Domain Adaptation Gaze Estimation by Embedding with Prediction Consistency

Zidong Guo[1], Zejian Yuan[1], Chong Zhang[2], Wanchao Chi[2], Yonggen Ling[2], and Shenghao Zhang[2]

[1] Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, China
gzd3118311122@stu.xjtu.edu.cn, yuan.ze.jian@xjtu.edu.cn
[2] Tencent Robotics X, China
aerentzhang@gmail.com, wanchaochi@tencent.com, ylingaa@connect.ust.hk, popshzhang@pku.edu.cn

Abstract. Gaze is the essential manifestation of human attention. In recent years, a series of work has achieved high accuracy in gaze estimation. However, the inter-personal difference limits the reduction of the subject-independent gaze estimation error. This paper proposes an unsupervised method for domain adaptation gaze estimation to eliminate the impact of inter-personal diversity. In domain adaption, we design an embedding representation with prediction consistency to ensure that linear relationships between gaze directions in different domains remain consistent on gaze space and embedding space. Specifically, we employ source gaze to form a locally linear representation in the gaze space for each target domain prediction. Then the same linear combinations are applied in the embedding space to generate hypothesis embedding for the target domain sample, remaining prediction consistency. The deviation between the target and source domain is reduced by approximating the predicted and hypothesis embedding for the target domain sample. Guided by the proposed strategy, we design Domain Adaptation Gaze Estimation Network(DAGEN), which learns embedding with prediction consistency and achieves state-of-the-art results on both the MPIIGaze and the EYEDIAP datasets.

1 Introduction

Gaze servers as an important visual cue of human attention. Accurate gaze estimation can provide critical support for many applications, such as human-computer interaction [1], virtual reality [2], and driver monitoring systems [3]. Although eye tracker can provide a precise gaze estimation [4], the high price and the demand for specific equipment limit its applications in the real world and more flexible environments. Unconstrained appearance-based gaze estimation methods can predict 2D gaze target position or 3D gaze angles based on patches cropped from RGB images. Thanks to the advancement of convolutional neural networks (CNN) and a large number of publicly available high-quality datasets, the error of gaze estimation has been dramatically decreased in recent years.

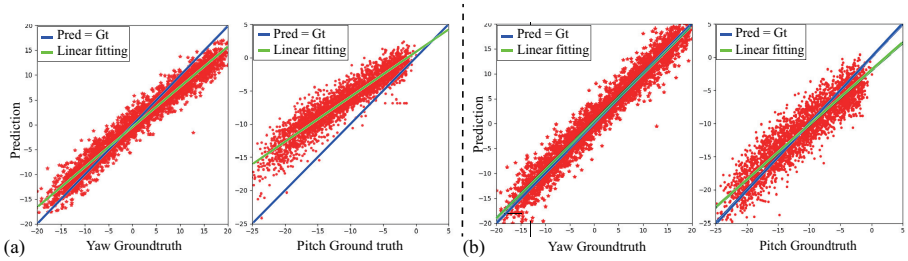

Fig. 1. Scatter plot of the groundtruth (X-axis) and the network gaze estimation (Y-axis) of the yaw and pitch angles in an evaluation set from the MPIIGaze dataset. The results are estimated by (a) Regression from eye region based on CNN and (b) DAGEN (Ours).

Appearance-based gaze estimation can decouple gaze direction from high-dimension images with various noises, but some challenges still restrict the further improvement of estimation precision. Obtaining gaze groundtruth requires specific equipment, a well-defined collection strategy, and highly concentrated attention of participants [5–10]. Under these strict conditions, current datasets violate the identical independent distribution(i.i.d) nature, that is, only tens of persons participating in the collection of thousands of gaze direction data per subject [11]. For gaze estimation that requires open set testing [12], the deviation between the distribution of the training set and the test set is reflected in the prediction as a person-specific bias. As shown in Fig. 1(a), the bias between the network regression and the groundtruth can often be observed, which is also mentioned in [13]. Some methods perform a person-specific gaze estimation through several new subject's labeled data to eliminate the bias [12,13]. However, in practice, even a bit of accurately labeled data is challenging to acquire.

In this work, we propose an unsupervised method for domain adaptation (DA) gaze estimation to eliminate the impact of inter-personal differences and fit new subject's data without groundtruth labels. In domain adaption, we design an embedding representation with prediction consistency to ensure that linear relationships between gaze directions remain consistent on gaze space and embedding space. We then build the Domain Adaptation Gaze Estimation Network (DAGEN) using EPC loss devised to measure this consistency. Moreover, a new training strategy is employed for domain adaptation.

The most crucial element of DAGEN is the embedding with prediction consistency (EPC), which is expected to eliminate the deviation between domains. Following the Locally Linear Embedding (LLE) representation method [14], the hypothesis label predicted on the target domain is linearly interpreted by its neighbor gaze directions from the source domain. Such linear combinations in gaze space would be migrated to embedding space to obtain hypothesis embedding, which ensures locally linear consistency between the embedding and prediction space. For the same gaze directions, we demand the embedding features encoding gaze should also be similar. However, due to the deviation between

domains, the embedding also retains some domain-specific features unrelated to gaze direction, which causes a fixed bias in the test. So EPC loss, which weighs the distance between target hypothesis embedding and predicted embedding for each target domain sample, can be used to illustrate the deviation between domains. We optimize the EPC loss to eliminate the deviation between domains, thus achieving domain adaptation. We present our DAGEN estimation results in Fig. 1(b) to exhibit the consequence of domain adaptation.

We evaluate our proposed method on two commonly used gaze datasets and indicate that our DAGEN can effectively improve the accuracy of gaze estimation. On both datasets, our estimation results exceeding the current state-of-the-art method. Specifically, the DAGEN achieves a 9.66% improvement (4.14° to 3.74°) on MPIIGaze, and an 18.9% improvement (5.3° to 4.3°) on EYEDIAP. Note that the input only uses the eye region patch, and the source and target domain are the train and evaluation set, respectively.

The major contributions of our work are summarized as follows:

1). We propose a new representation for the target domain embedding with prediction consistency, as a linear combination of neighbors from the source domain.

2). We design an innovative embedding with prediction consistency (EPC) loss for unsupervised domain adaptation gaze estimation, enabling it to measure the shift between the source and target domain.

3). Our method achieves state-of-the-art performance on MPIIGaze and EYEDIAP with only eye region as input.

2 Related Work

Gaze estimation methods are typically divided into appearance-based and model-based methods [15]. Model-based methods rely on the biological structure and reflection characteristics of the eyeball, and usually require high-resolution images with homogeneous illumination [16, 17]. Appearance-based methods can robustly decouple gaze angles from high-dimensional images with various noises. Recently, due to the application of many large data sets [6, 7, 10, 18] and the development of CNNs, the accuracy of appearance-based gaze estimation methods has been continuously improved. Zhang et al. first use LeNet [19] structure based on CNN and MPIIGaze dataset to process gaze estimation [7]. Subsequently, many works have improved the accuracy of gaze estimation through different methods. For example, multi-modal input was utilized in [18]; the key role of face was proved in [20]; a new convolution paradigm especially for gaze estimation was devised in [21]; timing information was used in [22]; the four models ensemble method was used to increase estimation accuracy in [8]; and a coarse-to-fine estimation strategy was designed in [23] .

However, recent work has discovered the fixed deviation in gaze estimation caused by person-specific diversity, as shown in Fig. 1(a). The diversity is reduced by learning gaze differences and applying calibration sets in [12]. Random effects, which actively learns the differences among-subjects during training, was

introduced in [11]. And a meta-learning method, performing person-specific calibration for each new subject and generating a person-specific network, is utilized in [13] to eliminate deviations.

Domain adaptation improves prediction performance in the target domain by aligning the distribution from the source domain [24,25]. Some work attempts to minimize the discrepancy between domains to obtain domain-invariant features directly [26, 27]. Recently, some methods found that aligning targets in both domains could significantly increase prediction performance. For instance, [28] uses the correlation between classes to perform domain adaptation by predicting the target hypothesis label and source groundtruth. For the gaze estimation problem, discriminator is applied to distinguish the source and target domains, thereby aligning the domains [9]. The differences was taken advantage in pairs of gaze directions and performs domain adaptation through gaze redirection and cycle consistency [29].

However, these methods do not address the deviations caused by subject-difference. We propose an unsupervised domain adaptation method to eliminate the inter-personal differences by introducing embedding with prediction consistency.

3 Proposed Method

Domain Adaptation (DA) is applied to solve the inter-personal differences reflected in the domain shift in the data distribution. Our method takes an eye region image I as input, regressing $g = (y, p)$ through a feature extractor and a linear mapping, where y and p means yaw and pitch in gaze direction. Given a source domain $S = \{(I_1^s, g_1^s), \cdots, (I_{N_s}^s, g_{N_s}^s)\}$ with several participants and groundtruth, and a target domain $T = \{I_1^t, \cdots, I_{N_t}^t\}$ using test set data without groundtruth, the proposed network adopts Domain Adaptation (DA) as the training strategy to align the embedding between S and T to increase its estimation performance.

Figure 2 provides an architecture of our Domain Adaptation Gaze Estimation Network (DAGEN). The feature extractor $\phi(\cdot)$ contains an ImageNet [30] pre-trained ResNet-18 [31] followed a multilayer perceptron. The embedding feature $\phi(I)$ will be constrained to keep consistency with predicted gaze direction during DA training. Finally, as the restrained prediction consistency embedding could decouple gaze-related information from the high-dimension image, the gaze direction $\hat{g}$ is calculated through a simple linear mapping operation h.

3.1 Target Domain Gaze Representation

We propose a Locally Linear Representation (LLR) for Target Domain Gaze that employs source domain gaze to represent the target hypothesis label in gaze space $\mathbb{G}$ linearly. For each sample in the target domain, the network prediction is considered as a hypothesis label. We linearly combine k source domain samples in its neighborhood in the $\mathbb{G}$ to describe it.

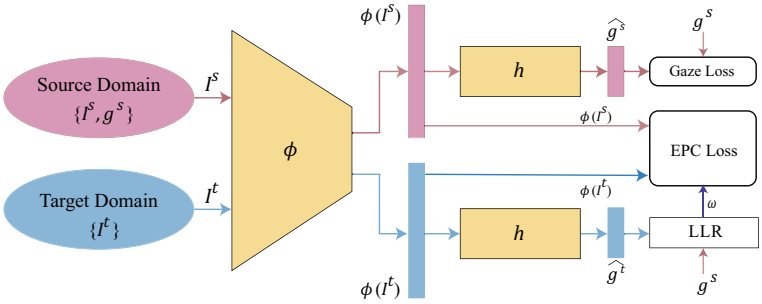

Fig. 2. The architecture of our proposed DAGEN. ResNet-18 and a fully connected layer are employed as feature exactor ϕ. A linear mapping h maps embedding feature $\phi(I)$ to gaze prediction $\hat{g}$. During DA training, LLR is utilized to generate linear weight w by source groundtruth and target prediction. Besides gaze loss for source domain, we apply EPC loss for unsupervised learning embedding features.

We first define the neighborhood for each target domain prediction in $\mathbb{G}$ to ensure the correct representation. Only when both angles in target hypothesis label $\hat{g_j^t}$ and source gaze direction g_i^s are not much different (less than μ), would g_i^s be set as a neighborhood of $\hat{g_j^t}$. We describe the set of all neighbors of $\hat{g_j^t}$ as $\mathcal{N}_j$, defined as,

$$\mathcal{N}_j = \left\{ g_i^s | \max\left(|y_i^s - \hat{y_j^t}|,\ |p_i^s - \hat{p_j^t}|\right) < \mu \right\}. \tag{1}$$

Every target domain prediction $\hat{g_j^t}$ in a mini-batch having over k neighbors would be randomly selected k neighbors to regenerate $\mathcal{N}_j$, which is employed to reconstruct the $\hat{g_j^t}$. We define the weight w_{ij} to summarize the contribution of the ith data in $\mathcal{N}_j$ to the $\hat{g_j^t}$ reconstruction, and the purpose is to find a suitable solution of each w_{ij}.

For 2D gaze direction g, the slightly larger number of neighbors means that it is challenging to find a suitable solution to minimize the reconstruction loss $E(w)$ during training. We consider involving more neighbors in the reconstruction of $\hat{g_j^t}$ and introduce an L2 regularization term to ensure a unique solution. So an L2 regularization term is suitable to solve this problem. The reconstruction loss $E(w)$ is formally expressed as,

$$E(W_j) = \|\hat{g_j^t} - \sum_{i=1}^{k} w_{ji} g_i^s\|_2^2 + \lambda \sum_{i=1}^{k} w_{ji}^2, \quad s.t.\ g_i^s \in \mathcal{N}_j \ and \ \sum_{i=1}^{k} w_{ji} = 1, \tag{2}$$

where $W_j = [w_{j1}, \cdots, w_{jk}]$. The Eq. (2) can be written in matrix form as,

$$\begin{aligned} E(W_j) &= W_j^T (\hat{G_j^t} - G_i^s)^T (\hat{G_j^t} - G_i^s) W_j + \lambda W_j^T W_j, \\ &= W_j^T (S_j + \lambda I) W_j, \end{aligned} \tag{3}$$

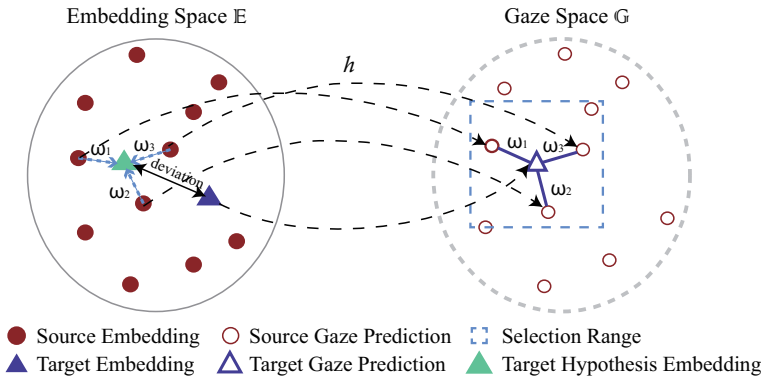

Fig. 3. Embedding with prediction consistency. The linear combination relationship in $\mathbb{G}$ is inherited to $\mathbb{E}$ to generate a hypothesis embedding $\hat{\phi}(I^t)$ for each target sample. The distance between $\hat{\phi}(I^t)$ and the target predicted embedding $\phi(I^t)$ measures the deviation between the source and target domain.

where $\hat{G}_j^t = [\hat{g_j^t}, \cdots, \hat{g_j^t}]_{1\times k}$, $G_i^s = [g_1^s, \cdots, g_k^s]$, and S_j is regarded as a local covariance matrix, defined as,

$$S_j = (\hat{G}_j^t - G_i^s)^T (\hat{G}_j^t - G_i^s). \tag{4}$$

The solution W_j^* that minimizes $E(W_j)$ obtained by the Lagrange multiplier method is,

$$W_j^* = \frac{(S_j + \lambda I)^{-1} 1_k}{1_k^T (S_j + \lambda I)^{-1} 1_k}. \tag{5}$$

With the optimized weight $W_j^* = [w_{j1}^*, \cdots, w_{jk}^*]$, LLR is formally described as,

$$\hat{g_j^t} = \sum_{i=1}^{k} w_{ji}^* g_i^s, \quad g_i^s \in \mathcal{N}_j. \tag{6}$$

3.2 Embedding with Prediction Consistency

Here we propose Embedding with Prediction Consistency (EPC) for domain adaptation. EPC transfers the same linear combination relationship in gaze space $\mathbb{G}$ to embedding space $\mathbb{E}$ to generate target hypothesis embeddings. For target domain sample I_j^t, the hypothesis embedding is declared as,

$$\hat{\phi}(I_j^t) = \sum_{i=1}^{k} w_{ji}^* \phi(I_i^s), \quad g_i^s \in \mathcal{N}_j. \tag{7}$$

As shown in Fig. 3, the LLR weight in gaze space $\mathbb{G}$ are inherited to the embedding space $\mathbb{E}$. For each target predicted embedding in $\mathbb{E}$, we generate the

target hypothesis embedding $\hat{\phi}(I^t)$ by Eq. (7). The linear relationship between target hypothesis embedding and source predicted embedding is the same as that between target and source gaze directions, which is the embedding with prediction consistency.

3.3 Loss Function

With the purpose of domain adaptation between the source and target domain, we propose DA loss consisting of two items, as shown in Eq. (8). Specifically, L_{EPC} measures the deviation between the source and target domain. Meanwhile, L_{gaze} supervises the predicted gaze directions of the source domain to guarantee that the network is always optimized towards reducing gaze estimation error.

$$L_{DA} = \lambda_{EPC} L_{EPC} + \lambda_{gaze} L_{gaze}, \tag{8}$$

where we empirically set$\lambda_{EPC} = 1$ and $\lambda_{gaze} = 1$.

We introduce an embedding with prediction consistency (EPC) loss for domain adaptation gaze estimation, which ensures same gaze directions should have the same embedding features unrelated to any interferences like appearance. Typically this constraint requires pairs of images in totally same gaze directions from different subjects. However, it is nearly unreachable to meet this condition for continuous gaze direction. As mentioned in section 3.1, for each target gaze hypothesis label, we employ LLR of adjacent source gaze to indicate it. The combination relationships in $\mathbb{G}$ are transferred to $\mathbb{E}$ to generate target hypothesis embedding remaining prediction consistency.

Given a batch of B_s source image samples and B_t target image samples during training, we formally compute the L_{EPC} using,

$$L_{EPC} = \frac{1}{B_t} \sum_{j=1}^{B_t} d(\phi\left(I_j^t\right), \sum_{i=1}^{k} w_{ji}^* \phi(I_i^s)), \quad h(\phi(I_i^s)) \in \mathcal{N}_j, \tag{9}$$

where $L1$ distance is employed as the function d. L_{EPC} measures the distance between the hypothesis and predicted embedding. Furthermore, since target hypothesis embedding is a linear combination of source predicted embedding, L_{EPC} also evaluates the deviation between the source and target domains. During training, as target hypothesis embedding and predicted embedding get closer and closer, the offset between domains is gradually eliminated.

Besides preserving L_{EPC} for domain adaptation, the source domain with groundtruth should also take part in parameter updating to guide training optimizing. The L_{gaze} is calculated based on cosine similarity as,

$$L_{gaze}\left(\hat{g^s},\ g^s\right) = \arccos \frac{\hat{g^s} \times g^s}{\left\|\hat{g^s}\right\| \cdot \|g^s\|}. \tag{10}$$

Algorithm 1 Training Procedure

Input:
Source Domain: $S = \{(I_1^s, g_1^s), \cdots, (I_{N_s}^s, g_{N_s}^s)\}$
Target Domain: $T = \{I_1^t, \cdots, I_{N_t}^t\}$
Output:
Model parameter θ^*

 # First Step: Pre-training in the Source Domain
for m in $[1, N_s]$ **do**
  for (I_i^s, g_i^s) in S **do**
    Forward I_i^s and obtain prediction $\hat{g_i^s}$.
    Back-propagation with Eq.(10) and update network parameters θ.
  end for
end for
 # Second Step: Joint Optimization
for m in $[1, M_t]$ **do**
  Sample a mini-batch B_s and B_t from S and T.
  Obtain prediction $\hat{g^s}$ and $\hat{g^t}$ with forwarding I^s and I^s.
  for b in $[1, B_t]$ **do**
    Select qualified sample set $\mathcal{N}_b$ from B_s for $\hat{g_t^b}$. (Eq. (1))
    if $\|\mathcal{N}_b\| <$ k **then**
      Continue
    else
      Randomly choosing k samples from $\mathcal{N}_b$.
      Obtain LLR representation W^* of $\hat{g_b}$ using k samples (Eq.(5))
      Calculate hypothesis embedding $\hat{\phi}(I_b^t)$ by Eq.(7)
      Compute L_{EPC} using Eq.(9)
    end if
    Compute L_{gaze} for B_s using Eq.(10)
    Back-propagation with Eq.(8) and update network parameters θ.
  end for
end for

3.4 Training

Since the network prediction decides the neighborhood and locally linear gaze representation in L_{EPC}, a well-trained model is necessary to generate credible target hypothesis labels. We pre-train the network using only source domain with groundtruth and the L_{gaze} for N_s epochs at first.

In the joint training procedure, we need to optimize the L_{gaze} and L_{EPC} simultaneously for M_t iterations. We employ an alternative optimization strategy [28] to perform each iteration. Specifically, we first update target hypothesis labels $\hat{g^t}$ with network parameters fixed in each loop and meanwhile estimate the prediction of the source domain. Then given the target label $\hat{g^t}$, we construct $\mathcal{N}_j$ and estimate L_{EPC}. It is worth mentioning that we use the source domain groundtruth for LLR to obtain higher estimation accuracy. Furthermore, network parameters are updated by back-propagation to minimize L_{EPC} and L_{gaze} finally.

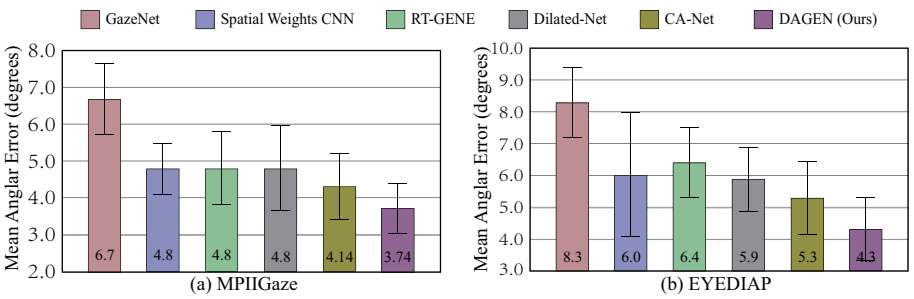

Fig. 4. Performance of gaze estimation on (a)MPIIGaze and (b)EYEDIAP using a leave-one-subject-out strategy. Bars represent the MAE, and the specific value in degrees is on the bottom of each bar; error bars indicate standard deviations.

Algorithm 1 summarizes the entire optimization precessing of our DAGEN. First step performs the essential pre-training, and Second step shows the joint optimization procedure. We asynchronously update the target label and optimize the network to ensure the effectiveness and efficiency during training. We use SGD with momentum $= 0.9$ as the optimizer and a base learning rate of 0.001, l2 weight regularization of 5×10^{-4}. B_s and B_t are both set to 64 during training.

4 Experiments

4.1 Datasets

We implement the proposed Domain Adaptation Gaze Estimation Network on two current gaze datasets: MPIIGaze [7] and EYEDIAP [6].

MPIIGaze is a very challenging dataset for appearance-based in-the-wild gaze direction estimation, because it has high within-subject variations in facial appearance and environments, for instance, make-up, hair change, illumination intensity and direction. We only use the standard evaluation subset MPIIFaceGaze provided by MPIIGaze, which contains 37667 images captured from 15 subjects and has facial keypoints label for image pre-processing.

EYEDIAP contains 94 video sequences of 16 subjects, who were looking at screen targets or physical targets in the collection. Only the videos collected with screen target sessions are used in our training and evaluation set. Note that, since two participants lack the videos in the screen target session, we sample one image every fifteen frames from the other 14 subjects.

4.2 Data Pre-processing

We manipulate the pre-processing procedure similar to [10, 20, 32] to normalize two datasets, and utilize the Surrey Face Model as the reference 3D face model.

Table 1. Comparison of Appearance-Based Gaze Estimation Methods.

Methods	Input	Data	GT	MPIIGaze	EYEDIAP
GazeNet [7]	left eye + head pose	×	×	6.7°	8.3°
SWCNN [20]	face	×	×	4.8°	6.0°
RT-GENE [8]	two eyes + face	×	×	4.8°	6.4°
Dilated-Net [21]	two eyes + face	×	×	4.8°	5.9°
MeNet [11]	face	×	×	4.9°	—
CA-Net [23]	two eyes + face	×	×	4.14°	5.3°
FAZE (3-shot) [13]	eye area	✓	✓	4.1°	—
FAZE (256-shot) [13]	eye area	✓	✓	3.75°	—
DAGEN (ours)	eye area	✓	×	**3.74°**	**4.30°**

In the appearance-based gaze estimation task, the head pose has a significant influence on the accuracy since its six freedoms bring calculational complexity and time-consuming. Consequently, we select four eye corners and two mouth corners described in [10] for PnP-based head pose estimation. Then transfer and rotate the virtual camera according to the head pose to eliminate the impact of position and roll angle.

In our work, considering applying a single image as input cover both eyes, we select the mean of four 3D eye corner landmarks as the gaze origin point to produce groundtruth for source domain. We normalize the camera's intrinsic parameters with a focal length of 960 mm, and a distance of 410 mm from the face to generate image patches of size 256×64 as input for training. In each test period, in order to better verify the effect of domain adaptation, we use the newcomer's entire data without groundtruth as the target domain.

4.3 Comparison with Appearance-Based Methods

We first compare the performance of the proposed method with the state-of-art appearance-based gaze estimation methods. The experiment is carried out in both MPIIGaze and EYEDIAP. For the evaluation protocol, we use leave-one-subject-out strategy on both MPIIGaze and EYEDIAP.

We choose several CNN-based methods proposed from 2015 to 2020 as comparisons, including GazeNet [7], Spatial weights CNN (SWCNN) [20], RT-GENE [8], Dilated-Net [21], MeNet [11], Faze [13] and CA-Net [23].

Although four models ensemble can increase the accuracy of RT-GENE, we do not show the result of that for fairness. Since initializing the model pre-trained on ImageNet can effectively improve accuracy, we apply this strategy for GazeNet, Spatial weights CNN and Dilated-Net refer to [21]. We only present

the results in the author's paper for cases where source codes are not provided, or the evaluation protocol is different from us.

Fig. 4(a) shows the results of MPIIGaze. The Mean Angular Error (MAE) of most work in recent years has become about 4.8° without any person calibration. These methods all have characteristics, such as the use of multi-modal input, the introduction of attention mechanism, the implementation of new training methods, or a new convolution strategy suitable for gaze estimation. CA-Net more cleverly used the coarse-to-fine information from faces to eyes to achieve a breakthrough of about 0.66°. Our method achieves 9.66% to 3.74° comparing to state-of-the-art method CA-Net with only eye area as input.

Fig. 4(b) shows the results in EYEDIAP. Due to the lower image resolution, the performance of EYEDIAP is generally worse than that of MPIIGaze. Many innovations in recent years still bring a significant breakthrough in performance, and the best accuracy obtained in [23] has reached 5.3°. We get an 18.9% increase with the state-of-the-art method to 4.3°.

Table 1 summarizes some differences and results of recent methods for reference, including that not illustrated in Fig. 4. The header *Data* and *GT* show whether the methods need data or groundtruth for a new subject before evaluation. It is noteworthy that some person-specific methods like few-shot (FAZE) have achieved a great improvement for gaze estimation. We show the result of FAZE [13] based on 3-shot and 256-shot within-MPIIGaze leave-one-person-out evaluation. With test images without labels, our method can obtain results close to 256-shot Faze, proving the effectiveness of domain adaptation.

4.4 Ablation Study

We further evaluate our method under different settings to better demonstrate the effectiveness of our various design choices in the DAGEN. For all ablation experiments, the source domain and test set's selection follows the leave-one-subject-out strategy on the MPIIGaze dataset.

Contribution of Domain Adaptation We first perform an ablation study to demonstrate the effect of domain adaptation. Specifically, we evaluate the consequence of adding domain adaptation, the impact of different target domain data, and the influence of domain adaptation objects. Table 2 shows the experimental results and the only change is *DA* is the choice of target domain.

Without DA shows the baseline model supervised by the L_{gaze} during training, having the MAE of 4.84°. In order to better assess the impact of target domain data on accuracy, we compare the estimation accuracy using GazeCapture [18] as the target domain. For *GC*, we randomly sample 20 images for each participant in GazeCapture. With a total of 1366 subjects and 27320 images, we get the MAE of 4.17°. Moreover, we randomly select 100 participants to discuss the influence of diversity in the target domain, named *GCsubset*. *Eval* uses the evaluation set as the target domain. The results show that utilizing diverse and targeted samples as the target domain can effectively improve the estimation performance, which may have reference significance for practical application.

Table 2. Comparison on different DA configurations.

Without DA	DA			
	GCsubset	GC	Pred	Eval
4.84°	4.66°	4.17°	3.99°	**3.74°**

For our proposed method described in Section 3, we use the target hypothesis label and the source groundtruth as the domain adaptation targets. *Pred* takes the source predicted value instead of groundtruth as the domain adaptive target, with the MAE of 3.99°. Since in this case, errors in the source domain data would also affect the domain adaptation process. In other words, using groundtruth as the DA target produces more substantial constraints for the updating direction of the parameters.

Effect of Feature Representation LLR utilizes k source groundtruth to represent a target hypothesis label. We evaluate the different choices of k, shown in Fig. 5(a). Generally, a higher k means more stable and robust LLR. However, because we select the appropriate sample from a mini-batch, a higher k brings a smaller probability of reaching the selection condition. In our experimental protocol, the calculating speed of EPC loss is from $32.17 - 32.53ms/iter$ in one Nvidia 1080Ti for different k, and the training speed is $76.5ms/iter$ in training.

Our embedding $\phi(I)$ has the dimension of F_g. Considering $\phi(I)$ perform prediction consistency, different F_g would lead to changes in characterization ability and robustness. We evaluate the accuracy of DAGEN for different dimensions $F_g = \{8, 16, 32, 64\}$ to select the most suitable one. Fig. 5(b) shows the result of dimension selection. In our experiments, our method is not sensitive to F_g, indicating that our method is very robust for F_g.

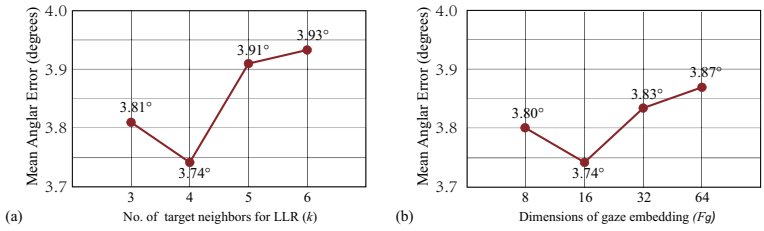

Fig. 5. Impact of different Feature Representation choice

Empirically we find $k = 4$ and $F_g = 16$ to be optimal hence select it.

Effect of Pre-trained Model We use ResNet-18 pre-trained on ImageNet as the backbone. And before the domain adaptation training, the network is

Table 3. Impact of Pre-trained Methods.

ImageNet	Source	MAE
×	×	4.63°
✓	×	4.2°
×	✓	4.51°
✓	✓	**3.74°**

Table 4. Impact of Selection Interval μ.

μ	MAE
0.05	4.04°
0.15	**3.74°**
0.3	3.84°
all	3.86°

first trained for five epochs in the source domain. We evaluated the estimation accuracy of whether the two pre-trainings participate, shown in Table 3, to show the contribution of the two pre-training strategies.

For the model pre-trained on ImageNet, it can effectively avoid the parameters falling into the local optimum, thereby improving the gaze estimation accuracy. In the case of pre-training in the source domain, obtaining a more accurate hypothesis label can significantly improve prediction accuracy. However, while the parameters fall into the local optimum, the quality of the hypothesis label is not improved, so the error is not substantially reduced.

Impact of Selection Range μ The target hypothesis label needs to be represented by the appropriate source groundtruth. We have defined this selection strategy in Eq. (1), where parameter μ indicates the select interval. We perform the impact of different choices of μ on estimation accuracy, shown in Table 4.

We can see that the estimation accuracy has not changed much when $\mu \geq 0.15$. The results reveal that although we established a locally linear relationship in the gaze space $\mathbb{G}$ and the embedding space $\mathbb{E}$, due to the linear mapping h, the network tends to exhibit a global linear relationship in $\mathbb{G}$ and $\mathbb{E}$. For the case where μ is small, few target samples can participate in DA training. Therefore, the network is straightforward to fall into overfitting, which significantly increases estimation error and even is challenging to converge.

4.5 Visual Results

We display some results in Fig. 6 to show the effectiveness of our method. Fig. 6(a-b) performs the scatter plot and linear fit of the pitch angles, which are predicted by the baseline model and our proposed DAGEN method on the evaluation set. Obviously, the fixed bias between the prediction and the groundtruth is significantly reduced in our method. Furthermore, We randomly pick several samples close to the fitted line (yellow points in Fig 6(a-b) in the baseline results and visualize the result of both baseline and DAGEN models in Fig 6(c-d). We can see that our DAGEN can produce accurate gaze directions with tiny deviations for the evaluated subject in different appearances and illumination.

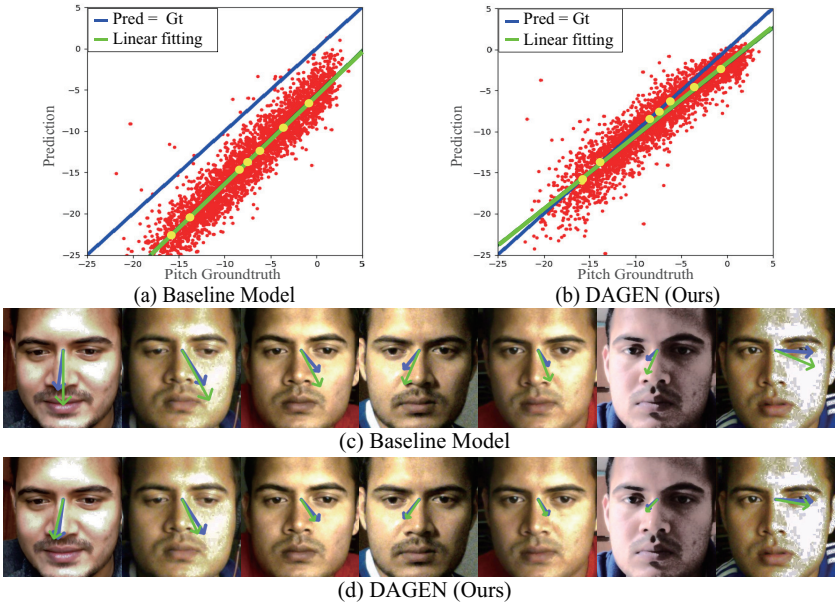

Fig. 6. Visible results of the evaluation set. (a) and (b) show the scatter (red) and linear fit line of pitch angles predicted with the baseline and our DAGEN. Yellow points in (a-b) are samples randomly selected from the linear fit line in (a). The groundtruth(blue) and prediction(green) of the chosen samples are displayed orderly in (c) and (d).

5 Conclusion

In this paper, we propose an unsupervised method for domain adaptation gaze estimation by embedding with prediction consistency. We utilize source groundtruth to perform a locally linear representation for target gaze estimation. The linear relationships are then inherited from gaze space to embedding space to perform prediction consistency. Moreover, we minimize the distance between the target hypothesis embedding and predicted embedding, which measures the deviation between the source and target domain. We experimentally showed that our approach dramatically reduces the impact of inter-personal differences and achieves state-of-the-art performance in MPIIGaze and EYEDIAP.

Acknowledgement

This work was supported by the National Key R&D Program of China (2016YFB 1001001), the National Natural Science Foundation of China (61976170, 91648121, 61573280), and Tencent Robotics X Lab Rhino-Bird Joint Research Program (201902, 201903). (Portions of) the research in this paper used the EYEDIAP dataset made available by the Idiap Research Institute, Martigny, Switzerland.

References

1. Fridman, L., Reimer, B., Mehler, B., Freeman, W.T.: Cognitive load estimation in the wild. In: CHI. (2018) 652
2. Konrad, R., Angelopoulos, A., Wetzstein, G.: Gaze-contingent ocular parallax rendering for virtual reality. ACM Trans. Graph. **39** (2020) 10:1–10:12
3. Vicente, F., Huang, Z., Xiong, X., la Torre, F.D., Zhang, W., Levi, D.: Driver gaze tracking and eyes off the road detection system. IEEE Trans. Intell. Transp. Syst. **16** (2015) 2014–2027
4. Kassner, M., Patera, W., Bulling, A.: Pupil: an open source platform for pervasive eye tracking and mobile gaze-based interaction. In: UbiComp Adjunct. (2014) 1151–1160
5. Sugano, Y., Matsushita, Y., Sato, Y.: Learning-by-synthesis for appearance-based 3d gaze estimation. In: CVPR. (2014) 1821–1828
6. Mora, K.A.F., Monay, F., Odobez, J.: EYEDIAP: a database for the development and evaluation of gaze estimation algorithms from RGB and RGB-D cameras. In: ETRA. (2014) 255–258
7. Zhang, X., Sugano, Y., Fritz, M., Bulling, A.: Appearance-based gaze estimation in the wild. In: CVPR. (2015) 4511–4520
8. Fischer, T., Chang, H.J., Demiris, Y.: RT-GENE: real-time eye gaze estimation in natural environments. In: ECCV (10). Volume 11214. (2018) 339–357
9. Kellnhofer, P., Recasens, A., Stent, S., Matusik, W., Torralba, A.: Gaze360: Physically unconstrained gaze estimation in the wild. In: ICCV. (2019) 6911–6920
10. 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** (2019) 162–175
11. Xiong, Y., Kim, H.J., Singh, V.: Mixed effects neural networks (menets) with applications to gaze estimation. In: CVPR. (2019) 7743–7752
12. Liu, G., Yu, Y., Mora, K.A.F., Odobez, J.: A differential approach for gaze estimation with calibration. In: BMVC. (2018) 235
13. Park, S., Mello, S.D., Molchanov, P., Iqbal, U., Hilliges, O., Kautz, J.: Few-shot adaptive gaze estimation. In: ICCV. (2019) 9367–9376
14. Roweis, T., S., Saul, K., L.: Nonlinear dimensionality reduction by locally linear embedding. Science (2000)
15. Hansen, D.W., Ji, Q.: In the eye of the beholder: A survey of models for eyes and gaze. IEEE Trans. Pattern Anal. Mach. Intell. **32** (2010) 478–500
16. Morimoto, C.H., Amir, A., Flickner, M.: Detecting eye position and gaze from a single camera and 2 light sources. In: ICPR (4). (2002) 314–317
17. Yoo, D.H., Chung, M.J.: A novel non-intrusive eye gaze estimation using cross-ratio under large head motion. Comput. Vis. Image Underst. **98** (2005) 25–51
18. Krafka, K., Khosla, A., Kellnhofer, P., Kannan, H., Bhandarkar, S.M., Matusik, W., Torralba, A.: Eye tracking for everyone. In: CVPR. (2016) 2176–2184
19. Lecun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE **86** (1998) 2278–2324
20. Zhang, X., Sugano, Y., Fritz, M., Bulling, A.: It's written all over your face: Full-face appearance-based gaze estimation. In: CVPR Workshops. (2017) 2299–2308
21. Chen, Z., Shi, B.E.: Appearance-based gaze estimation using dilated-convolutions. In: ACCV (6). Volume 11366. (2018) 309–324
22. Palmero, C., Selva, J., Bagheri, M.A., Escalera, S.: Recurrent CNN for 3d gaze estimation using appearance and shape cues. In: BMVC. (2018) 251

23. Cheng, Y., Huang, S., Wang, F., Qian, C., Lu, F.: A coarse-to-fine adaptive network for appearance-based gaze estimation. In: AAAI. (2020) 10623–10630
24. Tzeng, E., Hoffman, J., Saenko, K., Darrell, T.: Adversarial discriminative domain adaptation. In: CVPR. (2017) 2962–2971
25. Wang, J., Feng, W., Chen, Y., Yu, H., Huang, M., Yu, P.S.: Visual domain adaptation with manifold embedded distribution alignment. In: ACM Multimedia. (2018) 402–410
26. Long, M., Cao, Y., Wang, J., Jordan, M.I.: Learning transferable features with deep adaptation networks. In: ICML. Volume 37. (2015) 97–105
27. Long, M., Zhu, H., Wang, J., Jordan, M.I.: Deep transfer learning with joint adaptation networks. In: ICML. Volume 70. (2017) 2208–2217
28. Kang, G., Jiang, L., Yang, Y., Hauptmann, A.G.: Contrastive adaptation network for unsupervised domain adaptation. In: CVPR. (2019) 4893–4902
29. Yu, Y., Liu, G., Odobez, J.: Improving few-shot user-specific gaze adaptation via gaze redirection synthesis. In: CVPR. (2019) 11937–11946
30. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: NIPS. (2012) 1106–1114
31. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. (2016) 770–778
32. Zhang, X., Sugano, Y., Bulling, A.: Revisiting data normalization for appearance-based gaze estimation. In: ETRA. (2018) 12:1–12:9