Pattern Recognition 161 (2025) 111244

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Collaborative contrastive learning for cross-domain gaze estimation

Lifan Xia [a,1], Yong Li [a,b,1,*], Xin Cai [c], Zhen Cui [a], Chunyan Xu [a], Antoni B. Chan [b]

[a] School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, Jiangsu, China
[b] Department of Computer Science City University of Hong Kong, Kowloon Tong, Hong Kong, China
[c] Department of Information Engineering Chinese University of Hong Kong, Shatin, New Territories, Hong Kong, China

ARTICLE INFO

Keywords:
Gaze estimation
Domain generalization
Domain adaption

ABSTRACT

Gaze estimation methods commonly rely on single-camera facial appearance analysis to estimate the direction of a person's gaze. Recently, there has been significant interest in exploring gaze estimation techniques across different domains. Despite the advances, distinguishing authentic gaze-relevant features primarily relies on detecting subtle changes in the eye region. Moreover, non-gaze-related features are intricately entangled with the gaze-relevant components in a nonlinear manner. In response to these challenges, our work addresses the cross-domain gaze estimation problem through a feature decontaminating approach. Specifically, we propose a Cross Gaze Generalization (CGaG) method that explicitly encodes gaze- and domain-dedicated feature and leverages cross-identity features swapping to generate novel images where changes are exclusively confined to either the gaze or domain attributes. To consolidate feature purification and domain generalization, we utilize the equivariance between images and gaze labels to facilitate collaborative contrastive learning (CCL), which faithfully ensures that the generated novel images align with the expected outputs. Extensive experiments are conducted on various cross-domain tasks, demonstrating the effectiveness of CGaG. Meanwhile, CGaG achieves superior or comparable gaze estimation accuracy on both domain generalization and domain adaptation experiments. CGaG also shows promising cross-identity gaze or domain transfer visualizations.

1. Introduction

Appearance-based gaze estimation employs deep neural networks to infer the direction of gaze from monocular images, enabling the determination of a person's point of focus [1,2]. Recent advancements in deep learning have significantly bolstered the efficacy of gaze estimation [3–6]. However, trained gaze estimators often experience noticeable performance degradation when confronted with new scenarios, primarily due to the disparities that exist across different domains. These domain differences manifest in aspects such as different individuals, recording equipment disparities, fluctuations in illumination, and variations in resolution. Simultaneously, gaze annotation demands specialized equipment, like eye trackers, which makes it laborious and resource-intensive to collect. This limitation hinders acquiring a significant volume of labeled gaze data in diverse real-world scenarios.

To generalize the gaze estimation models to new scenarios, researchers have conducted a series of studies to develop cross-domain gaze estimation methods [7–10]. However, most of these methods require target domain data for model adaptation or fine-tuning. In real-world scenarios, accessing and labeling target domain data for model adaptation may be time-consuming or even infeasible due to privacy concerns or unavailability of specialized eye-tracking equipment. In this paper, we address this cross-domain gaze estimation issue via proposing a *Cross Gaze Generalization* (CGaG) method, aiming to generalize gaze estimation to various target domains effectively without requiring target domain data (i.e., domain generalization), which is different from most existing gaze domain adaptation works that use unsupervised training on the target domain. Specifically, we aim to encode gaze-related features that are not intertwined with domain variations, e.g., illumination, appearance, etc. Within the confines of a strictly domain generalization framework, we posit that generalizable gaze features should manifest invariance to a spectrum of subject-related variations present in the source domain. These variations encompass diverse individual characteristics such as disparities in lighting conditions, head orientations, age, gender, and other pertinent factors. Consequently, we anticipate that the trained gaze estimator that possesses fewer domain-related elements should be more amenable to domain generalization without training with target domain data.

* Corresponding author at: School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, Jiangsu, China.
E-mail addresses: xialifan@njust.edu.cn (L. Xia), yong.li@njust.edu.cn (Y. Li), caixin@link.cuhk.edu.hk (X. Cai), zhen.cui@njust.edu.cn (Z. Cui), cyx@njust.edu.cn (C. Xu), abchan@cityu.edu.hk (A.B. Chan).
[1] Lifan Xia and Yong Li contribute equally to this manuscript.

https://doi.org/10.1016/j.patcog.2024.111244
Received 25 March 2024; Received in revised form 11 September 2024; Accepted 25 November 2024
Available online 16 December 2024

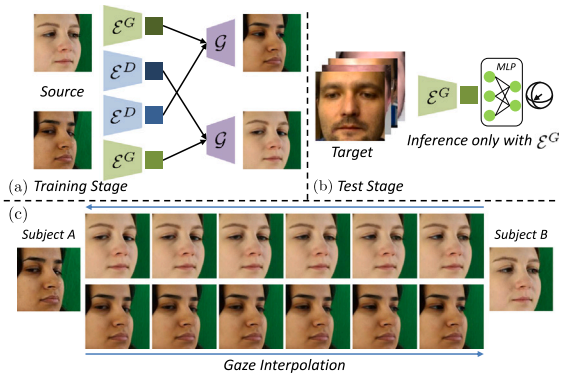


Fig. 1. (a) shows the overview of the proposed Cross Gaze Generalization (CGaG) method during the training stage. CGaG disentangles gaze and domain features and facilitates target-free cross-domain gaze estimation. (b) In the testing stage, only the trained gaze encoder is utilized for gaze estimation for inference. (c) We synthesize images produced by interpolating between gaze features from two randomly sampled subjects.

Fig. 1(a) illustrates the overview of our proposed CGaG. To acquire the desired features that exclusively encode gaze or domain characteristics, we employ two distinct encoders, each dedicated to extracting gaze or domain features. These features are subsequently randomly interchanged within the latent embedding space to synthesize novel face images, facilitating the transfer of gazes while fully preserving other domain-related factors.

The quality of the generated novel face images inherently serves as an indicator of the effectiveness with which the gaze and domain features are disentangled. Therefore, we utilize global and local contrastive learning (CL) mechanisms to rigorously constrain and ensure the faithful representation of gaze and domain attributes within the generated face images. By employing global CL, our focus lies in emphasizing the necessity for discernible disparities within the eye region of artificially generated face images compared to vanilla images. Furthermore, we leverage the equivariance between images and gaze labels and devise an elegant local CL mechanism. Local CL ensures domain features are trained to encompass domain-specific variations that pose difficulties in manual annotation. As demonstrated in Fig. 1(b), with the collaboration of global and local CL, CGaG can be directly applied to new scenarios without significantly compromising the accuracy of gaze estimation.

Fig. 1(c) demonstrates that our learned gaze features are disentangled from the domain features through visualization of synthesized images of two faces while linearly interpolating their gaze features. This effectively showcases the capability of CGaG to facilitate smooth transformations between these critical attributes. We will subsequently elaborate on how CGaG aids in the encoding of gaze-specific features, which exhibit promising potential for generalization across diverse target domains.

Our contributions in this paper are summarized as follows:

1. We propose a Cross Gaze Generalization (CGaG) method to mitigate the challenge of cross-domain gaze estimation without relying on any target data, i.e., domain generalization. CGaG aims to learn the gaze-related features merely using the source images and to directly generalize the trained gaze estimation model to various target domains.
2. In CGaG, we propose to seamlessly unify global and local CL to achieve the subtle gaze feature decoupling. We utilize the prior equivariance between images and gaze labels to facilitate domain-level CL, which is collaboratively combined with image-level CL for elegant subtle gaze feature separation.
3. Compared to existing methods, CGaG achieves state-of-the-art or comparable gaze estimation performance under the domain generalization setting. Visualization experiments show CGaG consistently disentangles gaze- and domain-related features across different subjects.

The remainder of this paper is organized as follows. In Section 2 we review related work in gaze estimation and domain generalization. In Section 3, we present the details of our CGaG method, while in Section 4, we demonstrate its efficacy through experiments.

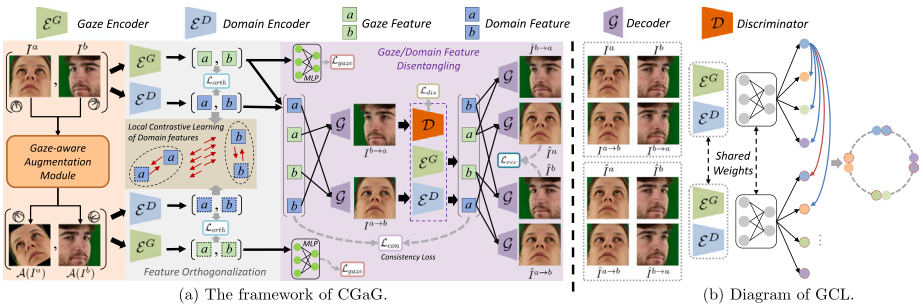
2. Related work

Gaze Estimation: Recently, gaze estimation has attracted widespread attention [11–17]. Early research typically focused on physical structure of the eyeball and used geometry models to estimate gaze [18,19]. However, due to individual differences in eye structure [18], early methods always rely on personal calibration and require the support of specific devices, such as infrared cameras [20] or depth cameras [21–23]. This also hinders their practicality since additional devices are needed.

With the development of deep learning, gaze estimation methods have achieved considerable improvements [3,24–26]. Zhang et al. [27] first established a gaze estimation model based on a CNN network. Chen et al. [4] utilized dilated convolution to increase the feature resolution for gaze estimation. Park et al. [25] proposed to use rotation consistency between gaze and head pose and enhanced gaze estimation adaptability using meta-learning paradigm. Cheng et al. [28] improved gaze estimation accuracy using a coarse-to-fine strategy. As the performance of gaze estimators often experiences noticeable degradation when confronted with new scenarios, Bao et al. [8] proposed a self-training strategy to adapt source domain data through gaze rotation consistency. Similar cross-domain methods include adversarial learning [29], outlier guidance [7], contrastive regression [9]. Sun et al. [30] used a cross-encoder structure for unsupervised gaze estimation based solely on eye regions. Also, Cai et al. [10] proposed to reduce the uncertainty of samples and enhance the adaptability of models in unsupervised source-free domain adaptation. Notably, innovative multitask methods [31,32] using multiview cameras and synthesized high-resolution data have further enhanced gaze estimation accuracy. Additionally, recent work has explored remote gaze estimation in retail environments, demonstrating the scalability and robustness of deep learning methods [33,34]. These advancements not only enhance gaze estimation accuracy but also open up new avenues for real-world applications in diverse and unconstrained settings.

Domain Generalization: Currently, cross-domain methods can be roughly divided into *domain adaptation* (DA) [35] and *domain generalization* (DG) [36]. The difference between DA and DG lies in the fact that during the training process of DA, both source domain and target domain data can be used, while DG can only use the source data. This distinction highlights that domain generalization is more challenging and has more practical applications [36]. DG methods have been widely applied in various computer vision tasks such as image classification [37], semantic segmentation [38], person re-identification [39], as well as natural language processing [40] and reinforcement learning [41]. Existing domain adaptation algorithms include techniques such as data augmentation [37,42], data generation [41,43], feature disentanglement [44,45], ensemble learning [46], and meta-learning [47] to enhance the model's generalization ability.

For unsupervised DA (UDA) gaze estimation, where the labels of target data are not used, ADL [48] is a representative unsupervised gaze adaptation method by adversarial learning and pinball loss. DAGEN [49] exploits gaze embedding and prediction consistency property to eliminate the impact of inter-personal diversity. PnP-GA [7] is a state-of-the-art unsupervised gaze adaptation approach via outlier-guided collaborative adaptation. GazeAdv [29] incorporates adversarial learning and Bayesian inference to learn gaze-related features.



(a) The framework of CGaG. (b) Diagram of GCL.

Fig. 2. (a) The framework of CGaG. (Orange and gray boxes) Given two randomly sampled images I^a and I^b, we perform image augmentation and encode gaze- and domain-related features via gaze encoder $\mathcal{E}^G$ and domain encoder $\mathcal{E}^D$, while applying an orthogonalization loss between the gaze and domain features. (Purple box) For vanilla input images, we conduct feature interchanging to generate two gaze-swapped images $I^{b\to a}$ and $I^{a\to b}$. Subsequently, the gaze- and domain-related features of $I^{b\to a}$ and $I^{a\to b}$ are re-encoded for image re-generation (Section 3.2), which facilitates image-level global contrastive learning (GCL). (Brown box) The domain features of both the vanilla and the augmented images facilitate feature-level local CL (Section 3.3). (b) In the GCL framework, an image and its reconstruction counterpart are considered as positive pairs, *e.g.* , I^a and $\hat{I}^a$, other images are used as negative samples.

RUDA [8] assigns target domain images with sub-labels which are derived from rotation consistency and facilitates domain adaptation with distribution loss. CRGA [50] learns the stable representation across the source and target domains via self-training adaptation and contrastive regression constraint. UnReGA [10] achieves source-free domain adaptation via reducing image and model uncertainties.

For DG gaze estimation, Cheng et al. [51] proposed PureGaze, which introduces DG for the first time in gaze estimation. Lee et al. [52] proposed LatentGaze, aims to map a target image to the source space by power of GAN inversion. Xu et al. [53] introduced GazeCF that purifies gaze-related features via training with diversified synthesized data. In contrast to these approaches, our CGaG does not need to explicitly define the numerous domain factors and learns to encode the gaze-related features in a collaborative contrastive learning manner.

3. Method

3.1. Overview

We introduce a domain generalization framework for gaze estimation called **Cross Gaze Generalization (CGaG)**, designed to improve the model's ability to generalize across different domains without accessing target domain data or labels. The data from the source and target domain can be represented as $\mathcal{D}_s = \{(I_i^s, g_i^s)\}_{i=1}^{N_s}$ and $\mathcal{D}_t = \{(I_i^t)\}_{i=1}^{N_t}$, where I_i^s and with I_i^t represent the ith image in the source and target domain. g_i^s denotes the gaze label for I_i^s.

Under the domain generalization setting, $\mathcal{D}_t$ is only available at the testing stage. The optimization process for domain generalization can be formulated as:

$$\min_{\phi,\theta} \mathbb{E}_{(I_i^s, g_i^s)\in\mathcal{D}_s} \|\mathcal{F}_\phi(\mathcal{E}_\theta(I_i^s)) - g_i^s\|_1 + \mathcal{L}_{general}, \tag{1}$$

where $\mathcal{E}_\theta$ is the gaze feature extractor and $\mathcal{F}_\phi$ is the gaze regressor. This equation seeks to minimize the discrepancy between the predicted gazes and true labels, as well as a regularization term $\mathcal{L}_{general}$ to improve the generalization of the gaze estimation model. For the target domain, gaze estimation is given by: $g_i^t = \mathcal{F}_\phi(\mathcal{E}_\theta(I_i^t))$, where the model trained on the source domain estimates gaze for the unseen images in the target domain.

Traditional gaze estimators usually show heavy performance degradation when applied to new domains [51–53]. To mitigate this issue, CGaG is formulated to systematically encode two discernible categories of features extracted from face images, namely the gaze-/domain-related components. The core of CGaG resides in its capacity to encapsulate the inherent gaze patterns within the gaze-related features, thereby mitigating the deleterious impact of domain-related factors and consequently facilitating cross-domain generalization.

Fig. 2(a) shows the framework of CGaG. It consists of two main components: the Gaze-disentangled Encoder, Decoder (GaED) and Collaborative Contrastive Learning (CCL). With GaED, we encode the gaze-/domain-related features and interchange these features between different subjects. This facilitates generating novel gaze-swapped face images and the subsequent re-generation, as well as reconstruction, of these images. Meanwhile, CCL endows CGaG with the capacity to explicitly supervise domain-related factors, aiming to preserve diverse nuances within the domain-related components. Below, we present the details for each component.

3.2. Gaze-disentangled encoder and decoder

To effectively disentangle gaze- and domain-related features, we devise a system composed of two specialized encoders: $\mathcal{E}^G$ for encoding gaze-related features and $\mathcal{E}^D$ for domain-related features, as illustrated in Fig. 2(a).

The input to GaED consists of images I^a and I^b from different subjects. Prior to the encoding stage, these images are processed through our Gaze-aware Augmentation Module (GaAM), which augments the gaze while maintaining the domain-related features. GaAM serves two key purposes. First, it creates augmented images to improve the robustness of the gaze encoder through more training data. Second, it precisely provides the relationship between original and augmented gaze/domain features which helps to guide the feature disentanglement. To augment the images, GaAM applies either random horizontal flips or rotations to a specified degree range, both of which modify the associated gaze labels. Flipping an image horizontally makes the yaw component of the gaze direction inverted while leaving the pitch unchanged. When it comes to rotations, as detailed in [8], an image rotated by an angle α results in an accordingly adjusted gaze direction as follows:

$$\mathcal{A}(\mathbf{g}) = R^{\mathbf{g}}\mathbf{g}, \quad R^{\mathbf{g}} = \begin{bmatrix} \cos\alpha & -\sin\alpha & 0 \\ \sin\alpha & \cos\alpha & 0 \\ 0 & 0 & 1 \end{bmatrix}, \tag{2}$$

where $\mathcal{A}$ denotes the augmentation operation and $\mathbf{g}$ is the gaze annotation before the augmentation.

Both the vanilla and the augmented images are subsequently sent to $\mathcal{E}^G$ and $\mathcal{E}^D$ for encoding the gaze-/domain-related features. To ensure accuracy in representing the gaze, we apply a constraint that aligns the gaze features with the corresponding ground truth $\mathbf{g}'$, formulated as:

$$\mathcal{L}_{gaze} = \|\mathcal{F}^G(\mathcal{E}^G(I)) - \mathbf{g}'\|_1 + \|\mathcal{F}^G(\mathcal{E}^G(\mathcal{A}(I))) - \mathcal{A}(\mathbf{g}')\|_1, \tag{3}$$

where $\mathcal{E}^G$ is the gaze encoder and $\mathcal{F}^G$ is the gaze regressor.

Additionally, to enhance the separation of the extracted features into distinct gaze and domain components, we impose orthogonal regularization. Drawing inspiration from prior works [54,55], we seek to

orthogonalize the domain and gaze features within a high-dimensional feature space. The orthogonal regularization loss is formalized as:

$$\mathcal{L}_{orth} = \cos(\mathcal{E}^G(I), \mathcal{E}^D(I)) + \cos(\mathcal{E}^G(\mathcal{A}(I)), \mathcal{E}^D(\mathcal{A}(I))), \tag{4}$$

where $\cos(\cdot,\cdot)$ calculates the cosine similarity between two feature vectors. After the first encoding stage, we swap the gaze features of I^a and I^b, then concatenate the swapped gaze features with their original domain features. This concatenated features are fed into the decoder $\mathcal{G}$ to obtain the gaze-swapped images $I^{a\to b}$ and $I^{b\to a}$, realizing initial disentanglement. Subsequently, the generated images are fed into the $\mathcal{E}^G$ and $\mathcal{E}^D$ again to obtain new gaze- and domain-related features. To maintain the consistency of these features, we formulate an embedding consistency loss $\mathcal{L}_{ec}$:

$$\begin{aligned}\mathcal{L}_{con} = &\|\mathcal{E}^G(I^{a\to b}) - \mathcal{E}^G(I^b)\|_2 + \|\mathcal{E}^D(I^{a\to b}) - \mathcal{E}^D(I^a)\|_2 \\ &+ \|\mathcal{E}^G(I^{b\to a}) - \mathcal{E}^G(I^a)\|_2 + \|\mathcal{E}^D(I^{b\to a}) - \mathcal{E}^D(I^b)\|_2.\end{aligned} \tag{5}$$

To encourage photo-realistic output from decoder $\mathcal{G}$, we incorporate a PatchGAN-based discriminator [56] D to adversarially encourage the synthesized face images to be as realistic as possible. We formulate the adversarial loss as:

$$\mathcal{L}_{dis} = \mathbb{E}[\log D(I) + \log(1 - D(\vec{I}))], \tag{6}$$

where I denotes the original input face images. $\vec{I}$ represents the gaze-swapped images, e.g., $I^{a\to b}$ and $I^{b\to a}$.

In the last stage, we regenerate the face images with the features of the gaze-swapped images $I^{b\to a}$ and $I^{a\to b}$, as shown in Fig. 2(a). To be more specific, we perform concatenation of specific feature pairs which are: the gaze features $\mathcal{E}^G(I^{a\to b})$ with the domain features $\mathcal{E}^D(I^{a\to b})$, and vice versa, as well as the recombination of $\mathcal{E}^G(I^{b\to a})$ with $\mathcal{E}^D(I^{a\to b})$, and $\mathcal{E}^G(I^{a\to b})$ with $\mathcal{E}^D(I^{b\to a})$. These concatenated features are then fed into the decoder $\mathcal{G}$, which outputs the reconstructed images $\hat{I}^{a\to b}$, $\hat{I}^{b\to a}$, $\hat{I}^a$, and $\hat{I}^b$ corresponding to each feature pair. To ensure the correctness of feature concatenation and successful training, we apply a pixel-level reconstruction loss $\mathcal{L}_{rec}$ as follows:

$$\mathcal{L}_{rec} = \|\hat{I}^a - I^a\|_1 + \|\hat{I}^b - I^b\|_1. \tag{7}$$

3.3. Collaborative contrastive learning

To guarantee reliable feature disentanglement, merely leveraging an encoder–decoder design is insufficient for distinguishing gaze and domain characteristics. Accordingly, we introduce a collaborative contrastive learning (CCL) mechanism to faithfully consolidate feature separation. Specifically, we implement two contrastive learning strategies: (1) Global contrastive learning that is applied across pairs of input or generated images, aiming to ensure the gaze-swapped images are expected; (2) Local contrastive learning is exploited to encapsulate the various domain factors in the domain-related features self-supervisedly.

Global contrastive learning (GCL). Assuming a training minibatch comprising N pairs of images, and considering the two sequential stages of image generation delineated in Section 3.2, the outcome yields a collection encompassing a total of $8N$ images, as illustrated in Fig. 2(a) and (b). We define the term "positive pairs" to refer to the original images and their corresponding reconstructed counterparts. For example, we consider two specific images (I^a and $\hat{I}^a$) as a positive pair, while all other images are treated as negative pairs. The optimization goal for the GCL is:

$$\mathcal{L}gcl(i,j) = -\log \frac{\exp(\cos(z_i, z_j)/\tau)}{\sum_{k=1}^{8N} \mathbf{1}_{[k\neq i]} \exp(\cos(z_i, z_k)/\tau)}. \tag{8}$$

Here, z_* represents the fused feature vector obtained by concatenating gaze-/domain-related features, followed by two fully-connected layers. τ is a temperature parameter that scales the similarity (cosine) scores, and the pair (i,j) denotes any given positive pair of images. The proposed GCL can differentiate both fine-grained gaze features and domain features such as skin color and illumination of different pairs. Moreover, pulling similar features closer together and pushing different features farther apart helps achieve better generation of samples.

Local contrastive learning (LCL). To ensure that the gaze encoder and domain encoder accurately produce gaze and domain features, direct supervision is needed for both encoders. Therefore, we introduce LCL to provide auxiliary supervision for the domain-related features.

LCL operates by treating the domain-related features from the original image and its augmented version as a positive pair, while all others are considered as negative pairs, as illustrated in Fig. 2 (a). LCL loss is:

$$\mathcal{L}_{lcl} = -\log \frac{\exp(\cos(q_i, \hat{q}_i)/\tau)}{\exp(\cos(q_i, \hat{q}_i)/\tau) + \sum_{j=1}^{8N} \exp(\cos(q_i, q_j)/\tau)}, \tag{9}$$

where $q_i = \mathcal{H}(\mathcal{E}^D(I))$ denotes the projected domain features of image I. $\mathcal{H}$ is a two-layer prediction head. $\hat{q}_i$ denotes the projected domain features of augmented image $\mathcal{A}(I)$. Accordingly, q_j means the projected domain features of other input or augmented images during training. With LCL, we are capable of explicitly supervising the domain-related features and consolidating the feature separation without extra manual domain-related annotations.

3.4. Overall optimization objectives

We integrate the above losses in GaED and CCL to reach the full objective:

$$\mathcal{L}_{tot} = \mathcal{L}_{gaze} + \lambda_1\mathcal{L}_{orth} + \lambda_2\mathcal{L}_{con} + \lambda_3\mathcal{L}_{dis} + \lambda_4\mathcal{L}_{rec} + \lambda_5\mathcal{L}_{gcl} + \lambda_6\mathcal{L}_{lcl}, \tag{10}$$

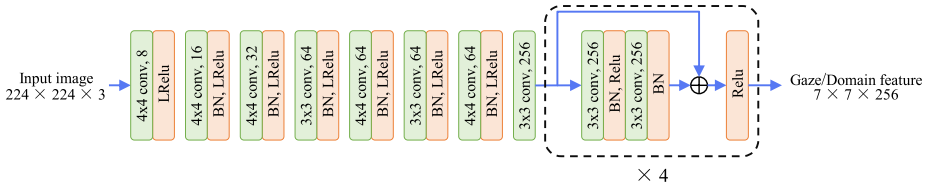
where $\lambda_1 \sim \lambda_6$ control the importance of different constraints in Sections 3.2 and 3.3.

4. Experiments

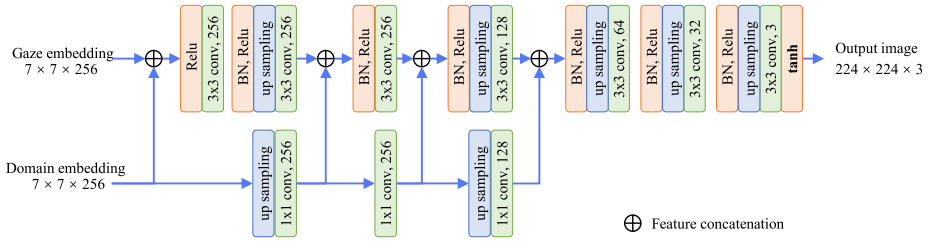
4.1. Experiment setup

Data Preparation. To verify the effectiveness of CGaG, we conduct experiments on four commonly used datasets: ETH-XGaze [57], Gaze360 [48], MPIIFaceGaze [26] and EyeDiap [58]. Among them, the training dataset provided by ETH-XGaze [57] contains 756,540 high-resolution face images of 80 subjects. The face images show various head poses and illumination conditions. Gaze360 [48] consists of 172K images from 238 subjects, both indoor and outdoor scenes, taken using a 360° panoramic camera. Following methodologies from [51, 59], we utilize selected frontal face images as the source data. MPIIFaceGaze [26] includes 45,000 images of 15 individuals with various gaze directions in different illumination conditions and head poses. The EyeDiap [58] dataset provides 94 video clips from 16 individuals. We use gaze screen targets as the target set, capturing an image every 15 frames and incorporating manually verified data provided by the original authors. About 6400 images in total. In this manuscript, we focus on gaze estimation under the cross-domain setting. Following [51], we exploit Gaze360 or ETH-XGaze as training sets and test our trained CGaG models on MPIIGaze and EyeDiap. It is because Gaze360 and ETH-XGaze datasets both show promising data diversity. For brevity, we denote ETH-XGaze as **E**, MPIIGaze as **M**, EyeDiap as **D**, and Gaze360 as **G**, respectively.

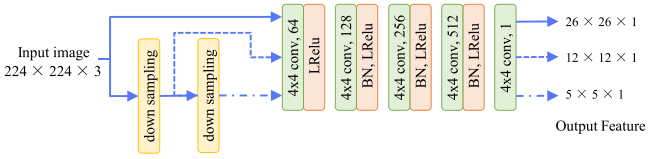
Implementation Details. Our proposed CGaG was implemented using the PyTorch library, and the model parameters were optimized via the Adam optimizer [60] with the learning rate set as 8×10^{-4} for the discriminator and 2×10^{-4} for the rest network components. For the hyper-parameter in Eq. (10), we set the $\lambda_1 \sim \lambda_6$ as 1, 0.1, 0.1, 10, 0.2, 0.1 via manual grid search. In the context of image-level contrastive learning, we configured the hyperparameter τ as 0.07. The probability of rotating images within GaAM was set as 0.5. All experiments were conducted using PyTorch on two RTX 3090 GPUs, each equipped with 24 GB of memory. The batch size was specified as 256. The training



(a) Gaze and feature encoder architectures in CGaG.



(b) Feature decoder architecture in CGaG.



(c) PatchGAN architecture in CGaG.

Fig. 3. The neural network architecture of (a) the encoder, (b) decoder and (c) discriminator in our CGaG framework.

Table 1
The parameters, FLOPs (Floating Point Operations) and Inference Time for different backbones used for gaze estimation.

Backbone	Method	Parameters	FLOPs	Inference time
Ours	Ours	5.07 M	294.85 M	199.24 ms
ResNet-18	[8,10,11,50,51,53]	11.18 M	1823.52 M	239.19 ms
ResNet-50	[7,8,11,50,51]	23.51 M	4131.69 M	370.26 ms

duration for CGaG on the ETH-XGaze dataset was approximately 60 h, while on the Gaze360 dataset, it was completed within 20 h.

CGaG consists of several neural network components, including the gaze encoder $\mathcal{E}^G$, the domain encoder $\mathcal{E}^D$, the image decoder $\mathcal{G}$, the discriminator $\mathcal{D}$. The gaze $\mathcal{E}^G$ and domain $\mathcal{E}^D$ encoder share an identical neural network configuration comprising seven convolutional layers followed by four residual blocks, as shown in Fig. 3(a). For the image decoder $\mathcal{G}$, we exploit a fully convolutional structure, as illustrated in Fig. 3(b). Specifically, $\mathcal{G}$ is composed of seven convolutional layers utilizing 3×3 convolutional kernels and a stride of 1×1. Within the convolutional layers of $\mathcal{G}$, five up-sampling layers are inserted, employing bilinear operations to obtain double-sized feature maps. Regarding the discriminator, we utilize the PatchGAN architecture, visually represented in Fig. 3(c). This architecture serves to discern the authenticity of local image patches.

ResNet-18 and ResNet-50 are the most commonly used backbones for gaze estimation. So we use the popular *thop*[2] library to measure the number of parameters and FLOPs (Floating Point Operations) and inference time for ResNet-18, ResNet-50, and the encoder in our proposed CGaG. When calculating the inference time, we tested the models on RTX 3090 and set batch size as 128 to reduce fluctuations. The comparisons are shown in Table 1. This comparison reveals that CGaG not only requires fewer parameters and lower FLOPs compared to ResNet-18 and ResNet-50, but it also exhibits reduced inference time.

[2] PyTorch-OpCounter: https://github.com/Lyken17/pytorch-OpCounter.

Table 2
Comparison of cross-domain performance with (top) typical- and (bottom) domain-generalization-based methods. *Here no target domain data is used for training.* **Bold** and underline denote best and second best results. **E**, **M**, **D**, and **G** are shorthand for the datasets **E**TH-XGaze, **M**PIIGaze, Eye**D**iap, and **G**aze360.

Methods	E → M	E → D	E → G	G → M	G → D	G → E
Full-Face [26]	12.35	30.15	34.70	11.13	14.42	26.99
RT-Gene [61]	–	–	35.33	21.81	38.60	22.92
Dilated-Net [4]	–	–	28.48	18.45	23.88	29.15
CA-Net [28]	–	–	–	27.13	31.41	–
ADL [48]	7.23	8.02	–	11.36	11.86	–
PureGaze [51]	7.08	7.48	27.51	9.28	9.32	22.61
LatentGaze [52]	7.98	9.81	–	–	–	–
GazeCF [53]	6.50	7.44	–	7.55	9.00	–
CGaG (ours)	**6.47**	**7.03**	**23.08**	**7.50**	**8.67**	**18.80**

4.2. Comparison with state-of-the-art methods

We conducted a comparative analysis of CGaG against several typical gaze estimation methods, the results are presented in Table 2. Note that all the compared methods in Table 2 will not use the images or annotations in the target domain during training. The compared methods include Full-Face [26], RT-GENE [61], Dilated-Net [4], CA-Net [28], ADL [48]. For the **E** → **G** and **G** → **E** experiments in Table 2, we re-implemented the compared methods [4,26,51,61] using the same neural network backbone as in CGaG. We also compare CGaG with the current state-of-the-art domain generalization methods: PureGaze [51], LatentGaze [52], GazeCF [53]. As illustrated in Table 2, our approach consistently surpasses both the typical gaze

Fig. 4. Visualization of gaze-swapping effects of CGaG. For each column, Subject A/B represent the two randomly sampled images from different subjects. The intermediary images between Subject A and B illustrate the two gaze-swapped images generated by CGaG.

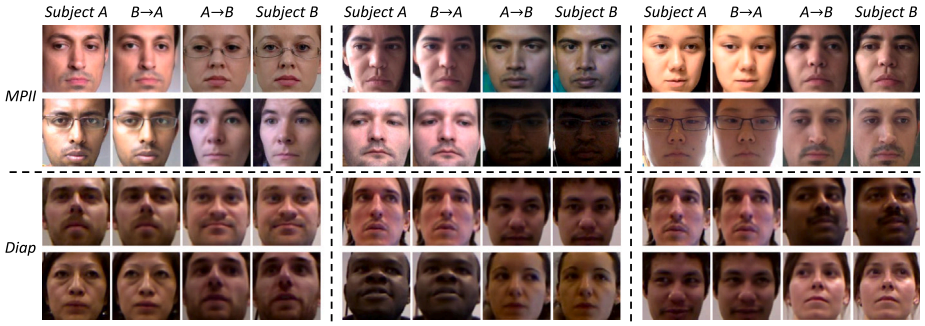


Fig. 5. Visualization of gaze-swapping effects of CGaG on MPIIFaceGaze and EyeDiap datasets. For each column, Subject A/B represent the two images from different subjects. The intermediary images between Subject A and B illustrate the two gaze-swapped images generated by CGaG.

estimation methods and the domain generalization methods across the six cross-dataset settings. We have two observations according to the comparisons: **(1)** The underwhelming *cross-domain* performance of the typical gaze estimation methods suggests their susceptibility to overfitting on factors unrelated to gaze. As indicated in [7,53], robust gaze estimation relies on the extraction of subtle features adept at capturing the specific attributes of the tiny eyeball regions within face images. Achieving this in cross-domain scenarios presents a considerable challenge, emphasizing the necessity of explicitly acquiring gaze-related features to ameliorate this issue. **(2)** Compared with the methods that are designed for cross-domain gaze estimation without touching target images, e.g., PureGaze [51], LatentGaze [52], GazeCF [53], CGaG consistently obtains lower estimation errors. On one hand, CGaG involves interchanging gazes across subjects during training, leading to the creation of numerous augmented images for training purposes. This approach serves to reduce the potential for overfitting. On the other hand, to guarantee an accurate representation of the synthesized gaze-swapped images, incorporating both the intended gaze and domain components, we implement a two-level contrastive learning framework. This collaborative approach enforces the separation of gaze and domain features at both local and global levels, ensuring a more comprehensive and faithful separation of the gaze-related and gaze-irrelevant factors.

Comparison with domain adaption methods: We additionally conducted experiments to validate the effectiveness of our framework after fine-tuning with a few target domain images and gaze annotations. To achieve this, we selected several representative state-of-the-art cross-domain gaze estimation methods, including ADDA [62], ADL [48], DAGEN [49], PnP-GA [7], GazeAdv [29], GVBGD [63], UMA [64], RUDA [8], CRGA [50], UnReGA [10]. Besides the methods introduced in Section 2. ADDA [62] learns a discriminative mapping from target images to the source feature space. GVBGD [63] equips adversarial domain adaptation with gradually vanishing bridge mechanism. UMA [64] adapts a pre-trained model to new domains via uncertainty prior.

Table 3
Performance comparison with unsupervised DA (UDA) and fine-tuned methods. The UDA methods train on the target images, while the fine-tuned DG methods use target images and annotations.

Methods	Target samples	E → M	E → D	G → M	G → D
Unsupervised DA methods.					
ADDA [62]	500	5.77	11.12	7.18	12.56
ADL [48]	100	6.23	7.80	7.00	8.77
GazeAdv [29]	100	7.26	8.37	7.88	9.81
GVBGD [63]	1000	6.68	7.27	7.64	12.44
UMA [64]	100	7.52	12.37	8.51	19.32
DAGEN [49]	500	5.68	7.92	8.02	11.08
PnP-GA [7]	10	5.53	5.87	6.18	7.92
RUDA [8]	100	5.70	7.52	6.20	7.02
CRGA [50]	> 0	5.48	5.66	5.89	6.49
UnReGA [10]	100	5.08	5.70	5.42	5.80
Fine-tuned DG methods.					
PureGaze-FT [51]	100	5.30	6.42	5.20	7.36
LatentGaze [52]	100	5.21	7.81	–	–
GazeCF [53]	100	4.76	5.43	5.34	**5.66**
CGaG (ours)	100	**4.59**	**5.07**	**5.13**	5.71

Table 3 presents the comparative analysis between CGaG and the methodologies under comparison. The findings demonstrate that following the fine-tuning of CGaG with a limited number of samples from the target domain, our approach outperforms other methods in most cases. An exception arises in the **G → D** scenario, where our performance is slightly inferior to GazeCF. The overall improvements indicate the CGaG-learned gaze-related features should be more distilled from the gaze-irrelevant factors and show better generalization ability. Compared with GazeCF [53] and LatentGaze [52], our approach does not necessitate an explicit definition of domain factors. It adeptly encompasses diverse gaze-irrelevant factors within the domain component. Furthermore, CGaG leverages a collaborative contrastive learning paradigm to strengthen the separation of features, ensuring

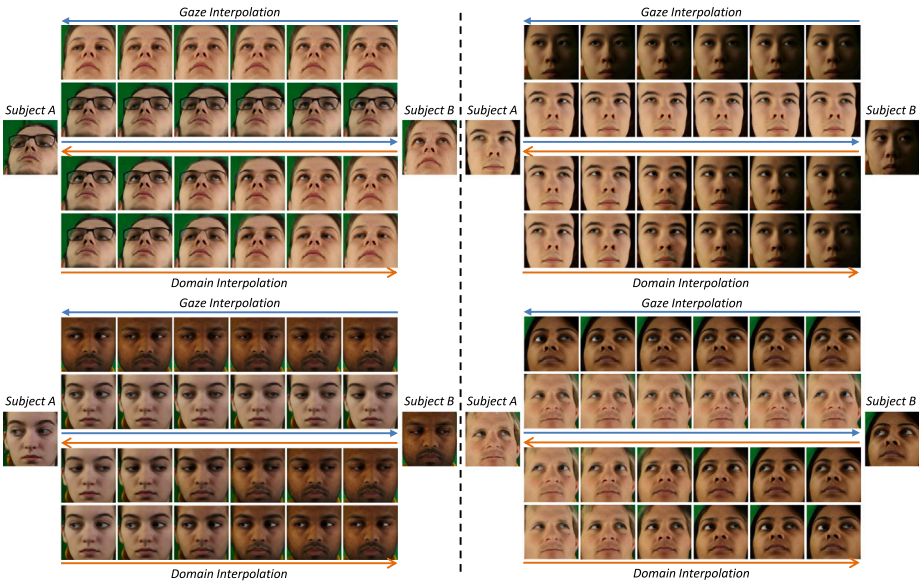
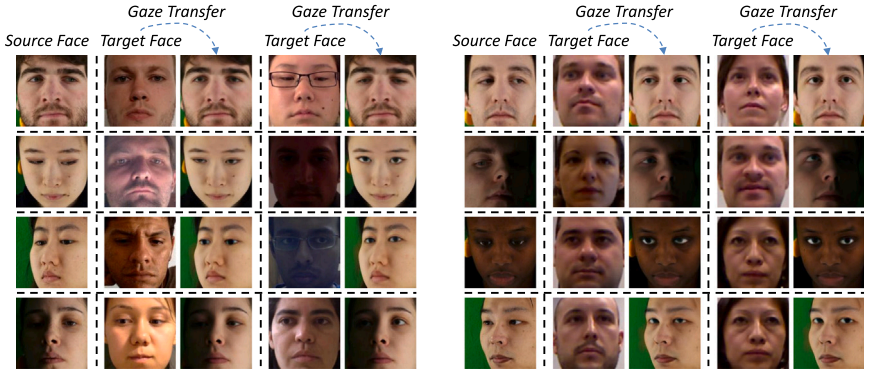


Fig. 6. Gaze and domain feature interpolation between different subjects. CGaG shows smooth transitions in the synthesized images when either using gaze or domain features for interpolation.



(a) The Source Face is from ETH-XGaze dataset. The Target Face is from MPIIFaceGaze dataset.

(b) The Source Face is from ETH-XGaze dataset. The Target Face is from EyeDiap dataset.

Fig. 7. Visualization of gaze transferring between two datasets. CGaG demonstrates the capability to directly transfer the gaze of untrained target samples to the source samples.

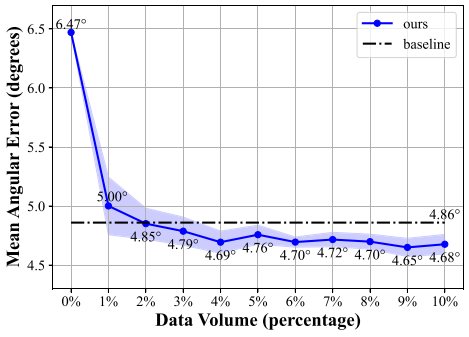


Fig. 8. The comparison between the baseline method trained on MPIIFaceGaze and fine-tuned CGaG (pre-trained on ETH-XGaze, and fine-tuned on MPIIFaceGaze) with different amounts of target training data. Using the pre-trained CGaG, we achieve comparable performance to supervised baseline method with only a small amount of target domain data.

that gaze-specific attributes are purified from undesired extraneous factors.

4.3. Qualitative analysis

Gaze-swapping. For each pair of images from different subjects, CGaG generates two gaze-swapped images. The visualization outcomes are depicted in Fig. 4. Subject A and Subject B denote the original images, while the intermediate images portray the outcomes synthesized images after gaze-swapping. Note that CGaG is not just copying the eye regions into the synthesized images, but rather synthesizing the correct eye appearance for the corresponding person, e.g., the input image pair in the 1st row in Fig. 4 have different eye colors, and CGaG can still faithfully perform gaze-swapping using the disentangled gaze features. The visualization results suggest the successful retention of diverse domain-specific features within the generated images, while the features related to gaze have effectively interchanged between distinct subjects. This outcome indicates the successful disentanglement of gaze-related features from the domain-specific components. Moreover, as illustrated in Fig. 5, we present the gaze-swapping effects observed on both the MPIIFaceGaze and EyeDiap datasets. This demonstration serves as additional validation of CGaG's effectiveness, showcasing its capability to swap gaze while preserving the domain factors with minimal alterations.

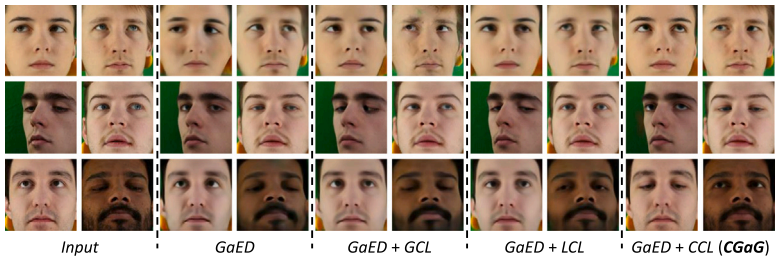


Fig. 9. Ablation study for gaze-swapping effects on ETH-XGaze dataset. Both the GCL and the LCL have shown positive impacts. And the combination of the GCL and the LCL achieves the best visualization results.

Table 4
Ablation studies of (top) GCL and LCL components in CGaG; (bottom) using domain features for gaze estimation.

Method	E → M	E → D	G → M	E → D
Baseline (GaED)	8.54	8.71	10.51	10.17
GaED + LCL	7.23▼15.34%	7.17▼17.68%	8.87▼15.60%	9.24▼9.14%
GaED + GCL	7.62▼10.77%	7.61▼12.63%	9.01▼14.27%	9.07▼10.82%
GaED + CCL (**CGaG**)	6.47▼24.24%	7.03▼19.29%	7.50▼28.64%	8.67▼14.75%
Using domain features	10.78▲−26.23%	11.84▲−35.71%	14.31▲−36.16%	14.25▲−40.12%

Gaze and subject interpolation. To further verify the efficacy of feature disentanglement, we visualize the gaze-/domain-related feature interpolation between two subjects, as shown in Fig. 6. A sequence of gaze- and domain-swapped images is generated by blending the representations from two different subjects. Fig. 6 demonstrates the seamless transitions of gaze and domain between different subjects via the disentangled gaze and domain factors learned by CGaG. These observations provide sufficient evidence for the separation of gaze-related and domain-realated features.

Gaze-transfer between datasets. To confirm the generalization capabilities of CGaG, we conducted gaze transfer from the target domain to the source domain, as shown in Fig. 7. We combined the encoded gaze features from the target domain images with the domain-specific features from the source domain images. The visualization results illustrate that gaze features originating from the target domain can indeed be transferred to the source domain. This observation strongly suggests the effective generalization of gaze-related features encoded by CGaG.

4.4. Ablation studies

Fine-tuning. We aim to reduce the reliance on annotated gaze images while maintaining gaze estimation accuracy in the target domain through our pre-trained CGaG model. To test this hypothesis, we first trained a baseline gaze estimator, utilizing the identical gaze encoder used in the CGaG model, with the MPIIFaceGaze dataset. Following this, we conducted a fine-tuning process on the pre-trained CGaG model, which was originally pre-trained on the ETH-XGaze dataset by incrementally increasing the number of annotated images sourced from the target domain. Both experiments were conducted following a three-fold subject-independent cross-validation protocol. For the fine-tuning process, we employed varying percentages of annotated images from the target domain, ranging from 1% to 10%. Besides, we repeated these experiments 10 times to calculate the mean and variance to facilitate a clear comparative analysis. Fig. 8 shows the comparison. It is clear that our model achieves comparable performance with the supervised baseline method when utilizing approximately 2% of annotated images. As the quantity of target domain data increases to 5%, the performance stabilizes and nearly reaches saturation.

Contrastive learning. We conducted an experiment to verify the contribution of global contrastive learning (GCL) and local contrastive learning (LCL), as presented in Table 4 (top). Two primary observations emerge: **(1)** Both GCL and LCL demonstrate a positive impact

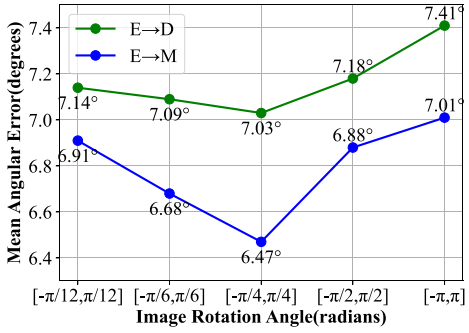


Fig. 10. The impact of different image rotation angles on cross-domain gaze estimation. CGaG achieves its best when the rotation angle range set as $[-\pi/4, \pi/4]$.

on domain feature generalization. The combination of both components yields the highest accuracy, underscoring the collaborative synergy between them. The results verify that CL plays a pivotal role in consolidating gaze feature distillation and laying the groundwork for cross-domain generalization. **(2)** Only using LCL exhibits marginal performance improvements compared to only using GCL. This underscores the significance of explicitly constraining domain features that are intricately interwoven with the gaze component and pose challenges in their disentanglement from gaze features. The collaborative property of LCL and GCL can be further verified in Fig. 9, where the combination of them yields the best visual quality.

Rotation consistency. For LCL, we exploited the intrinsic "rotation consistency" property of human gaze. We conducted ablation experiments to illustrate the impact of different rotation angle ranges, and the results are presented in Fig. 10. CGaG achieves increasing benefits when the rotation angle lies in $[-\pi/4, \pi/4]$. This phenomenon can be explained that the more powerful image rotation operation will induce confusion as the input faces will be severely rotated. Besides, slighter image rotation will induce minor domain differences and might not be beneficial for feature-level CL. Thus, we used $[-\pi/4, \pi/4]$ as the image rotation range throughout our experiments.

Gaze-feature and domain-feature disentanglement. We next ran an experiment to show that the domain features are disentangled from the gaze features. Table 4 (bottom) shows the results when using the *domain* feature for cross-domain gaze estimation. The domain features achieve the worst performance, suggesting that CGaG has removed most of the gaze information from the domain features.

Fig. 11. Failed cases of CGaG. The intermediate two images in each group are the failed gaze-swapping results, where the faces of (a) and (c) are not aligned, (b) shows extreme illumination and (d) shows eye occlusion due to huge head pose.

4.5. Limitations of CGaG

We demonstrate several representative failure cases in Fig. 11. CGaG may fail under few abnormal conditions, e.g., face misalignment in Fig. 11(a) and (c), eye occlusion due to large change in head pose in Fig. 11(d). Although CGaG shows normal gaze-swapping results in Fig. 11(b), CGaG may fail to encode gaze-related features under heavier illumination changes. Thus, the images in the source domain should encompass a broad spectrum of diverse appearances to guarantee CGaG's robustness.

5. Conclusion

In this paper, we introduce a novel framework called Cross Gaze Generalization (CGaG) for cross-domain gaze estimation. CGaG utilizes a gaze-disentangled encoder and decoder for feature encoding and interchanging. Meanwhile, CGaG introduces a collaborative contrastive learning mechanism to enhance feature separation. Whether through quantitative or qualitative evaluations, CGaG consistently demonstrates superior or comparable performance to existing methods. CGaG holds significant potential for deployment on mobile devices, making it particularly valuable for applications such as collecting data in educational settings to monitor students' attention and engagement, or in healthcare for tracking the gaze behavior of Alzheimer's disease (AD) patients. For future work, we plan to explore the integration of prevalent foundation models to enhance the generalizability of gaze estimation across a wider range of scenarios.

CRediT authorship contribution statement

Lifan Xia: Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Investigation. **Yong Li:** Writing – review & editing, Writing – original draft, Supervision, Methodology, Investigation. **Xin Cai:** Writing – review & editing, Writing – original draft, Methodology, Investigation. **Zhen Cui:** Supervision. **Chunyan Xu:** Supervision. **Antoni B. Chan:** Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research was funded by the Research Grants Council of Hong Kong (Collaborative Research Fund No. C7055-21GF) and by the Hong Kong Scholars Program.

Data availability

Data will be made available on request.

References

[1] C.L. Kleinke, Gaze and eye contact: a research review, Psychol. Bull. 100 (1) (1986) 78.
[2] S. Hoppe, T. Loetscher, S.A. Morey, A. Bulling, Eye movements during everyday behavior predict personality traits, Front. Hum. Neurosci. (2018) 105.
[3] X. Zhang, Y. Sugano, M. Fritz, A. Bulling, Appearance-based gaze estimation in the wild, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 4511–4520.
[4] Z. Chen, B.E. Shi, Appearance-based gaze estimation using dilated-convolutions, in: Asian Conference on Computer Vision, Springer, 2018, pp. 309–324.
[5] Y. Cheng, F. Lu, X. Zhang, Appearance-based gaze estimation via evaluation-guided asymmetric regression, in: Proceedings of the European Conference on Computer Vision, ECCV, 2018, pp. 100–115.
[6] Y. Yu, G. Liu, J.-M. 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, 2019, pp. 11937–11946.
[7] Y. Liu, R. Liu, H. Wang, F. Lu, Generalizing gaze estimation with outlier-guided collaborative adaptation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 3835–3844.
[8] Y. Bao, Y. Liu, H. Wang, F. Lu, Generalizing gaze estimation with rotation consistency, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 4207–4216.
[9] Y. Wang, Y. Jiang, J. Li, B. Ni, W. Dai, C. Li, H. Xiong, T. Li, Contrastive regression for domain adaptation on gaze estimation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 19376–19385.
[10] X. Cai, J. Zeng, S. Shan, X. Chen, Source-free adaptive gaze estimation by uncertainty reduction, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 22035–22045.
[11] Y. Sun, J. Zeng, S. Shan, Gaze estimation with semi-supervised eye landmark detection as an auxiliary task, Pattern Recognit. 146 (2024) 109980.
[12] H. Cheng, Y. Liu, W. Fu, Y. Ji, L. Yang, Y. Zhao, J. Yang, Gazing point dependent eye gaze estimation, Pattern Recognit. 71 (2017) 36–44.
[13] J. Liu, J. Chi, H. Yang, X. Yin, In the eye of the beholder: A survey of gaze tracking techniques, Pattern Recognit. 132 (2022) 108944.
[14] A. Recasens, A. Khosla, C. Vondrick, A. Torralba, Where are they looking? in: C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, R. Garnett (Eds.), Advances in Neural Information Processing Systems, Vol. 28, Curran Associates, Inc., 2015.
[15] Q. Huang, A. Veeraraghavan, A. Sabharwal, Tabletgaze: dataset and analysis for unconstrained appearance-based gaze estimation in mobile tablets, Mach. Vis. Appl. 28 (2017) 445–461.
[16] Y. Sugano, Y. Matsushita, Y. Sato, Appearance-based gaze estimation using visual saliency, IEEE Trans. Pattern Anal. Mach. Intell. 35 (2) (2013) 329–341, http://dx.doi.org/10.1109/TPAMI.2012.101.
[17] P. Pathirana, S. Senarath, D. Meedeniya, S. Jayarathna, Eye gaze estimation: A survey on deep learning-based approaches, Expert Syst. Appl. (ISSN: 0957-4174) 199 (2022) 116894, http://dx.doi.org/10.1016/j.eswa.2022.116894, URL https://www.sciencedirect.com/science/article/pii/S0957417422003347.
[18] K. Alberto Funes Mora, J.-M. Odobez, Geometric generative gaze estimation (g3e) for remote rgb-d cameras, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1773–1780.
[19] N.H. Cuong, H.T. Hoang, Eye-gaze detection with a single webcam based on geometry features extraction, in: 2010 11th International Conference on Control Automation Robotics & Vision, 2010, pp. 2507–2512.
[20] C. Lu, P. Chakravarthula, Y. Tao, S. Chen, H. 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, IEEE, 2020, pp. 320–331.
[21] L. Sun, Z. Liu, M.-T. Sun, Real time gaze estimation with a consumer depth camera, Inform. Sci. 320 (2015) 346–360.
[22] K. Wang, Q. Ji, Real time eye gaze tracking with 3d deformable eye-face model, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 1003–1011.

[23] D. Lian, Z. Zhang, W. Luo, L. Hu, M. Wu, Z. Li, J. Yu, S. Gao, RGBD based gaze estimation via multi-task CNN, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33, 2019, pp. 2488–2495.
[24] Y. Yu, J.-M. Odobez, Unsupervised representation learning for gaze estimation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 7314–7324.
[25] S. Park, S.D. Mello, P. Molchanov, U. Iqbal, O. Hilliges, J. Kautz, Few-shot adaptive gaze estimation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 9368–9377.
[26] X. Zhang, Y. Sugano, M. Fritz, A. 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, 2017, pp. 51–60.
[27] X. Zhang, Y. Sugano, M. Fritz, A. Bulling, MPIIGaze: Real-world dataset and deep appearance-based gaze estimation, Cornell Univ. - arXiv IEEE Trans. Pattern Anal. Mach. Intell. (2017).
[28] Y. Cheng, S. Huang, F. Wang, C. Qian, F. Lu, A coarse-to-fine adaptive network for appearance-based gaze estimation, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34, 2020, pp. 10623–10630.
[29] K. Wang, R. Zhao, H. Su, Q. Ji, Generalizing eye tracking with bayesian adversarial learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 11907–11916.
[30] Y. Sun, J. Zeng, S. Shan, X. Chen, Cross-encoder for unsupervised gaze representation learning, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 3702–3711.
[31] D. Lian, L. Hu, W. Luo, Y. Xu, L. Duan, J. Yu, S. Gao, Multiview multitask gaze estimation with deep convolutional neural networks, IEEE Trans. Neural Netw. Learn. Syst. 30 (10) (2019) 3010–3023, http://dx.doi.org/10.1109/TNNLS.2018.2865525.
[32] Y. Sugano, Y. Matsushita, Y. Sato, Learning-by-synthesis for appearance-based 3D gaze estimation, in: 2014 IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1821–1828, http://dx.doi.org/10.1109/CVPR.2014.235.
[33] S. Senarath, P. Pathirana, D. Meedeniya, S. Jayarathna, Customer gaze estimation in retail using deep learning, IEEE Access 10 (2022) 64904–64919, http://dx.doi.org/10.1109/ACCESS.2022.3183357.
[34] P. Pathirana, S. Senarath, D. Meedeniya, S. Jayarathna, Single-user 2D gaze estimation in retail environment using deep learning, in: 2022 2nd International Conference on Advanced Research in Computing, ICARC, 2022, pp. 206–211, http://dx.doi.org/10.1109/ICARC54489.2022.9754167.
[35] M. Wang, W. Deng, Deep visual domain adaptation: A survey, Neurocomputing 312 (2018) 135–153.
[36] J. Wang, C. Lan, C. Liu, Y. Ouyang, T. Qin, W. Lu, Y. Chen, W. Zeng, P. Yu, Generalizing to unseen domains: A survey on domain generalization, IEEE Trans. Knowl. Data Eng. (2022).
[37] X. Yue, Y. Zhang, S. Zhao, A. Sangiovanni-Vincentelli, K. Keutzer, B. Gong, Domain randomization and pyramid consistency: Simulation-to-real generalization without accessing target domain data, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 2100–2110.
[38] D. Li, J. Yang, K. Kreis, A. Torralba, S. Fidler, Semantic segmentation with generative models: Semi-supervised learning and strong out-of-domain generalization, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 8300–8311.
[39] X. Jin, C. Lan, W. Zeng, Z. Chen, Feature alignment and restoration for domain generalization and adaptation, 2020, arXiv preprint arXiv:2006.12009.
[40] Y. Balaji, S. Sankaranarayanan, R. Chellappa, Metareg: Towards domain generalization using meta-regularization, Adv. Neural Inf. Process. Syst. 31 (2018).
[41] K. Zhou, Y. Yang, Y. Qiao, T. Xiang, Domain generalization with mixstyle, 2021, arXiv preprint arXiv:2104.02008.
[42] S. Shankar, V. Piratla, S. Chakrabarti, S. Chaudhuri, P. Jyothi, S. Sarawagi, Generalizing across domains via cross-gradient training, 2018, arXiv preprint arXiv:1804.10745.
[43] Y. Shu, Z. Cao, C. Wang, J. Wang, M. Long, Open domain generalization with domain-augmented meta-learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 9624–9633.
[44] X. Peng, Y. Li, K. Saenko, Domain2vec: Domain embedding for unsupervised domain adaptation, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16, Springer, 2020, pp. 756–774.
[45] D. Li, J. Yang, K. Kreis, A. Torralba, S. Fidler, Semantic segmentation with generative models: Semi-supervised learning and strong out-of-domain generalization, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 8300–8311.
[46] W. He, H. Zheng, J. Lai, Domain attention model for domain generalization in object detection, in: Pattern Recognition and Computer Vision: First Chinese Conference, PRCV 2018, Guangzhou, China, November 23-26, 2018, Proceedings, Part IV 1, Springer, 2018, pp. 27–39.
[47] Q. Dou, D. Coelho de Castro, K. Kamnitsas, B. Glocker, Domain generalization via model-agnostic learning of semantic features, Adv. Neural Inf. Process. Syst. 32 (2019).
[48] P. Kellnhofer, A. Recasens, S. Stent, W. Matusik, A. Torralba, Gaze360: Physically unconstrained gaze estimation in the wild, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, ICCV, 2019.
[49] Z. Guo, Z. Yuan, C. Zhang, W. Chi, Y. Ling, S. Zhang, Domain adaptation gaze estimation by embedding with prediction consistency, in: Proceedings of the Asian Conference on Computer Vision, 2020.
[50] Y. Wang, Y. Jiang, J. Li, B. Ni, W. Dai, C. Li, H. Xiong, T. Li, Contrastive regression for domain adaptation on gaze estimation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 19376–19385.
[51] Y. Cheng, Y. Bao, F. Lu, PureGaze: Purifying gaze feature for generalizable gaze estimation., in: Proceedings of the AAAI Conference on Artificial Intelligence, 2022, pp. 436–443, http://dx.doi.org/10.1609/aaai.v36i1.19921.
[52] I. Lee, J.-S. Yun, H.H. Kim, Y. Na, S.B. Yoo, LatentGaze: Cross-domain gaze estimation through gaze-aware analytic latent code manipulation, in: Proceedings of the Asian Conference on Computer Vision, 2022, pp. 3379–3395.
[53] M. Xu, H. Wang, F. Lu, Learning a generalized gaze estimator from gaze-consistent feature, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 37, 2023, pp. 3027–3035.
[54] K. Bousmalis, G. Trigeorgis, N. Silberman, D. Krishnan, D. Erhan, Domain separation networks, Adv. Neural Inf. Process. Syst. 29 (2016).
[55] P. Liu, X. Qiu, X. Huang, Adversarial multi-task learning for text classification, 2017, arXiv preprint arXiv:1704.05742.
[56] J.-Y. Zhu, T. Park, P. Isola, A.A. Efros, Unpaired image-to-image translation using cycle-consistent adversarial networks, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2223–2232.
[57] X. Zhang, S. Park, T. Beeler, D. Bradley, S. Tang, O. Hilliges, Eth-xgaze: A large scale dataset for gaze estimation under extreme head pose and gaze variation, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16, Springer, 2020, pp. 365–381.
[58] K.A. Funes Mora, F. Monay, J.-M. 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, 2014, http://dx.doi.org/10.1145/2578153.2578190.
[59] Y. Cheng, H. Wang, Y. Bao, F. Lu, Appearance-based gaze estimation with deep learning: A review and benchmark, IEEE Trans. Pattern Anal. Mach. Intell. (2024).
[60] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, 2014, arXiv preprint arXiv:1412.6980.
[61] T. Fischer, H.J. Chang, Y. Demiris, Rt-gene: Real-time eye gaze estimation in natural environments, in: Proceedings of the European Conference on Computer Vision, ECCV, 2018, pp. 334–352.
[62] E. Tzeng, J. Hoffman, K. Saenko, T. Darrell, Adversarial discriminative domain adaptation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 7167–7176.
[63] S. Cui, S. Wang, J. Zhuo, C. Su, Q. Huang, Q. Tian, Gradually vanishing bridge for adversarial domain adaptation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 12455–12464.
[64] M. Cai, F. Lu, Y. Sato, Generalizing hand segmentation in egocentric videos with uncertainty-guided model adaptation, in: Proceedings of the Ieee/Cvf Conference on Computer Vision and Pattern Recognition, 2020, pp. 14392–14401.

Lifan Xia received the B.S. degree from Qingdao University, Qingdao, China in 2022. He is currently pursuing the master's degree with the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China. His research interests include computer vision and gaze estimation.

Yong Li received the Ph.D. degree from Institute of Computing Technology, Chinese Academy of Sciences in 2020. He worked as a software engineer in Baidu company from 2015 to 2016. He has been a assistant professor at School of Computer Science and Engineering, Nanjing University of Science and Technology since 2020.

Xin Cai received the B.S. degree in computer science from University of Chinese Academy of Sciences in 2020 and the M.S. degree from the Institute of Computing Technology , Chinese Academy of Sciences, Beijing, China. He is currently pursuing the Ph.D. degree in The Chinese University of Hong Kong.

Zhen Cui received the B.S., M.S., and Ph.D. degrees from Shandong Normal University, Sun Yat-sen University, and Institute of Computing Technology (ICT), Chinese Academy of Sciences in 2004, 2006, and 2014, respectively. Currently, he is a Professor of Nanjing University of Science and Technology, China.

Chunyan Xu received the Ph.D. degree from the School of Computer Science and Technology, Huazhong University of Science and Technology, 2015. Now she works in the School of Computer Science and Engineering from Nanjing University of Science and Technology, Nanjing, 210094, China. Her research interests include computer vision, manifold learning and deep learning.

Antoni B. Chan received the B.S. and M.Eng. degrees in electrical engineering from Cornell University, Ithaca, NY, in 2000 and 2001, and the Ph.D. degree in electrical and computer engineering from the University of California, San Diego, in 2008. He is currently a Professor in the Department of Computer Science, City University of Hong Kong.