UAGE: A Supervised Contrastive Method for Unconstrained Adaptive Gaze Estimation

Enfan Lan[1,2,3], Zhengxi Hu[1,2,3], and Jingtai Liu[1,2,3]✉

[1] Institute of Robotics and Automatic Information System, Nankai University
[2] Tianjin Key Laboratory of Intelligent Robotics
[3] TBI center, Nankai University
{lef, hzx}@mail.nankai.edu.cn, liujt@nankai.edu.cn

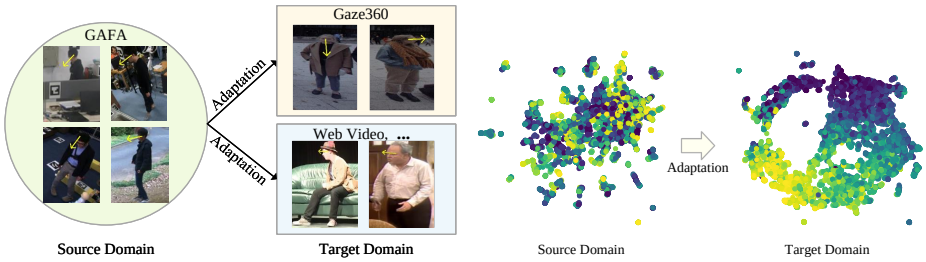
Abstract. Gaze estimation, which involves perceiving human gaze directions, is the foundation of gaze analysis. It provides crucial clues for understanding human attention and intention. However, most existing methods are designed for constrained environments, which leads to a significant performance drop in unconstrained practical applications. In this work, we propose a supervised contrastive method for Unconstrained Adaptive Gaze Estimation (UAGE), which consists of an unconstrained gaze estimation method and a Gaze-guided Contrastive Domain Adaptation (GCDA) framework. Our method leverages the entire human body states and the uncertainty of gaze behaviors to robustly estimate gazes in unconstrained environments, rather than solely relying on head states. Additionally, we employ the GCDA framework to adapt the model to new domains, thereby improving its generalization ability. Experiment results show that our UAGE method has achieved state-of-the-art within-domain performance on the unconstrained GAFA dataset and has reduced the angular error by 14% compared to the baseline in cross-domain gaze estimation, with GAFA as the source domain and Gaze360 as the target domain. The code is available at https://github.com/youthhfor/UAGE.git.

Keywords: Gaze estimation · Domain adaptation · Unconstrained environment

1 Introduction

Gaze is a crucial non-verbal way for humans to transmit information. It directly reflects human attention and helps in identifying objects of interest [8,20], understanding as well as inferring human intention [33]. Humans, even infants, naturally possess the ability to estimate and analyze gaze, a capability that machines inherently lack. Therefore, in order to better understand human behaviors and

✉Corresponding author.
This work is supported by the National Natural Science Foundation of China under Grant 62173189.



(a) Unconstrained adaptive gaze estimation **(b)** Feature alignment

Fig. 1: We introduce a novel UAGE method for adaptive gaze estimation in unconstrained environments. It utilizes the entire human body states and the uncertainty of gaze behaviors to estimate gazes and leverages the GCDA framework to adapt to new domains. Our UAGE greatly performs (a) unconstrained gaze estimation in both the source and target domains and (b) feature distribution alignment in latent space.

enhance human-computer interaction, machines should also be equipped with the ability of gaze estimation. This involves automatically perceiving accurate human gaze directions and serves as the foundation for further gaze analysis. In practice, gaze estimation has been widely applied in various human-computer interaction scenarios, such as driver anomaly detection [9], VR/AR [3,23], robot navigation [19,38] and medical analysis [42].

With the development of deep learning, existing gaze estimation methods have achieved excellent performance in constrained environments. Appearance-based methods [28,39–41] leverage the appearance features of human eyes and faces to obtain accurate gaze predictions in within-domain evaluation. To further enhance the generalization ability of models, some works [1,4,7,27,32] have also explored cross-domain gaze estimation in constrained environments. These works adapt models pretrained on the source domain with a few labeled or unlabeled target data to achieve better performance in the target domain.

Despite the significant advances in constrained gaze estimation methods, most of them cannot be applied to unconstrained new domains. This is because, in unconstrained environments, obtaining high-quality frontal human images is difficult and gaze behaviors exhibit considerable uncertainty. To address this issue, recent works [10,12,20,26] have explored within-domain unconstrained gaze estimation methods and collected large-scale datasets. However, these methods are limited by local states of humans or extra annotations and there are no well-elaborated methods designed for cross-domain gaze estimation in unconstrained environments. This hinders the generalization capability of models in practice.

In this paper, we propose a UAGE method for adaptive gaze estimation in unconstrained environments. As shown in Fig. 1, our UAGE achieves effective gaze estimation in unconstrained environments and adapts to new domains, such as Gaze360 [20] and web videos, through feature alignment. Our UAGE consists of a novel unconstrained gaze estimation method and a GCDA framework for few-shot domain adaptation. For gaze estimation in unconstrained environments,

we find that humans typically focus on the entire body states of others to estimate gazes, and human gaze behaviors exhibit uncertainty. When humans are engaged in similar states in the environment, their gaze directions may vary widely within the gaze distribution latent space. Inspired by this, we propose to leverage the entire human body states and the uncertainty of gaze behaviors to perform unconstrained gaze estimation. Specifically, four branches are employed to acquire head, body, pose, and velocity features respectively, from which entire human body states can be derived. Then, the concatenated features are fed into a conditional variational autoencoder (CVAE) [29] to generate multi-modal results in latent space, which are empirically decoded into robust gaze predictions.

For domain adaptation in unconstrained environments, our key idea is that in both the source and target domains, features extracted from inputs with similar gaze labels should cluster together in latent space. Therefore, given a model pretrained on the source domain and a few labeled target data, we adopt the GCDA framework to adjust the feature extractor to align the features in latent space. This facilitates the adjustment of the mapping between latent space and gaze distribution. Experiment results demonstrate the effectiveness of our UAGE method in both unconstrained within-domain and cross-domain evaluations.

In summary, our main contributions are as follows:

- We propose a UAGE method for gaze estimation in unconstrained environments, which utilizes the entire human body states in the environment and the uncertainty of gaze behaviors to robustly estimate gaze directions.
- UAGE also contains a GCDA framework for few-shot domain adaptation, which leverages gaze-guided contrastive learning to align the features in latent space and then adjusts the mapping to the gaze distribution.
- Experiment results demonstrate that our UAGE has achieved state-of-the-art within-domain gaze estimation performance on GAFA dataset and has significantly outperformed the baselines by 14% in cross-domain gaze estimation, with GAFA as the source domain and Gaze360 as the target domain.

2 Related Work

2.1 Within-Domain Gaze Estimation

Gaze estimation methods can be roughly categorized into geometry-based and appearance-based. Geometry-based methods usually use dedicated devices to extract geometric features, such as infrared corneal reflection [43] and pupil center [31]. These features are leveraged to build subject-specific eye models, from which gazes can be obtained. Appearance-based methods often take eye/head images as input and learn the image-gaze mapping. Early methods applied algorithms like support vector regression [34] and random forests [18] to estimate gaze. Recently, deep learning methods [28, 30, 37, 40, 41] based on convolutional neural networks (CNNs) have demonstrated outstanding performance.

However, most of these methods are designed for constrained environments that contain close-up face/eyes images in front views, which limits their performance in unconstrained applications. To tackle this problem, some works

[12, 17, 20, 26] have investigated gaze estimation in unconstrained environments, where images are acquired from diverse distances in various scenes. Gaze360 [20] created a large-scale unconstrained dataset that includes 360° horizontal coverage of gaze and proposed a gaze estimation method that considers the error bound. Furthermore, GAFA [26] proposed a dataset in surveillance views that contains freely moving individuals and more cases of occlusion and low resolution. Their method innovatively leverages the intrinsic gaze, head, and body coordination to estimate gaze. GFIE [17] introduced the concept of gaze candidate space and combined gaze salience with scene salience to achieve accurate gaze predictions. These methods typically utilize spatial features of the head and body, as well as temporal features between frames, to regress gaze directions. Our UAGE additionally incorporates pose features and concatenates them with head, body, and velocity features to obtain the entire human body states in the environment. Moreover, considering that people in similar states within unconstrained environments may have different gaze directions, we integrate the probabilistic approaches CVAE [11, 22, 29] into the network to generate multimodal results in latent space, which enhances the robustness of final estimation.

2.2 Cross-Domain Gaze Estimation

Despite existing gaze estimation methods having achieved outstanding within-domain performance, they often suffer from performance degradation on cross-domain evaluation due to large differences in gaze distributions between source and target domains. Therefore, recent studies have explored cross-domain gaze estimation, which can be categorized into generalization-based and adaptation-based. Generalization-based methods [2,7,36] aim to enhance model's generalization ability across all domains rather than a specific one. Cheng et al. [7] proposed to use adversarial learning to extract domain-agnostic gaze features. Adaptation-based methods [1, 4, 13, 25, 27, 32] attempt to align the feature distributions of the source and target domains. Bao et al. [1] discovered the rotation-consistency in gaze estimation and utilized it to generate sub-labels for adaptation. Wang et al. [32] proposed a contrastive regression method for adaptation. Cai et al. [4] achieved adaptation by reducing sample and model uncertainty.

However, there are no well-elaborated cross-domain gaze estimation methods designed for unconstrained environments, which limits the performance of models in unconstrained practical applications. To address this issue, we integrate a GCDA framework into our UAGE method, which achieves feature alignment in the target domain. This framework clusters samples with similar gaze labels together and separates samples with dissimilar gaze labels. As a result, stable gaze feature representations can be obtained, making the adjustment of mapping from the latent space to gaze distribution easier.

3 Method

In this section, we introduce the detailed network architecture of our UAGE, as shown in Fig. 2, which consists of an unconstrained gaze estimation method and

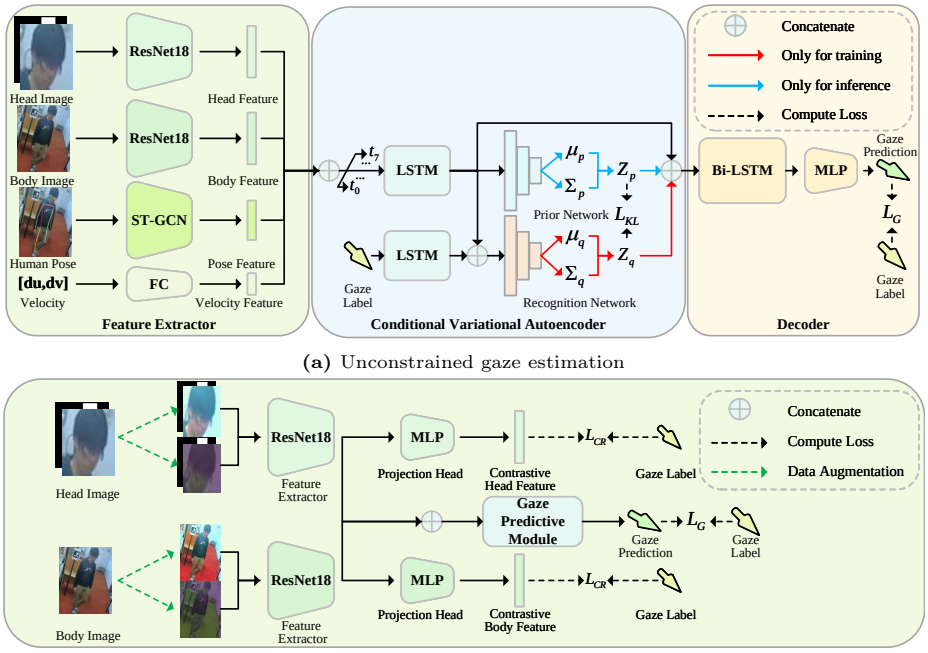


Fig. 2: The overall network architecture of UAGE. The model takes 7-frame videos as input. (a) For unconstrained gaze estimation, four branches are employed to acquire head, body, pose, and velocity features respectively. These features are then concatenated and fed into the CVAE module to generate multi-modal latent results, which are finally decoded into gaze directions. (b) For GCDA framework, data augmentation and gaze-guided contrastive loss are additionally applied in the head and body branches.

a GCDA framework for few-shot domain adaptation. In Sec. 3.1, we introduce our unconstrained gaze estimation method, which can be divided into three parts: a feature extractor to estimate the entire human states, a conditional variational autoencoder to generate multi-modal results in latent space, and a decoder to obtain gaze predictions. Based on this, in Sec. 3.2, we introduce the gaze-guided contrastive learning method for further adaptive gaze estimation and then describe the few-shot domain adaptation procedure of GCDA.

3.1 Gaze Estimation in Unconstrained Environment

Feature Extractor The feature extractor module is shown on the left side of Fig. 2a, which consists of the head, body, pose, and velocity branches. The model takes 7-frame videos as input. For the head and body branches, body images and head crops are fed into two separate ResNet-18 [15] backbones to obtain high-dimensional spatial appearance features of the head and body. The head branch also takes binary masks of the head bounding box as input and

acquires head position features through average pooling layers. In addition, to model the dynamic changing process, we exploit a velocity branch to model the 2D body velocity of the person in the video. Specifically, given the body bounding boxes of the person in the video, we compute the variation of body centers in the image coordinate system and obtain 2D body velocities, which are then fed into a fully connected layer to extract velocity features.

In unconstrained environments, it is challenging to acquire the entire human body states to accurately estimate gaze, due to occlusion, various illumination conditions, and the lack of clear face images. To address this issue, we incorporate human pose and take an additional branch to extract pose features. Specifically, a bottom-up pose estimator Openpifpaf [24] is employed to detect the 2D body joint coordinates of individuals in videos, which are used to form the skeletons in OpenPose [5] format. The human skeletons naturally present a graph structure, so we construct a spatial-temporal graph with joints as nodes and connections in both the human skeleton structures and frames as edges. Then, STGCN [35] is adopted to propagate information between neighboring nodes through edges and update the feature vectors on each node, thereby generating pose features. Finally, the outputs of the four branches are concatenated to comprehensively estimate the entire human body states in unconstrained environments.

Conditional Variational Autoencoder We find that human gaze behaviors exhibit significant uncertainty in unconstrained environments. Specifically, when people are engaged in similar body states, they may have entirely different gaze directions. Therefore, we employ a probabilistic approach CVAE [29] to learn a one-to-many mapping from image observations X to gazes Y through latent variables Z. It generates multi-modal latent results Z, which correspond to the multiple possible gaze patterns for similar body states. Thus, it can model the uncertainty of gaze behaviors and generate more robust gaze predictions.

As shown in the middle of Fig. 2a, our CVAE module consists of three submodules: an encoder to embed feature and gaze vectors, a prior network $p_\theta(Z|X)$ to model Z from X, and a recognition network $q_\phi(Z|X,Y)$ to capture dependencies between Z and Y, where θ and ϕ represent network parameters.

Firstly, concatenated features and ground truth gaze labels are separately encoded by an LSTM [16] layer to obtain the embedding of image observations h_x and gazes h_y. Then, h_x is fed into the prior network $p_\theta(Z|X)$ to get the mean μ_p and variance Σ_p of the normal distribution $\mathcal{N}(\mu_p, \Sigma_p)$, following the original VAE [22]. In addition, the recognition network $q_\phi(Z|X,Y)$ takes both h_x and h_y as input to predict the mean μ_q and variance Σ_q of the normal distribution $\mathcal{N}(\mu_q, \Sigma_q)$. Both the prior network and recognition network are composed of multi-layer perceptrons (MLPs). The prior and recognition networks project h_x and (h_x, h_y) to latent space respectively. We apply the Kullback-Leibler divergence (KLD) loss between $\mathcal{N}(\mu_p, \Sigma_p)$ and $\mathcal{N}(\mu_q, \Sigma_q)$ to align the two spaces.

$$\mathcal{L}_{\mathrm{KLD}} = \frac{1}{2}[(\mu_p - \mu_q)^\top {\Sigma_q}^{-1}(\mu_p - \mu_q) - \log\det({\Sigma_q}^{-1}\Sigma_p) + \mathrm{Tr}({\Sigma_q}^{-1}\Sigma_p) - k] \quad (1)$$

Here, k represents the dimension of the normal distributions. Then, prior network can generate multi-modal gaze-related latent variables h_z during inference.

Decoder Finally, the multi-modal gaze-related latent variables h_z and image observations h_x are decoded into robust gaze predictions through decoder $p_\omega(G|X, Z)$. As shown in the right side of Fig. 2a, the decoder consists of a Bi-LSTM module and an MLP. Firstly, the h_z is sampled from the normal distribution($\mathcal{N}(\mu_q, \Sigma_q)$ in training and $\mathcal{N}(\mu_p, \Sigma_p)$ in testing). Then, h_z and h_x are concatenated and fed into the decoder to produce the final gaze predictions.

We adopt cosine similarity loss, which is widely used when data is in the form of vectors, as the loss function for unconstrained gaze estimation.

$$\mathcal{L}_{\mathrm{C}} = 1 - \boldsymbol{x}^\top \boldsymbol{\mu} \tag{2}$$

Here, $\boldsymbol{x}$ indicates gaze estimation results, $\boldsymbol{\mu}$ indicates the ground truth gaze labels. Both of them are 3D vector in Cartesian coordinates. In summary, the loss function of our unconstrained gaze estimation method can be expressed as:

$$\mathcal{L}_{\mathrm{G}} = \lambda_1 \mathcal{L}_{\mathrm{KLD}} + \lambda_2 \mathcal{L}_{\mathrm{C}} \tag{3}$$

Here, λ_1, λ_2 are hyperparameters that weight each components in loss function.

3.2 Gaze-Guided Contrastive Domain Adaptation Framework

Preliminary: Few Shot Domain Adaptation For few-shot domain adaptation tasks, fully labeled source domain data $\mathcal{D}_{\mathcal{S}} = \{(I_n^{\mathcal{S}}, g_n^{\mathcal{S}})\}_{n=1}^{N_s}$ and a small amount of labeled $D_{\mathcal{T}} = \{(I_n^{\mathcal{T}}, g_n^{\mathcal{T}})\}_{n=1}^{N_t}$ target domain data are given. Here, $(I_n^{\mathcal{S}}, g_n^{\mathcal{S}})$ and $(I_n^{\mathcal{T}}, g_n^{\mathcal{T}})$ denote the n-th pair of image and gaze label in the source domain and target domain respectively. N_s and N_t denote the number of samples in the source and target domain. The goal of domain adaptation is to adapt the model $\mathcal{F}_\theta$ trained on the source domain $\mathcal{D}_{\mathcal{S}}$ to achieve minimum error on the unseen target domain $\mathcal{D}_{\mathcal{T}}$, with the help of a small subset of labeled target domain data. In the following, we introduce our GCDA framework for few-shot cross-domain gaze estimation.

Gaze-Guided Contrastive Learning Cross-domain gaze estimation is challenge in unconstrained environments, due to the large variation between the source and target domains, such as differences in scenes, human behaviors, and gaze distributions. Consequently, to achieve better performance, it is crucial to extract underlying gaze-related information and obtain stable feature representations that are not affected by domain shifts.

We find that gaze labels can reflect the relevance of their corresponding features. For a well-trained model, samples with similar gaze labels cluster together in latent space, while those with dissimilar labels are pushed far apart. To some extent, this characteristic diminishes when performing cross-domain gaze estimation without adaptation. Therefore, to maintain this characteristic and then

obtain stable gaze feature representations in the target domain, we propose a gaze-guided contrastive learning method to adjust the feature extractors.

Contrastive learning methods [6, 14] have demonstrated extraordinary performance in self-supervised classification tasks, which discover underlying patterns of data by comparing paired samples. However, our gaze-guided contrastive learning still faces two issues. SimCLR [6] performs data augmentation on each sample in a batch, only sample i and its augmented pair j are considered as positive, even if other samples k that share the same class with i are considered as negative. After processed by the loss function, positive pairs will stay close to each other, while negative pairs will be pushed far apart. This raises the first problem: for conventional contrastive learning methods in a self-supervised manner, samples with the same class may be embedded far apart. To tackle this issue, we adopt [21] to perform contrastive learning in a supervised manner, where samples have the same class as i will also be considered as positive.

Secondly, gaze is a continuously changing signal, making all gaze labels in datasets unique. Consequently, there are as many classes as the size of the dataset, which degrades supervised contrastive learning to self-supervised one. To address this, we propose to discretize the gaze direction space for each sample in a batch. Given a sample i and its gaze label g_i, samples whose cosine similarity with g_i is greater than 0.99 (an angular difference of 8°) are considered to belong to the same class as i, while the remaining are considered as another class.

In summary, our gaze-guided contrastive learning method is described as follows. Given a batch of N randomly sampled data, we apply data augmentation to each sample to obtain $2N$ paired data. Let $i \in \mathcal{I} = \{1...2N\}$ be the index of an arbitrary augmented sample, its positive pair set $P(i)$ can be defined as:

$$P(i) = \{j | \frac{g_i \cdot g_j}{\|g_i \cdot g_j\|} > 0.99, j \in \mathcal{I} \backslash \{i\}\} \tag{4}$$

Here, g_i and g_j are gaze labels of samples i and j. The augmented pair of sample i has the same gaze labels as g_i, thus it is also included in $P(i)$. Then, the loss function of gaze-guided contrastive learning for sample i can be defined as:

$$\mathcal{L}_{\text{CR}} = \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp(\text{sim}(z_i, z_p)/\tau)}{\sum_{k=1}^{2N} \mathbb{I}_{k \neq i} \exp(\text{sim}(z_i, z_k)/\tau)} \tag{5}$$

Here, z_i is the normalized feature embedding of sample i in latent space, $\text{sim}(.,.)$ is the cosine similarity function, τ is a temperature parameter that scales the similarities, and $\mathbb{I}_{k \neq i}$ is an indicator function that is 1 only if $k \neq i$.

Unlike [32], our method is designed for unconstrained adaptive gaze estimation rather than a constrained one. Thus, we perform contrastive learning in a supervised manner and modify the loss function to acquire more stable features.

Few-Shot Domain Adaptation Framework As shown in Fig. 2b, our GCDA framework applies gaze-guided contrastive learning to the head and body branches

Algorithm 1 Gaze-guided Contrastive Domain Adaptation

Input: A small amount of target domain data $\mathcal{D}_{\mathcal{T}} = \{(I_n^{\mathcal{T}}, g_n^{\mathcal{T}})\}_{n=1}^{N_t}$, contrastive projection head r, and gaze estimation model $\mathcal{F}_\theta$ pretrained on the source domain $\mathcal{D}_{\mathcal{S}}$. $\mathcal{F}_\theta$ consists of feature extractors f and gaze prediction modules h.
Output: Adapted network $\hat{\mathcal{F}}_\theta$

for Sampled minibatch $\{I_k^{\mathcal{T}}\}_{k=1}^{N}$ **do**
  for all $k \in \{1, ..., N\}$ **do**
    Randomly augment data: $\hat{x}_{2k-1} = \mathcal{A}(I_k^{\mathcal{T}})$ and $\hat{x}_{2k} = \mathcal{A}^{'}(I_k^{\mathcal{T}})$
    Obtain contrastive features: $\hat{z}_{2k-1} = r(f(\hat{x}_{2k-1}))$ and $\hat{z}_{2k} = r(f(\hat{x}_{2k}))$
    Obtain gaze predictions: $\hat{g}_{2k-1} = h(f(\hat{x}_{2k-1}))$ and $\hat{g}_{2k} = h(f(\hat{x}_{2k}))$
  end for
  for all $i \in \{1, ..., 2N\}$ **do**
    Compute positive pair set $P(i)$ of i.
    Compute gaze-guided contrastive loss $\mathcal{L}_{\text{CR}}$.
    Compute adaptation loss $\mathcal{L}_{\text{DA}}$.
  end for
  Update the parameters of networks $\mathcal{F}_\theta$ to minimize $\mathcal{L}_{\text{DA}}$.
end for

to adapt feature extractors f. As a result, inputs with similar gaze labels will be projected to cluster together in latent space, while samples with dissimilar labels will be pushed far apart, leading to stable gaze feature representations in the target domain. Our GCDA also employs a gaze estimation loss to the gaze prediction modules h to adjust the mapping between the latent space and gaze distribution. The overall domain adaptation loss function can be expressed as:

$$\mathcal{L}_{\text{DA}} = \mu_1 \mathcal{L}_{\text{G}} + \mu_2 \mathcal{L}_{\text{CR}} \tag{6}$$

Here, μ_1 and μ_2 are hyperparameters that weight each loss function.

Our GCDA framework adapts the gaze estimation model $\mathcal{F}_\theta$ pretrained on the source domain $\mathcal{D}_{\mathcal{S}}$ to the target domain $\mathcal{D}_{\mathcal{T}}$. The adaptation procedure is summarized in Algorithm 1. Firstly, we sample a minibatch $\{I_k^{\mathcal{T}}\}_{k=1}^{N}$ from the target domain $\mathcal{D}_{\mathcal{T}}$, and apply two separate stochastic data augmentations, $\mathcal{A}$ and $\mathcal{A}^{'}$, to transform any given sample $I_k^{\mathcal{T}}$. This results in two related views, denoted as $\hat{x}_{2k-1}$ and $\hat{x}_{2k}$. Then, following SimCLR [6], the augmented data $\hat{x}_{2k-1}$ and $\hat{x}_{2k}$ are fed into the feature extractors $f(.)$, followed by the projection heads $r(.)$, to obtain contrastive features $\hat{z}_{2k-1}$ and $\hat{z}_{2k}$ in latent space. In addition, $\hat{x}_{2k-1}$ and $\hat{x}_{2k}$ are fed into gaze prediction modules h to obtain gaze directions $\hat{g}_{2k-1}$ and $\hat{g}_{2k}$. This process results in $2N$ data points.

We apply domain adaptation loss to these $2N$ data points to obtain stable gaze feature representations and adapt the mapping to the gaze distribution. Specifically, for each sample i, we acquire its positive pair set $P(i)$ according to Eq. (4), and compute gaze estimation loss over gaze prediction $\hat{g}_i$ and gaze label g_i according to Eq. (3). Then, contrastive loss $\mathcal{L}_{\text{CR}}$ is calculated over contrastive features $\hat{Z} = \{\hat{z}_i\}_{i=1}^{2N}$ with Eq. (5). Finally, the adaptation loss $\mathcal{L}_{\text{DA}}$ is obtained with Eq. (6). We average the $\mathcal{L}_{\text{DA}}$ of the $2N$ data points and update the pa-

rameters of $\mathcal{F}_\theta$ to minimize $\mathcal{L}_{\mathrm{DA}}$. In summary, our UAGE method integrates the unconstrained gaze estimation method with the GCDA framework, achieving adaptive gaze estimation in unconstrained environments.

4 Experiment

4.1 Experiment Setup

Dataset The **GAFA** [26] dataset is collected by multiple cameras from a surveillance view in five unconstrained scenes(i.e. office, kitchen, laboratory, living room and courtyard). It contains freely moving people, diverse gaze behaviors, and a wide range of challenging head poses. The annotations consist of 3D gaze directions in the camera coordinate system, head and body orientations, head and body bounding boxes, as well as body keypoints detected by OpenPose [5]. We use 710K frames for training, 79K for validation and 93K for testing.

The **Gaze360** [20] dataset is captured by a panoramic camera in unconstrained environments. It contains a gaze distribution that covers the entire horizontal range of 360°. The annotations in the Gaze360 dataset consist of 3D gaze directions in the eye coordinate system, 3D eye and gaze target positions in the panoramic camera system, and head bounding boxes. We use 129K frames for training, 17K for validation and 26K for testing.

Images in GAFA typically have low resolution and are taken from a back or overhead view, making the eyes of people rarely visible, while images in Gaze360 usually have high resolution and are taken from a front view.

Models for comparison For within-domain gaze estimation, we experimentally compare our UAGE method with several representative unconstrained gaze estimation methods. **Random** and **Fixed bias** serve as baselines, representing the lower bound for testing results. **Random** arbitrarily samples 3D gaze directions from a normal distribution and takes them as estimation results, while **Fixed bias** takes the mean direction of the training set as its prediction results. **Dias et al.** [10] utilize OpenPose [5] to detect human facial keypoints and estimate 2D gaze directions from pose features. **Gaze360** [20] takes a sequence of 7 head image frames as input, considers the error bounds of gaze estimation, and outputs 3D gaze directions. **X-Gaze** [39] takes high-resolution facial images as input and outputs 3D results. **GAFA** [26] takes a sequence of 7 frames as input, which are composed of whole body images, head position masks, and body velocities. It predicts intermediate head and body orientations and leverages the intrinsic gaze, head and body coordination to derive the final 3D gaze directions.

For cross-domain gaze estimation, there are no well-elaborated methods specifically designed for unconstrained environments, making direct comparisons with other methods difficult. Therefore, we first compare our method with GAFA by directly testing on the target domain. Secondly, given a small amount of labeled target domain data, we compare our UAGE method with direct fine-tuning. The evaluation metric for all experiments is the 3D/2D angular error.

Table 1: Results on GAFA [26] dataset comparing with unconstrained gaze estimation methods. The table below reports the 3D/2D mean angular errors(MAE) in five scenes. The last three columns indicate the mean MAE for all scenes in front, back and all 360° views. Our UAGE achieves comparable improvement against the baseline models.

Method	Office	LR	Kitchen	Library	Courtyard	Front180°	Back180°	Mean
Random	89.5/89.9	90.6/90.5	90.5/90.2	90.1/90.1	90.3/90.1	90.3/90.1	90.1/90.2	90.2/90.2
Fixed bias	88.0/76.0	85.5/76.7	86.0/82.4	89.0/85.1	89.7/87.8	86.3/99.4	90.3/55.0	88.1/79.7
Dias et al. [10]	-/27.2	-/25.2	-/19.8	-/24.9	-/36.1	-/22.89	-/34.8	-/27.1
Gaze360 [20]	24.0/19.2	41.1/31.3	32.4/21.2	27.5/20.7	28.2/28.3	21.8/19.6	36.3/26.7	30.4/24.5
X-Gaze [39]	24.2/23.0	42.0/40.9	23.3/22.9	24.6/22.3	30.2/31.9	26.2/23.5	31.5/31.7	29.2/28.4
GAFA [26]	**14.4**/14.3	25.1/22.6	20.4/19.6	19.8/18.4	25.4/26.9	20.7/17.4	**23.2**/21.9	21.7/20.9
UAGE	15.3/**14.0**	**23.5**/**21.5**	**18.1**/**17.0**	**18.7**/**16.6**	**23.8**/**25.8**	**18.8**/**15.7**	23.7/**21.4**	**20.5**/**19.4**

4.2 Performance Evaluation and Analysis

To verify the effectiveness of our UAGE method for within-domain gaze estimation in unconstrained environments, we train and test our method on the GAFA dataset. In addition, we evaluate our UAGE method with GAFA as the source domain and Gaze360 as the target domain to demonstrate its validity for cross-domain gaze estimation in unconstrained environments. Finally, we conduct ablation experiments on UAGE to verify the significance of each component.

Within-Domain Gaze Estimation Performance The quantitative results of gaze estimation in unconstrained environments are shown in Tab. 1. Our UAGE method outperforms all comparison methods. The method by Dias et al. [10] lacks 3D results because it's designed for 2D gaze estimation, and its performance is significantly affected by the low accuracy of detected facial keypoints on the GAFA dataset. Both Gaze360 [20] and X-Gaze [39] take only head images as input, which limits their robustness on the unconstrained GAFA dataset, where the eyes of people are rarely visible. GAFA [26] leverages the intrinsic gaze, head, and body coordination to predict gaze, achieving competitive results. However, their method additionally requires annotated head and body orientations, which are difficult to obtain in practice and in other datasets.

Our UAGE method considers pose features and estimates the entire human body states in the environment with concatenated features. We also take into account the uncertainty of gaze estimation in unconstrained environments and utilize a CVAE block to generate multi-modal results in latent space, which are decoded into robust gaze estimations. Supervised solely by gaze labels, our UAGE method achieves the best performance in terms of mean MAE across all scenes. Notably, our method shows significant improvements in front views.

Cross-Domain Gaze Estimation Performance Tab. 2 shows the MAE results of cross-domain gaze estimation, with GAFA as the source domain and Gaze360 as the target domain. Here, "UAGE w/o GCDA" denotes our unconstrained gaze estimation method without the adaptation framework. Firstly, our

Table 2: Cross-domain gaze estimation results on Gaze360 dataset. The table below reports the mean 3D/2D MAE for all scenes in front, back and all 360° views. The first two rows show the direct cross-dataset evaluation results. The last two rows report the cross-domain results after adaptation by direct fine-tuning and our UAGE.

Method	Front180°	Back180°	Mean
GAFA [26]	86.4/86.4	65.5/78.7	81.8/84.6
UAGE w/o GCDA	58.5/70.9	54.7/61.4	57.7/68.9
Direct fine-tune	20.5/31.6	41.2/38.2	25.0/33.1
UAGE	**17.8**/**25.7**	**34.5**/**30.0**	**21.5**/**26.7**

Table 3: Ablation study results on (a) within-domain gaze estimation and (b) cross-domain gaze estimation in unconstrained environments. The tables shows the mean 3D/2D MAE for all scenes in front, back and all 360° views. Each row reports the results of the model with the removal of one component.

Method	Front180°	Back180°	Mean
No Pose	19.1/15.7	23.5/21.2	20.6/**19.2**
No CVAE	18.9/16.0	**23.3**/**21.0**	**20.4**/19.5
No Head	20.1/17.2	24.2/21.8	21.6/20.7
No Body	20.1/16.8	23.9/23.1	21.4/21.1
UAGE	**18.8**/**15.7**	23.7/21.4	20.5/19.4

(a) Within-Domain Gaze Estimation

Method	Front180°	Back180°	Mean
No Pose	18.9/28.3	35.4/31.4	22.4/28.9
No CVAE	18.8/26.8	40.3/29.2	23.5/27.3
No Head	29.4/43.9	50.0/40.9	33.9/43.2
No Body	21.0/30.1	57.0/36.1	28.8/31.4
UAGE	**17.8**/**25.7**	**34.5**/**30.0**	**21.5**/**26.7**

(b) Cross-Domain Gaze Estimation

method outperforms GAFA by a large margin in cross-dataset evaluation, which directly inferring on the Gaze360 testing set with models pretrained on the GAFA training set. This demonstrates the generalization ability of our method.

In addition, compared with direct fine-tuning, our UAGE achieves a substantial improvement of 14%, indicating that our method has discovered more effective gaze feature representations through gaze-guided contrastive learning, thereby enhancing generalization with only a small amount of labeled data from the target domain. The GAFA dataset contains low-resolution images captured from a surveillance view with freely moving people, while the Gaze360 dataset consists of high-resolution images captured from a front view with stationary people. Due to the significant differences in gaze distributions and environments, conventional zero-shot methods struggle to perform well for unconstrained cross-domain gaze estimation. Therefore, we opt to perform domain adaptation in a few-shot manner. Specifically, both direct fine-tuning and UAGE utilize 10% labeled data of the Gaze360 training set to perform few-shot domain adaptation.

Ablation Study We conduct ablation studies on within-domain gaze estimation and cross-domain gaze estimation, respectively. The 3D and 2D MAE results are shown in Tab. 3. "No Pose" removes the pose branch and exploits the remaining features to predict gaze. It shows a slight performance decrease in within-domain evaluation, since the human pose is obtained by the exist-

ing pose estimator [24], and the accuracy of keypoints detection decreases in unconstrained environments, making the extracted pose features less effective. However, in cross-domain evaluation, model without pose features experiences a more significant performance decline, demonstrating the importance of pose features in estimating the entire human body states, which enhances the generalization ability of model.

In "No CVAE" , the concatenate features are directly fed into the decoder to predict gazes, leading to degraded performance in the cross-domain task. Our CVAE module considers the uncertainty of gaze estimation, recognizing that people in similar states may have different gaze behaviors. It prevents the model from overfitting and improves generalization in domain adaptation. However, in within-domain evaluation, this characteristic is less pronounced because the distributions of the training and testing sets are usually made different, resulting in a slight performance drop.

"No Head" and "No Body" remove the head branch and body branch respectively, both of them cause significant performance drops. This indicates that it's important to estimate gaze from the entire body states in unconstrained environments. Through ablation studies, we demonstrate the effectiveness of each component in our UAGE method.

4.3 Qualitative Result

We conduct qualitative experiments for within-domain and cross-domain gaze estimation in unconstrained environments. For within-domain gaze estimation, we perform inference on the GAFA testing set. The results are shown in the left three columns of Fig. 3. The first row shows gaze estimation results in front views. When the eyes are clearly visible, the angular error is very small, while the error increases but remains below 5° when the human face is less visible. The second row presents predictions in back views. Evidently, there is an increase in angular error due to the face being rarely visible. The last row presents gaze estimation results under occlusion, where our method still achieves relatively accurate results, with an angular error of approximately 25° even in severely occluded situations. Overall, despite various scenes, human behaviors and shooting views, our method still performs well in unconstrained environments.

For cross-domain gaze estimation, we perform inference on the Gaze360 testing set. The results are shown in the last column of Fig. 3. Despite the significant differences between the two datasets, our UAGE method still achieves excellent results when adapted with 10% Gaze360 training data. More experiments can be referred to the **supplementary materials**.

Future Work The disparity in gaze direction distribution caused by camera position is also one of the fundamental issues in unconstrained adaptive gaze estimation. However, it's difficult to obtain camera position in most existing datasets. We will explore the virtual camera calibration and gaze redirection to align disparity in the future. In addition, we perform domain adaptation

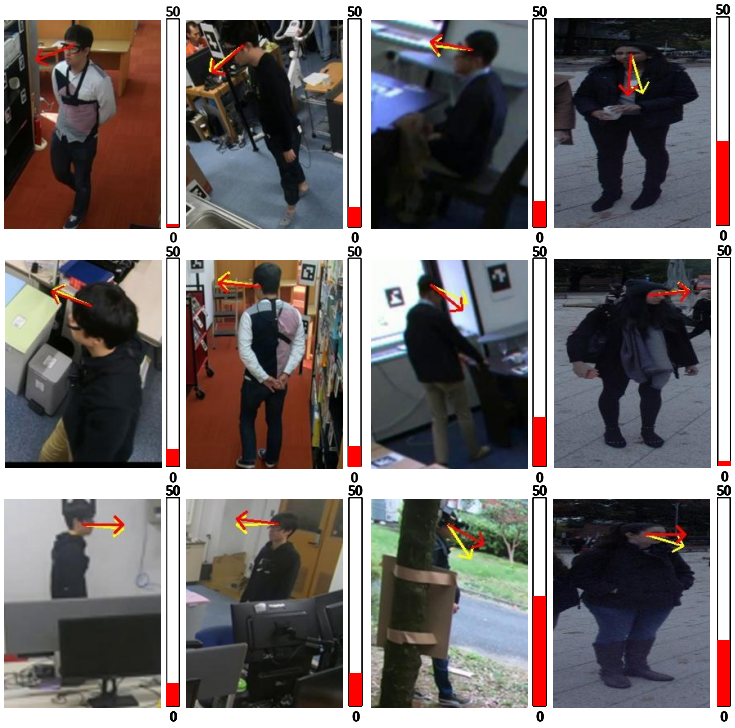


Fig. 3: Qualitative results. The last column is results on the Gaze360 testing set, the other columns are results on the GAFA testing set. Red arrows represent the ground truth gaze labels, yellow arrows represent the predicted gaze directions, and the bars on the right indicates the magnitude of angular errors.

in a few-shot manner because of the large variation between the unconstrained source and target domains. More challenging zero-shot domain adaption is worth exploring for gaze estimation in unconstrained environments. Finally, collecting a large-scale, diverse, and unconstrained gaze estimation dataset can significantly advance the application of gaze estimation methods in real-world environments.

5 Conclusion

In this paper, we introduce a novel UAGE method for adaptive gaze estimation in unconstrained environments. Our UAGE leverages the entire human body states and the uncertainty of gaze behaviors to estimate gaze directions. In addition, our UAGE can adapt to new domains with GCDA framework, which aligns the features in latent space and adjusts the mapping from the latent space to gaze distribution. Experiment results demonstrate that our UAGE has achieved state-of-the-art performance in within-domain evaluation and has reduced the angular error by 14% compared to the baseline in cross-domain evaluation.

References

1. Bao, Y., Liu, Y., Wang, H., Lu, F.: Generalizing gaze estimation with rotation consistency. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 4207–4216 (2022)
2. Bao, Y., Lu, F.: Pcfgaze: Physics-consistent feature for appearance-based gaze estimation. arXiv preprint arXiv:2309.02165 (2023)
3. Burova, A., Mäkelä, J., Hakulinen, J., Keskinen, T., Heinonen, H., Siltanen, S., Turunen, M.: Utilizing vr and gaze tracking to develop ar solutions for industrial maintenance. In: Proceedings of the 2020 CHI conference on human factors in computing systems. pp. 1–13 (2020)
4. Cai, X., Zeng, J., Shan, S., Chen, X.: Source-free adaptive gaze estimation by uncertainty reduction. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 22035–22045 (2023)
5. Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2d pose estimation using part affinity fields. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 7291–7299 (2017)
6. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: Int. Conf. Mach. Learn. pp. 1597–1607. PMLR (2020)
7. Cheng, Y., Bao, Y., Lu, F.: Puregaze: Purifying gaze feature for generalizable gaze estimation. In: AAAI. pp. 436–443 (2022)
8. Chong, E., Wang, Y., Ruiz, N., Rehg, J.M.: Detecting attended visual targets in video. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 5396–5406 (2020)
9. Deng, T., Yang, K., Li, Y., Yan, H.: Where does the driver look? top-down-based saliency detection in a traffic driving environment. IEEE Transactions on Intelligent Transportation Systems **17**(7), 2051–2062 (2016)
10. Dias, P.A., Malafronte, D., Medeiros, H., Odone, F.: Gaze estimation for assisted living environments. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 290–299 (2020)
11. Doersch, C.: Tutorial on variational autoencoders. arXiv preprint arXiv:1606.05908 (2016)
12. Fischer, T., Chang, H.J., Demiris, Y.: Rt-gene: Real-time eye gaze estimation in natural environments. In: Eur. Conf. Comput. Vis. pp. 334–352 (2018)
13. Guo, Z., Yuan, Z., Zhang, C., Chi, W., Ling, Y., Zhang, S.: Domain adaptation gaze estimation by embedding with prediction consistency. In: Asian Conf. Comput. Vis. (2020)
14. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 9729–9738 (2020)
15. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 770–778 (2016)
16. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8), 1735–1780 (1997)
17. Hu, Z., Yang, Y., Zhai, X., Yang, D., Zhou, B., Liu, J.: Gfie: A dataset and baseline for gaze-following from 2d to 3d in indoor environments. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 8907–8916 (2023)
18. Huang, Q., Veeraraghavan, A., Sabharwal, A.: Tabletgaze: dataset and analysis for unconstrained appearance-based gaze estimation in mobile tablets. Machine Vision and Applications **28**, 445–461 (2017)

19. Jin, T., Yu, Q., Zhu, S., Lin, Z., Ren, J., Zhou, Y., Song, W.: Depth-aware gaze-following via auxiliary networks for robotics. Engineering Applications of Artificial Intelligence **113**, 104924 (2022)
20. Kellnhofer, P., Recasens, A., Stent, S., Matusik, W., Torralba, A.: Gaze360: Physically unconstrained gaze estimation in the wild. In: Int. Conf. Comput. Vis. pp. 6912–6921 (2019)
21. Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D.: Supervised contrastive learning. Adv. Neural Inform. Process. Syst. **33**, 18661–18673 (2020)
22. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
23. Konrad, R., Angelopoulos, A., Wetzstein, G.: Gaze-contingent ocular parallax rendering for virtual reality. ACM Trans. Graph. **39**(2), 1–12 (2020)
24. Kreiss, S., Bertoni, L., Alahi, A.: Pifpaf: Composite fields for human pose estimation. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 11977–11986 (2019)
25. Liu, Y., Liu, R., Wang, H., Lu, F.: Generalizing gaze estimation with outlier-guided collaborative adaptation. In: Int. Conf. Comput. Vis. pp. 3835–3844 (2021)
26. Nonaka, S., Nobuhara, S., Nishino, K.: Dynamic 3d gaze from afar: Deep gaze estimation from temporal eye-head-body coordination. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 2192–2201 (2022)
27. Park, S., Mello, S.D., Molchanov, P., Iqbal, U., Hilliges, O., Kautz, J.: Few-shot adaptive gaze estimation. In: Int. Conf. Comput. Vis. pp. 9368–9377 (2019)
28. Park, S., Spurr, A., Hilliges, O.: Deep pictorial gaze estimation. In: Eur. Conf. Comput. Vis. pp. 721–738 (2018)
29. Sohn, K., Lee, H., Yan, X.: Learning structured output representation using deep conditional generative models. Adv. Neural Inform. Process. Syst. **28** (2015)
30. Sugano, Y., Matsushita, Y., Sato, Y.: Learning-by-synthesis for appearance-based 3d gaze estimation. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 1821–1828 (2014)
31. Valenti, R., Sebe, N., Gevers, T.: Combining head pose and eye location information for gaze estimation. IEEE Trans. Image Process. **21**(2), 802–815 (2011)
32. Wang, Y., Jiang, Y., Li, J., Ni, B., Dai, W., Li, C., Xiong, H., Li, T.: Contrastive regression for domain adaptation on gaze estimation. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 19376–19385 (2022)
33. Wei, P., Liu, Y., Shu, T., Zheng, N., Zhu, S.C.: Where and why are they looking? jointly inferring human attention and intentions in complex tasks. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 6801–6809 (2018)
34. Xu, P., Ehinger, K.A., Zhang, Y., Finkelstein, A., Kulkarni, S.R., Xiao, J.: Turkergaze: Crowdsourcing saliency with webcam based eye tracking. arXiv preprint arXiv:1504.06755 (2015)
35. Yan, S., Xiong, Y., Lin, D.: Spatial temporal graph convolutional networks for skeleton-based action recognition. In: AAAI. vol. 32 (2018)
36. Yu, Y., Odobez, J.M.: Unsupervised representation learning for gaze estimation. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 7314–7324 (2020)
37. Yu, Z., Yoon, J.S., Lee, I.K., Venkatesh, P., Park, J., Yu, J., Park, H.S.: Humbi: A large multiview dataset of human body expressions. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 2990–3000 (2020)
38. Zhang, Q., Hu, Z., Song, Y., Pei, J., Liu, J.: The human gaze helps robots run bravely and efficiently in crowds. In: 2023 IEEE International Conference on Robotics and Automation (ICRA). pp. 7540–7546. IEEE (2023)

39. Zhang, X., Park, S., Beeler, T., Bradley, D., Tang, S., Hilliges, O.: Eth-xgaze: A large scale dataset for gaze estimation under extreme head pose and gaze variation. In: Eur. Conf. Comput. Vis. pp. 365–381. Springer (2020)
40. Zhang, X., Sugano, Y., Fritz, M., Bulling, A.: Appearance-based gaze estimation in the wild. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 4511–4520 (2015)
41. Zhang, X., Sugano, Y., Fritz, M., Bulling, A.: Mpiigaze: Real-world dataset and deep appearance-based gaze estimation. IEEE Trans. Pattern Anal. Mach. Intell. **41**(1), 162–175 (2017)
42. Zhao, Z., Wang, S., Wang, Q., Shen, D.: Mining gaze for contrastive learning toward computer-assisted diagnosis. In: AAAI. pp. 7543–7551 (2024)
43. Zhu, Z., Ji, Q.: Eye gaze tracking under natural head movements. In: IEEE Conf. Comput. Vis. Pattern Recog. vol. 1, pp. 918–923. IEEE (2005)