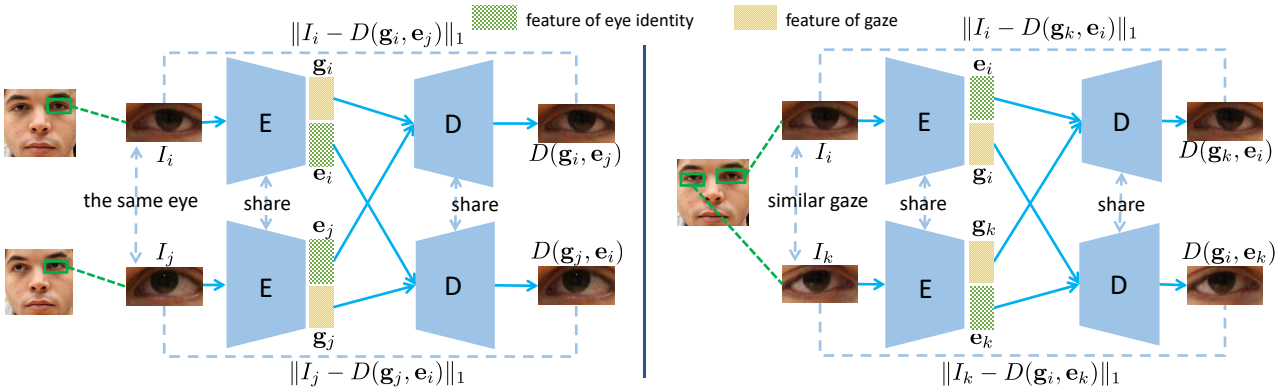


Figure 3. Unsupervised gaze presentation learning framework using Cross-Encoder. The encoder and decoder are updated using both the input pairs with the same eye (left) and the input pairs with similar gaze (right). On the left, the eye features are switched during the reconstructing. On the right, the gaze features are switched.

coded in $\mathbf{s}_i$ or $\mathbf{s}_j$, it is not ensured that the specific feature $\mathbf{d}_i$ and $\mathbf{d}_j$ convey no shared information. There exists a degenerative solution that the encoder keeps all the information in $\mathbf{d}_i$ or $\mathbf{d}_j$ and leaves some noises in $\mathbf{s}_i$ and $\mathbf{s}_j$. To address this issue, we propose to train the Cross-Encoder with a complementary pair of inputs simultaneously. In the new pair, the shared information becomes the difference. Considering the example in Fig. 2, the complementary pair could be a pair of slashed circle and solid circle, where the shared information is the shape and the difference is the texture. Thus, for the new pair, $\mathbf{d}$ should be the shared feature and $\mathbf{s}$ should be the specific feature. By minimizing a similar loss as in (1), we ensure that feature $\mathbf{d}$ does not convey the information that should be in $\mathbf{s}$.

3.2. Unsupervised Gaze Representation Learning

The feature of gaze is intrinsically intertwined with those of the eye identity. To make the features about the gaze and the eye identity be separately encoded into two embeddings, we train the Cross-Encoder using dual input pairs: the eye-consistent pair of the same eye, and the gaze-similar pair with similar gaze.

Figure 3 illustrates the unsupervised gaze presentation learning framework using Cross-Encoder. The parameters of the encoder and decoder are updated using two types of input pairs simultaneously. As can be seen, the eye-consistent pair (I_i, I_j) are selected as images of the same eye from the same person in different video frames. Note that here the left and the right eye are not considered as *the same eye*. Each image is encoded as the gaze feature (yellow rectangle) and the eye feature (green rectangle). Since the two images are of the same eye, their eye features $\mathbf{e}_i$ and $\mathbf{e}_j$ should be consistent. Thus, we could reconstruct the input pair according to their own gaze feature and the switched eye feature. The gaze-similar pair (I_i, I_k) are the images of the two eyes from the person in one video frame. Because when someone is looking at a distant object, the gaze direction of two eyes are nearly parallel [27], we assume that the gaze features $\mathbf{g}_i$ and $\mathbf{g}_j$ are close in the feature space. Thus, we could reconstruct the input pair according to their own eye feature and the switched gaze feature. One may argue that apart from the eye's deformable shape (encoded in eye feature) and gaze (encoded in gaze feature), other environmental factors such as illumination would affect the eye's appearance on the images. However, this issue is not crucial. First, to avoid strong illumination factors, we turn the images to gray scale and perform histogram equalization. Furthermore, the dimension of the gaze feature can be sightly extended to tolerate environmental changes among frames. This would not affect performance as can be seen in the experiment.

Mathematically speaking, we train the Cross-Encoder for gaze representation by minimizing

$$\mathcal{L} = \sum_{(i,j)\in\mathcal{E}} \|I_i - D(\mathbf{g}_i, \mathbf{e}_j)\|_1 + \|I_j - D(\mathbf{g}_j, \mathbf{e}_i)\|_1 + \alpha\mathcal{R}_\mathcal{E} + \beta \sum_{(i,k)\in\mathcal{G}} \|I_i - D(\mathbf{g}_k, \mathbf{e}_i)\|_1 + \|I_k - D(\mathbf{g}_i, \mathbf{e}_k)\|_1 + \gamma\mathcal{R}_\mathcal{G}, \tag{2}$$

where $\mathcal{E}$ is the set of pairs with the same eyes and $\mathcal{G}$ is the set of pairs with similar gaze. $D(\mathbf{g}_i, \mathbf{e}_j)$ denotes the reconstructed image according to gaze feature $\mathbf{g}_i$ and eye feature $\mathbf{e}_j$. $\mathcal{R}_\mathcal{E}$ and $\mathcal{R}_\mathcal{G}$ are the residual losses for the two pairs respectively. α, β, γ are the coefficients to balance the items. Considering that the gaze are similar but not equivalent for the pairs in $\mathcal{G}$, β and γ are smaller than 1 and α, respectively.

4. Experiments

We thoroughly evaluated Cross-Encoder through its learned representation and the pre-trained model, by comparing them with the-state-of-the-art methods on public datasets. We analyze the disentangled features and discuss

Table 1. Angular errors (mean±*std*) of 100-shot gaze estimation within Columbia, UTMultiview and MPIIGaze datasets. GS stands for gaze-similar pair. d_g and d_e: the dimension of gaze feature and eye feature. For *eye feature, gaze feature(no GS pair), gaze feature(no residual loss)* setting, d_e is 32, d_g is 9, 12 and 12 for UTMultiview, Columbia and MPIIGaze respectively. Note that for MPIIGaze, we use the Cross-Encoder pretrained unsupervisedly on Columbia and only trained for 10 epochs on MPIIGaze.

	w/ head pose			w/o head pose		
methods	Columbia	UTMultiview	MPIIGaze	Columbia	UTMultiview	MPIIGaze
ImageNet-Pretrained ResNet18	12.1±0.1	20.2±0.5	10.6±0.2	11.9±0.2	24.9±0.5	10.6±0.2
auto-encoder	10.5±0.2	18.0±0.5	9.5±0.2	10.6±0.3	18.5±0.5	9.5±0.1
auto-encoder (EFC)	9.2±0.3	13.5±0.3	9.2±0.2	9.4±0.3	22.1±0.5	8.9±0.1
SimCLR [6]	7.2±0.1	12.1±0.2	10.0±0.3	8.2±0.03	21.3±0.7	9.8±0.2
BYOL [12]	9.9±0.1	14.4±0.2	11.1±0.5	10.2±0.03	23.5±0.2	11.0±0.6
Yu et al. [37]	8.95	8.56	-	-	-	-
Cross-Encoder (proposed)						
- eye feature	12.8±0.1	15.5±0.4	9.8±0.1	12.6±0.2	31.9±0.3	9.7±0.1
- gaze feature (no GS pair)	7.6±0.1	10.6±0.3	8.2±0.1	8.5±0.2	17.2±0.6	8.1±0.2
- gaze feature (no residual loss)	6.7±0.1	**7.4**±0.1	**7.2**±0.2	7.4±0.1	**8.2**±0.2	**7.2**±0.2
- gaze feature (d_g=9,d_e=32)	6.7±0.1	7.7±0.3	8.1±0.2	7.6±0.1	8.8±0.2	8.0±0.2
- gaze feature (d_g=12,d_e=32)	6.6±0.1	8.0±0.2	7.5±0.1	7.3±0.1	8.9±0.2	7.6±0.2
- gaze feature (d_g=15,d_e=32)	**6.4**±0.1	8.0±0.2	7.5±0.2	**7.1**±0.1	9.2±0.2	7.3±0.2
- gaze feature (f_g=12, f_e=16)	6.7±0.1	7.6±0.2	7.2±0.2	7.4±0.2	8.6±0.2	7.2±0.1
- gaze feature (f_g=12, f_e=64)	6.5±0.1	7.8±0.2	7.5±0.2	7.2±0.1	8.9±0.1	7.4±0.1

the effectiveness of the feature switching mechanism, two types of input pairs, feature dimension, and residual loss.

4.1. Experimental settings

Implementation details: We implemented the Cross-Encoder using PyTorch. In our experiments, we used ResNet18 [13] as the encoder, and four DenseNet [15] deconvolution blocks as the decoder. It is worth noting that the encoder and decoder can be of any other architecture. The eyes images were cropped according to the detected facial landmarks [14] around the eyes. All input images were gray scale and were histogram equalized to eliminate the illumination effects. We trained Cross-Encoder on one TITAN RTX GPU, using Adam [21] optimizer for 200 epochs. The learning rate was 0.0001. In each batch, the two types of training pairs were half and half. For the eye-consistent pair, we randomly selected a subject's right or left eyes from two random frames of a clip. For the gaze-similar pair, we randomly selected a frame and crop the subject's two eyes.

Datasets: We evaluated the methods on public gaze datasets Columbia Gaze [32], UTMultiview [33], and MPIIGaze [41]. All of the datasets contain various head pose and gaze directions. **Columbia Gaze(C)** consists of 6000 face images from 56 subjects. **UTMultiview(U)** consists of 64000 face images from 50 subjects. We used the real world part of UTMultiview in our experiments. **MPIIGaze(M)** contains 213659 images of 15 subjects and were collected in front of the screen of a laptop in daily life. 5-fold, 3-fold, and leave-one-out cross validation evaluation was used for Columbia, UTMultiview, and MPIIGaze respectively. We also used XGaze [39], TabletGaze [16], and FreeGaze as the unsupervised data, ignoring their annotations, **XGaze(X)** was collected from 18 high-resolution Canon 250D digital SLR cameras when the subjects were looking at the points on a screen. We used the 756540 images of 80 subjects in the traing set of XGaze. **TabletGaze(T)** was collected when the subjects were watching the Tablet with 4 different poses. We sampled 171971 images in 817 videos of 51 subjects. **FreeGaze(F)** is a self-collected dataset with 138 Asians looking around freely in front of 4 cameras. It consists of 867808 images.

Table 2. Mean angular errors of 100-shot gaze estimation under cross-dataset settings.

	C	U	M
supervised			
- trained on C	-	10.84	8.35
- trained on U	7.19	-	8.11
- trained on X	**5.67**	**8.79**	**7.28**
unsupervised			
- Yu et al. [37](trained on U)	8.82	-	-
Cross-Encoder			
- trained on C	-	9.79	8.32
- trained on U	7.48	-	9.09
- trained on X	7.76	10.30	9.04
- trained on X, T, and F	**7.09**	**9.58**	**8.20**

4.2. Evaluation of the learned representation

We evaluated the learned representation by a few shot gaze estimation task, as in [37], under both within-dataset and cross-dataset setting.

Table 3. Angular errors (mean±*std*) of 50/200-shot gaze estimation within Columbia ,UTMultiview and MPIIGaze datasets.

		AE(EFC)	simCLR	BYOL	ours
50	C	11.1±0.3	8.1±0.04	10.8±0.1	**7.0**±0.2
	U	14.9±0.5	14.4±0.5	15.1±0.5	**8.8**±0.4
	M	9.8±0.2	10.7±0.4	11.9±0.4	**8.5**±0.2
200	C	7.8±0.1	6.3±0.02	9.4±0.03	**6.2**±0.1
	U	11.9±0.3	11.0±0.3	14.1±0.2	**7.3**±0.2
	M	8.8±0.1	9.2±0.3	10.4±0.4	**7.3**±0.1

Within-dataset evaluation: We compared the representations of the *ImageNet-pretrained ResNet18*, the vanilla *auto-encoder* with the same encoder and decoder architectures as Cross-Encoder, the only previous self-supervised gaze representation learning method (*Yu et al. [37]*), two recent popular contrastive learning methods *SimCLR* [6] and *BYOL* [12], and variations of the proposed Cross-Encoder. Auto-encoder with equal feature constraint, *Auto-encoder(EFC)*, forces the features to be equal by adding a constraint(L1 loss) under the auto-encoder framework. Given a pair of inputs, EFC encodes each input into two features. Instead of swapping the to-be-consistent features, EFC forces the to-be-consistent feature to be equal by minimizing the L1 loss on them.

Table 1 shows the mean and standard deviation of angular errors of 100-shot gaze estimation within Columbia, UTMultiview and MPIIGaze datasets. In each fold, we first trained Cross-Encoder using the unlabelled training sets. Then, we trained the gaze estimator according to 100 random training samples with labels and repeated for 10 times to show the mean and standard deviation. Since head poses can affect gaze estimation, we reported the results with and without head pose information. We listed Yu et al. [37] as one with head pose because the head pose is necessary when they regressed the gaze. For other methods, in the *w/ head pose* setting, the representation was regarded as the concatenated head pose and the learned gaze feature.

In Table 1, the Cross-Encoder-learned gaze feature outperforms other representations on varied datasets and settings. Pretrained ResNet18 is the worst because its features are learned for image classification. Cross-Encoder surpasses auto-encoder, because the auto-encoder-learned representation preserves all the information to reconstruct the eyes' image, including gaze-irrelevant information. SimCLR [6] and BYOL [12] methods focus on learning general visual features rather than representation specially for gaze estimation, which results in their poor performance. Compared with Yu et al. [37], Cross-Encoder achieved consistent improvements with different settings of feature dimension on two datasets. Rather than directly concatenating the features to head pose, Yu et al. [37] incorporated the head pose information to normalize the input images and to transform the regressed gaze vector from the camera coordinate system to head coordinate system. Thus, they explicitly eliminated the influence of head pose. Although without concatenating with head pose or transforming to head pose coordinating system, the proposed Cross-Encoder can get lower MAEs than [37] and is simpler than [37].

Table 4. Mean angular errors of Cross-Encoder and the state-of-the-art methods on Columbia and UTMultiview datasets.

	Columbia	UTMultiview
Yu et al. [37]	**3.42**	5.52
Park et al. [31]	3.59	-
Zhang et al. [38]	-	5.9
Wang et al. [35]	-	5.4
Cross-Encoder(proposed)	3.52	**4.81**

We also show 50/200-shot performance in Table 3, where Cross-Encoder consistently outperforms the others. It indicates that Cross-Encoder is more stable with the number of few-shot samples than other methods.

Cross-dataset evaluation: To investigate the learned features' generalization ability, we used different datasets to train the representation and conduct the 100-shot gaze estimation. We compared the supervised features of ResNet18 trained on three public gaze estimation dataset(*i.e.,* Columbia, UTMultiview, XGaze), the state-of-the-art unsupervised gaze representation [37], and the unsupervised gaze features by Cross-Encoder trained on different datasets. To see the power of unlabeled data, the union of datasets (XGaze, TabletGaze, and FreeGaze) was also used unsupervisedly to train Cross-Encoder. Table 2 shows the mean angular errors of 100-shot gaze estimation under cross-dataset settings. Note that the representation is concatenated with the head pose. 5-fold, 3-fold, leave-one-out evaluation protocols were used on Columbia, UTMultiview, and MPIIGaze, respectively.

In Table 2, we have three observations. First, Cross-Encoder outperforms the state-of-the-art unsupervised gaze representation [37]. Second, with the same training data, supervised features are better than unsupervised ones except for ones trained by Columbia, as the supervised methods are easily to overfit the Columbia dataset, which only contains 6000 face images. Third, more training data contributes to better representation learned by Cross-Encoder. The Cross-Encoder trained on the union of XGaze, TabletGaze, and FreeGaze is the best among all unsupervised models, because it contains the most images and subjects.

4.3. Comparison with the state-of-the-art gaze estimation methods

We further evaluated Cross-Encoder by comparing it with the state-of-the-art gaze estimation methods, including three supervised ones [31] [38] [35] and a self-supervised one [37]. Table 4 shows the mean angular errors of the compared methods on Columbia and UTMultiview datasets. 5-

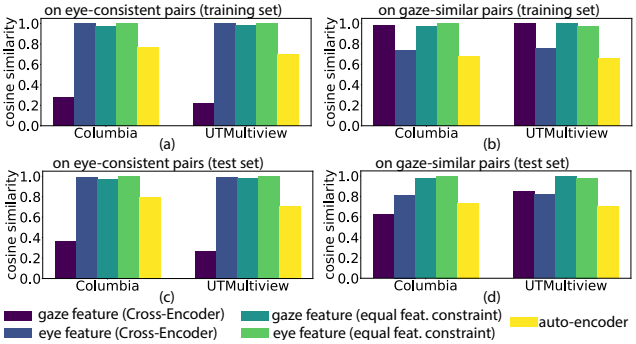


Figure 4. Average cosine similarity of representations between images in eye-consistent pairs and gaze-similar pairs.

fold and 3-fold evaluation protocols were used one the two datasets, respectively. We first trained the Cross-Encoder using the unlabelled training data. Then we put a two-layer MLP after the encoder to build a gaze estimator on the annotated training data. Performance of the estimator on the test set is reported. The results for other methods were originally reported by the authors in their papers.

In Table 4, Cross-Encoder achieves competitive performance as the state-of-the-art gaze estimation methods. Cross-Encoder gets smaller errors than the other methods on UTMultiview. On Columbia, Cross-Encoder performs similarly to Park [31] but slightly worse than Yu [37]. The reason is that Columbia contains much fewer samples than UTMultiview. Pretraining on a larger amount of data makes the encoder better captures the distribution of the data.

4.4. Ablation Study

Are the gaze and eye features disentangled? First, we compared the gaze feature and eye feature in few-shot gaze estimation task. In Table 1, using the eye feature is much worse than that using the gaze feature. It indicates that little information about the gaze is encoded in the eye feature.

Second, we visualized the learned features in the two types of input pairs. Fig. 5 shows examples of the features learned by Cross-Encoder, the vanilla auto-encoder and auto-encoder(EFC), on both the training set and test set of Columbia and UTMultiview. For each feature, the x-axis corresponds to the index of dimension, and the y-axis corresponds to the value. For the images in gaze-similar pair (int the same column), their Cross-Encoder learned gaze features are similar but their eye features have differences. And vise verse for the images in eye-consistent pairs(in the same row). It indicates that the Cross-Encoder does encode the information about gaze in the gaze features and encode what the eye looks like in the eye features. The auto-encoder-learned features varies with either different eyes or gaze, because the information are intertwined in the features.

Third, Fig. 4 plots the averaged cosine similarity of the learned features between images in pairs. To compute the averaged similarity, we randomly chose eye-consistent pairs

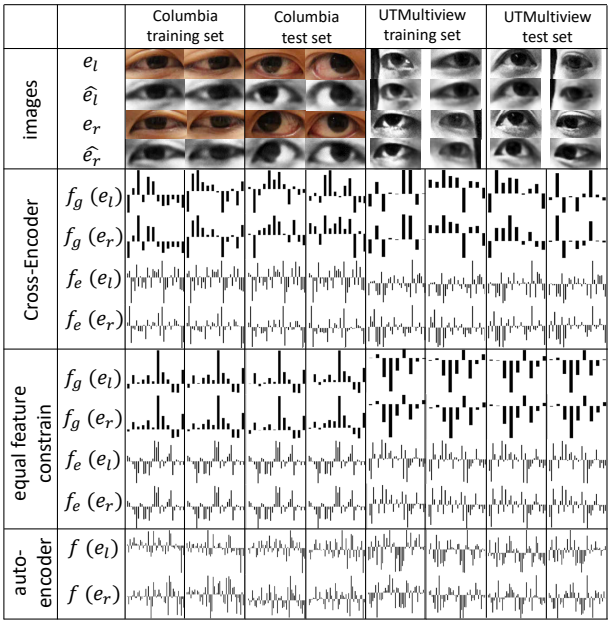
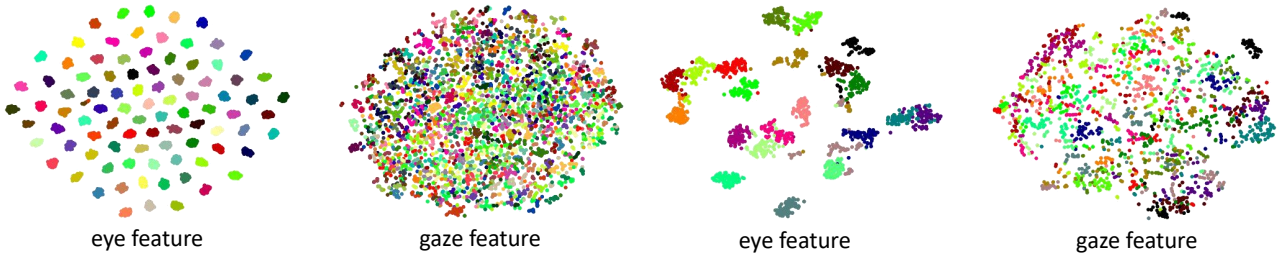


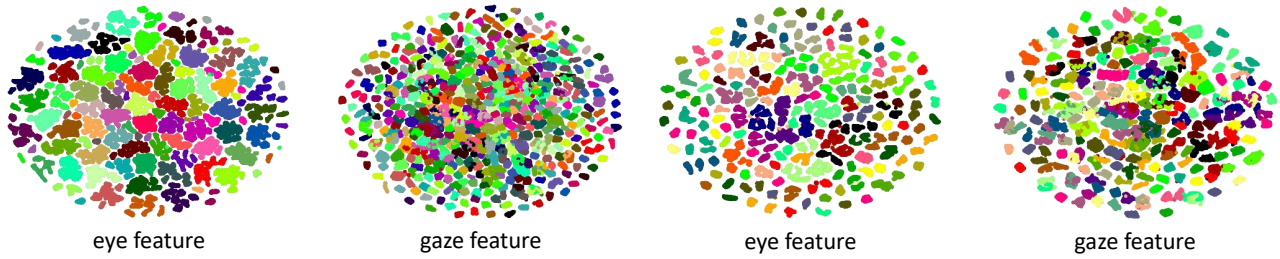
Figure 5. Examples of eye images from different datasets and their eye feature (f_e) and gaze feature (f_g). In each dataset, the eyes (in rows) are from the same person but different frames. The eye in columns are the left one(e_l) and right one(e_r) from one frame. $\hat{e_l}$ and $\hat{e_r}$ are the reconstructed images by Cross-Encoder.

that covered all the subjects. For the gaze-similar pairs, we used the left and right eyes in every image. It is observed that neither the gaze nor the eye feature learned by other methods show differences between the two types of pairs. Yet, the gaze feature (Cross-Encoder) on gaze-similar pairs are significantly more similar than those on eye-consistent pairs. The eye feature (Cross-Encoder) on gaze-similar pairs are less similar than those on eye-consistent pairs. It indicates the advantages of Cross-Encoder in disentangling the gaze and eye features. It is worth noting that the similarities of the eye features are large (more than 0.5) on both the pairs, the value is even higher than that of gaze feature on the gaze-similar pair (test). The reason is that in gaze-similar pair, although the left and the right eye have different eye-identity, they are of the same person and gain a moderately large similarity. We also observe that similarity of the gaze feature (Cross-Encoder) drops on the test set on gaze-similar pairs. The values are nearly 1 for on the training set, and are around 0.6 and 0.8 on the test set. It is reasonable because the gazes are not absolutely equivalent, but are assumed to be similar in our method.

Forth, we visualized the eye features and gaze features in 2-dimensional space using t-SNE [34] in Fig. 6. Each point denotes the feature of an eye's image. Different colors denote different eye identities. Note that one *eye-identity* corresponds to one eye, which makes the left and the right eye of a same person has different eye-identities, so the number of colors are twice the number of subjects. It is observed



(a) Columbia training set. (b) Columbia validation set.



(c) UTMultiview training set. (d) UTMultiview validation set.

Figure 6. Visualization of Cross-Encoder-learned representation using t-SNE [34]. Each point corresponds to an eye's image and is colored by its eye's identity. Better viewed in color.

that the eye features are clustered by eye identities but the the clusters are not so obvious for gaze features. The gaze features are better mixed than the eye features, although a few person-specific clusters exist. It is worth noting that despite the imperfect clustering, the Cross-Encoder-learned gaze feature is effective for gaze estimation.

Cross-Encoder versus equal feature constraint: Since Cross-Encoder is designed to make the features be similar, an alternate is to force the features to be equal by adding a constraint (*e.g.,* L_1 loss) under the conventional auto-encoder framework. We demonstrated the advantage of the Cross-Encoder over auto-encoder using equal feature constraint. First, the constraint-learned gaze features performed worse than the Cross-Encoder-learned ones in gaze estimation task. In Table 1, the errors of auto-encoder(equal feature constraint) are much larger than those of the best Cross-Encoder among all datasets. Second, using equal feature constraint cannot disentangle the eye features and gaze features. As shown in Fig. 5, the constraint-learned gaze features and eye features are almost the same in spite of varied gaze and eyes. It is also observed in Fig 4 that the cosine similarities of both the two constraint-learned features are near 1, on both eye-consistent pairs and gaze-consistent pairs. The possible reason is that it reaches a degenerative solution which encodes all the input images into two similar features. The features are enough to reconstruct the original images by losing subtle information.

Two types of training pairs: It is necessary to use both eye-consistent pair and gaze-consistent pair. Table 1 shows a big gap between the performance of the Cross-Encoders with only eye-consistent pair and with both pairs.

Dimension of the gaze feature and the eye feature: Table 1 reports the performance of Cross-Encoders with different gaze feature and eye feature dimensions. The optimal gaze feature dimensions are 15, 15, and 9 on Columbia, MPIIGaze and UTMultiview, respectively. Columbia and MPIIGaze has a larger feature dimension probably because it has more variance, *e.g.,* subjects who wear glass. The optimal eye feature dimension is 16 for MPIIGaze and 64 for the other two datasets. In general, the differences are subtle and Cross-Encoder exceeds other methods regardless of the dimension. This shows that Cross-Encoder is not sensitive to the dimension of features.

Residual loss: Table 1 reports the performance of Cross-Encoders without residual loss. Adding the residual loss improves the performance on Columbia but reduces that on UTMultiview and MPIIGaze. This might ascribe to the different distribution of the datasets. We concluded that residual loss is optional and the reconstruction loss is the cutting edge loss.

5. Conclusions

In this paper, we have presented an unsupervised learning method to learn a gaze representation. Our key contributions are to propose an unsupervised representation learning method Cross-Encoder, and applying it to gaze estimation by introducing two strategies to select the training pairs. We conducted experiments to prove the ability and disentanglement of the learned representations, and the performance of fine-tuned Cross-Encoder. They all show the validity of our method. Cross-Encoder is a general approach for unsupervised representation learning. Future work can explore other applications of Cross-Encoder.

References

[1] Christer Ahlstrom, Katja Kircher, and Albert Kircher. A gaze-based driver distraction warning system and its effect on visual behavior. *IEEE Transactions on Intelligent Transportation Systems*, 14(2):965–973, 2013.

[2] Ioana Bacivarov, Mircea Ionita, and Peter Corcoran. Statistical models of appearance for eye tracking and eye-blink detection and measurement. *IEEE Transactions on Consumer Electronics*, 54:1312–1320, 08 2008.

[3] Christopher JC Burges. A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery*, 2(2):121–167, 1998.

[4] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European Conference on Computer Vision*, pages 132–149, 2018.

[5] Junxiang Chen and Kayhan Batmanghelich. Weakly supervised disentanglement by pairwise similarities. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 3495–3502, 2020.

[6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pages 1597–1607, 2020.

[7] 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*, volume 34, pages 10623–10630, 2020.

[8] Yihua Cheng, Feng Lu, and Xucong Zhang. Appearance-based gaze estimation via evaluation-guided asymmetric regression. In *Proceedings of the European Conference on Computer Vision*, pages 100–115, 2018.

[9] Eric Crawford and Joelle Pineau. Spatially invariant unsupervised object detection with convolutional neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3412–3420, 2019.

[10] Haoping Deng and Wangjiang Zhu. Monocular free-head 3d gaze tracking with deep learning and geometry constraints. In *2017 IEEE International Conference on Computer Vision*, 2017.

[11] Kunihiko Fukushima. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, 36(4):193–202, 1980.

[12] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Ávila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent - A new approach to self-supervised learning. In *Advances in Neural Information Processing Systems*, 2020.

[13] 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.

[14] Zhenliang He, Jie Zhang, Meina Kan, Shiguang Shan, and Xilin Chen. Robust fec-cnn: A high accuracy facial landmark detection system. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 98–104, 2017.

[15] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[16] Qiong Huang, Ashok Veeraraghavan, and Ashutosh Sabharwal. Tabletgaze: dataset and analysis for unconstrained appearance-based gaze estimation in mobile tablets. *Machine Vision and Applications*, 28(5):445–461, 2017.

[17] Tomas Jakab, Ankush Gupta, Hakan Bilen, and Andrea Vedaldi. Unsupervised learning of object landmarks through conditional image generation. In *Advances in Neural Information Processing Systems*, volume 31, pages 4016–4027, 2018.

[18] Ananya Harsh Jha, Saket Anand, Maneesh Singh, and VSR Veeravasarapu. Disentangling factors of variation with cycle-consistent variational auto-encoders. In *Proceedings of the European Conference on Computer Vision*, September 2018.

[19] Anuradha Kar and Peter Corcoran. A review and analysis of eye-gaze estimation systems, algorithms and performance evaluation methods in consumer platforms. *IEEE Access*, 5:16495–16519, 2017.

[20] 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*, October 2019.

[21] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

[22] Chris L Kleinke. Gaze and eye contact: a research review. *Psychological bulletin*, 100(1):78, 1986.

[23] Yong Li, Jiabei Zeng, and Shiguang Shan. Learning representations for facial actions from unlabeled videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[24] Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *International Conference on Machine Learning*, pages 4114–4124, 2019.

[25] Feng Lu, Yue Gao, and Xiaowu Chen. Estimating 3d gaze directions using unlabeled eye images via synthetic iris appearance fitting. *IEEE Transactions on Multimedia*, 18(9):1772–1782, 2016.

[26] Päivi Majaranta and Andreas Bulling. Eye tracking and eye-based human–computer interaction. In *Advances in physiological computing*, pages 39–65. Springer, 2014.

[27] Mary-Ellen Meadows. *Conjugate Gaze*, pages 674–675. Springer New York, 2011.

[28] Takayasu Moriya, Holger R Roth, Shota Nakamura, Hirohisa Oda, Kai Nagara, Masahiro Oda, and Kensaku Mori.

Unsupervised segmentation of 3d medical images based on clustering and deep representation learning. In *Biomedical Applications in Molecular, Structural, and Functional Imaging*, volume 10578, page 1057820. International Society for Optics and Photonics, 2018.

[29] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European conference on computer vision*, pages 69–84, 2016.

[30] 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.

[31] Seonwook Park, Adrian Spurr, and Otmar Hilliges. Deep pictorial gaze estimation. In *Proceedings of the European Conference on Computer Vision*, pages 721–738, 2018.

[32] B.A. Smith, Q. Yin, S.K. Feiner, and S.K. Nayar. Gaze Locking: Passive Eye Contact Detection for Human-Object Interaction. In *ACM Symposium on User Interface Software and Technology*, pages 271–280, Oct 2013.

[33] Yusuke Sugano, Yasuyuki Matsushita, and Yoichi Sato. Learning-by-synthesis for appearance-based 3d gaze estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[34] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[35] Kang Wang, Rui Zhao, Hui Su, and Qiang Ji. Generalizing eye tracking with bayesian adversarial learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11907–11916, 2019.

[36] Svante Wold, Kim Esbensen, and Paul Geladi. Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3):37–52, 1987.

[37] 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*, June 2020.

[38] Xucong Zhang. Appearance-based gaze estimation in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[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, 2020.

[40] 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.

[41] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. Mpiigaze: Real-world dataset and deep appearance-based gaze estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(1):162–175, 2019.