Generalizing Gaze Estimation with Rotation Consistency

Yiwei Bao[1] Yunfei Liu[1] Haofei Wang[2] Feng Lu[1, 2 *]
[1] State Key Laboratory of VR Technology and Systems, School of CSE, Beihang University
[2] Peng Cheng Laboratory, Shenzhen, China
{baoyiwei, lyunfei, lufeng}@buaa.edu.cn wanghf@pcl.ac.cn

Abstract

Recent advances of deep learning-based approaches have achieved remarkable performance on appearance-based gaze estimation. However, due to the shortage of target domain data and absence of target labels, generalizing gaze estimation algorithm to unseen environments is still challenging. In this paper, we discover the rotation-consistency property in gaze estimation and introduce the 'sub-label' for unsupervised domain adaptation. Consequently, we propose the Rotation-enhanced Unsupervised Domain Adaptation (RUDA) for gaze estimation. First, we rotate the original images with different angles for training. Then we conduct domain adaptation under the constraint of rotation consistency. The target domain images are assigned with sub-labels, derived from relative rotation angles rather than untouchable real labels. With such sub-labels, we propose a novel distribution loss that facilitates the domain adaptation. We evaluate the RUDA framework on four cross-domain gaze estimation tasks. Experimental results demonstrate that it improves the performance over the baselines with gains ranging from 12.2% *to* 30.5%*. Our framework has the potential to be used in other computer vision tasks with physical constraints.*

1. Introduction

Gaze is one of the most important cues for human intention prediction. It has been used in a variety of applications such as virtual/augmented reality [21, 30], human-computer interaction [18, 35, 37], and medical analysis [3, 20]. To obtain accurate gaze estimations, various systems have been developed. Appearance-based gaze estimation is one of the most promising approaches, since it has the lowest hardware requirements.

With the advancement of deep learning techniques, Convolutional Neural Networks (CNN) have achieved significant performance improvement in many computer vision tasks. Gaze estimation task is no exception, various CNN-based gaze estimation methods have been proposed over the last decades [8]. These systems usually have different inputs: eye images [9, 24, 29, 39, 42], face images [19, 22, 43] or both face/eye images [1, 7, 23]. However, existing methods suffer from severe performance degradation when adapting to new domains, which is mainly caused by the difference between the domains, e.g., subject appearance, image quality, shooting angle and illumination.

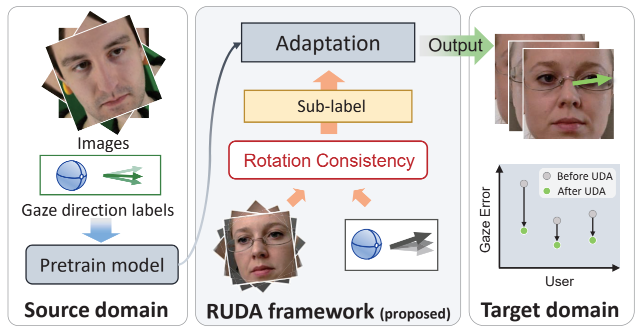

Figure 1. The overall structure of the proposed rotation-enhanced unsupervised domain adaptation (RUDA) framework for gaze estimation. RUDA adapts the pre-trained model to target domain without requiring any gaze labels in target domain.

One of the major challenge of gaze domain adaptation is that we usually do not have access to target domain labels in real world scenarios, and we cannot directly train the gaze estimator in target domain. To address this problem, unsupervised domain adaptation approaches aim to find a gaze-relevant constraint generalizing the model to target domain without label. Kellnhofer *et al*. propose to supervise gaze estimation model with an domain discriminator by adversarial learning [19]. Similarly, Wang *et al*. employ an appearance discriminator and a head pose classifier for adaptation [39]. More recently, Liu *et al*. propose to guide the model with outliers [25]. Although some unsupervised domain adaptation approaches for gaze estimation have been proposed, it is still a challenging task.

*Corresponding Author. This work was supported by the National Natural Science Foundation of China (NSFC) under Grant 61972012.

To build a gaze-relevant constraint to supervise the model without requiring ground truth labels, we dive into the physical nature of gaze. We find that the human gaze, as a 3D direction vector, is rotation-consistent. Rotating the face image results in the same rotation angle of the gaze direction, we call this rotation-consistency property. And we define the relative rotation angle as the *sub-label*, meaning that it is not an absolute angle, but the relative difference angle before and after the rotation. This rotation consistency could serve as the desired gaze-relevant constraint without ground truth. Although training with rotated images in source domain does not improve gaze estimation accuracy because user faces are already aligned by normalization [42], we argue that the rotation consistency property provides a gaze-relevant optimization target for adaptation.

In light of this, we present the Rotation-enhanced Unsupervised Domain Adaptation (RUDA) framework for gaze estimation. Our approach creates sub-labels between original and randomly rotated images. The estimator is generalized to target domain via rotation consistency of estimation results with no target domain label required and low computation cost. The contributions of this work are as follow:

- We propose the Rotation-enhanced Unsupervised Domain Adaptation (RUDA) framework for gaze estimation. The RUDA first trains a rotation-augmented model in source domain, then adapts the model to target domain using the synthesized images with physically-constrained gaze directions.
- We found the rotation consistency property, which can be used to generate *sub-labels* for unsupervised gaze adaptation tasks. To facilitate adaptation, we design a novel distribution loss which supervise the model with rotation consistency and sub-labels.
- Experimental results demonstrate that the RUDA framework achieves consistent improvement over the baseline model on four cross-domain gaze estimation tasks, ranging from 12.2% to 30.5%. It achieves surprisingly good results, even outperforms some state-of-the-art methods trained on target domain with labels.

2. Related work

Gaze Estimation. Early studies estimate gaze by reconstructing a 3D eyeball model and calculate gaze from the anatomical eye structure. These methods usually offer accurate gaze estimates, while they require personal calibration and dedicated devices such as depth camera [34,38,40], infrared camera [28] and infrared lights [15].

Calibration-free appearance-based gaze estimation with single web camera received favor of researchers in the last decades. In 2015, Zhang *et al*. first propose to estimate gaze from eye images using CNN [42]. Following this work, a number of gaze estimation dataset have been released [10, 19, 23, 31, 33, 41, 43]. Based on them, various deep learning-based approaches using different inputs have been proposed: using eye images [9, 24, 29, 39, 42], using face images [19, 22, 43] or using both [1, 7, 23].

More recently, cross domain gaze estimation task attracted more and more attention. Park *et al*. proposed to learn a person-specific gaze estimation network with few samples by meta-learning [29]. Guo *et al*. eliminated the inter-personal diversity by ensuring prediction consistency [16]. Cheng *et al*. proposed to improve cross dataset accuracy without target domain data by eliminating gaze-irrelevant feature [6]. Liu *et al*. [25] proposed a plug-and-play cross-domain gaze estimation framework with the guidance of outliers. Although it significantly outperforms the existing methods, their method requires as many as 20 models for collaborative learning. Zheng *et al*. [45] propose to redirect head and gaze in a self-supervised manner by embedding transformation including rotation, which helps down stream tasks like gaze estimation. In other tasks like 3D hand pose estimation, rotation has also been used as a constrain for self-supervised learning [32].

Unsupervised Domain Adaptation. Unsupervised domain adaptation (UDA) is one of the common tasks in computer vision, which has been extensively studied for a long time. Early UDA methods use geodesic distance as the subspace distance to learn domain-invariant representations [12, 14]. Inspired by this, some researchers proposed to reduce domain gap by matching the statistics of source and target domain [2, 26]. Chen *et al*. propose a representation subspace distance (RSD) that aligns features from two domains specifically for regression tasks [4].

Inspired by the generative adversarial net [13], adversarial learning have been adopted for UDA tasks. For example, a min-max game between feature extractor and domain discriminator is built to close the domain gap [27, 36, 44].

Although the above-mentioned methods achieve considerable improvement, most of them are designed for classification tasks, instead of regression task. The RSD proposed by Chen *et al*. [4] is specifically designed for regression tasks, however, we found that their approach dose not perform well on gaze estimation task. Therefore, the UDA for gaze estimation still remains to be explored.

3. Rotation Consistency in Gaze Estimation

There are two main challenges in unsupervised gaze adaptation tasks: 1) the shortage of target domain samples for adaptation, and 2) the absence of ground-truth labels in target domain. Various data augmentation approaches have been proposed to generate training data in source domain, e.g., color jittering, introducing noise, flipping, translation and rotation. However, existing data augmentation

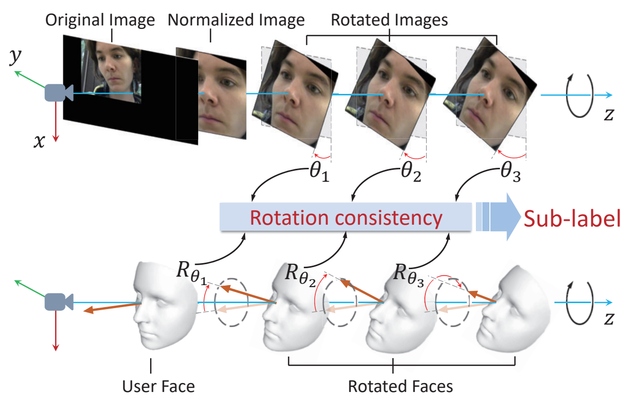

Figure 2. Illustration of the rotation consistency property in gaze estimation. When we rotate the face image with an angle of θ, the gaze direction is rotated with $\mathbf{R}_\theta$ correspondingly, where $\mathbf{R}_\theta$ is the 3D rotation matrix with angle θ.

approaches only bring limited performance improvement if directly adopted in unsupervised gaze adaptation tasks.

To cope with the absence of ground-truth labels in UDA tasks, we define a *sub-label*, which is a relative angle and can be used as constraints in gaze adaptation tasks. Due to the nature of the gaze estimation task, here we rotate the raw images with different angles to synthesize more images. Since the ground-truth labels are usually absent, we rotate the image with different angles and assign a sub-label to each image based on the rotation consistency property.

Note that the sub-label is not an absolute angle, instead, it is a relative angle between the original image and the rotated image. For example, the gaze direction is $\mathbf{g}$ in the original image, we rotate the image with angles of θ_1, θ_2, and θ_3, the sub-label of the rotated image is $\mathbf{R}_{\theta_1}$, $\mathbf{R}_{\theta_2}$ and $\mathbf{R}_{\theta_3}$, respectively. The core idea of rotation-consistency can be summarized in Eq. (1):

$$(\mathbf{R}^{\mathbf{g}})^{-1} * (F(\mathbf{R}I)) = F(I), \tag{1}$$

where I is the input face image, F is the gaze mapping function from the image to gaze direction, $\mathbf{R}$ is the rotation matrix of the input image, and $\mathbf{R}^{\mathbf{g}}$ is the rotation matrix of the gaze direction. $\mathbf{R}I$ represents the rotated image, and $F(\mathbf{R}I)$ indicates the estimated gaze direction of the rotated image. In practice, image pixels cannot be rotated by rotation matrix directly. We formulate them in this way for simplicity. The rotation consistency formula suggests that ideally, the rotation angle of the image is equal to the rotation angle of the estimated gaze.

Why rotation consistency? Rotation is a commonly-used data augmentation approach in computer vision. However, in gaze estimation tasks, training with rotated images brings little performance gain in both within- and cross-dataset tasks. In fact, it is more often used for data normalization: by rotating and scaling the virtual camera, the user's face is changed to the same size and location, while the x-axis of the camera coordinate system and the user head coordinate system are aligned and the z-axis of the camera coordinate system is perpendicular to the image plane. As a result, rotation operations help the camera look at different faces in a unified way (top-left in Fig. 2).

On the other hand, our proposed rotation consistency-based strategy plays a different role. It aims at solving the shortage of target domain data and absence of target label problem in cross-domain gaze estimation, and it indeed boosts the performance. Fig. 2 illustrates the idea of rotation consistency. It bridges the relative rotation angles between the image and the 3D gaze. In this way, for unsupervised domain adaptation, although the real gaze directions are unknown, the relative rotation angles can serve as the *sub-labels* to train the network. In addition, we can generate as many target images as we want with different sub-labels if we rotate the image with different angles.

Conversion between the image and gaze rotation angle. Given a normalized image I, we use the center of image as rotation center O, and rotate the image with θ (clockwise), the rotation matrix $\mathbf{R}$ can be defined as follows:

$$\mathbf{R} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}. \tag{2}$$

For each pixel location $I_i \in I$, the rotated pixel location is $\mathbf{R}I_i^T$. Gaze is a 3D direction vector $\mathbf{g}$ defined in camera coordinate system. Therefore, the corresponding rotation matrix for gaze direction is

$$\mathbf{R}^{\mathbf{g}} = \begin{bmatrix} \mathbf{R} & 0 \\ 0 & 1 \end{bmatrix}. \tag{3}$$

As a result, the rotated gaze direction is $\mathbf{R}^{\mathbf{g}}\mathbf{g}^T$. In actual training, the gaze direction is denoted as a 2D Euler angles $\mathbf{g} = [y, p]$, where y is the yaw angle and p is the pitch angle. Thus, conversion between 2D Euler angles and 3D direction vector is needed before and after rotation.

4. Method

4.1. Task Definition

For UDA tasks, we are given a fully labeled source domain and a small amount of unlabeled samples from target domain. Let $\mathcal{D}_s = \{I_i^s, \mathbf{g}_i^s\}_{i=1}^{N_s}$ represents N_s images with gaze label $\mathbf{g}^s$ in source domain, and $\mathcal{D}_t = \{I_i^t\}_{i=1}^{N_t}$ represents N_t images without gaze labels in target domain. Our goal is to generalize a gaze estimation network F_θ with parameter θ that performs well in $\mathcal{D}_t$. Only a small subset of unlabeled target domain samples $\mathcal{D}'_t$ is used for adaptation. Before that, F_θ is pretrained on $\mathcal{D}_s$. In the following, we will introduce the details of our proposed method.

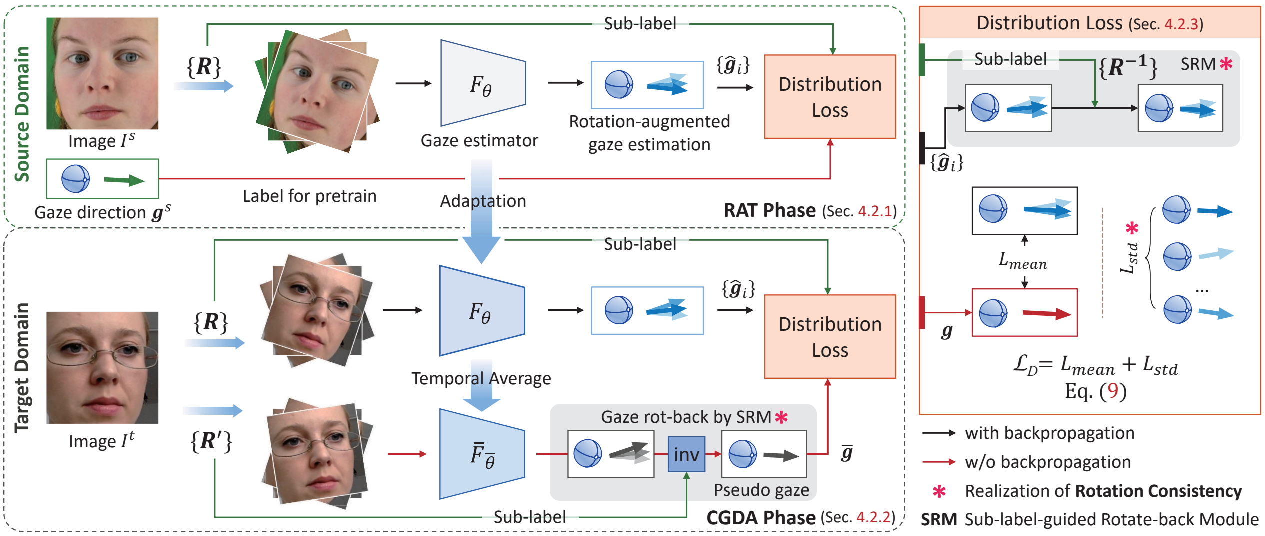

Figure 3. Overview of the proposed RUDA framework, which consists of two phases: 1) the Rotation-Augmented Training (RAT) phase and 2) the Consistency-Guided Domain Adaptation (CGDA) phase. We first train a rotation-augmented model to predict gaze on rotated images. Then, based on rotation consistency, sub-labels generated by image rotation are used to guide the SRM module and calculate the proposed Distribution Loss for unsupervised domain adaptation.

4.2. Rotation-Enhanced Unsupervised Domain Adaptation for Gaze Estimation

Fig. 3 shows the overview of the proposed RUDA framework, which consists of two steps: 1) the Rotation-Augmented Training (RAT) phase, and 2) the Consistency-Guided Domain Adaptation (CGDA) phase. To estimate gaze from rotated images, we train a model F_θ with rotated images in source domain in RAT phase. We adapt F_θ to the target domain with the guidance of sub-label (produced by rotation) and pseudo label (produced by the temporal average model) in CGDA phase. We further propose a distribution loss ($\mathcal{L}_D$) to supervise the model with mean value (gaze label or pseudo label) and standard deviation (rotation consistency and sub-label) in both RAT and CGDA phase.

4.2.1 Rotation-Augmented Training

In order to adapt the model to target domain guided by rotation consistency property, the model should be able to estimate the gaze on rotated images. Thus, in RAT, we train a rotation augmented model in source domain.

Gaze estimators are usually trained with labeled source domain data $\{I_i^s, \mathbf{g}_i^s\}$ with $\mathcal{L}_1$ loss function:

$$\arg\min_\theta \mathcal{L}_1(\hat{\mathbf{g}}_i^s, \mathbf{g}_i^s), \tag{4}$$

where $\hat{\mathbf{g}}_i^s = F_\theta(I_i^s)$ is the estimation result. To predict gaze from rotated images, we also train the model F_θ with rotation augmented source domain samples. For each image I^s in the training set of source domain, we randomly rotate it K times to obtain a new set $\mathcal{I}^s$:

$$\mathcal{I}^s = \{\mathbf{R}I^s | k = 1, 2, \ldots, K\}. \tag{5}$$

Here we record the rotation matrix $\mathbf{R}$ as the sub-label for the rotated image set $\mathcal{I}^s$. A group of estimation results of rotated images $\mathcal{I}^s$ is denoted as $\{\hat{\mathbf{g}}^s\} = F_\theta(\mathcal{I}^s)$.

To maintain stable estimation across different rotation angles, we train the model with our proposed distribution loss function $\mathcal{L}_D$ and $\mathcal{L}_1$ loss. In $\mathcal{L}_D$, the mean value of estimation results is supervised by gaze label $\mathbf{g}$ and the STD of $\{\hat{\mathbf{g}}\}$ is supervised by sub-label. We explain the details of $\mathcal{L}_D$ in Sec. 4.2.3.

In a nutshell, the RAT phase can be formalized as:

$$\arg\min_\theta (\mathcal{L}_1(\hat{\mathbf{g}}^s, \mathbf{g}^s) + \mathcal{L}_D(\{\mathbf{R}\}, \{\hat{\mathbf{g}}^s\}, \mathbf{g}^s)). \tag{6}$$

4.2.2 Consistency-Guided Domain Adaptation

In RAT phase, we generalize the rotation-augmented model F_θ to the target domain with the guidance of sub-label based on the rotation consistency property. We also introduce a temporal average model $\bar{F}_{\bar{\theta}}$, which produces pseudo labels to prevent the estimation collapse.

First, we obtain the sub-label by randomly rotating the unlabeled sample $I^t \in \mathcal{D}_t$ by K times:

$$\mathcal{I}^t = \{\mathbf{R}I^t | k = 1, 2, \ldots, K\}. \tag{7}$$

Ideally, the rotation angle between the estimations of $\mathcal{I}^t$ and the estimation of original image I^t should be equal to sub-labels $\{\mathbf{R}\}$ according to Eq. (1). F_θ is supervised by rotation consistency with sub-label $\{\mathbf{R}\}$ instead of true label.

If rotation consistency is the only constraint applied to F_θ, the estimation results collapse to the z-axis of camera coordinate system since it is the rotation axis. Inspired by [11], we introduce a temporal average model $\bar{F}_{\bar{\theta}}$ that produces stable pseudo labels to avoid collapse.

At the beginning of CGDA phase, $\bar{F}_{\bar{\theta}}$ is initialized as a copy of F_θ. After training with T iterations, parameters of temporal average model $\bar{\theta}$ are updated from θ by Exponential Moving Average (EMA) algorithm:

$$\bar{\theta}^T = \alpha \bar{\theta}^{T-1} + (1-\alpha)\theta^T, \tag{8}$$

where α is a momentum coefficient. $\bar{F}_{\bar{\theta}}$ first estimates the gazes from a group of rotated images. Then, we design a Sub-label-guided Rotate-back Module (SRM) to recover the estimations that corresponds to the original image. According to the rotation consistency property, we rotate the estimation results of rotated images with the inverse matrix of sub-label. The pseudo label $\mathbf{g}'$ is defined as the mean direction of recovered gaze estimations:

$$\mathbf{g}'^t = Mean(\{(\mathbf{R}^\mathbf{g})^{-1}\}\bar{F}_{\bar{\theta}}(\mathcal{I}^t)). \tag{9}$$

The estimation from $\bar{F}_{\bar{\theta}}$, *i.e.*, pseudo label is much more stable than F_θ during adaptation [11], while still being capable of fine adjustment. F_θ is punished if the estimation deviates far from pseudo label because of collapse.

In CGDA phase, the model is also supervised by $\mathcal{L}_D$ while the gaze label is replaced by pseudo label. The adaptation process is summarized as:

$$\begin{aligned} &\arg\min_\theta(\mathcal{L}_D(\{\mathbf{R}\},\{\hat{\mathbf{g}}^t\},\mathbf{g}'^t), \\ &\bar{\theta}^T = \alpha\bar{\theta}^{T-1} + (1-\alpha)\theta. \end{aligned} \tag{10}$$

4.2.3 Distribution Loss Function

To supervise the model with rotation consistency and sub-label, we propose the Distribution Loss $\mathcal{L}_D$, which consists of two terms $\mathcal{L}_{mean}$ and $\mathcal{L}_{std}$. $\mathcal{L}_{mean}$ constrains the estimation to be accurate by gaze label in RAT phase and prevent the estimation from collapse in CGDA phase. $\mathcal{L}_{std}$ constrains the estimations to be consistent with each other by the sub-label R. $\mathcal{L}_D$ is defined as follow:

$$\begin{aligned} \mathcal{L}_D(\{\mathbf{R}\},\{\hat{\mathbf{g}}\},\mathbf{g}) &= \mathcal{L}_{mean} + \mathcal{L}_{std}, \\ \mathcal{L}_{mean}(\{\mathbf{R}\},\{\hat{\mathbf{g}}\},\mathbf{g}) &= \frac{1}{K}\sum_{k=1}^{K}\mathcal{L}_1(\mathbf{g}',\mathbf{g}), \\ \mathcal{L}_{std}(\{\mathbf{R}\},\{\hat{\mathbf{g}}\}) &= \sqrt{\frac{\sum_{k=1}^{K}(\mathbf{g}'-\overline{\{\mathbf{g}'\}})^2}{K}}, \\ \{\mathbf{g}'\} &= \{(\mathbf{R}^\mathbf{g})^{-1}\}\{\hat{\mathbf{g}}\}, \end{aligned} \tag{11}$$

where $\{\mathbf{g}'\}$ stands for a group of recovered estimation results by SRM module based on rotation consistency, $\overline{\{\mathbf{g}'\}}$ stands for the average direction of $\{\mathbf{g}'\}$. $\mathcal{L}_D$ treats a group of recovered gaze estimation as a distribution. $\mathcal{L}_{mean}$ supervise the model by requiring every sample of the distribution to be equal to the desired mean value $\mathbf{g}$. $\mathcal{L}_{std}$ requires the standard deviation of the distribution to be 0, which is the proposed rotation consistency in Eq. (1). The whole procedure of RUDA framework is summarized in Algorithm 1.

Algorithm 1 Rotation-enhanced unsupervised domain adaptation algorithm for gaze estimation.

Input: $\mathcal{D}_s$, $\mathcal{D}'_t \subset \mathcal{D}_t$ and F_θ
Output: F_θ

```
# Rotation augmented training
for i ← 1 to N_s do
    Get I^s, {R} by augmentation with Eq. (5)
    {ĝ^s} ← F_θ(I^s)
    L_1 ← ĝ^s, g^s
    L_D ← {R}, {ĝ^s}, g^s with Eq. (11)
    Train F_θ with Eq. (6)
end for
# Rotation consistent domain adaptation
F̄_θ̄ ← F_θ
for i ← 1 to N'_t do
    Get I_1^t, I_2^t by {R_1}, {R_2} with Eq. (7)
    {ĝ^t} ← F_θ(I_1^t)
    g'^t ← Mean({(R_2^g)^-1} F̄_θ̄(I_2^t)) with Eq. (9)
    L_D ← {R_1}, {ĝ^t}, g'^t with Eq. (11)
    Train F_θ by L_D with Eq. (10)
    Update θ̄ with Eq. (8)
end for
```

4.3. Implementation Details

Our method is implemented using PyTorch framework. ResNet18 is used as backbone. K is set to 5 in RAT phase and is set to 20 in CGDA phase. Momentum coefficient α in EMA algorithm is set to 0.99. Batch size is set to 80 and 10 during source domain training and domain adaptation phase respectively. We randomly chose 100 unlabeled images from target domain for adaptation. The model is trained for 10 epochs in source domain and for adaptation. We employed the Adam optimizer with a learning rate of 10^{-4} and $\beta = (0.5, 0.95)$.

5. Experiments

5.1. Data Preparation

To verify the effectiveness of the RUDA framework, we conducted experiments on four commonly used gaze estimation datasets: ETH-XGaze ($\mathcal{D}_E$) [41], Gaze360 ($\mathcal{D}_G$) [19], MPIIFaceGaze ($\mathcal{D}_M$) [43] and EyeDiap ($\mathcal{D}_D$) [10].

- **ETH-XGaze:** ETH-XGaze dataset is collected under laboratory environment with high-resolution cameras. We

Table 1. Unsupervised domain adaptation results of our proposed RUDA framework with different backbone models. The results are angular error in degrees.

Method	$\mathcal{D}_E \rightarrow \mathcal{D}_M$	$\mathcal{D}_E \rightarrow \mathcal{D}_D$	$\mathcal{D}_G \rightarrow \mathcal{D}_M$	$\mathcal{D}_G \rightarrow \mathcal{D}_D$
ResNet18	8.20	7.16	7.74	7.64
ResNet18+RAT	7.92	7.44	7.60	7.10
ResNet18+RUDA	**5.70**	**6.29**	**6.20**	**5.86**
ResNet50	7.15	6.43	8.35	7.86
ResNet50+RAT	7.40	6.91	7.69	7.08
ResNet50+RUDA	**5.78**	**5.10**	**6.88**	**6.73**

follow the original paper and take 750,000 face crops from 80 participants as training set.

- **Gaze360:** Gaze360 dataset is collected in arbitrary environment by a 360°camera. It has a wide distribution over the horizontal axis of gaze. We only use 84900 images with frontal faces.
- **MPIIFaceGaze:** MPIIFaceGaze is collected during daily usage of laptops. We chose 3000 images for 15 subjects respectively as the standard protocol suggests.
- **EyeDiap:** EyeDiap dataset is collected under laboratory environment with screen and floating targets. Note that due to the misalignment of timeline, some of the labels are not reliable. We selected 6400 sample images that are manually checked by original authors.

We perform gaze normalization proposed by [42] for all datasets except $\mathcal{D}_G$, as it does not provide head pose labels. Rotations are performed after gaze normalization. Face images are cropped and resized to 224x224. We further normalize the image pixels to $[0, 1]$ as the final input. More details can be found in [8].

5.2. Performance of RUDA Framework

To test the performance of RUDA framework, we implement it based on two state-of-the-art backbone network: ResNet18 and ResNet50 [17]. We train the backbone network on source domain with $\mathcal{L}_1$ loss as baseline. As shown in Tab. 1, RUDA framework improves the performance of both backbone network by a big margin. For ResNet18, RUDA framework improves the performance by 30.5%, 12.2%, 19.9% and 23.3% on four cross domain tasks, respectively. For ResNet50, RUDA framework brings 19.2%, 20.7%, 17.6% and 14.4% performance gain. Thanks to the reasonable design of RUDA framework and wide data distribution of ETH-XGaze dataset, the performance of ResNet50+RUDA model on $\mathcal{D}_E \rightarrow \mathcal{D}_D$ task even outperforms state-of-the-art within dataset gaze estimation methods, e.g., [5, 7]. The results show that RAT strategy alone does not improve cross domain performance, as expected. The ability to estimate gaze from rotated images does not improve estimation accuracy on normalized face images with upright orientation. After rotation consistency guided adaptation, the performance improves significantly. This proves our point in Sec. 3 that rotation consistency contain much more significant relation with physical model of gaze than data augmentation like rotation.

Table 2. Comparison with state-of-the-art unsupervised domain adaptation methods. Results are angular error in degrees.

Method	$\mathcal{D}_E \rightarrow \mathcal{D}_M$	$\mathcal{D}_E \rightarrow \mathcal{D}_D$	$\mathcal{D}_G \rightarrow \mathcal{D}_M$	$\mathcal{D}_G \rightarrow \mathcal{D}_D$
ResNet18	8.20	7.16	7.74	7.64
Fine-tune	5.12	5.50	5.36	5.22
ADDA [36]	8.55	10.63	8.59	16.68
DAGEN [16]	7.53	8.46	9.31	12.05
GazeAdv [39]	8.48	7.70	9.15	11.15
Gaze360 [19]	7.15	6.87	7.45	9.73
RSD [4]	8.74	7.46	9.17	10.61
RUDA(ours)	**5.70**	**6.29**	**6.20**	**5.86**

Table 3. Ablation study for different loss function in source domain training, different pre-trained models in domain adaptation phase and different loss functions in domain adaptation phase. Results are angular error in degrees.

	Method	$\mathcal{D}_E \rightarrow \mathcal{D}_M$	$\mathcal{D}_E \rightarrow \mathcal{D}_D$	$\mathcal{D}_G \rightarrow \mathcal{D}_M$	$\mathcal{D}_G \rightarrow \mathcal{D}_D$
1	ResNet18	8.20	7.16	7.74	7.64
2	ResNet18+$\text{R}_{\mathcal{L}_1}$	7.46	7.17	9.11	7.60
3	ResNet18+$\text{R}_{\mathcal{L}_2}$	8.10	8.09	7.69	7.08
4	ResNet18+$\text{R}_{\mathcal{L}_D}$	7.92	7.44	7.60	7.10
5	ResNet18+$\text{DA}_{\mathcal{L}_D}$	5.73	6.58	7.55	7.27
6	ResNet18+$\text{R}_{\mathcal{L}_1}$+$\text{DA}_{\mathcal{L}_D}$	6.01	**6.03**	8.67	5.93
7	ResNet18+$\text{R}_{\mathcal{L}_2}$+$\text{DA}_{\mathcal{L}_D}$	6.89	7.10	6.58	6.03
8	ResNet18+$\text{R}_{\mathcal{L}_D}$+$\text{DA}_{\mathcal{L}_1}$	6.96	6.68	**6.06**	6.33
9	ResNet18+$\text{R}_{\mathcal{L}_D}$+$\text{DA}_{\mathcal{L}_2}$	6.38	6.74	6.20	6.48
10	ResNet18+$\text{R}_{\mathcal{L}_D}$+$\text{DA}_{\mathcal{L}_D}$	**5.70**	6.29	6.20	**5.86**

5.3. Comparison with SOTA UDA methods

To demonstrate the performance of RUDA framework, we compare it with state-of-the-art unsupervised domain adaptation methods on four cross domain tasks: $\mathcal{D}_E \rightarrow \mathcal{D}_M$, $\mathcal{D}_E \rightarrow \mathcal{D}_D$, $\mathcal{D}_G \rightarrow \mathcal{D}_M$, $\mathcal{D}_G \rightarrow \mathcal{D}_D$. We choose four typical methods for comparison:

- **ADDA [36]:** Reduce domain gap between source and target domain features by adversarial learning. A discriminator which classifies feature to source or target domain is introduced. 500 target domain images are used in our implementation for better performance.
- **DAGEN [16]:** A SOTA unsupervised domain adaptation method for gaze estimation by embedding representation design. 500 target domain images are used in our implementation for better performance.
- **GazeAdv [39]:** A SOTA unsupervised domain adaptation method for gaze estimation by adversarial learning. Appearance classification and head pose classification are designed as adversarial tasks.
- **Gaze360 [19]:** A SOTA unsupervised domain adaptation

Table 4. Gaze estimation error in degrees for different image rotation angles. For a given degree r, we randomly rotate the image in a range of $[-r, r]$. Rotation angles in source domain training and target domain adaptation remain the same.

	RAT				RAT+CGDA			
Rotation	$\mathcal{D}_E\!\rightarrow\mathcal{D}_M$	$\mathcal{D}_E\!\rightarrow\mathcal{D}_D$	$\mathcal{D}_G\!\rightarrow\mathcal{D}_M$	$\mathcal{D}_G\!\rightarrow\mathcal{D}_D$	$\mathcal{D}_E\!\rightarrow\mathcal{D}_M$	$\mathcal{D}_E\!\rightarrow\mathcal{D}_D$	$\mathcal{D}_G\!\rightarrow\mathcal{D}_M$	$\mathcal{D}_G\!\rightarrow\mathcal{D}_D$
15°	8.44	7.31	8.45	8.11	8.18	6.65	8.24	8.18
40°	8.22	6.62	8.71	7.67	7.63	6.93	7.52	7.11
65°	7.93	8.59	8.09	7.73	6.68	6.81	6.72	7.62
90°	7.92	7.44	7.60	7.10	**5.70**	**6.29**	**6.20**	**5.86**

method for gaze estimation by combination of adversarial learning, image flip and pinball loss.

- **RSD [4]:** A SOTA unsupervised domain adaptation method specially designed for regression tasks. It closes domain gap through orthogonal bases of the representation spaces without changing the feature scale.

For a fair comparison, we replace the backbone of all methods with ResNet18. The result is shown in Tab. 2. Our method outperforms SOTA methods by a big margin. RUDA framework significantly improves the performance on all tasks. Note that general unsupervised domain adaptation methods do not bring any performance improvement, which shows the difficulty of cross domain gaze estimation tasks. Methods designed for gaze estimation may bring performance gain on certain cross domain tasks, while make other tasks worse. Our RUDA framework performs stably and improves the performance in all four tasks.

5.4. Ablation Study

To prove the effectiveness of each component in RUDA framework, we conducted ablation study on all four cross domain tasks. In Tab. 3, we show the results for different combination of source domain training strategy, domain adaptation strategy and loss function:

- R$_{\mathcal{L}_1}$, R$_{\mathcal{L}_2}$: Training with $\mathcal{L}_1$, $\mathcal{L}_2$ loss respectively on rotation augmented source domain.
- R$_{\mathcal{L}_D}$: The proposed RAT strategy with $\mathcal{L}_D$ loss function.
- DA$_{\mathcal{L}_1}$, DA$_{\mathcal{L}_2}$: The proposed CGDA strategy in which $\mathcal{L}_D$ loss is replaces by $\mathcal{L}_1$, $\mathcal{L}_2$ loss function respectively.
- DA$_{\mathcal{L}_D}$: The proposed CGDA strategy with $\mathcal{L}_D$ loss.

In Tab. 3, row 1-4 show that training with rotated images on source domain does not improve cross domain accuracy. But training with proposed $\mathcal{L}_D$ appears to be more stable than $\mathcal{L}_1$, $\mathcal{L}_2$ thanks to the $\mathcal{L}_{std}$ term. Results from row 5-7 prove the effectiveness of DA$_{\mathcal{L}_D}$, *i.e.*, the proposed CGDA strategy. CGDA improves accuracy when combined with models from row 1 to 3 on all cross domain tasks. In row 5 to 7, although some of the combination reaches compatible ([row 5, $\mathcal{D}_E\!\rightarrow\mathcal{D}_M$], [row 6, $\mathcal{D}_G\!\rightarrow\mathcal{D}_D$]) or even better ([row 6, $\mathcal{D}_E\!\rightarrow\mathcal{D}_D$]) performance than RUDA in certain tasks, they performs similar or even worse than the baseline ResNet18 in other tasks ([row 5, $\mathcal{D}_G\!\rightarrow\mathcal{D}_M$], [row 6, $\mathcal{D}_G\!\rightarrow\mathcal{D}_M$], [row 7, $\mathcal{D}_E\!\rightarrow\mathcal{D}_D$]). Without $\mathcal{L}_D$ loss in source domain training, combinations from row 5 to row 7 suffer from obvious performance gap for different cross domain tasks. This is a fatal flow for unsupervised domain adaptation tasks as we have no target domain label to verify whether performance is improved or dropped. Compared with row 8 to row 10, methods with RAT strategy shows apparent stability and improves accuracy in all four tasks.

In row 8 and row 9, we test the combination of proposed CGDA strategy with different loss function in domain adaptation. Compare to row 10, $\mathcal{L}_D$ achieves better overall performance gain than $\mathcal{L}_1$ and $\mathcal{L}_2$ loss function.

Above experiments validate the effectiveness of RAT and CGDA strategy. With the help of RAT and CGDA, the proposed RUDA framework achieves the most stable and satisfactory improvement in all four tasks.

5.5. Hyper Parameters and Further Analysis

5.5.1 Hyper Parameters

In this section, we carried out experiments to investigate the impact of hyper parameters. Rotation degree is the most important hyper parameter as we create sub-label by rotation. In Tab. 4, we show the results of different rotation range. For a given degree r, we randomly rotate the image in range $[-r, r]$. In RAT phase, models perform similarly under different rotation degree. After adaptation by CGDA, the accuracy increases with the rotation range. During adaptation, the model is supervised by the rotation consistency of estimation results from rotated images, which corresponds to the $\mathcal{L}_{std}$ term in distribution loss. Hence, we test models without CGDA on the subset of target domain $\mathcal{D}_t'$ and count STD. As shown in Fig. 4, STD drops as the range of rotation shrinks. In consequence, smaller the range of rotation is, smaller the $\mathcal{L}_D$ is, less uncertainty signal for the model to learn during adaptation.

We also evaluate the impact of the number of rotation for each image during CGDA phase. We set the number of rotation to 10, 15, 20, 25 during adaptation respectively while keeps rotation number in RAT phase at 5. The result jitters when number of rotation changes. But the overall disturbance is relatively subtle compared to range of rotation.

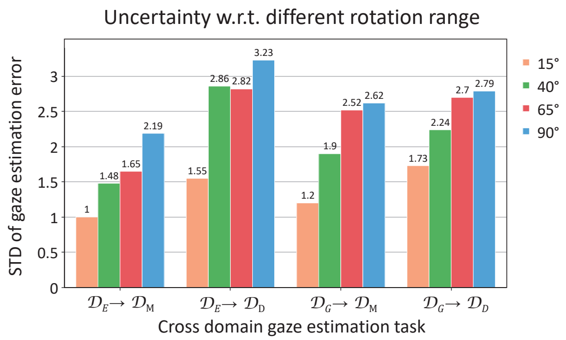

Figure 4. Standard deviation (STD) of gaze error obtained in the RAT phase, for the 100 target domain images used for adaptation later. Larger rotation ranges produce larger STDs, which provide sufficient uncertainties for domain adaptation.

Table 5. Gaze estimation error in degrees for different image rotation numbers in domain adaptation phase.

Number of Rotation	$\mathcal{D}_E \rightarrow \mathcal{D}_M$	$\mathcal{D}_E \rightarrow \mathcal{D}_D$	$\mathcal{D}_G \rightarrow \mathcal{D}_M$	$\mathcal{D}_G \rightarrow \mathcal{D}_D$
10	6.18	6.71	6.31	5.78
15	6.40	**6.23**	6.24	**5.60**
20	**5.70**	6.29	**6.20**	5.86
25	6.21	6.49	6.24	5.98

5.5.2 Importance of Rotation Consistency

We design the RUDA framework around the rotation consistency for it connects deeply with the physical nature of gaze. To prove the importance of rotation consistency, we replace rotation consistency with other data augmentation methods in RUDA framework.

Specifically, we choose two commonly-used image augmentations: 1) geometry augmentation, we apply random scaling and random translating to the normalized face images, and 2) noise augmentation, we randomly apply four kinds of different noise to the image including random noise, Gaussian noise and Poisson noise. Examples of three kinds of operation are shown in Fig. 5.

The results are shown in Tab. 6. Although geometry augmentation and noise augmentation are proven to be effective in other computer vision tasks such as classification and object detection, they do not bring any improvement in cross-domain gaze estimation tasks. We argue that it might be because that geometry consistency and noise consistency are easier to achieve as these two augmentation only disturb the appearance of images, do not touch the physical nature of gaze. Rotation brings more uncertainty information, *i.e.*, it changes not only the appearance but also gaze direction.

5.5.3 System Limitations

The proposed RUDA framework have successfully addressed one of the critical problems in unsupervised domain adaptation, i.e., the shortage of training data and the absence of target labels. On the other hand, another common challenge of appearance-based gaze domain adaptation tasks is that the data distribution of source domain and target domain can be different. When the range of source domain is significantly smaller, the adaptation capability decreases. Such a problem has not been well addressed by existing methods. In the future, we can try to handle this problem and combine the resulting technique into our RUDA framework to further improve the system robustness.

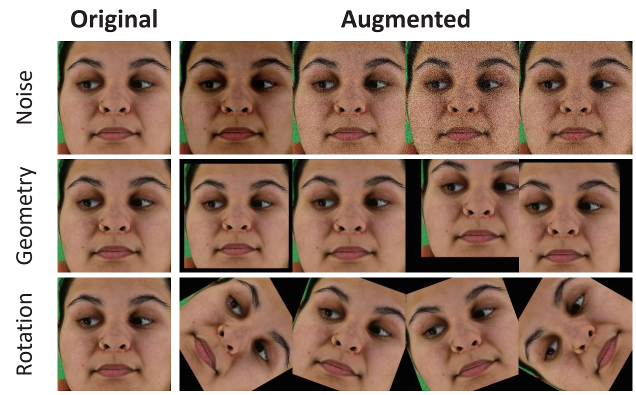

Figure 5. Different image augmentations compared in Tab. 6.

Table 6. Comparison with other data augmentation methods. Results are angular error in degrees.

Method	$\mathcal{D}_E \rightarrow \mathcal{D}_M$	$\mathcal{D}_E \rightarrow \mathcal{D}_D$	$\mathcal{D}_G \rightarrow \mathcal{D}_M$	$\mathcal{D}_G \rightarrow \mathcal{D}_D$
ResNet18	8.20	7.16	7.74	7.64
Geometry+RAT	9.75	8.50	7.88	7.41
Geometry+RUDA	9.71	10.17	7.40	7.33
Noise+RAT	8.70	8.12	7.80	7.65
Noise+RUDA	9.02	7.43	6.94	8.40
Rotation+RAT	7.92	7.44	7.60	7.10
Rotation+RUDA(ours)	**5.70**	**6.29**	**6.20**	**5.86**

6. Conclusions

In this paper, we present the rotation-enhanced unsupervised domain adaptation framework for gaze estimation tasks. Based on the rotation consistency property, the proposed RUDA framework adapts the model to unlabeled target domain. It first trains a rotation-augmented model with RAT strategy in source domain, then generalized to target domain via the guidance of sub-labels, *i.e.*, estimation consistency across different rotation angles in CGDA phase. Experimental results show that the RUDA framework achieves stable and significant improvement in four different cross-domain tasks. The idea of rotation consistency may be applied in other physical related regression tasks such as pose estimation.

References

[1] Yiwei Bao, Yihua Cheng, Yunfei Liu, and Feng Lu. Adaptive feature fusion network for gaze tracking in mobile tablets. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 9936–9943. IEEE, 2021. 1, 2

[2] Fabio Maria Carlucci, Lorenzo Porzi, Barbara Caputo, Elisa Ricci, and Samuel Rota Bulo. Autodial: Automatic domain alignment layers. In *2017 IEEE international conference on computer vision (ICCV)*, pages 5077–5085. IEEE, 2017. 2

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

[4] Xinyang Chen, Sinan Wang, Jianmin Wang, and Mingsheng Long. Representation subspace distance for domain adaptation regression. In *International Conference on Machine Learning*, pages 1749–1759. PMLR, 2021. 2, 6, 7

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

[6] Yihua Cheng, Yiwei Bao, and Feng Lu. Puregaze: Purifying gaze feature for generalizable gaze estimation. *arXiv preprint arXiv:2103.13173*, 2021. 2

[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. 1, 2, 6

[8] Yihua Cheng, Haofei Wang, Yiwei Bao, and Feng Lu. Appearance-based gaze estimation with deep learning: A review and benchmark. *arXiv preprint arXiv:2104.12668*, 2021. 1, 6

[9] Yihua Cheng, Xucong Zhang, Feng Lu, and Yoichi Sato. Gaze estimation by exploring two-eye asymmetry. *IEEE Transactions on Image Processing*, 29:5259–5272, 2020. 1, 2

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

[11] Yixiao Ge, Dapeng Chen, and Hongsheng Li. Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification. *arXiv preprint arXiv:2001.01526*, 2020. 5

[12] Boqing Gong, Yuan Shi, Fei Sha, and Kristen Grauman. Geodesic flow kernel for unsupervised domain adaptation. In *2012 IEEE conference on computer vision and pattern recognition*, pages 2066–2073. IEEE, 2012. 2

[13] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 2

[14] Raghuraman Gopalan, Ruonan Li, and Rama Chellappa. Domain adaptation for object recognition: An unsupervised approach. In *2011 international conference on computer vision*, pages 999–1006. IEEE, 2011. 2

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

[16] Zidong Guo, Zejian Yuan, Chong Zhang, Wanchao Chi, Yonggen Ling, and Shenghao Zhang. Domain adaptation gaze estimation by embedding with prediction consistency. In *Proceedings of the Asian Conference on Computer Vision*, 2020. 2, 6

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

[18] Christina Katsini, Yasmeen Abdrabou, George E Raptis, Mohamed Khamis, and Florian Alt. The role of eye gaze in security and privacy applications: Survey and future hci research directions. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–21, 2020. 1

[19] Petr Kellnhofer, Adria Recasens, Simon Stent, Wojciech Matusik, and Antonio Torralba. Gaze360: Physically unconstrained gaze estimation in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6912–6921, 2019. 1, 2, 5, 6

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

[21] Robert Konrad, Anastasios Angelopoulos, and Gordon Wetzstein. Gaze-contingent ocular parallax rendering for virtual reality. *ACM Transactions on Graphics (TOG)*, 39(2):1–12, 2020. 1

[22] Rakshit Kothari, Shalini De Mello, Umar Iqbal, Wonmin Byeon, Seonwook Park, and Jan Kautz. Weakly-supervised physically unconstrained gaze estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9980–9989, 2021. 1, 2

[23] Kyle Krafka, Aditya Khosla, Petr Kellnhofer, Harini Kannan, Suchendra Bhandarkar, Wojciech Matusik, and Antonio Torralba. Eye tracking for everyone. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2176–2184, 2016. 1, 2

[24] Dongze Lian, Lina Hu, Weixin Luo, Yanyu Xu, Lixin Duan, Jingyi Yu, and Shenghua Gao. Multiview multitask gaze estimation with deep convolutional neural networks. *IEEE transactions on neural networks and learning systems*, 30(10):3010–3023, 2018. 1, 2

[25] Yunfei Liu, Ruicong Liu, Haofei Wang, and Feng Lu. Generalizing gaze estimation with outlier-guided collaborative adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3835–3844, 2021. 1, 2

[26] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *International conference on machine learning*, pages 97–105. PMLR, 2015. 2

[27] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Conditional adversarial domain adaptation. *arXiv preprint arXiv:1705.10667*, 2017. 2

[28] Conny Lu, Praneeth Chakravarthula, Yujie Tao, Steven Chen, and Henry Fuchs. Improved vergence and accommodation via purkinje image tracking with multiple cameras for ar glasses. In *2020 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 320–331. IEEE, 2020. 2

[29] 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. 1, 2

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

[31] Brian A Smith, Qi Yin, Steven K Feiner, and Shree K Nayar. Gaze locking: passive eye contact detection for human-object interaction. In *Proceedings of the 26th annual ACM symposium on User interface software and technology*, pages 271–280, 2013. 2

[32] Adrian Spurr, Aneesh Dahiya, Xi Wang, Xucong Zhang, and Otmar Hilliges. Peclr: Self-supervised 3d hand pose estimation from monocular rgb via contrastive learning. *arXiv preprint arXiv:2106.05953*, 2021. 2

[33] Yusuke Sugano, Yasuyuki Matsushita, and Yoichi Sato. Learning-by-synthesis for appearance-based 3d gaze estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1821–1828, 2014. 2

[34] Li Sun, Zicheng Liu, and Ming-Ting Sun. Real time gaze estimation with a consumer depth camera. *Information Sciences*, 320:346–360, 2015. 2

[35] Yunus Terzioğlu, Bilge Mutlu, and Erol Şahin. Designing social cues for collaborative robots: the role of gaze and breathing in human-robot collaboration. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, pages 343–357, 2020. 1

[36] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7167–7176, 2017. 2, 6

[37] Haofei Wang, Xujiong Dong, Zhaokang Chen, and Bertram E Shi. Hybrid gaze/eeg brain computer interface for robot arm control on a pick and place task. In *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 1476–1479. IEEE, 2015. 1

[38] Kang Wang and Qiang Ji. Real time eye gaze tracking with 3d deformable eye-face model. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1003–1011, 2017. 2

[39] 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. 1, 2, 6

[40] Quan Wen, Derek Bradley, Thabo Beeler, Seonwook Park, Otmar Hilliges, Junhai Yong, and Feng Xu. Accurate real-time 3d gaze tracking using a lightweight eyeball calibration. In *Computer Graphics Forum*, volume 39, pages 475–485. Wiley Online Library, 2020. 2

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

[42] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. Appearance-based gaze estimation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4511–4520, 2015. 1, 2, 6

[43] 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. 1, 2, 5

[44] Yuchen Zhang, Tianle Liu, Mingsheng Long, and Michael Jordan. Bridging theory and algorithm for domain adaptation. In *International Conference on Machine Learning*, pages 7404–7413. PMLR, 2019. 2

[45] Yufeng Zheng, Seonwook Park, Xucong Zhang, Shalini De Mello, and Otmar Hilliges. Self-learning transformations for improving gaze and head redirection. *Advances in Neural Information Processing Systems*, 33:13127–13138, 2020. 2