Contrastive Regression for Domain Adaptation on Gaze Estimation

Yaoming Wang[1,2†], Yangzhou Jiang[1,2†], Jin Li[1], Bingbing Ni[1], Wenrui Dai[1],
Chenglin Li[1], Hongkai Xiong[1], and Teng Li[2,3*]
[1]Shanghai Jiao Tong University [2]Huawei Inc [3]Anhui University
{wang_yaoming, jiangyangzhou, deserve_lj, nibingbing,
daiwenrui, lcl1985, xionghongkai}@sjtu.edu.cn; tenglwy@gmail.com

Abstract

Appearance-based Gaze Estimation leverages deep neural networks to regress the gaze direction from monocular images and achieve impressive performance. However, its success depends on expensive and cumbersome annotation capture. When lacking precise annotation, the large domain gap hinders the performance of trained models on new domains. In this paper, we propose a novel gaze adaptation approach, namely Contrastive Regression Gaze Adaptation (CRGA), for generalizing gaze estimation on the target domain in an unsupervised manner. CRGA leverages the Contrastive Domain Generalization (CDG) module to learn the stable representation from the source domain and leverages the Contrastive Self-training Adaptation (CSA) module to learn from the pseudo labels on the target domain. The core of both CDG and CSA is the Contrastive Regression (CR) loss, a novel contrastive loss for regression by pulling features with closer gaze directions closer together while pushing features with farther gaze directions farther apart. Experimentally, we choose ETH-XGAZE and Gaze-360 as the source domain and test the domain generalization and adaptation performance on MPIIGAZE, RT-GENE, GazeCapture, EyeDiap respectively. The results demonstrate that our CRGA achieves remarkable performance improvement compared with the baseline models and also outperforms the state-of-the-art domain adaptation approaches on gaze adaptation tasks.

1. Introduction

With the development of deep learning, gaze estimation techniques have been widely applied in human-computer interaction systems, such as intelligent cockpits [11], VR/AR games [2, 21, 38], medical analysis [4], etc. Recently, appearance-based approaches [23, 37] are attracting more and more attention, as they regress gaze direction from monocular images alone and get rid of expensive and limited eye model devices. Despite the success of appearance-based gaze estimation, expensive and cumbersome annotation capture constrains its application in daily life. Large-scale gaze datasets [9, 10, 19, 23, 35, 39] along with related gaze estimation approaches have been proposed to alleviate this problem. These approaches yield promising performance in the within-dataset test (training and test data are from a same dataset) but are degraded dramatically in the cross-dataset test (training and test data are from different datasets), due to the gap between different domains, such as the differences of subjects, background environments, and illuminations.

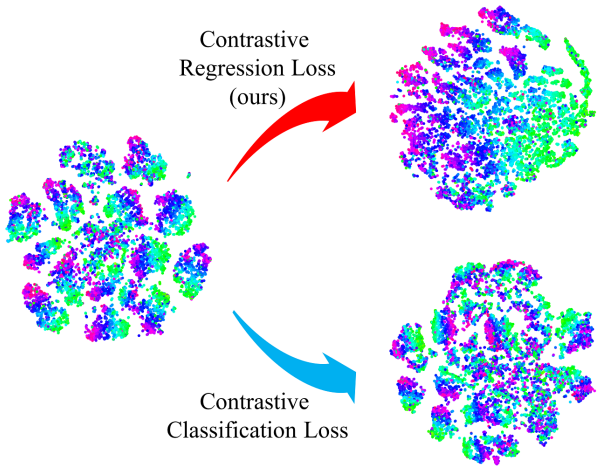

Figure 1. Illustration of the feature distribution learned from domain adaptation task $\mathcal{D}_G \rightarrow \mathcal{D}_D$ (with close gaze directions share similar colors) indicates that the original contrastive classification loss function exhibits no effect on regression problem, while our derived contrastive regression loss pulls features with close gaze labels together and pushes features with remote gaze labels apart.

Recently, collaborative model ensembles [25] and additional annotations [22] are leveraged to narrow the cross-dataset gap. They require additional models or annotations for domain adaptation and lead to extra complexity of the

*Corresponding author: Teng Li. †: Equal contribution.

learning pipeline. For the same dataset, inter-person gap (i.e., personal calibration) can be alleviated by learning the personal error between the visual axis and optical axis with adversarial training [28, 33] and few shot learning [27, 34]. However, there still lacks a self-supervised approach to address the cross-dataset gap without introducing additional labels or models.

Contrastive learning is dominating recent advances in self-supervised learning [24] and has been transferred to various downstream tasks including classification, segmentation and detection [16]. However, existing methods [3, 13, 16] for classification tasks on datasets like ImageNet and CIFAR cannot be directly extended to regression tasks. Fig. 1 illustrates an example where the standard contrastive learning for classification tasks fails to learn useful representations for gaze regression tasks. In fact, existing unsupervised and supervised constrastive learning for classification cannot accommodate to gaze regression tasks.

- Unsupervised contrastive learning treats different views of an image as the positives and the views of other images as the negatives. It tends to extract global semantic information that benefits classification tasks, e.g., information for face recognition. However, global semantic information might mislead regression tasks, especially appearance-based gaze direction regression, and compromises the accuracy of gaze estimation.

- Supervised contrastive learning [20] deems images with the same label as the positives and degenerates to unsupervised contrastive learning given continuous gaze annotations (labels are different from each other in a batch). Moreover, in classification tasks, different labels indicate different categories and does not reveal meaningful information. In contrast, the relationship between labels reveal the relationship between the features for regression tasks.

In this paper, we propose a novel gaze adaptation approach, namely Contrastive Regression Gaze Adaptation, for generalizing gaze estimation on the target domain in an unsupervised manner. We first derive a novel contrastive regression loss for regression tasks by assuming the similarity between labels is proportional to the ratio of the related features. Subsequently, we develop two modules, i.e., Contrastive Domain Generalization (CDG) and Contrastive Self-training Adaptation (CSA), based on the contrastive regression loss for Contrastive Regression Gaze Adaptation. CDG introduces the contrastive regression loss into the domain generalization task to learn stable representation from the source domain, whereas CSA incorporates the pseudo label generated from the source domain model and the CDG loss to further improve the adaptation performance in the target domain. The contributions of this paper are summarized as below.

- We develop a novel gaze adaptation method, namely Contrastive Regression Gaze Adaptation (CRGA), for self-supervised cross-domain gaze estimation without introducing additional models or labels.

- We propose a novel Contrastive Regression framework based on the derived Contrastive Regression (CR) loss to learn robust domain-invariant representation for regression tasks.

To our best knowledge, we are the first to introduce contrastive learning into regression tasks to improve the domain generalization and adaptation performance dramatically. Experimental results demonstrate that CRGA achieves remarkable performance improvements compared with the baseline models and outperforms the state-of-the-art domain adaptation approaches on gaze adaptation tasks. Specifically, CRGA achieves performance improvements over the baseline of 40.4%, 34.7%, 55.8%, 34.3%, from source domain ETH-XGAZE to MPIIGaze, RT-GENE, GazeCapture, and EyeDiap respectively. Besides, CRGA achieves improvements over the baseline of 31.7%, 30.5%, 32.9% and 23.8% from source domain Gaze360 to MPII, RT-GENE, GazeCapture, and EyeDiap respectively.

2. Related Works

2.1. Domain Adaptive Gaze Estimation

While deep neural networks learn image-to-gaze mapping effectively, performance degrades severely on the new domain. Many efforts are dedicated to alleviating this problem. From the perspective of data, large-scale and diverse gaze datasets are collected to meet the real-world setting, such as GazeCapture [23], ETH-XGaze [35] and Gaze360 [19] etc. To align the input data distribution cross domain, standard data preprocessing methods are proposed to map the input data to a normalized space [30,36]. Specifically, work [36] uses a virtual camera to warp the face patch according to 3d head pose. Besides, multiple GAN methods are leveraged to align the input data distribution between different domain [28, 32].

From the perspective of learning, some methods attempt to learn a more general representation for gaze or align the feature distribution between two domains. Work [27] learns a rotation-aware latent representation of gaze with meta-learning. Adversarial training is commonly used for aligning the feature distribution [19, 33] and purify gaze feature [7]. For instance, Kellnhofer *et al.* [19] finetune a mixture of the labeled Gaze360 images and unlabeled images along with a discriminator to identify the domain. Besides, Liu *et al.*. [25] propose to use an ensemble of networks to learn collaboratively with the guidance of outliers.

2.2. Contrastive Learning

Recently, contrastive learning shows superior performance in self-supervised and semi-supervised learning, and even surpasses supervised methods when transferring the representation to cross-domain and downstream tasks [3, 16, 18]. The idea of contrastive learning is to learn representation by contrasting multi-views of samples as positive pairs against other negative samples [5, 15]. It could also be interpreted as maximizing mutual information between latent representations [1, 17, 26], and noise contrastive estimation [14] could be leveraged. He *et al.* [16] extend negative samples in mini-batch to large momentum update memory bank. Chen *et al.* [5, 6] then find a non-linear projection head matters and large batch also helps. [29] proposes to use the contrastive loss to capture the equivariance with respect to geometric transformations in 3D head pose estimation. Grill *et al.* [12] even manage to remove negative samples totally. Contrastive learning could also be expanded to the supervised scenes. Khosla *et al.* [20] propose to build positive pairs with data augmentation together with annotation label. Note that the aforementioned methods are designed for a general classification dataset, which could hardly apply to gaze regression scene.

3. Methodology

3.1. Preliminary: Domain Adaptation

Given the source domain data as $\mathcal{D}_{\mathcal{S}} = \{(x_n^{\mathcal{S}}, g_n^{\mathcal{S}})\}_{n=1}^{N_s}$, where $(x_n^{\mathcal{S}}, g_n^{\mathcal{S}})$ denotes the n-th pair of observation and the corresponding gaze direction, $N_{\mathcal{S}}$ is the number of pairs. Similarly, the target domain data is denoted as $\mathcal{D}_{\mathcal{T}} = \{(x_n^{\mathcal{T}}, g_n^{\mathcal{T}})\}_{n=1}^{N_{\mathcal{T}}}$, where $(x_n^{\mathcal{T}}, g_n^{\mathcal{T}})$ represents the n-th pair and $N_{\mathcal{T}}$ is the number of pairs. For domain adaptation, our goal is to learn a predictive function $f : x \rightarrow g$ on the source domain $\mathcal{S}$ to achieve a minimum error on the target domain $\mathcal{T}$ as:

$$\min_f \mathbb{E}_{(x^{\mathcal{T}}, y^{\mathcal{T}})}[\mathcal{L}(f(x^{\mathcal{T}}), g^{\mathcal{T}})] \tag{1}$$

3.2. Contrastive Regression

Recently, contrastive learning exhibits strong power in learning stable representation for domain adaptation on classification tasks. However, no well-elaborated contrastive method is proposed for domain adaptation on regression tasks. Thus, we propose a novel contrastive regression framework to learn robust and invariant representation for regression tasks. For regression model, different from classification tasks, the relationship between labels reveal the relationship between the features. Then we can make a assumption that the ratio between predict distribution $p(y_i|x)$ and $p(y_k|x)$ is proportional to the similarity between label distribution $p(g_i)$ and $p(g_j)$.

Proposition 1. *We derive the new contrastive loss function for regression task as*

$$-\log \frac{\sum_k \mathbb{S}[p(g_i); p(g_k)] \cdot f_k(y_k, x)}{\sum_j f_j(y_j, x)} \tag{2}$$

where $f_i(y_i, x)$ is the density ratio.

Proof. Please refer to the supplementary material. □

For simplicity, we abbreviate the similarity function $\mathbb{S}[p(g_i); p(g_k)]$ as $\mathbb{S}_{i,k}$ in the rest paper. This loss still encounters some problems in practice. Specifically, despite that the loss function encourages $\sum_k \mathbb{S}_{i,k} \cdot f_k(y_k, x)$ to get larger to approximate $\sum_j f_j(y_j, x)$, negative values may appear in the beginning and results in NAN in the loss computation. Then, we introduce the variant of Eq. 2 as:

$$-\log \frac{\sum_k \sigma(\mathbb{S}_{i,k}) \cdot f_k(y_k, x)}{\sum_j f_j(y_j, x)} \tag{3}$$

where $\sigma(\cdot)$ is the relu function, used to zero out negative values. Besides, this loss function would not be bounded if $\mathbb{S}_{i,k}$ tends to infinity. Thus, we further introduce the normalization and rewritten Eq. 3 as:

$$-\log \frac{\sum_k \sigma(\mathbb{S}_{i,k}) \cdot f_k(y_k, x)}{\sum_j |\mathbb{S}_{i,j}| \cdot f_j(y_j, x)} \tag{4}$$

where $|\mathbb{S}_{i,j}|$ is the absolute value of our similarity $\mathbb{S}$. As $f_j(y_j, x)$ always takes the exponential distribution and greater than zero, the loss function has the lower bound as $\mathcal{L} \geq -\log 1 = 0$. We name the loss function in Eq. 4 as the Contrastive Regression (CR) loss.

Proposition 2. *The two forms of loss function in Eq. 2 and Eq. 4 have the same effect, i.e., they pull features with closer gaze directions closer together while pushing features with farther gaze directions farther apart.*

Proof. Please refer to the supplementary material. □

Similarity function Considering that the gaze direction is mainly concentrated in front of the face and the gradient near zero of cosine similarity is too small, we derive a -log KL function as the similarity:

$$\mathbb{S}_{i,j} = -\log \frac{|g_i - g_j|}{0.07} = \log \frac{0.07}{|g_i - g_j|} \tag{5}$$

The detail of why we derive this function and the property of this similarity function can be referred to in the supplementary materials.

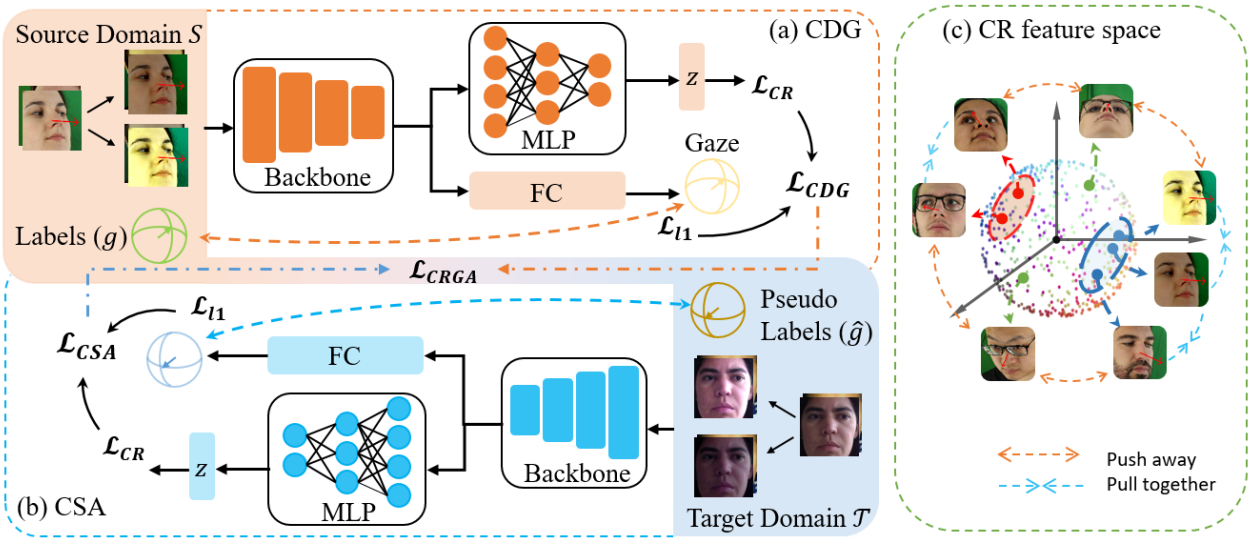

Figure 2. The overall framework of our proposed CRGA (best viewed in color). (a)Contrastive Domain Generalization (CDG) module: Input images are first augmented into two different view, and then be fed into the network for gaze and CR feature, which could be used for computing L_{CDG} (b) The Contrastive Self-training Adaptation (CSA) Module: Pseudo labels are generated for unlabeled images in target domain and used for computing L_{CSA} similarly. L_{CSA} and L_{CDG} as regulation compose the final loss L_{CRGA}. (c) Visualization of the CR normalized feature space. Features with blue arrow would be pulled together, while features with orange arrow would be pushed apart.

3.3. Contrastive Regression Gaze Adaptation

In this section, we will elaborate on the proposed Contrastive Regression Gaze Adaptation (CRGA). Fig. 2 illustrates the overall framework of CRGA that consists of two modules. The first module is Contrastive Domain Generalization (CDG), which leverages CDG loss to learn a stable representation from the source domain. Subsequently, the Contrastive Self-training Adaptation (CSA) module leverages the contrastive self-training with pseudo labeling to improve the adaptation performance on the target domain.

3.3.1 Contrastive domain generalization

Given source domain data, we follow the convention in contrastive learning and leverages two separate data augmentation operators A, $\tilde{A}$ to get two views of input image as $I = A(input), \tilde{I} = \tilde{A}(input)$. The two separate data augmentation operators are sampled from the same family $\mathcal{A}$ of augmentations as $A \sim \mathcal{A}, \tilde{A} \sim \mathcal{A}$. The detail of the data augmentation family $\mathcal{A}$ can be referred to in the Sec. 4.2. Then, I and $\tilde{I}$ are fed to the model $f(\cdot)$ to get the features $V = f(I)$ and $\tilde{V} = f(\tilde{I})$. Usually, a parametric gaze predictive head $h(\cdot)$ is employed to predict the gaze distribution as $y = h(v)$. Considering a minibatch of N examples and we get pairs of augmented examples derived from the minibatch, resulting in 2N data points. Following [5], a projection head $r(\cdot)$, usually a multi-layer perceptron (mlp), is employed to map features to the space where contrastive loss is applied and get $z = r(v)$. We follow [5] and further introduce the cosine similarity for the ℓ_2 normalized z with the temperature parameter τ. Thus, we could turn $f_k(u_k, x)$ into $exp(sim(z_i, z_k)/\tau)$. When $k = j$, we have $sim(z_k, z_j) = 1$. To further encourage the simplify and the fast convergence, we follow [5] and introduce the indicator function $\mathbb{1}_{\mathrm{j}\neq\mathrm{i}}$ to omit the i_th sample. Finally, the CR loss defined in Eq. 4 can be converted as:

$$-\log \frac{\sum_k \mathbb{1}_{\mathrm{k}\neq\mathrm{i}} \sigma(\mathbb{S}_{i,k}) \cdot exp(sim(z_i, z_k)/\tau)}{\sum_j \mathbb{1}_{\mathrm{j}\neq\mathrm{i}} |\mathbb{S}_{i,j}| \cdot exp(sim(z_i, z_j)/\tau)} \quad (6)$$

where the indicator function$\mathbb{1}_{\mathrm{j}\neq\mathrm{i}}$ evaluating to 1 if $j \neq i$, while evaluating to 0 when $j = i$. The replacement of the parametric predictive head by the non-parametric predictive head would lose supervised information from the gaze labels. Thus we leverage additive $\ell 1$ loss to our CR loss. For clarification, we simplify $exp(sim(z_i, z_k)/\tau)$ to $e_\tau(z_i, z_k)$. Then we derive our final loss function L_{CDG} as:

$$-\log \frac{\sum_k \mathbb{1}_{\mathrm{k}\neq\mathrm{i}} \sigma(\mathbb{S}_{i,k}) \cdot e_\tau(z_i, z_k)}{\sum_j \mathbb{1}_{\mathrm{j}\neq\mathrm{i}} |\mathbb{S}_{i,j}| \cdot e_\tau(z_i, z_j)} + \frac{\gamma}{2N} \sum_i |y_i - g_i| \quad (7)$$

Here g_i is the gaze label, γ is the hyperparameter and the related ablation study can be found in Sec. 4.5.

3.3.2 Contrastive self-training adaptation

In this section, we elaborate on Contrastive Self-training Adaptation (CSA). Firstly, we consider the source-free domain adaptation (SFDA), where source data is unavailable and only the pre-trained source model could be accessed. Given the target domain data $\mathcal{D}_\mathcal{T} = \{(x_n^\mathcal{T}, g_n^\mathcal{T})\}_{n=1}^{N_\mathcal{T}}$, the target model $f^\mathcal{T}(\cdot)$ and the target gaze predictive head $r^\mathcal{T}(\cdot)$ (initialized by the pre-trained source model and head). Then the target model generates the pseudo gaze direction as $\hat{g}^\mathcal{T} = r^\mathcal{T}(f^\mathcal{T}(x^\mathcal{T}))$. We apply the pseudo gaze direction $\hat{g}^\mathcal{T}$

Algorithm 1 Domain Adaptation (CRGA)

Input: Source Data $\mathcal{D}_{\mathcal{S}} = \{(x_n^{\mathcal{S}}, g_n^{\mathcal{S}})\}_{1:N_{\mathcal{S}}}$, target data $\mathcal{D}_{\mathcal{T}} = \{x_n^{\mathcal{T}}\}_{1:N_{\mathcal{T}}}$, and pretrained network M on $\mathcal{D}_{\mathcal{S}}$ with weights θ, where M contains feature extractor f, gaze predictive head h and CR projection head r.

Output: Adapted network M.

Part 1: Train M on $\mathcal{D}_{\mathcal{S}}$ with CDG loss 7. (CDG)
while not converged **do**
  Sample batch data $(x^{\mathcal{S}}, g^{\mathcal{S}})$ from $\mathcal{D}_{\mathcal{S}}$.
  Random augment data: $I^{\mathcal{S}} = A(x^{\mathcal{S}})$ and $\tilde{I}^{\mathcal{S}} = \tilde{A}(x^{\mathcal{S}})$
  Feed $x_1^{\mathcal{S}}$ and $x_2^{\mathcal{S}}$ to model M, and get gaze prediction $g^{\mathcal{S}} = h(f(I^{\mathcal{S}}, \tilde{I}^{\mathcal{S}}))$ and CR feature $r(f(I^{\mathcal{S}}, \tilde{I}^{\mathcal{S}}))$.
  Compute CDG Loss over gaze prediction and CR feature according to Eq. 7, and optimize $M(\theta)$.
end while
Part 2: Adapt M to $\mathcal{D}_{\mathcal{T}}$ with CRGA loss in Eq. 9 and pseudo labels $\hat{g}_n^{\mathcal{T}}$. (CSA)
while not converged **do**
  In the beginning of each epoch: generate pseudo labels $\hat{g}_n^{\mathcal{T}}$ for target data with network M.
  Sample batch data $(x^{\mathcal{T}}, \hat{g}^{\mathcal{T}})$ from $\mathcal{D}_{\mathcal{T}}$
  Random augment data: $I^{\mathcal{T}} = A(x^{\mathcal{T}})$ and $\tilde{I}^{\mathcal{T}} = \tilde{A}(x^{\mathcal{T}})$
  Feed $I^{\mathcal{T}}$ and $\tilde{I}^{\mathcal{T}}$ to model M, and update $M(\theta)$ by descending CRGA loss in Eq. 9 .
end while

as the label of the target data and leverage our CR loss combined with the L1 loss proposed in Eq. 7 to learn a stable representation and more precise prediction for target data. The related loss function L_{CSA} is:

$$-\log \frac{\sum_k \mathbb{1}_{\mathrm{k}\neq\mathrm{i}} \sigma(\hat{\mathbb{S}}_{i,k}) \cdot e_\tau(z_i, z_k)}{\sum_j \mathbb{1}_{\mathrm{j}\neq\mathrm{i}} \left|\hat{\mathbb{S}}_{i,j}\right| \cdot e_\tau(z_i, z_j)} + \frac{\epsilon}{2N} \sum_i |y_i - \hat{g}_i| \quad (8)$$

where $\hat{\mathbb{S}}_{i,k} = \mathbb{S}[p(\hat{g}_i); p(\hat{g}_k)]$, ϵ is the hyperparameter (the ablation study is presented in Secondly, we further consider the scene where source data is available. Then the source data can be used as the regularization term at the beginning of self-training. With an annealed temperature γ (gradually degrade to 0 from 1), we derive the final source data available domain adaptation loss as:

$$L_{CRGA} = L_{CSA} + \gamma \cdot L_{CDG} \quad (9)$$

The ablation study on whether source data can be accessed is shown in Sec. 4.4.

4. Experiments

4.1. Datasets

We employ six gaze datasets as six different domains: ETH-XGaze ($\mathcal{D}_E$), Gaze360 ($\mathcal{D}_G$), MPIIGaze ($\mathcal{D}_M$), RT-GENE ($\mathcal{D}_R$), GazeCapture ($\mathcal{D}_C$), and EyeDiap ($\mathcal{D}_D$). We choose ETH-XGaze and Gaze360 as the source domain and test the generalization and adaptation performance on MPIIGaze, RT-GENE, GazeCapture, and EyeDiap respectively. We follow [8, 25] to pre-process the gaze datasets and eliminate the influence of different head poses through rotating the virtual camera and wrapping the images. The details of the six datasets are in the supplementary materials. The visualization of different datasets is shown in Fig. 3.

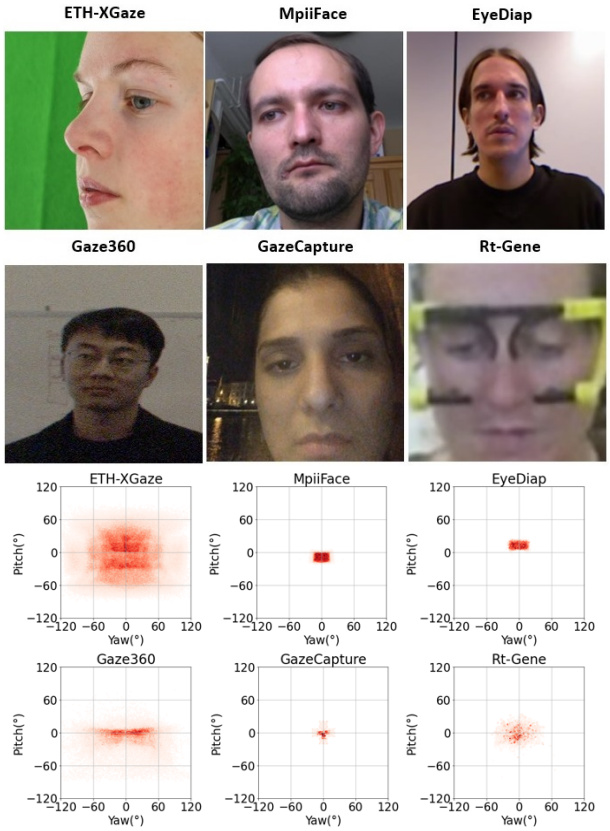

Figure 3. Illustration of the gaze direction distributions of six gaze datasets (best viewed in color). The top 2 rows are the images samples from six gaze datasets. The bottom two rows are the gaze direction distribution statistics.

4.2. Experimental Details

Please refer to the supplementary materials.

4.3. Domain Generalization

For domain generalization, the target images are not available during the training. We train our baseline model and our CDG model only using the source domain data. Results are exhibited in Tab. 1. When we take ETH-XGAZE as the source domain, we train our baseline model following the pipeline of [35] and reach 4.47° evaluation error consistent with 4.5° reported in [35]. When we take Gaze360 as the source domain, we train our baseline model following the pipeline of [19] and reach 10.9° evaluation error

Method	Source	CDG	CSA	$\mathcal{D}_E \rightarrow \mathcal{D}_M$	$\mathcal{D}_E \rightarrow \mathcal{D}_R$	$\mathcal{D}_E \rightarrow \mathcal{D}_C$	$\mathcal{D}_E \rightarrow \mathcal{D}_D$
Baseline.orig [35]	4.50	-	-	7.5	31.2	10.5	11.0
Baseline.our impl	4.47	-	-	9.19	18.23	13.43	8.62
CDG	4.56	✓	×	6.73 ∇26.7%	16.45 ∇9.8%	9.23 ∇31.2%	7.95 ∇7.8%
CSA	-	×	✓	**5.37** ∇41.6%	14.06 ∇22.9%	8.25 ∇38.6%	6.77 ∇21.5%
CRGA	-	✓	✓	5.48 ∇40.4%	**11.91** ∇34.7%	**5.94** ∇55.8%	**5.66** ∇34.3%
Method	Source	CDG	CSA	$\mathcal{D}_G \rightarrow \mathcal{D}_M$	$\mathcal{D}_G \rightarrow \mathcal{D}_R$	$\mathcal{D}_G \rightarrow \mathcal{D}_C$	$\mathcal{D}_G \rightarrow \mathcal{D}_D$
Baseline.orig [35]	11.1	-	-	10.3	26.6	12.9	11.3
Baseline.our impl	10.9	-	-	8.63	23.36	9.23	8.52
CDG	11.0	✓	×	7.03 ∇18.5%	20.79 ∇11.0%	8.28 ∇10.3%	7.27 ∇14.7%
CSA	-	×	✓	7.30 ∇15.4%	21.32 ∇9.6%	7.99 ∇13.4%	7.73 ∇9.3%
CRGA	-	✓	✓	**5.89** ∇31.7%	**16.23** ∇30.5%	**6.19** ∇32.9%	**6.49** ∇23.8%

Table 1. Domain adaptation results compared with baselines. Angular gaze error (°) is used as evaluation metric.

consistent with 11.1° reported in [19]. We leverage CDG loss to train our CDG module in the source dataset and Our CDG module achieves remarkable performance improvement compared with the baseline models. Specifically, CDG achieves performance improvements over the baseline of 26.7%, 9.8%, 31.2%, 7.8%, from source domain ETH-XGAZE to MPIIGaze, RT-GENE, GazeCapture, and EyeDiap respectively. Besides, CDG also achieves improvements over the baseline of 18.5%, 11.0%, 10.3%, 14.7% from source domain Gaze360 to MPII, RT-GENE, GazeCapture, and EyeDiap respectively.

4.4. Domain Adaptation

We experiment with two scenarios for domain adaptation as we elaborated in Sec. 3.3.2: Source-free domain adaptation (SFDA) and vanilla domain adaptation. In the source-free domain adaptation scene, where source domain data is not available, we begin from a pre-trained gaze estimation model and adapt it to the target domain with Contrastive Self-training Adaptation (CSA). In the vanilla domain adaptation scene, source domain data is available, so we perform CSA on the target domain with the model pre-trained on the source domain with CDG, and we denote the whole two-stage cross-domain adaptation framework as Contrastive Regression Gaze Adaptation (CRGA).

For CRGA, we first use the L^S_{CDG} with the annealed temperature γ as the constrain as we described in Sec. 3.3.2. Then, we update the pseudo-label with our adaptive model and use the new pseudo label to perform several iterations of self-training with the $\gamma = 0$ (one epoch each iteration) and finally achieve our domain adaptation results. The ablation study on the number of iteration is shown in Sec. 4.5. Results in Tab.1 demonstrate the effectiveness of our CSA and CRGA. Specifically, comparing with baseline, CSA reduces gaze estimation error by 41.6%, 22.9%, 38.6% and 21.5% when adapt from ETH-GAZE to MPIIGaze, RT-GENE, GazeCapture and EyeDiap. And CRGA further reduce the error by 40.4%, 34.7%, 55.8%, 34.3%. When we take Gaze360 as the source domain, CSA improves performance over the baseline of 15.4%,9.6%, 13.4%, and 9.3% respectively and CRGA further improve the performance of 31.7%, 30.5%, 32.9% and 23.8% compared to the baseline.

Method	Source	$\rightarrow \mathcal{D}_M$	$\rightarrow \mathcal{D}_R$	$\rightarrow \mathcal{D}_C$	$\rightarrow \mathcal{D}_D$
Baseline‡	$\mathcal{D}_E$	9.19	18.23	13.43	8.62
GazeAdv† [33]	$\mathcal{D}_E$	6.75	-	-	8.10
PureGaze† [7]	$\mathcal{D}_E$	7.08	-	-	7.48
PnP-GA† [25]	$\mathcal{D}_E$	5.53	-	-	5.87
PnP-GA‡ [25]	$\mathcal{D}_E$	6.00	-	-	6.17
CRGA‡	$\mathcal{D}_E$	**5.48**	**11.91**	**5.94**	**5.66**
Baseline‡	$\mathcal{D}_G$	8.63	23.36	12.55	8.52
GazeAdv† [33]	$\mathcal{D}_G$	8.19	-	-	12.27
Gaze360† [19]	$\mathcal{D}_G$	9.9	21.9	-	-
PureGaze† [7]	$\mathcal{D}_G$	9.28	-	-	9.32
PnP-GA† [25]	$\mathcal{D}_G$	6.18	-	-	7.92
PnP-GA‡ [25]	$\mathcal{D}_G$	**5.74**	-	-	7.04
CRGA‡	$\mathcal{D}_G$	5.89	**16.23**	**6.19**	**6.49**

Table 2. Cross-dataset gaze estimation performance compared with the state-of-art approaches. † indicates that the model employs ResNet-18 as the backbone while ‡ indicates that the model employs ResNet-50 as the backbone. Angular gaze error (°) is used as the evaluation metric.

To demonstrate the superiority of our method CRGA, we also compare with other state-of-the-art methods on unsupervised gaze domain estimation, and the results are presented in Tab. 2. Our CRGA outperforms all the state-of-the-art methods on seven domain adaptation tasks, except for $\mathcal{D}_G \rightarrow \mathcal{D}_M$, slightly inferior compared to PnP-GA [25], which employs an ensemble of networks and requires additional computational resources and memory consumption.

4.5. Extension Experiments

We conduct several extension experiments to further test the effectiveness of our proposed approach, including ablation studies on hyperparameters, backbones, different loss functions, and iterations for self-training. For simplicity, not all the experiments are performed on eight tasks from two source domains to four target domains, and the details are shown in each extension experiment respectively.

4.5.1 Ablation study on hyperparameters.

We evaluate how the CDG performance varies with the change of hyperparameter γ. γ controls the ratio of CR loss and $\ell 1$ loss in derived CDG loss. To provide a simple and intuitive presentation, we choose $\mathcal{D}_M$ and $\mathcal{D}_G$ as the target domain following [25]. We test three ratio hyperparameters which are commonly used in statistical analysis, 0.1, 1, and 10. The results are shown in rows 1-3 of Tab. 3, where we find that the best system performance occurs at the hyperparameter $\gamma = 1$. Moreover, we keep the best hyperparameter $\gamma = 1$ and test different choices for ϵ. We choose $\epsilon = 0.1, 1, 10$. The results are shown in row 4-6 of Tab. 3. We find the best performance also appears when $\epsilon = 1$. We set $\gamma = 1$ and $\epsilon = 1$ for remaining experiments.

CDG	$\mathcal{D}_E \rightarrow \mathcal{D}_M$	$\mathcal{D}_E \rightarrow \mathcal{D}_D$	$\mathcal{D}_G \rightarrow \mathcal{D}_M$	$\mathcal{D}_G \rightarrow \mathcal{D}_D$
$\gamma = 0.1$	7.72	7.86	7.46	7.50
$\gamma = 1$	**6.73**	7.95	**7.03**	**7.27**
$\gamma = 10$	7.92	**7.50**	7.37	8.07
CRGA	$\mathcal{D}_E \rightarrow \mathcal{D}_M$	$\mathcal{D}_E \rightarrow \mathcal{D}_D$	$\mathcal{D}_G \rightarrow \mathcal{D}_M$	$\mathcal{D}_G \rightarrow \mathcal{D}_D$
$\epsilon = 0.1$	5.84	6.16	6.48	7.00
$\epsilon = 1$	**5.48**	**5.66**	**5.89**	**6.49**
$\epsilon = 10$	6.15	6.00	6.31	6.72

Table 3. Ablation study on different hyperparameters γ for CDG and ϵ for CRGA. Angular gaze error ($^\circ$) is used as the evaluation metric. Here, a lower error rate stands for better performance.

4.5.2 Ablation study on loss functions

To further prove the effectiveness of our proposed CR loss, we perform experiments on the domain generalization task from source $\mathcal{D}_G$ to target $\mathcal{D}_M$, $\mathcal{D}_R$, $\mathcal{D}_C$, $\mathcal{D}_D$. We employ three different loss functions for comparison. Firstly, we use the vanilla $\ell 1$ loss as the objective function and this is actually our baseline model. Then, we employ the supervised contrastive loss (SupCon) [20], which is used for classification contrastive learning tasks, combined with the $\ell 1$ loss using our derived optimal hyperparameter $\gamma = 1$. Finally, we use our CDG loss as the objective function, which is composed of CR loss and $\ell 1$ with the optimal hyperparameter $\gamma = 1$. The results in Tab. 4 demonstrated our CR loss is suitable for regression tasks and outperforms the SupCon loss usually adopted in classification tasks. By the way, the SupCon loss even performs worse than our baseline models on three tasks of four because that the contrastive loss derived from classification tasks encourages the model to pay attention to the global semantic information, which confuses the model in gaze regression tasks instead.

Method	source	$\rightarrow \mathcal{D}_M$	$\rightarrow \mathcal{D}_R$	$\rightarrow \mathcal{D}_G$	$\rightarrow \mathcal{D}_D$
$\ell 1$	$\mathcal{D}_G$	8.63	23.3	9.23	8.52
CR+$\ell 1$	$\mathcal{D}_G$	**7.03**	**20.79**	**8.28**	**7.27**
SupCon+$\ell 1$ [20]	$\mathcal{D}_G$	7.37	23.71	9.94	8.65

Table 4. Ablation study on different loss functions. CR+$\ell 1$ equals our CDG loss as we aforementioned. Angular gaze error ($^\circ$) is used as the evaluation metric.

4.5.3 Ablation study on iterations of self-training

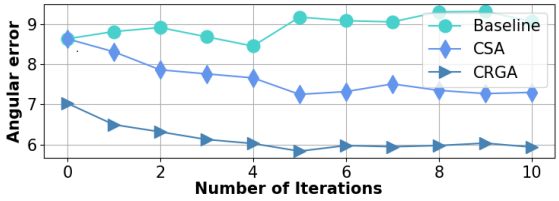

Figure 4. $\mathcal{D}_G \rightarrow \mathcal{D}_M$ using ResNet50

As we elaborated in Sec 4.4, we perform several iterations of one-epoch self-training and constantly update the pseudo labels after each iteration. To find the optimal number of the one-epoch self-training iterations ($i.e.$, I), we evaluate the CRGA performance under different numbers of self-training iterations. We perform experiments on the domain adaptation task $\mathcal{D}_G \rightarrow \mathcal{D}_M$ using ResNet-50 as the backbone Furthermore, controlled trials have been added to demonstrate whether the self-training without CRGA loss could bring vital performance improvement. In detail, we conduct 3 pipelines for comparison, one in which we perform our CRGA for different iterations I, another in which we perform our CSA for different iterations I, the third one in which we perform self-training with different iterations I on the baseline model without our derived CSA loss. The results are illustrated in Fig. 4, Both CSA and CRGA gradually improve performance as self-training iterations increase and level off. When compared with the baseline model without CSA loss, which oscillates at the original performance, our CSA loss proves to be effective.

4.5.4 Extension experiments on 100 samples.

We further perform the 100 images experiments on 4 domain adaptation tasks, following PnP-GA, as illustrated in Tab. 5. The experimental settings keep the same with experiments on full source images. Specifically, apart from the baseline model, CDG uses only 100 source images, CSA uses only 100 target images, CRGA uses 100 source + 100

Method	$\mathcal{D}_E \rightarrow \mathcal{D}_M$	$\mathcal{D}_E \rightarrow \mathcal{D}_D$	$\mathcal{D}_G \rightarrow \mathcal{D}_M$	$\mathcal{D}_G \rightarrow \mathcal{D}_D$
PnP-GA	6.00	6.17	**5.74**	7.04
CDG	7.05	7.88	7.62	7.47
CSA	5.87	5.95	6.12	6.81
CRGA	**5.68**	**5.72**	6.09	**6.68**

Table 5. Experiments on 100 images using ResNet50.

target images like PnP-GA. All of our methods only employ a single ResNet50 model, while PnP-GA employs an ensemble of 10+ models. Besides, our CSA outperforms PnP-GA in 3 tasks without source images.

4.5.5 Extension experiments on feature visualization

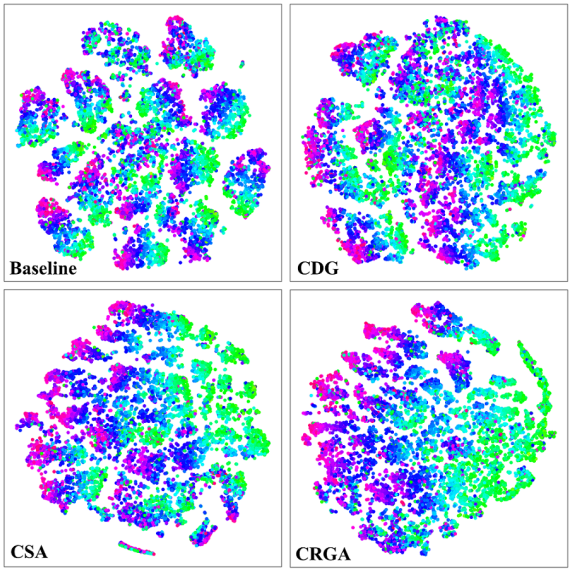

Figure 5. Illustration of the feature distribution, different colors indicate different true gaze directions. (best viewed in color).

To reveal the effectiveness of learning a good representation in an intuitive way, we visualize the distribution of features V on the domain adaptation task $\mathcal{D}_G \rightarrow \mathcal{D}_D$ with t-SNE [31]. Features generated from four different models are presented in Fig. 5, where feature points with close gaze directions share similar colors. For the baseline model, features have no obvious relationship with the gaze directions. While for CDG, the model is pre-trained on $\mathcal{D}_G$, then it can pull together features with the same gaze directions with some strength. CSA directly learns features on $\mathcal{D}_D$ without the constrain from the source domain. Despite the better representation compared with CDG, a green area abruptly appears in the blue and purple area in the bottom left corner of the figure. Compared with the other three models, CRGA shows the best performance, i.e., from left to right, showing a gradation of color from purple to green. This means the feature with close gaze directions are pulling together while those with remote gaze directions are pushing apart.

4.5.6 Extension experiments on head pose estimation.

To prove that our CR loss works well on other regression tasks, we choose head pose regression domain adaptation on $\mathcal{D}_E \rightarrow \mathcal{D}_R$ as the extensive experiment. Illustrated in Tab. 6, in the first line we conduct experiments using 100 samples like Tab. 5. The last line illustrates experiments using all the images like we did in our original paper.

Methods	Baseline	ST*	CDG	CSA	CRGA
$\mathcal{D}_E \rightarrow \mathcal{D}_R$	21.34	26.15	18.43	17.27	**16.50**
Methods	CDG-Sup†	CDG	CSA-Sup†	CSA	CRGA
$\mathcal{D}_E \rightarrow \mathcal{D}_R$	25.76	19.54	24.45	17.41	**16.12**

Table 6. Experiments on head pose domain adaptation on $\mathcal{D}_E \rightarrow \mathcal{D}_R$. * ST indicates self training without contrastive loss. † -Sup indicates using supervised contrastive classification loss.

More experiments on different backbones, the comparison between contrastive regression loss and contrastive classification loss on head pose estimation, ablation study on the prior λ in Eq. 5 and other extensive experiments are elaborated in the supplementary material to further prove the effectiveness of our proposed approach.

5. Conclusion

In this paper, we propose a novel gaze adaptation approach, namely CRGA, for generalizing gaze estimation on the target domain in an unsupervised manner. CRGA leverages the CDG module to learn the stable representation from the source domain and leverages the CSA module to learn from the pseudo labels on the target domain. The core of both CDG and CSA is the CR loss, a novel contrastive loss for regression by pulling features with closer gaze directions closer together while pushing features with farther gaze directions farther apart. Our approach demonstrates dramatic performance improvement on eight gaze domain adaptation tasks over the baseline, and also outperforms the state-of-the-art domain adaptation approaches on gaze adaptation tasks.

Acknowledgement

This work was supported in part by the National Natural Science Foundation of China under Grant 61932022, Grant 61720106001, Grant 61971285, Grant 61831018, Grant 61871267, Grant T2122024, and in part by the Program of Shanghai Science and Technology Innovation Project under Grant 20511100100.

References

[1] Philip Bachman, R Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. *Advances in Neural Information Processing Systems*, 32:15535–15545, 2019.

[2] Alisa Burova, John Mäkelä, Jaakko Hakulinen, Tuuli Keskinen, Hanna Heinonen, Sanni Siltanen, and Markku Turunen. 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*, pages 1–13, 2020. 1

[3] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *Thirty-fourth Conference on Neural Information Processing Systems (NeurIPS)*, volume 33, pages 9912–9924, 2020. 2, 3

[4] 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

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

[6] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. Big self-supervised models are strong semi-supervised learners. *arXiv preprint arXiv:2006.10029*, 2020. 3

[7] Yihua Cheng, Yiwei Bao, and Feng Lu. Puregaze: Purifying gaze feature for generalizable gaze estimation. *arXiv preprint arXiv:2103.13173*, 2021. 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. 5

[9] Tobias Fischer, Hyung Jin Chang, and Yiannis Demiris. Rt-gene: Real-time eye gaze estimation in natural environments. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 334–352, 2018. 1

[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. 1

[11] Michael A Gerber, Ronald Schroeter, Li Xiaomeng, and Mohammed Elhenawy. Self-interruptions of non-driving related tasks in automated vehicles: Mobile vs head-up display. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–9, 2020. 1

[12] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. *arXiv preprint arXiv:2006.07733*, 2020. 3

[13] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 21271–21284, 2020. 2

[14] Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 297–304. JMLR Workshop and Conference Proceedings, 2010. 3

[15] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742. IEEE, 2006. 3

[16] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9729–9738, 2020. 2, 3

[17] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. In *International Conference on Learning Representations*, 2018. 3

[18] Ashish Jaiswal, Ashwin Ramesh Babu, Mohammad Zaki Zadeh, Debapriya Banerjee, and Fillia Makedon. A survey on contrastive self-supervised learning. *Technologies*, 9(1):2, 2021. 3

[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] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33, 2020. 2, 3, 7

[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

[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] Xiao Liu, Fanjin Zhang, Zhenyu Hou, Li Mian, Zhaoyu Wang, Jing Zhang, and Jie Tang. Self-supervised learning: Generative or contrastive. *IEEE Transactions on Knowledge and Data Engineering*, 2021. 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, 5, 6, 7

[26] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 3

[27] 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. 2

[28] Ashish Shrivastava, Tomas Pfister, Oncel Tuzel, Joshua Susskind, Wenda Wang, and Russell Webb. Learning from simulated and unsupervised images through adversarial training. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2107–2116, 2017. 2

[29] Adrian Spurr, Aneesh Dahiya, Xucong Zhang, Xi Wang, and Otmar Hilliges. Self-supervised 3d hand pose estimation from monocular RGB via contrastive learning. *CoRR*, abs/2106.05953, 2021. 3

[30] 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

[31] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008. 8

[32] Kang Wang, Rui Zhao, and Qiang Ji. A hierarchical generative model for eye image synthesis and eye gaze estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 440–448, 2018. 2

[33] 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. 2, 6

[34] Yu Yu, Gang Liu, and Jean-Marc Odobez. Improving few-shot user-specific gaze adaptation via gaze redirection synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11937–11946, 2019. 2

[35] Xucong Zhang, Seonwook Park, Thabo Beeler, Derek Bradley, Siyu Tang, and Otmar Hilliges. Eth-xgaze: A large scale dataset for gaze estimation under extreme head pose and gaze variation. In *European Conference on Computer Vision*, pages 365–381. Springer, 2020. 1, 2, 5, 6

[36] Xucong Zhang, Yusuke Sugano, and Andreas Bulling. Revisiting data normalization for appearance-based gaze estimation. In *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications*, pages 1–9, 2018. 2

[37] 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

[38] 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

[39] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. Mpiigaze: Real-world dataset and deep appearance-based gaze estimation. *IEEE transactions on pattern analysis and machine intelligence*, 41(1):162–175, 2017. 1