IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 17 November 2023, accepted 3 December 2023, date of publication 7 December 2023,
date of current version 13 December 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3340446

RESEARCH ARTICLE

Improving Domain Generalization in Appearance-Based Gaze Estimation With Consistency Regularization

MOON-KI BACK, CHEOL-HWAN YOO, AND JANG-HEE YOO, (Senior Member, IEEE)
Electronics and Telecommunications Research Institute (ETRI), Daejeon 34129, South Korea

Corresponding author: Jang-Hee Yoo (jhy@etri.re.kr)

This work was supported by the Institute of Information and Communications Technology Planning and Evaluation (IITP) Grant funded by the Korean Government through MSIT (Development of AI Technology for Early Screening of Infant/Child Autism Spectrum Disorders Based on Cognition of the Psychological Behavior and Response) under Grant 2019-0-00330.

ABSTRACT Gaze estimation, a method for understanding human behavior by analyzing where a person is looking, has significant applications in various fields including advertising, driving assistance, medical diagnostics, and human-computer interaction. Although appearance-based methods have shown promising performance in uncontrolled environments, they often perform poorly when applied to similar but different domains due to variances in image quality, gaze distribution, and illumination. To overcome this limitation, this study aims to improve the domain generalization of appearance-based gaze estimation models using deep learning techniques. We propose an end-to-end deep learning approach that facilitates domain-agnostic feature learning and introduce a novel loss function, spherical gaze distance (SGD), and a regularization method, gaze consistency regularization (GCR). Our experiments, conducted using three commonly used datasets for appearance-based gaze estimation: ETH-XGaze, MPIIGaze, and GazeCapture, demonstrate the effectiveness of SGD and GCR. The results show that the proposed approach outperforms all the state-of-the-art methods on the domain generalization task and significantly improves performance when SGD and GCR are combined. These findings have important implications for the field of gaze estimation, suggesting that the proposed method could enhance the robustness and generalizability of gaze estimation models.

INDEX TERMS Consistency regularization, domain generalization, gaze estimation, feature learning.

I. INTRODUCTION

The human gaze provides useful information for understanding nonverbal communication, such as an individual's interests, intentions, and behaviors. Recently, gaze estimation has received considerable interest as a method for enhancing our understanding of human behavior and is rapidly expanding in various high-tech industries, such as advertising [1], [2], driving assistance [3], [4], medical diagnostics [5], [6], and human-computer interaction [7], [8], [9]. The capability of gaze estimation techniques significantly depends on the correctness of estimated gaze direction; thus, related studies have focused on the robustness and accuracy of gaze directions across diverse domains.

The associate editor coordinating the review of this manuscript and approving it for publication was Orazio Gambino.

Gaze estimation has mainly been achieved using two methods: model- and appearance-based approaches. Model-based approaches require specialized devices such as near-infrared cameras to determine the physical location of the pupil [10] or the corneal reflection [11]. These methods localize eye-related features, such as the eyelid, iris, and eyeball, and involve the manual construction of a geometric model of the eye to estimate the gaze direction. These methods have achieved reasonable accuracy in estimating gaze direction [12], whereas they require expensive dedicated devices and time-consuming calibration processes for each target individual. In contrast, appearance-based approaches require only conventional RGB cameras and automatically map facial images into gaze directions using a deep learning model. Furthermore, with the advancement of deep learning techniques and the release of large-scale gaze datasets [13],

[14], [15], appearance-based gaze estimation has achieved notably improved performance in uncontrolled environments.

Although appearance-based approaches have shown promising performance, an optimized model for a specific domain would encounter performance degradation in similar but different domains. This performance degradation is primarily due to variances in image quality, gaze distribution, and illumination among numerous domains [16]. This issue becomes noticeable when convolutional neural networks (CNNs) are employed to extract gaze-relevant features from full-face images because the region containing features unrelated to gaze (e.g., background, skin tone, and hair color) is much larger than the eye region, which is a critical factor in estimating gaze direction.

To overcome these limitations, some researchers have utilized domain adaptation methods that train deep learning models with a small subset of samples from the target domain. Liu et al. [16] introduced a domain adaptation method for gaze estimation. The authors proposed a plug-and-play gaze adaptation framework composed of ensemble networks that learned collaboratively under the guidance of outliers. Kothari et al. [17] proposed a method that performs a domain adaptation process by extracting additional geometric information, such as facial landmarks. However, these methods require additional models or annotations that not only require extra computational resources, but also increase the complexity of the model's estimation procedure as the number of target domains increases.

Other approaches have been studied to improve the generalization ability of deep learning models without utilizing samples from the target domains. Cheng et al. [18] adopted an adversarial learning to eliminate gaze-irrelevant features, and proposed a feature purification framework to integrate with existing gaze estimation methods. Bao et al. [19] proposed the rotation-augmented training to address the absence of the target domain's gaze distribution during the training phase in the source domain. This method trains a model using facial images randomly rotated by affine transformation from the original images. Wang et al. [20] developed the contrastive domain generalization based on contrastive learning to facilitate stable representation learning for the source domain. While the aforementioned methods are carefully designed to ensure robust performance across a range of domains, fundamental training methods such as regularization techniques and loss functions are largely performed based on methods commonly used in typical regression tasks. Accordingly, methods specifically tailored for gaze estimation tasks have yet to be proposed.

In this paper, we propose an end-to-end deep learning approach to improve the generalization ability of appearance-based gaze estimation models. Initially, to balance the model's generalizability with the risk of overfitting [21], we propose the gaze consistency regularization (GCR), which is a regularization method that guides the model to consistently estimate gaze directions, even in the presence of intentional input perturbations such as adjustments to brightness or saturation of pixel values. The GCR performs with two feed-forward steps: 1) training a model with labeled data to accurately estimate the gaze direction, and 2) training a model in an unsupervised setting to maintain similar gaze directions for similar facial images. Therefore, GCR mitigates overfitting of gaze-irrelevant features and improves generalizability by focusing on features related to the human gaze. Second, we introduce a novel loss function called spherical gaze distance (SGD) for the GCR to further improve the accuracy of the gaze direction. SGD maps different gaze directions onto points in a three-dimensional (3D) sphere and then measures the distance between them along the surface. Because the human gaze is intrinsically a concept that exists in a specific direction in the real world, our motivation is to consider this aspect in the loss function to quantify the error more precisely between the estimated gaze direction and the ground truth. Although appearance-based gaze estimation typically does not require 3D information corresponding to a 2D facial image as input, back-propagating the error measured by SGD could conceivably facilitate gaze estimation grounded in 3D perception. The contributions of this paper are summarized as follows:

- We propose a regularization method that leverages intentional image augmentation to assist the model in learning domain-agnostic features and ensures consistent estimation results across variations derived from a single facial image.
- We introduce a novel loss function utilizing spherical distance, enhancing the performance of our regularization method by accurately quantifying geometrical differences in gaze directions.

II. RELATED WORK

With the release of numerous publicly available gaze datasets, various methods have been proposed to solve cross-dataset challenges. These methods can be categorized into two approaches: domain adaptation and domain generalization. In this section, we briefly review previous literature on appearance-based gaze estimation in terms of domain generalization and adaptation.

A. APPEARANCE-BASED GAZE ESTIMATION

Appearance-based gaze estimation methods using deep learning have been studied over the past decade [22] and demonstrated to perform well in uncontrolled environments. Zhang et al. [23] first proposed a gaze estimation method based on deep learning. They use a CNN model to estimate 2D gaze angles from multimodal inputs, which consist of a pair of eye images and corresponding head pose angles. Since then, numerous gaze estimation methods based on deep learning have emerged [24]. Murthy and Biswas [25] introduced a feature manipulation method to reduce the estimation error. This method utilizes two independent deep neural networks: one is designed to exclude person-dependent features and the other discovers a set of points

that contain gaze-relevant features extracted from eye images. Cheng et al. [26] proposed a training strategy based on a coarse-to-fine approach for a more accurate gaze estimation. This strategy consists of two training phases: 1) extracting coarse-grained features from facial images using a CNN to estimate basic gaze directions and 2) refining these basic gaze directions with residuals computed from the corresponding eye images. Bao et al. [27] designed a neural network architecture that focuses on the correlation between eyes and the face. The network extracts features from the left and right eye images through channel-wise fusion and then integrates these features with facial images to recalibrate the network's attention toward crucial gaze-related features. In line with the studies described above, Zhang et al. [15] recently released a large-scale gaze dataset called ETH-XGaze. The authors also presented a standardized experimental protocol and evaluation metrics based on this gaze dataset, which has become a benchmark for performance comparison in gaze estimation studies.

B. DOMAIN GENERALIZATION ON GAZE ESTIMATION

In domain adaptation, the deep learning model is permitted to access a few samples from the target domains during the adaptation process, which is usually performed in an unsupervised setting (i.e., without labels). Wang et al. [28] focused on the differences in data distribution between the source and target domains and proposed a method to reduce the distribution distance using adversarial training. Wang et al. [20] introduced a method that utilizes pseudo-labels generated by a model trained in the source domain. Lee et al. [29] proposed a domain-shift method that maps target images to source domain images using a generative adversarial network. Domain adaptation is generally focused on improving the performance in the target domain, often without considering the potential performance degradation in the source domain. Additionally, it lacks flexibility because it requires samples from the target domain [18].

In contrast, domain generalization aims to train a model to generalize well to any target domain, even if data from the target domain are never observed during training. To the best of our knowledge, Cheng et al. [18] were the first to propose a domain generalization method that is well-suited for appearance-based gaze estimation tasks. The authors designed a framework based on self-adversarial training to eliminate gaze-irrelevant features such as personal appearance and illumination conditions from the input facial images. Jiang et al. [30] proposed a simple training method that focused on reducing the overfitting problem without additional computational resources or model parameters. Xu et al. [21] specifically defined gaze-irrelevant factors such as identity, expression, illumination, and tone, and proposed a method for generating synthesized gaze data based on adversarial attacks and data augmentation. Unlike the approach used in [18], which depended on an adversarial network for eliminating gaze-irrelevant features, the approach used in [21] achieved more explainable domain generalization by defining specific features and experimentally verifying their relevance. While the aforementioned methods have demonstrated significant improvements in gaze domain generalization, they necessitate pre-trained models, such as facial expression classification, or involve separate pre- and post-processing steps. Consequently, there can be a dependency on pre-trained models and inflexibility, as these processing steps are repeatedly executed when the source domain changes. Given the uniqueness of the gaze estimation task, this study revisits the basic learning methods that have been overlooked in appearance-based gaze estimation and explores more flexible learning schemes.

III. METHODOLOGY

We propose a simple yet effective learning scheme that enables gaze estimation models to learn domain-agnostic features. To optimize the performance of our learning scheme, we introduced a novel loss function that leverages the geometrical differences between distinct gaze directions. In this section, we describe the overall training procedure of appearance-based gaze estimation in Section III-A and then provide more details for the proposed method in Section III-B and III-C.

A. GAZE ESTIMATION VIA REGRESSION

A gaze dataset collected from a specific domain is defined as $D_* = \{(x_i, y_i)|_{i=1}^N\}$, where x_i denotes the i-th facial image, y_i is the corresponding gaze direction, and N is the total number of facial images. For simplicity, following other studies [15], [18], [19], [20], we represent the gaze direction as two angles (*pitch* and *yaw*) instead of a 3D vector representation. Because the *roll* angle is largely assumed to be constant in appearance-based gaze estimation, this simplification assists in reducing the model parameters consider and allows for a more intuitive analysis of the estimation results.

The goal of the appearance-based gaze estimation is to optimize a gaze estimation model $G(\cdot)$ with the gaze dataset. In general, the gaze estimation model can be decomposed into two modules: a backbone network $F(\cdot)$ that extracts feature vectors $\mathbf{z}$ from facial images, and a multi-layer perceptron $R(\cdot)$ to regress these feature vectors into gaze directions. The estimated gaze direction $\hat{y}$ on the basis of the gaze estimation model can be represented as $\hat{y} = G(x) = R(F(x))$. Given that both $\hat{y}$ and $\mathbf{y}$ are represented as 2D or 3D vectors composed of continuous real numbers, typical regression losses used in machine learning can be employed. In previous studies [15], [18], [21], [25], vector norms such as L_1 and L_2 are employed as their loss functions to optimize G. In other words, the L_1 or L_2 distance is used to measure the discrepancy between the estimated gaze direction and the ground truth, and the parameters θ of G are optimized by minimizing the distance:

$$\theta = \arg\min_{\theta} \sum_{i} L[G(x_i), y_i] \tag{1}$$

where L denotes the loss function used to quantify the difference between estimated and actual gaze directions.

B. SPHERICAL GAZE DISTANCE

Vector norms are commonly used as loss functions in regression tasks. Considering that the human gaze can intrinsically be represented as a vector pointing in a specific direction within a 3D space, traditional vector norms may not be optimal loss functions for gaze estimation tasks. Because vector norms are calculated by individually aggregating each element in both the estimated and actual directions, they cannot represent geometric information based on the correlation between the elements. Drawing inspiration from the analogy that the representation of gaze direction is akin to longitude and latitude on a sphere, we designed a loss function using the Haversine formula [31], a well-established method for calculating the great-circle distance between two points on a sphere. Thus, we hypothesized that it would be more intuitive and suitable to utilize the angle between the gaze directions as a more accurate measure. To reflect the spherical nature of eye movement and effectively quantify the geometrical differences between distinct gaze directions, we propose an SGD that retains the conventional 2D angular representation (*pitch* and *yaw*) for consistency with vector norms, while also using the angle to precisely quantify the distance between the estimated gaze direction and the ground truth. The SGD can be formulated as:

$$L_{SGD} = d = 2r \arcsin([\sin^2(\frac{\beta_g - \beta_e}{2}) + \cos(\beta_e)\cos(\beta_g)\sin^2(\frac{\alpha_g - \alpha_e}{2}) + \epsilon]^{\frac{1}{2}}) \quad (2)$$

where β_g and β_e are the actual and estimated vertical gaze angles (*pitches*) respectively, α_g and α_e are the actual and estimated horizontal gaze angles (*yaws*). Additionally, r denotes the radius of the sphere, d denotes the shortest distance between two points on the sphere's surface, and ϵ is a constant for numerical stability. In our experiments, we set $\epsilon = 10^{-8}$ and $r = 1$ to maintain a scale similar to vector norms.

As shown in Fig. 1, two distinct gaze directions, represented by *pitch* and *yaw* angles, can be mapped to points on a sphere. The shortest distance between these points on the great circle varied according to the interior angle of the gaze direction. SGD not only serves as an alternative to vector norms but can also be back-propagated to optimize the gaze estimation model in accordance with Equation (1).

C. GAZE CONSISTENCY REGULARIZATION

In deep learning, regularization is commonly used to improve domain generalization and prevent models from overfitting to a specific dataset. This approach is also beneficial in appearance-based gaze estimation, where simple regularization techniques such as weight decay, feature normalization, and weight moving average can enhance domain generalization [30]. However, these conventional methods do not fully address the unique aspects of appearance-based gaze estimation. Estimating gaze directions in 3D space from 2D images without additional geometric information is a challenging task in this context. Moreover, the eye region, which is crucial for gaze estimation, is smaller than areas unrelated to gaze, such as the background or skin color, increasing the risk of overfitting. To address this, we constrain the model to output consistent estimates for perturbed variations derived from the same face image, driving the model to focus on invariant features when estimating gaze direction. As shown in Fig. 2, GCR leverages intentional image augmentation to guide the model in learning domain-agnostic features and ensures consistent estimation results across variations derived from a single facial image. Additionally, we employ SGD to enable the gaze estimation model to indirectly utilize geometric information.

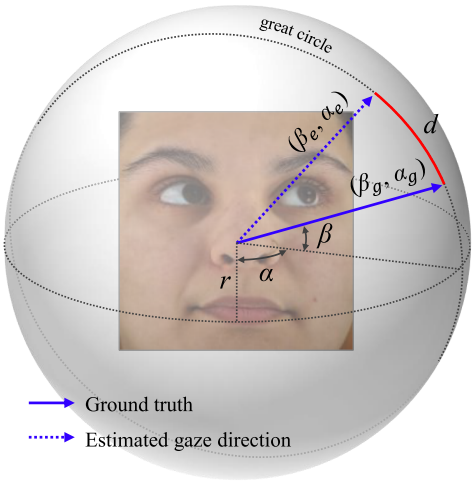

FIGURE 1. Illustration of the SGD that maps two distinct gaze directions to points on a sphere and measures the shortest distance along the great circle. α and β represent the horizontal and vertical gaze angles, respectively, and subscripts g and e denote the ground truth and estimated gaze angles, respectively.

Algorithm 1 Training Procedure of the GCR

Input: Training Dataset in a Domain D_*
Parameter: $r, \epsilon, \lambda_{gaze}, \lambda_{con}$
Output: $G_\theta(\cdot)$

for $i \leftarrow 1$ *to* N **do**
  $(\mathbf{x}, \mathbf{y}) \leftarrow D_*$
  $x^g, x^c \leftarrow A(\mathbf{x})$
  $\hat{y}^g \leftarrow G_\theta(x^g)$
  $\hat{y}^c \leftarrow G_\theta(x^c)$
  $L_{gaze} \leftarrow \hat{y}^g, \mathbf{y}, r, \epsilon$ with Eq. (2).
  $L_{con} \leftarrow \hat{y}^g, \hat{y}^c, r, \epsilon$ with Eq. (2).
  $L_{total} \leftarrow \lambda_{gaze}, L_{gaze}, \lambda_{con}, L_{con}$ with Eq. (3).
  Optimize G_θ with Eq. (1)
end for
return $G_\theta(\cdot)$

The overall training procedure for GCR consists of four phases: augmenting input image, estimating basic

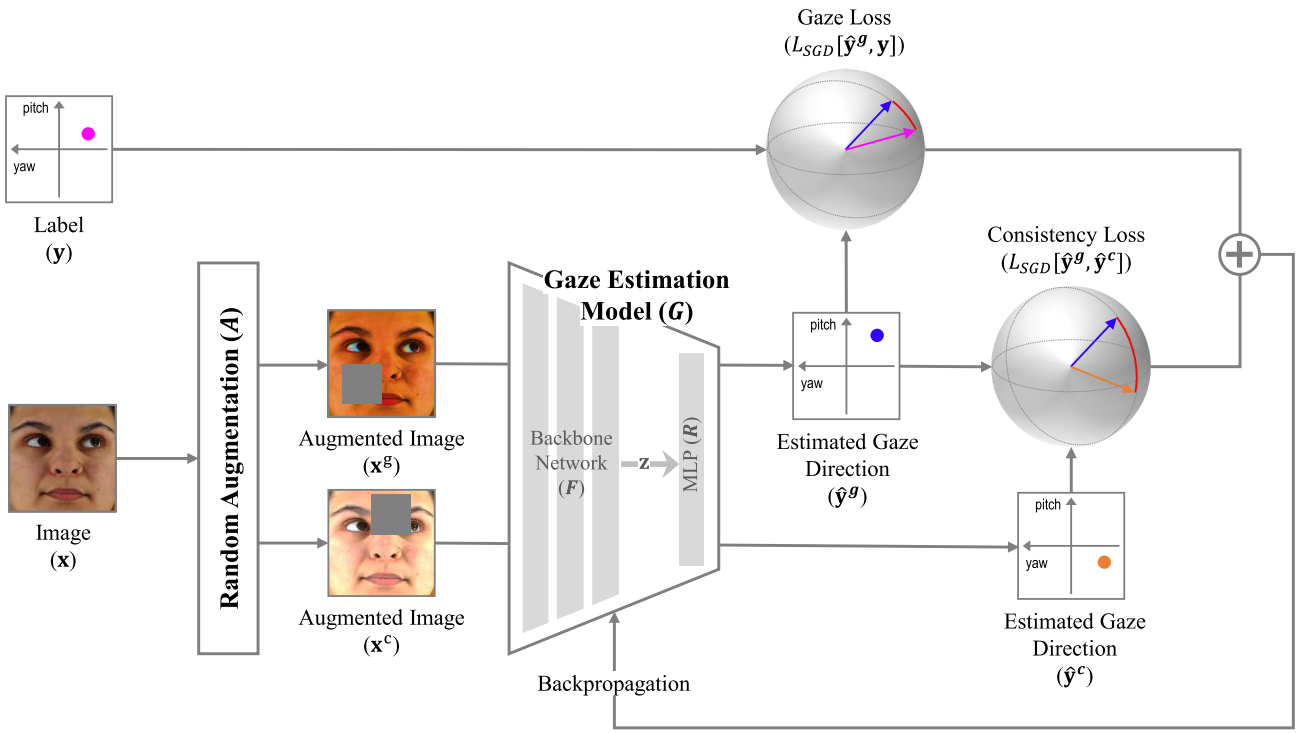

FIGURE 2. Illustration of the GCR procedure. A randomly augmented image (x^g) is input into a gaze estimation model to compute gaze loss, which allows the model to learn essential features related to gaze and estimate the basic gaze direction. Simultaneously, another randomly augmented image (x^c), which is semantically identical to x^g but has different pixel values, is used to focus the model on the features that appear common to x^g and x^c. As training progresses, the consistency loss continuously compensates for the gaze loss, reducing the risk of overfitting to a given domain and guiding the model to learn domain-invariant features.

gaze direction, ensuring consistent estimation results, and optimizing the model. The step-by-step procedure of these phases is further summarized in Algorithm 1.

1) RANDOM AUGMENTATION

Data augmentation techniques manipulate given data within a range in which the original and augmented data are nearly identical in terms of human perception. These techniques can be employed to enhance the accuracy of the estimated gaze direction and generalization ability of the gaze estimation model [32]. Before inputting the facial image into the gaze estimation model, as depicted in Fig. 2, random augmentation is applied twice to the original facial image using the augmentation operator $A(\cdot)$. Specifically, A comprises a combination of cutout [33] and color space augmentation [34], which includes adjustments to the brightness, saturation, and contrast of the facial image. However, adjustments affecting to face alignment, such as affine transformation, were excluded to guarantee the invariance of the actual gaze direction $\mathbf{y}$.

2) GAZE ESTIMATION

The estimated gaze direction $\hat{y}^g$ is obtained by feeding the augmented facial image x^g into G. Gaze loss L_{gaze} is then computed using $\hat{y}^g$ and its corresponding label $\mathbf{y}$ based on Equation (2). As the training progresses, the backbone network focuses on the eye region to extract essential gaze-relevant features. Subsequently, the multi-layer perceptron regresses the gaze direction based on these extracted features by minimizing gaze loss. While most previous studies used vector norms to compute gaze loss, we introduced SGD to enable the gaze estimation model to achieve a more optimal performance.

3) GAZE REGULARIZATION

In training G, the gaze estimation phase described above uses the augmented facial images generated by A. Although it is less susceptible to overfitting and promotes better generalization than training with the original images, there is a risk that excessive or inappropriate augmentation can act as noise in the gaze estimation model, thereby hindering its ability to learn essential features related to gaze. Motivated by recent studies on self- and semi-supervised learning [35], [36], [37], which utilize common learning strategies aimed at learning more robust feature representations by attracting semantically positive samples, we propose employing an auxiliary constraint wherein $\hat{y}^g$ and $\hat{y}^c$ should point in nearly identical directions, as they are derived from a single-source image. To rectify the outlier directions that deviate from the basic gaze directions, the consistency loss L_{con} is used to minimize the distance between $\hat{y}^g$ and $\hat{y}^c$. This ensures the G has the ability to output consistent gaze directions for

comparable facial images x^g and x^c by learning the shared gaze-relevant features. In particular, the proposed SGD, which precisely compares two distinct gaze directions, can be employed to rectify the estimated gaze directions during this phase. By adopting this learning scheme, the proposed GCR enhances the ability of the gaze estimation model to produce consistent outputs for semantically identical inputs, thereby facilitating the learning of domain-invariant gaze features.

4) MODEL OPTIMIZATION

The total loss of the GCR is a combination of gaze and consistency losses. The total loss is formulated as follows:

$$L_{total} = \lambda_{gaze} L_{gaze} + \lambda_{con} L_{con} \tag{3}$$

where λ_{gaze} and λ_{con} are tunable parameters. We set $\lambda_{gaze} = 1$ and $\lambda_{con} = 1$. The θ-parameterized gaze estimation model, denoted as G_θ, is trained to improve both the accuracy and robustness of estimated gaze directions through back-propagation with Equation (2). In the total loss, L_{con} can be regarded as a directional penalty for L_{gaze}, because the total loss converges to approximately L_{gaze} when $\hat{y}^g$ and $\hat{y}^c$ are nearly identical.

IV. EXPERIMENTAL RESULTS

To verify the impact of the proposed method on gaze domain generalization, we applied our method to a gaze estimation model and trained it using public gaze datasets collected from different environments. We then compared its performance with that of state-of-the-art (SOTA) methods in terms of domain generalization. In this section, we elaborate on the gaze datasets, training procedures, and benchmark comparisons. Also, the experimental results obtained using the proposed method are discussed.

A. DATASETS

We conducted experiments using three datasets commonly used for appearance-based gaze estimation: ETH-XGaze (D_E) [15], MPIIGaze (D_M) [13], and GazeCapture (D_C) [38]. The ETH-XGaze dataset comprises 1.1 M facial images with gaze and head pose labels collected from 110 participants in a laboratory environment. MPIIGaze and GazeCapture contain 45 K and 2.4 M facial images with the corresponding gaze labels, respectively. Both datasets were constructed using mobile devices in an uncontrolled manner. As summarized in Table 1, ETH-XGaze provides a broader range of gaze directions than MPIIGaze and GazeCapture. Therefore, we used it as our training dataset and reserved MPIIGaze and GazeCapture for evaluation. In particular, given that MPIIGaze and GazeCapture are in-the-wild datasets obtained from daily life scenarios, they can be deemed suitable for evaluating the generalization performance. Consequently, we conducted two domain generalization experiments, denoted as $D_E \rightarrow D_M$ and $D_E \rightarrow D_C$. Our experimental setup aligns with previous studies [18], [19], [20], [29], [30], ensuring consistent and fair performance comparison.

TABLE 1. Summary of gaze datasets used in the study.

Dataset	#Subjects	#Images	Max. Gaze
ETH-XGaze [15]	110	1,083,492	$\pm 70°$, $\pm 120°$
MPIIGaze [13]	15	45,000	$\pm 20°$, $\pm 20°$
GazeCapture [38]	1,473	2,445,504	$\pm 20°$, $\pm 20°$

B. IMPLEMENTATION DETAILS

The experiments were conducted using the PyTorch framework. Following the evaluation protocol defined in [15], we employed ResNet-50 [39], which was pretrained on ImageNet-1K [40], as the backbone for all experiments. A gaze estimation model is built to output a 2D gaze angle by mapping the N-dimensional features extracted from a facial image by the backbone into a 2D vector using a linear transformation. Utilizing the normalization method proposed in [41], we obtained normalized facial images from the original images and resized them to 224 × 224 resolution for the input of the gaze estimation model as training data. We executed all experiments on a single NVIDIA RTX A6000 GPU and trained the gaze estimation model for 25 epochs. We set the batch size to 64 and used the Adam [42] optimizer with a learning rate of 10^{-4}. Considering the rounding errors that occur when calculating multiple trigonometric functions, we applied 64-bit floating-point arithmetic to compute the loss.

C. EXPERIMENTAL RESULTS AND ANALYSIS

We used angular error as the evaluation metric, consistent with most previous studies. Angular error represents the internal angle between the estimated and actual gaze directions. A smaller angular error indicates a better performance. The angular error can be formulated as:

$$E_{angular} = \arccos\left(\frac{g \cdot \hat{g}}{\|g\| \|\hat{g}\|}\right) \tag{4}$$

where both g and $\hat{g}$ denote 3D unit vectors derived from 2D gaze angles.

The experimental results presented in Table 2 demonstrate the effectiveness of the proposed SGD and GCR. Following the evaluation protocol in [15], we trained our baseline model using the ResNet-50 backbone and the L_1 loss function on D_{E_train}. The mean angular errors were 4.77° for D_{E_test}, 7.10° for D_M and 10.96° for D_C, respectively. Compared with the experimental results in [15] and [20], our baseline model can be considered a reliable benchmark and was used as a performance indicator in this study. As shown in rows 2-4 of Table 2, SGD is well suited for appearance-based gaze estimation. Specifically, compared to our baseline, SGD reduced the mean angular error by 4.14%, 0.28%, and 3.39% when evaluated using D_{E_test}, D_M, and D_C, respectively. In particular, SGD outperformed the L_1 and L_2

TABLE 2. Experimental results compared with baselines. Angular error (degree) is used as evaluation metric.

Method	Data Aug.		Angular Error (°)								
	Color	Cutout	Test($D_{E\_train} \rightarrow D_{E\_test}$)			$D_E \rightarrow D_M$			$D_E \rightarrow D_C$		
			Mean±Std	Min.	Max.	Mean±Std	Min.	Max.	Mean±Std	Min.	Max.
Baseline(ResNet-50 + L_1) [15]	-	-	4.5	-	-	7.56	-	-	10.5	-	-
ResNet-50 + L_1 (Our Baseline)	-	-	4.77±3.62	0	70.01	7.10±3.96	0	59.15	10.96±6.38	0	69.86
ResNet-50 + L_2	-	-	5.06±3.51	0	63.46	7.85±4.38	0	66.90	11.15±6.47	0	74.30
ResNet-50 + L_{SGD}	-	-	**4.58**±3.30	0	65.39	**7.08**±4.14	0.03	63.05	**10.60**±6.33	0	73.85
ResNet-50 + L_1 + *GCR*	✓	✓	4.36±3.01	0	56.61	6.33±3.88	0.04	62.21	9.91±6.02	0	79.40
ResNet-50 + L_2 + *GCR*	✓	✓	4.63±3.15	0	52.56	6.73±3.95	0.03	61.80	9.50±5.97	0	73.31
ResNet-50 + L_{SGD} + *GCR*	✓	✓	**4.36**±2.95	0	51.39	**6.03**±3.65	0.02	63.74	**8.55**±5.92	0	72.14

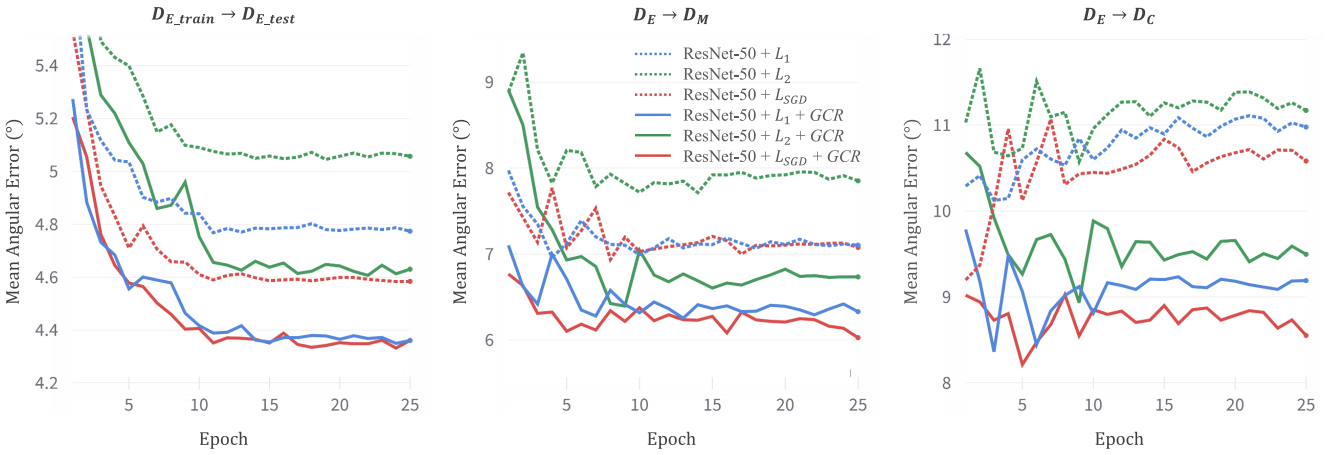

FIGURE 3. Comparison of mean angle errors during training using the proposed GCR combined with different loss functions. Different colors repre- sent different loss functions, and the dashed lines indicate experiments without the GCR for each corresponding solid line.

loss functions on the given domain D_E without compromising its generalization capability. Thus, it can be used instead of vector norms to optimize existing gaze estimation models.

To validate how well GCR improves the generalization capability, we conducted three experiments using different loss functions. As shown in rows 5-7 of Table 2, GCR significantly improved the model's generalizability across all experiments. Compared with the above-mentioned results, GCR consistently alleviates overfitting in the source domain (D_E) and improves the performance in unseen similar domains (D_M and D_C). Moreover, the GCR helps reduce both the deviation and maximum angular error. This appears to be due to gaze regularization, which shifts the distribution of estimated gaze directions closer to the actual distribution by decreasing the likelihood of outliers.

As shown in row 7 of Table 2, the combination of GCR and SGD achieves the best performance, and its performance consistently exhibits low angular errors during training, as illustrated in Fig. 3. Specifically, compared to our baseline, GCR combined with SGD reduced the mean angular error by 9.4%, 17.74%, and 28.18% when evaluated for D_{E_test}, D_M, and D_C, respectively. These results indicate that SGD is more compatible with GCR than with either L_1 or L_2. We presume that this superiority stems from the fact that L_{con}, when measured by the SGD, is quantified more accurately than when measured using L_1 or L_2. This accuracy imposes a distinct directional penalty on L_{gaze}, leading to an optimized performance of the gaze estimation model. To elucidate how GCR and SGD synergistically enhance the generalization of the gaze estimation model, we superimposed a heat map onto facial images using Grad-CAM [43]. As highlighted in the prominent regions of Fig. 4, the gaze estimation model trained with GCR and SGD appears to estimate the final gaze direction by identifying features correlated with the gaze, in addition to eye-specific features. Even with the same gaze direction, the shape of pupils and eyeballs can change depending on the head pose. Thus, it is plausible that gaze estimation considers facial contours and the nose shape.

D. COMPARISON WITH SOTA METHODS

For a more comprehensive comparison, we summarize the SOTA methods for appearance-based gaze estimation and list their performance in Table 3. To assess the proposed method alongside the domain adaptation methods, we used a checkmark to indicate whether each method utilized samples from the target domain. The results in Table 3 show that GCR outperforms all the SOTA methods on

TABLE 3. Performance comparison with SOTA domain generalization methods. D_T indicates the gaze dataset of the target domain, and it is marked with ✓ if a subset of D_T is used for domain adaptation.

Method	D_T	Mean Angular Error (°)		
		Test	$D_E \rightarrow D_M$	$D_E \rightarrow D_C$
Baseline [15]	-	4.50	7.56	10.5
LatentGaze [29]	-	3.94	7.98	-
RAT [19]	-	-	7.40	-
PureGaze [18]	-	-	7.08	-
Reg-Gaze [30]	-	-	6.75	-
CDG [20]	-	4.56	6.73	9.23
Ours(L_1+*GCR*)	-	**4.36**	**6.33**	**9.19**
Ours(L_{SGD}+*GCR*)	-	**4.36**	**6.03**	**8.55**
GazeAdv [28]	✓	-	6.75	-
RUDA [19]	✓	-	5.70	-
PnP-GA [16]	✓	-	5.53	-
CRGA [20]	✓	-	5.48	-

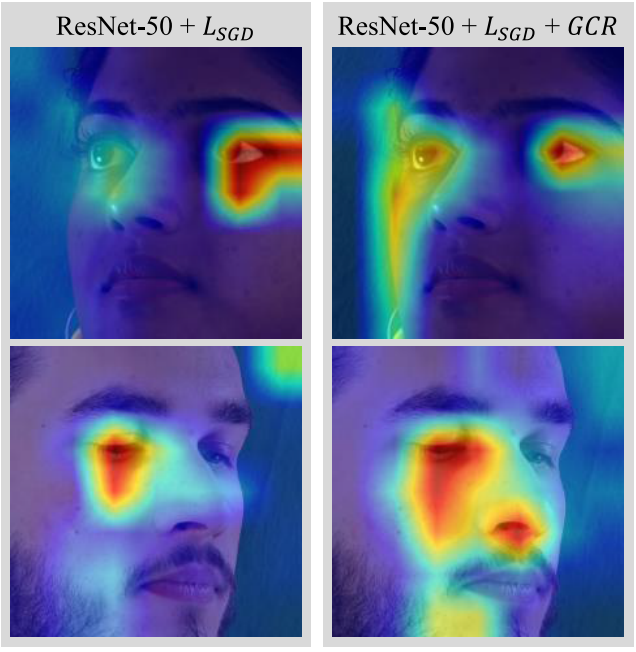

FIGURE 4. Grad-CAM visualization of the gaze estimated by the model trained with the GCR and/or the SGD. The prominent regions indicate the most influential features used by the model to estimate gaze direction.

domain generalization task and significantly improves the performance when combined with SGD. In addition, the results show competitive performance compared to methods based on domain adaptation. The proposed method can be easily employed in existing gaze estimation models to achieve optimal performance, as loss functions based on vector norms are compatible with SGD, and GCR does not require additional deep learning models. However, there are some limitations. GCR could potentially slow down training due to the repeated generation of augmented images, particularly in large-scale applications, and SGD requires a system capable of handling high computational precision. Thus, the proposed method is currently less suited for online learning in systems with limited floating-point capabilities, such as embedded systems. Despite these limitations, significant potential exists for further research and refinement, including optimizing SGD and enhancing learning efficiency to expand the method's applicability.

V. CONCLUSION

In this paper, we present a method designed to enhance the domain generalization of appearance-based gaze estimation models using deep learning. To strike a balance between the generalizability of the model and the risk of overfitting, we proposed a learning scheme that ensures consistent gaze estimations for semantically identical inputs. Moreover, we introduced a novel loss function to further enhance the accuracy of the estimated gaze direction and to quantify the geometrical differences between distinct gaze directions. Our experimental results demonstrate that the proposed method achieves leading performance in gaze domain generalization. The proposed method can be leveraged to enhance the generalizability of existing gaze estimation models, as the loss function is compatible with vector norms and the learning scheme does not require additional deep learning models or computational resources. In the future, more experiments will be conducted in diverse environments to further verify the effectiveness of the proposed method. Additionally, we plan to explore the application of our method in diagnosing autism spectrum disorder by analyzing atypical gaze patterns in patients.

REFERENCES

[1] C. Bermejo, D. Chatzopoulos, and P. Hui, "EyeShopper: Estimating Shoppers' gaze using CCTV cameras," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 2765–2774.

[2] R. N. To and V. M. Patrick, "How the eyes connect to the heart: The influence of eye gaze direction on advertising effectiveness," *J. Consum. Res.*, vol. 48, no. 1, pp. 123–146, Feb. 2021.

[3] Q. Zhuang, Z. Kehua, J. Wang, and Q. Chen, "Driver fatigue detection method based on eye states with pupil and iris segmentation," *IEEE Access*, vol. 8, pp. 173440–173449, 2020.

[4] G. Yuan, Y. Wang, H. Yan, and X. Fu, "Self-calibrated driver gaze estimation via gaze pattern learning," *Knowl.-Based Syst.*, vol. 235, Jan. 2022, Art. no. 107630.

[5] J. Kerr-Gaffney, A. Harrison, and K. Tchanturia, "Eye-tracking research in eating disorders: A systematic review," *Int. J. Eating Disorders*, vol. 52, no. 1, pp. 3–27, Jan. 2019.

[6] J. Li, Z. Chen, Y. Zhong, H.-K. Lam, J. Han, G. Ouyang, X. Li, and H. Liu, "Appearance-based gaze estimation for ASD diagnosis," *IEEE Trans. Cybern.*, vol. 52, no. 7, pp. 6504–6517, Jul. 2022.

[7] P. Li, X. Hou, X. Duan, H. Yip, G. Song, and Y. Liu, "Appearance-based gaze estimator for natural interaction control of surgical robots," *IEEE Access*, vol. 7, pp. 25095–25110, 2019.

[8] C. Katsini, Y. Abdrabou, G. E. Raptis, M. Khamis, and F. Alt, "The role of eye gaze in security and privacy applications: Survey and future HCI research directions," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, Apr. 2020, pp. 1–21.

[9] Z. A. Bature, S. B. Abdullahi, S. Yeamkuan, W. Chiracharit, and K. Chamnongthai, "Boosted gaze gesture recognition using underlying head orientation sequence," *IEEE Access*, vol. 11, pp. 43675–43689, 2023.

[10] R. Valenti, N. Sebe, and T. Gevers, "Combining head pose and eye location information for gaze estimation," *IEEE Trans. Image Process.*, vol. 21, no. 2, pp. 802–815, Feb. 2012.

[11] E. D. Guestrin and M. Eizenman, "General theory of remote gaze estimation using the pupil center and corneal reflections," *IEEE Trans. Biomed. Eng.*, vol. 53, no. 6, pp. 1124–1133, Jun. 2006.
[12] Y. Cheng, H. Wang, Y. Bao, and F. Lu, "Appearance-based gaze estimation with deep learning: A review and benchmark," 2021, *arXiv:2104.12668*.
[13] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling, "MPIIGaze: Real-world dataset and deep appearance-based gaze estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 1, pp. 162–175, Jan. 2019.
[14] P. Kellnhofer, A. Recasens, S. Stent, W. Matusik, and A. Torralba, "gaze360: Physically unconstrained gaze estimation in the wild," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6911–6920.
[15] X. Zhang, S. Park, T. Beeler, D. Bradley, S. Tang, and O. Hilliges, "ETH-XGaze: A large scale dataset for gaze estimation under extreme head pose and gaze variation," in *Proc. Eur. Conf. Comput. Vis.*, Glasgow, U.K., Aug. 2020, pp. 365–381.
[16] Y. Liu, R. Liu, H. Wang, and F. Lu, "Generalizing gaze estimation with outlier-guided collaborative adaptation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 3815–3824.
[17] R. Kothari, S. De Mello, U. Iqbal, W. Byeon, S. Park, and J. Kautz, "Weakly-supervised physically unconstrained gaze estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 9975–9984.
[18] Y. Cheng, Y. Bao, and F. Lu, "PureGaze: Purifying gaze feature for generalizable gaze estimation," in *Proc. AAAI Conf. Artif. Intell.*, Jun. 2022, pp. 436–443.
[19] Y. Bao, Y. Liu, H. Wang, and F. Lu, "Generalizing gaze estimation with rotation consistency," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 4197–4206.
[20] Y. Wang, Y. Jiang, J. Li, B. Ni, W. Dai, C. Li, H. Xiong, and T. Li, "Contrastive regression for domain adaptation on gaze estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 19354–19363.
[21] M. Xu, H. Wang, and F. Lu, "Learning a generalized gaze estimator from gaze-consistent feature," in *Proc. AAAI Conf. Artif. Intell.*, Jun. 2023, pp. 3027–3035.
[22] A. Kar and P. Corcoran, "A review and analysis of eye-gaze estimation systems, algorithms and performance evaluation methods in consumer platforms," *IEEE Access*, vol. 5, pp. 16495–16519, 2017.
[23] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling, "Appearance-based gaze estimation in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 4511–4520.
[24] P. Pathirana, S. Senarath, D. Meedeniya, and S. Jayarathna, "Eye gaze estimation: A survey on deep learning-based approaches," *Expert Syst. Appl.*, vol. 199, Aug. 2022, Art. no. 116894.
[25] L. Murthy and P. Biswas, "Appearance-based gaze estimation using attention and difference mechanism," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 3137–3146.
[26] Y. Cheng, S. Huang, F. Wang, C. Qian, and F. Lu, "A coarse-to-fine adaptive network for appearance-based gaze estimation," in *Proc. AAAI Conf. Artif. Intell.*, Apr. 2020, pp. 10623–10630.
[27] Y. Bao, Y. Cheng, Y. Liu, and F. Lu, "Adaptive feature fusion network for gaze tracking in mobile tablets," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Jan. 2021, pp. 9936–9943.
[28] K. Wang, R. Zhao, H. Su, and Q. Ji, "Generalizing eye tracking with Bayesian adversarial learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 11899–11908.
[29] I. Lee, J.-S. Yun, H. H. Kim, Y. Na, and S. B. Yoo, "LatentGaze: Cross-domain gaze estimation through gaze-aware analytic latent code manipulation," in *Proc. Asian Conf. Comput. Vis.*, Dec. 2022, pp. 3379–3395.
[30] Y. Jiang, H. Zhang, and B. Ni, "Revisit regularization techniques for gaze estimation generalization," in *Proc. 6th Int. Conf. Video Image Process.*, Dec. 2022, pp. 91–95.
[31] C. C. Robusto, "The cosine-haversine formula," *Amer. Math. Monthly*, vol. 64, no. 1, p. 38, Jan. 1957.
[32] A. A. Akinyelu and P. Blignaut, "Convolutional neural network-based methods for eye gaze estimation: A survey," *IEEE Access*, vol. 8, pp. 142581–142605, 2020.
[33] T. DeVries and G. W. Taylor, "Improved regularization of convolutional neural networks with cutout," 2017, *arXiv:1708.04552*.
[34] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *J. Big Data*, vol. 6, no. 1, pp. 1–48, Jul. 2019.
[35] K. Sohn, D. Berthelot, N. Carlini, Z. Zhang, H. Zhang, C. A. Raffel, E. D. Cubuk, A. Kurakin, and C.-L. Li, "FixMatch: Simplifying semi-supervised learning with consistency and confidence," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, Dec. 2020, pp. 596–608.
[36] X. Chen and K. He, "Exploring simple Siamese representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 15745–15753.
[37] D. Kim, Y. Yoo, S. Park, J. Kim, and J. Lee, "SelfReg: Self-supervised contrastive regularization for domain generalization," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9599–9608.
[38] K. Krafka, A. Khosla, P. Kellnhofer, H. Kannan, S. Bhandarkar, W. Matusik, and A. Torralba, "Eye tracking for everyone," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2176–2184.
[39] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[40] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Apr. 2015.
[41] X. Zhang, Y. Sugano, and A. Bulling, "Revisiting data normalization for appearance-based gaze estimation," in *Proc. ACM Symp. Eye Tracking Res. Appl.*, Jun. 2018, pp. 1–9.
[42] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
[43] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.

MOON-KI BACK received the B.S. and Ph.D. degrees in computer engineering from Chungnam National University, Daejeon, South Korea, in 2013 and 2021, respectively. He is currently a Postdoctoral Researcher with the Electronics and Telecommunications Research Institute (ETRI), South Korea. His research interests include computer vision, learning-based gaze estimation, and human–computer interaction.

CHEOL-HWAN YOO received the B.S. and Ph.D. degrees in electrical engineering from Korea University, Seoul, South Korea, in 2014 and 2020, respectively. Since 2020, he has been with the Electronics and Telecommunications Research Institute (ETRI), South Korea, as a Senior Researcher. His research interests include deep learning, image processing, computer vision, and human–robot interaction.

JANG-HEE YOO (Senior Member, IEEE) received the B.Sc. degree in physics and the M.Sc. degree in computer science from the Hankuk University of Foreign Studies, South Korea, in 1988 and 1990, respectively, and the Ph.D. degree in electronics and computer science from the University of Southampton, U.K., in 2004. Since November 1989, he has been with the Electronics and Telecommunications Research Institute (ETRI), South Korea, as a Principal Researcher. He has also been a Professor with the Department of Artificial Intelligence, University of Science and Technology, South Korea. He was a Visiting Scientist with the University of Washington, Seattle, USA, from August 2014 to July 2015. His current research interests include computer vision, human motion analysis, biometric systems, HCI, and intelligent robots. He is a member of IEIE.