Fine Gaze Redirection Learning with Gaze Hardness-aware Transformation

Sangjin Park, Daeha Kim, and Byung Cheol Song
Inha University, Incheon, Republic of Korea
san9569@naver.com, kdhht5022@gmail.com, bcsong@inha.ac.kr

Abstract

The gaze redirection is a task to adjust the gaze of a given face or eye image toward the desired direction and aims to learn the gaze direction of a face image through a neural network-based generator. Considering that the prior arts have learned coarse gaze directions, learning fine gaze directions is very challenging. In addition, explicit discriminative learning of high-dimensional gaze features has not been reported yet. This paper presents solutions to overcome the above limitations. First, we propose the feature-level transformation which provides gaze features corresponding to various gaze directions in the latent feature space. Second, we propose a novel loss function for discriminative learning of gaze features. Specifically, features with insignificant or irrelevant effects on gaze (e.g., head pose and appearance) are set as negative pairs, and important gaze features are set as positive pairs, and then pair-wise similarity learning is performed. As a result, the proposed method showed a redirection error of only 2° for the GazeCapture dataset. This is a 10% better performance than a state-of-the-art method, i.e., STED. Additionally, the rationale for why latent features of various attributes should be discriminated is presented through activation visualization. Code is available at https://github.com/san9569/Gaze-Redir-Learning

1. Introduction

Gaze is a representative non-verbal cue that is detected first when a person concentrates on a specific object. Recently, gaze information has been used for assistant robots [30], driver's intention detection systems for avoiding safety-critical situations [36], gaze tracking in VR systems [21], and so on.

Classical approaches for gaze representation extracted hand-crafted descriptors from face (or eye) images and used them as gaze features [33, 23]. However, the simplistic nature of hand-crafted descriptors has been an obstacle to performance generalization. With the rapid development of the feature extraction capability of neural network(s), the recent methods were able to extract more powerful gaze features [40, 27, 26, 45]. In particular, generator-based methods [26, 45] showed the effect of gaze representation learning by directly manipulating the gaze direction of the eye or face images.

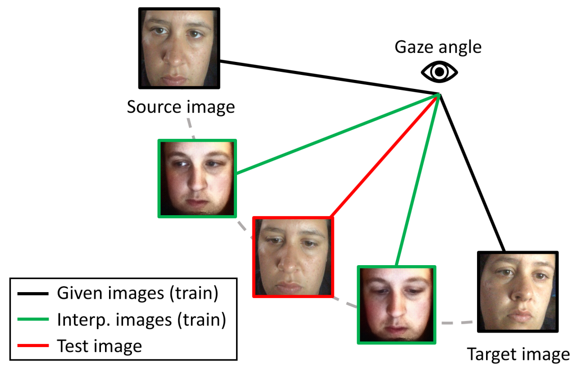


Figure 1: Conceptual illustration of our problem definition. The proposed method can learn various gaze directions compared to previous works [26, 45].

This paper argues that the following two issues should be resolved for learning a more robust representation of gaze. First, the gaze directions that cannot be represented by input images *must be properly reflected on the (latent) feature space*. As in Figure 1, prior arts [26, 45] used only the limited gaze directions of the input images, *i.e.*, the source and target images, as supervision during training, so it was difficult to learn the representation of unseen gaze directions. Second, gaze is tightly coupled with several human factors, such as head pose and appearance, which have little or no relevance to gaze [34, 19]. So, if gaze, head pose, appearance, etc. are entangled in the feature space, it will be very difficult to learn a feature that can fully represent the gaze [22]. The discriminative learning of gaze features and inessential features such as head pose features, that is, *learning of inter-feature relationship has not yet been attempted*.

This paper provides a novel concept of gaze understanding that tackles both of the above-mentioned issues. First, we propose so-called *Gaze Hardness-aware Transformation* (GHT) to generate various gaze features from a pair of source and target images. GHT is defined by linear interpolation of the source and target gaze features, i.e., $\mathbf{z}_s^g$ and $\mathbf{z}_t^g$ (cf. $\mathbb{T}$ in Fig. 2). Transformed feature(s) $\mathbf{z}_{tr}^g$ serves as a kind of *additional supervision* that increases the number of gaze directions that cannot be represented by source and target alone and is also input to the proposed gaze consistency loss function (cf. Sec. 3.2). Additionally, since GHT is designed to increase the learning difficulty of gaze representation, it prevents trivial solutions and alleviates the overfitting problem at the later stage of training (cf. Sec. 4.4). Second, this paper proposes a so-called structured gaze (SG) loss function for discriminative learning of gaze features and inessential features. We define gaze and inessential features as a negative pair, and different gaze features as a positive pair to form triplet tuples. The SG loss function based on the triplet tuple calculates structured feature similarity through various combinations between positive and negative pairs in a mini-batch. Here, to alleviate the inherent overfitting problem of metric learning, the hard negatives and positives of Zhu *et al.* [46] are additionally utilized. Therefore, the SG loss function learns the inter-feature relationship based on the so-called 'push and pull' strategy (cf. Sec. 3.3).

The contribution points of this paper are summarized as follows:

- GHT generates features of diverse gaze directions that are not limited to a given source and target. To the authors' knowledge, learning for gaze direction based on feature-level transformation has not been reported yet.
- Metric learning based on the SG loss function succeeded in learning the inter-feature relationship between gaze features and inessential features.
- For the GazeCapture [20] dataset, the proposed method achieved more than 10% improvement in quantitative performance compared to the state-of-the-art (SOTA) gaze redirection methods. In addition, the disentanglement property of the proposed method was demonstrated through activation visualization.

2. Related Work

Gaze redirection. Gaze redirection is a computer vision task that redirects the gaze direction of the face image toward the target gaze direction. Warp-based methods [8] warped an input eye image to the desired output appearance. GAN-based methods [11, 38] generated redirected images using Generative Adversarial Network (GAN) which has been widely used in the generation task. [1] used an auto-encoder based on numerical and pictorial guidance, and [16] used a style-based generator to generate redirected images.

Transforming auto-encoder (TA) [12] that learned an equivariant mapping between latent features and input/output spaces was applied to the latest gaze redirection methods [26, 45] and showed reliable performance. They learned the auto-encoding process that transforms the gaze direction of the source image into that of the target image. In [26, 45], the (geometric) transformation that adjusts the (source) gaze direction was called the redirection process ($\mathbb{R}$ in this paper), and was defined by the rotation operation (cf. **Appendix**). STED [45] defines gaze, head pose, and task-irrelevant attributes in the latent space and additionally generating pseudo-label-based images. However, existing methods could not precisely learn the gaze representation in the wild environment because they only used images with a limited number of gaze directions. Feature-level transformation proposed in Sec. 3.2 can be a solution to this.

Feature-level transformations. One of the methods to improve the generalization performance of neural networks is transformation on the feature space [5, 44, 46]. For example, DAML [5] generated hard negative features through a generator network and used them for similarity learning. HDML [44] produced synthetic features through feature interpolation that can adaptively adjust the hardness of similarity learning. Recently, a data-efficient transformation that produces features useful for discriminative learning has been developed to solve the computational and optimization problems of [5, 44]. Zhu *et al*. [46] alleviated the overfitting problem as well as the phenomenon that similarity learning of positive and negative pairs was stuck at a trivial solution by employing feature extrapolation and interpolation. Inspired by [44, 46], we propose a novel feature-level transformation that can adaptively control the hardness of gaze learning, thereby generating gaze features corresponding to various gaze directions.

Deep metric learning with multiple pairs. Deep metric learning uses a distance metric to understand the semantic relationship between latent features. Contrastive loss [9, 13] and triplet loss [3, 29] learn that the distance between pairs of different classes becomes farther within a predetermined margin and the distance between pairs of the same class becomes closer. The pair-based metric loss has been gradually extended to quadruplet [2, 14] or N-pair loss [32], that is, the generalized triplet based on N-pair negatives. Song *et al*. [25] proposed a lifted structured loss that designs the relationships between all positive and negative samples in a mini-batch as a structured formula. We apply the inter-feature relationship to gaze through metric learning.

3. Method

3.1. Problem Formulation

Our goal is to make the model learn the fine gaze representation by generating an image $\widetilde{\mathbf{x}}_t$ in which the gaze direction of source images $\mathbf{x}_s$ is redirected to that of the target image $\mathbf{x}_t$. Our base model, i.e., the transforming auto-encoder

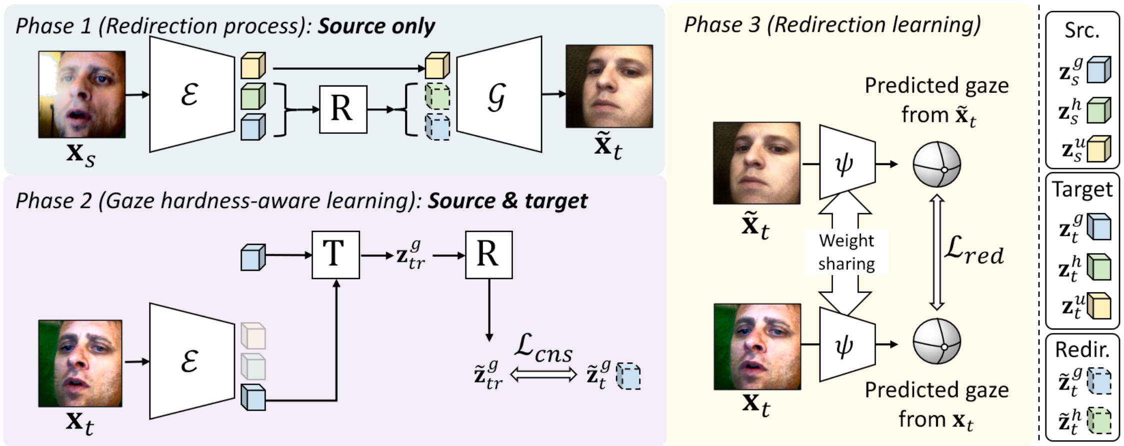


Figure 2: Overview of the proposed method. R is the conventional rotation process (cf. Sec. 3.1). T is gaze hardness-aware transformation and generates the new gaze feature $\mathbf{z}^g_{tr}$ (cf. Sec. 3.2). $\mathbf{z}^g_{tr}$ can be generated as many as the number of mini-batches and represents various gaze directions. ψ is a pre-trained network for redirection learning and is frozen during training (cf. Sec. 3.4). After all loss functions are calculated through all phases, the parameters of $\mathcal{E}$ and $\mathcal{G}$ are updated. In the inference, phase 1 is performed to generate the redirected image $\widetilde{\mathbf{x}}_t$, and phase 3 is used to calculate the redirection error for evaluation.

(TA) [12, 45, 26], defines gaze $\mathbf{z}^g$, head pose $\mathbf{z}^h$, and task-irrelevant features $\mathbf{z}^u$ in the latent space, respectively, and then learns the equivariant mapping between the feature space and the input space (cf. Sec. 2). However, given a test image with a gaze direction that is difficult to observe from the source and target images, we cannot accurately represent the gaze direction (cf. Fig. 1). That is, learning various gaze directions with a limited number of gaze directions is quite challenging. Therefore, we attempt to learn the unseen gaze direction by generating a feature $\mathbf{z}^g_{tr}$ representing a new gaze direction through linear interpolation between a given source and target gaze features ($\mathbf{z}^g_s$ and $\mathbf{z}^g_t$).

Overview. Figure 2 is the overview of the proposed method. In phase 1, given the source image $\mathbf{x}_s$, encoder $\mathcal{E}$ produce the (latent) feature $\mathbf{z}_s$ on the unit hypersphere: $\mathbf{z}_s = \mathrm{Nm}(\mathcal{E}(\mathbf{x}_s))$ where Nm indicates the L2 normalization. $\mathbf{z}_s$ is composed of a concatenation form of gaze $\mathbf{z}^g_s$, head pose $\mathbf{z}^h_s$, and task-irrelevant feature $\mathbf{z}^u_s$: $\mathbf{z}_s = \mathrm{Concat}(\mathbf{z}^g_s, \mathbf{z}^h_s, \mathbf{z}^u_s)$. For redirecting the gaze and head direction, $\mathbf{z}^g_s$ and $\mathbf{z}^h_s$ are rotated to $\widetilde{\mathbf{z}}^g_t$ and $\widetilde{\mathbf{z}}^h_t$, respectively, by a conventional rotation process R [45]. R rotates the source feature to the target feature using gaze and head pose ground-truths (GTs) of source and target (cf. **Appendix** for more details). Also, it is used to rotate the new feature of phase 2 to the target feature. To preserve identity and details, $\mathbf{z}^u_s$ is fed directly to generator $\mathcal{G}$, and $\mathcal{G}$ generates a redirected image $\widetilde{\mathbf{x}}_t$ using rotated features ($\widetilde{\mathbf{z}}^g_t$ and $\widetilde{\mathbf{z}}^h_t$) and $\mathbf{z}^u_s$: $\widetilde{\mathbf{x}}_t = \mathcal{G}(\widetilde{\mathbf{z}}_t)$ where $\widetilde{\mathbf{z}}_t = \mathrm{Concat}(\widetilde{\mathbf{z}}^g_t, \widetilde{\mathbf{z}}^h_t, \mathbf{z}^u_s)$.

In phase 2, the target image $\mathbf{x}_t$ is encoded in the same way as $\mathbf{x}_s$ by $\mathcal{E}$: $\mathbf{z}_t = \mathrm{Nm}(\mathcal{E}(\mathbf{x}_t)) = \mathrm{Concat}(\mathbf{z}^g_t, \mathbf{z}^h_t, \mathbf{z}^u_t)$. Then, GHT (denoted by T) generates the new gaze feature $\mathbf{z}^g_{tr}$ through the linear interpolation between source and target gaze feature. To learn the new direction of $\mathbf{z}^g_{tr}$, $\mathbf{z}^g_{tr}$ is redirected to $\widetilde{\mathbf{z}}^g_{tr}$ which represents the target gaze direction through R. Here, self-labels are used for redirection of $\mathbf{z}^g_{tr}$ (cf. Sec. 3.2). Finally, gaze consistency loss $\mathcal{L}_{cns}$ based on cosine distance between $\widetilde{\mathbf{z}}^g_{tr}$ and $\widetilde{\mathbf{z}}^g_t$ is minimized (cf. Eq. 2).

In phase 3, in order to supervise the gaze and head direction of $\widetilde{\mathbf{x}}_t$, the redirection loss $\mathcal{L}_{red}$, which is angular error between the gaze (or head) directions of $\widetilde{\mathbf{x}}_t$ and $\mathbf{x}_t$ estimated by the pre-trained networks ψ, is minimized (cf. Sec. 3.4). ψ was pre-trained with the gaze (and head) estimation task and was frozen during training.

3.2. Gaze Hardness-aware Learning

This section describes Gaze Hardness-aware Transformation (GHT) to create a new gaze feature. Generating additional supervision of gaze directions is the core of GHT. Specifically, GHT creates views that cannot be expressed with $\mathbf{z}^g_s$ and $\mathbf{z}^g_t$ alone. Inspired by hardness-aware interpolation [44], we define a transformed feature $\mathbf{z}^g_{tr}$ through linear interpolation as follows:

$$\mathbf{z}^g_{tr} = \alpha_{sim}\mathbf{z}^g_s + (1 - \alpha_{sim})\,\mathbf{z}^g_t, \tag{1}$$

where $\alpha_{sim} \in (0, 1)$ is an adaptive coefficient that is initialized to 0.5, and $\mathbf{z}^g_{tr}$ is generated as many as the number of mini-batches or more. α_{sim} in Eq. 1 increases as learning progresses, and the proportion of $\mathbf{z}^g_t$ decreases gradually. Weakening the influence of $\mathbf{z}^g_t$ including the GT gaze direction make it harder to learn the gaze direction of $\mathbf{z}^g_{tr}$. Therefore, $\mathbf{z}^g_{tr}$ serves as an additional supervision of gaze directions that the source and target cannot see, and contributes to learning the gaze consistency between gaze features. To update α_{sim} and learn the generated gaze feature, we define a loss function $\mathcal{L}_{cns}$ based on redirected gaze consistency as follows:

$$\mathcal{L}_{cns} = 1 - \alpha_{sim} \quad \text{s.t.} \quad \alpha_{sim} = \cos(\widetilde{\mathbf{z}}^g_{tr}, \widetilde{\mathbf{z}}^g_t), \tag{2}$$

where $\cos(\widetilde{\mathbf{z}}^g_{tr}, \widetilde{\mathbf{z}}^g_t) = \frac{\widetilde{\mathbf{z}}^g_{tr} \cdot \widetilde{\mathbf{z}}^g_t}{\|\widetilde{\mathbf{z}}^g_{tr}\| \|\widetilde{\mathbf{z}}^g_t\|}$ and $\|\cdot\|$ is L2 norm. $\widetilde{\mathbf{z}}^g_{tr}$ and $\widetilde{\mathbf{z}}^g_t$ are the gaze features in which $\mathbf{z}^g_{tr}$ and $\mathbf{z}^g_s$ are redirected toward the gaze direction of the target image using R, respectively. The self-label of $\mathbf{z}^g_{tr}$ required to obtain $\widetilde{\mathbf{z}}^g_{tr}$ is calculated by substituting the gaze labels of source and target images into Eq. 1. Note that $\mathbf{z}^g_{tr}$ and $\mathbf{z}^g_s$ are redirected in the gaze direction of the same $\mathbf{z}^g_t$, so the ideal value of α_{sim}, i.e. $\cos(\widetilde{\mathbf{z}}^g_{tr}, \widetilde{\mathbf{z}}^g_t)$, is 1. In the ideal case, $\mathbf{z}^g_{tr}$ is generated only from $\mathbf{z}^g_s$, which corresponds to the most difficult level of learning the gaze representation.

By doing this, generated gaze features allow neural networks to learn not only the gaze given the data but also various gaze directions. Actually, it was experimentally confirmed that the generalization performance of the proposed method improves in the cross-dataset setting as the number of $\mathbf{z}^g_{tr}$ increases (cf. Sec. 4.4).

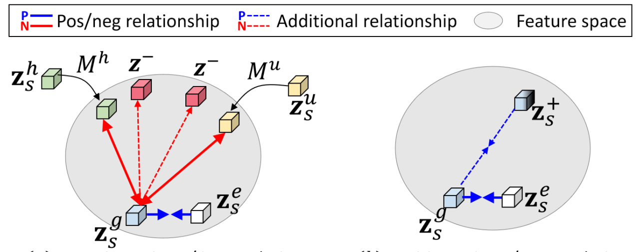
3.3. SG Loss Function

We want to make the change of gaze direction in the redirection process less affected by head pose and task-irrelevant features. For this disentanglement property, we propose similarity learning between features through metric loss.

The basic idea of triplet tuple-based similarity learning is to define the same classes as positive pairs, and define different classes as negative pairs. Inspired by psychological studies [34, 19] that gaze is actually associated with gaze-irrelevant factors such as head pose, we form negative pairs ($\mathbf{z}^g_s$,$\mathbf{z}^h_s$), ($\mathbf{z}^g_s$,$\mathbf{z}^u_s$) by defining $\mathbf{z}^h_s$ and $\mathbf{z}^u_s$ as negative attributes for $\mathbf{z}^g_s$, respectively.

However, a positive pair cannot be defined only with $\mathbf{z}^g_s$ of a single attribute. Inspired by a prior art [6] that the feature extracted from the eye image can represent *fine-grained gaze*, we define the eye feature $\mathbf{z}^e_s$ extracted from an encoder E_{eye} with the cropped eye image $\mathbf{x}^e_s$ as input. E_{eye} is pre-trained ResNet-18 with gaze estimation task and is frozen during training. That is, $\mathbf{z}^e_s = E_{eye}(\mathbf{x}^e_s)$. As a result, a positive pair is defined as $(\mathbf{z}^g_s, \mathbf{z}^e_s)$. Note that an eye image has a fine-grained gaze property although $\mathbf{z}^g_s$ *and* $\mathbf{z}^e_s$ *are extracted from the different networks*, so the proposed positive pair can contribute to the similarity learning.

However, comparing with previous studies [10, 17] handling dozens or hundreds of class labels, we have only two negative attributes $(\mathbf{z}^h_s, \mathbf{z}^u_s)$ to discriminate gaze features. Inspired by [46], we generate additional negative samples by linear interpolation of a negative pairs, i.e. $(\mathbf{z}^g_s, \mathbf{z}^h_s)$ and $(\mathbf{z}^g_s, \mathbf{z}^u_s)$.



(a) Negative pairs w/ interpolation (b) Positive pairs w/ extrapolation

Figure 3: Conceptual illustration of the proposed SG loss function. (a) Interpolated feature $\mathbf{z}^-_s$ is used for additional negative sample. (b) Extrapolated feature $\mathbf{z}^+_s$ provides a hard example for learning positive pairs.

$$\begin{aligned} \mathbf{z}^-_s &= \mathrm{Nm}\left(\mathrm{M}^h(\mathbf{z}^h_s) + (\mathbf{z}^g_s - \mathrm{M}^h(\mathbf{z}^h_s))\alpha^-\right) \\ \text{s.t.} &\quad \mathrm{M}^h(\mathbf{z}^h_s) = \mathrm{Nm}(\mathrm{ReLU} \circ \mathrm{Linear}(\mathbf{z}^h_s)), \end{aligned} \tag{3}$$

where $\alpha^- \sim \mathrm{Beta}(2.0, 2.0)$ has the range of [0,1] and the fine-grained $\mathbf{z}^e_s$ can replace $\mathbf{z}^g_s$ (cf. Sec. 4.4). Note that $\mathbf{z}^u_s$ is also used to generate an additional negative sample instead of $\mathbf{z}^h_s$ of Eq. 3. Here, M^u is used instead of M^h.

Here, since $\mathbf{z}^h_s$ and $\mathbf{z}^u_s$ represent heterogeneous attributes, $\mathbf{z}^-_s$ (of Eq. 3) generated through naive linear interpolation can be regarded as easy negative samples. So, we realign $\mathbf{z}^h_s$ and $\mathbf{z}^u_s$ to the surface of the unit hypersphere using two multi-layer perceptrons (MLPs), i.e., M^h and M^u. Through this additional alignment process, $\mathbf{z}^-_s$ can not only be located on the same level of feature space, but also can be utilized as useful samples for metric learning. Now, $\mathbf{z}^-_s$ includes semantic attributes that $\mathbf{z}^h_s$ and $\mathbf{z}^u_s$ cannot express, and is defined as a negative pair by binding to $\mathbf{z}^g_s$ (or $\mathbf{z}^e_s$), which acts as an anchor (see Fig. 3(a)). Also, since $\mathbf{z}^-_s$ are uniformly distributed, the bias problem of pair-based similarity learning can be alleviated (cf. **Appendix**).

On the other hand, similarity learning of positive pairs $(\mathbf{z}^g_s, \mathbf{z}^e_s)$ with relatively less constraints than negative pairs is tempted to have a trivial solution. To prevent this problem, we generate a hard positive (proxy) vector $\mathbf{z}^+_s$ through the feature extrapolation [46] as follows:

$$\mathbf{z}^+_s = \mathrm{Nm}\left(\mathbf{z}^g_s + (\mathbf{z}^e_s - \mathbf{z}^g_s)\alpha^+\right), \tag{4}$$

where $\alpha^+ = \alpha^- + 1$ sampled in the range of [1,2] represents the extrapolation coefficient. A proxy vector $\mathbf{z}^+_s$ located in the vicinity of a pair of positive relationships $(\mathbf{z}^g_s, \mathbf{z}^e_s)$ provides an additional constraint so that $(\mathbf{z}^g_s, \mathbf{z}^e_s)$ does not have a trivial solution (see Fig. 3(b)).

Finally, the SG loss function $\mathcal{L}_{sg}$ is defined based on the basic triplets $(\mathbf{z}^g_s, \mathbf{z}^e_s, \mathbf{z}^h_s$ or $\mathbf{z}^u_s)$ and the additional vectors defined above.

$$
\begin{aligned}
\mathcal{L}_{sg} =& \frac{1}{2|\mathcal{P}|} \sum_{(i,j)\in\mathcal{P}} \max\left(0, J_{i,j}\right)^2 \\
J_{i,j} =& S\left(\frac{D_{i,j}}{\tau}\right) + \sum_{(i,k)\in\mathcal{N}} S\left(\frac{\delta - D_{i,k}}{\tau}\right) \\
&+ \sum_{(j,l)\in\mathcal{N}} S\left(\frac{\delta - D_{j,l}}{\tau}\right),
\end{aligned}
\tag{5}
$$

where $\mathcal{P}$ and $\mathcal{N}$ represent the sets of all positive and negative pairs in a mini-batch, respectively. $D_{i,j} = \|\mathbf{z}_i - \mathbf{z}_j\|^2$ stands for the Euclidean distance between vectors. $S(\cdot)(= \ln(1+\exp(\cdot)))$ indicates the softplus function. δ is the margin for negative samples and was set to 1.3. τ is the temperature hyper-parameter and was set to 0.89. Note that indices i and j of Eq. 5 correspond to $\mathbf{z}_s^g$ and $\mathbf{z}_s^e$ (or $\mathbf{z}_s^+$), respectively, and the similarity of positive pair is calculated through $D_{i,j}$. Indices k, l correspond to $\mathbf{z}_s^h$ (or $\mathbf{z}_s^u$) and $\mathbf{z}_s^-$ having negative relationship with elements of positive pair, respectively. Therefore, the SG loss function follows a so-called *structured formula* in which all combinations of positive and negative pairs are considered for similarity learning. Refer to **Appendix** for further analysis of the SG loss function and generalized contrastive loss.

3.4. Total Loss Function

The total loss function of the proposed method is defined as follows:

$$
\mathcal{L} = \frac{1}{N}\sum_{n=1}^{N}(\lambda_{red}\mathcal{L}^n_{red} + \lambda_{cns}\mathcal{L}^n_{cns} + \mathcal{L}^n_{other}) + \lambda_{sg}\mathcal{L}_{sg}, \tag{6}
$$

where λ_{red}, λ_{cns}, and λ_{sg} are set to 5.0, 2.0 and 10.0, respectively. N is the size of the mini-batch. The first term $\mathcal{L}_{red}$ is a loss function calculated through the mean angular error (MAE) metric between $\widetilde{\mathbf{x}}_t$ and $\mathbf{x}_t$. That is, $\mathcal{L}_{red} = \mathrm{MAE}(\psi(\widetilde{\mathbf{x}}_t), \psi(\mathbf{x}_t))$, where $\mathrm{MAE}(\mathbf{a}, \mathbf{b}) = \cos^{-1}\frac{\mathbf{a}\cdot\mathbf{b}}{\|\mathbf{a}\|\|\mathbf{b}\|}$ and ψ is ResNet-18 pre-trained with gaze or head direction estimation task [45]. The second term $\mathcal{L}_{cns}$ is a loss function for consistency learning between two redirected gaze features (see Eq. 2). The third term $\mathcal{L}_{sg}$ is a loss function for discriminative learning of gaze features (see Eq. 5). The last term $\mathcal{L}_{other}$ consists of pixel-wise reconstruction and perceptual loss functions between $\widetilde{\mathbf{x}}_t$ and $\mathbf{x}_t$ for further feature regularization [45] (cf. **Appendix**). The loss functions except for $\mathcal{L}_{sg}$ are calculated with N samples, and $\mathcal{L}_{sg}$ is computed as much as the sizes of the positive and negative transformed features, i.e., $|\mathcal{P}|$ and $|\mathcal{N}|$.

4. Experiments

Configurations. We implemented the neural networks using the PyTorch library [28] and the following experiments were performed in the environment of AMD 7742 CPU and NVIDIA A100 GPU. Each experiment was repeated three times. This is more reliable compared to STED of only one-time experiment. In Fig. 2, an encoder $\mathcal{E}$ and a generator $\mathcal{G}$ are based on DenseNet-based architecture [45], and an input image is resized to 128×128. Eye features extraction network E_{eye} is the pre-trained ResNet-18 [6] with manually cropped eye images as input. As with other methods [31], we use the data normalization procedure [42] that pre-processes the gaze dataset to exclude the roll component of head orientation.

The learning parameters of the designed neural network are updated by repeating the forward and backward processes about 140K times. The initial learning rate (LR) of $\mathcal{E}$ and $\mathcal{G}$ was set to 10^{-3}, and a step LR scheme that decreases LR by 0.8 times every 25K iterations was used. The weight decay coefficient was 10^{-4}, and Adam optimizer [18] was employed. The mini-batch size was set to 32.

4.1. Dataset and Evaluation Metrics

We adopted open datasets that could be used for research purposes, and informed consent was obtained in the case of *EYEDIAP* [7]. We used a total of four gaze datasets: *GazeCapture* [20], *MPIIGaze* [43], *Columbia Gaze* [31], and *EYEDIAP* [7]. The datasets include annotated head pose and gaze direction information. *GazeCapture* consists of 2M images acquired from 1,474 subjects in an unconstrained setting. *MPIIGaze* consists of 213,569 images of 15 subjects acquired in daily life. *Columbia Gaze* contains 6,000 images from 56 subjects. *EYEDIAP* is a gaze dataset derived from 16 subjects. Our model was trained on the train split of the *GazeCapture* dataset, and the generalization performance was verified through cross-dataset evaluation for three different gaze datasets.

A total of four evaluation metrics were used to evaluate the proposed method. First, $\boldsymbol{err_g}$ represents the MAE between GT and the prediction of gaze direction, which were estimated from $\mathbf{x}_t$ and $\widetilde{\mathbf{x}}_t$ by ψ pre-trained on gaze (or head pose) estimation task [45], respectively. So does $\boldsymbol{err_h}$. *Disentanglement error* is a metric for measuring the mutual influence of factors such as gaze and inessential features. For example, the disentanglement error of gaze to head ($g \rightarrow h$) is the MAE between the head pose GT and the redirected image from $\mathbf{z}_s$ including the perturbed gaze feature $\hat{\mathbf{z}}_s^g$. Here, the perturbed gaze feature $\hat{\mathbf{z}}_s^g$ is the result of adding uniform distribution-based random perturbation $\varepsilon \sim U(-0.1\pi, 0.1\pi)$ to $\mathbf{z}_s^g$: $\hat{\mathbf{z}}_s^g = \mathbf{z}_s^g + \varepsilon$. In addition, various combinations of features and GTs were utilized for disentanglement errors: $h \rightarrow g$, the effect of change in head pose factor on gaze direction, and $u \rightarrow g(/h)$, the effect of changes in task-irrelevant factors on gaze (head pose) direction. Finally, *LPIPS* [15] is a metric that measures the perceptual similarity between $\mathbf{x}_t$ and $\widetilde{\mathbf{x}}_t$, and quantifies the visual quality of redirected image [11, 45].

Table 1: Quantitative results of within-dataset evaluation protocol. "†" denotes our reproduced result. (a) Comparision with the state-of-the-art methods on the GazeCapture dataset. The results of FAZE and STED were borrowed from [26] and [45]. Here, in the case of FAZE, $u \to g(/h)$ metric was excluded because it does not have a task-irrelevant feature. The percentage indicates the degree of improvement of the proposed method compared to STED. (b) Comparision with the STED on the MPIIGaze, Columbia and EYEDIAP datasets.

Method	err_g	$u \to g$	$h \to g$	err_h	$u \to h$	$g \to h$	LPIPS
StarGAN [4]	4.602	-	-	3.989	-	-	0.257
He *et al.* [11]	4.617	-	-	1.392	-	-	0.223
GazeFlow† [37]	5.314	-	-	4.122	-	-	0.255
FAZE [26]	7.114	-	4.882	2.470	-	0.542	0.279
STED [45]	2.195	0.507	2.072	0.816	0.211	0.388	0.205
Ours	**1.884** ▼14.2%	**0.372** ▼26.7%	**1.902** ▼6.7%	**0.72** ▼11.7%	**0.184** ▼12.8%	**0.342** ▼11.9%	**0.199** ▼2.9%

(a) GazeCapture [20]

Dataset	Method	err_g	$u \to g$	$h \to g$	err_h	$u \to h$	$g \to h$	LPIPS
MPIIGaze	STED†	2.133	0.605	2.312	0.724	0.314	0.442	0.204
	Ours	**1.814**	**0.512**	**1.994**	**0.684**	**0.211**	**0.339**	**0.202**
Columbia	STED†	3.134	0.902	3.307	**0.886**	0.334	1.002	0.233
	Ours	**2.872**	**0.782**	**2.902**	0.902	**0.314**	**0.987**	**0.212**
EYEDIAP	STED†	13.094	6.413	12.796	0.817	0.662	1.674	**0.224**
	Ours	**11.094**	**5.498**	**9.438**	**0.802**	**0.403**	**0.904**	0.232

(b) MPIIGaze [43], Columbia [31] and EYEDIAP[7]

Table 2: Quantitative results of cross-dataset evaluation protocol. All methods are trained on GazeCapture dataset. "†" denotes our reproduced result.

Test dataset	MPIIGaze			Columbia			EYEDIAP		
Method	err_g	$h \to g$	LPIPS	err_g	$h \to g$	LPIPS	err_g	$h \to g$	LPIPS
StarGAN	4.488	2.783	0.260	6.522	3.359	0.255	14.906	4.025	0.248
He *et al.*	5.092	3.411	0.241	7.345	3.831	0.227	13.548	3.831	0.218
GazeFlow†	6.024	4.917	0.244	8.933	4.120	0.234	18.344	4.953	0.231
FAZE†	6.894	4.114	0.221	9.233	4.324	0.247	19.563	5.122	0.24
STED	2.233	1.849	0.203	3.333	2.136	0.242	11.290	2.670	0.213
Ours	**1.998** ▼10.5%	**1.714** ▼7.3%	**0.194** ▼4.4%	**3.002** ▼9.9%	**1.974** ▼7.5%	**0.221** ▼8.6%	**10.231** ▼9.3%	**2.134** ▼20.0%	**0.204** ▼4.2%

4.2. Quantitative results

Within-Dataset Evaluation. Table 1 shows the performance of the proposed method according to the so-called within-dataset evaluation protocol. Table 1a compares the proposed method with other methods for the *GazeCapture* dataset. The proposed method outperformed the other SOTA methods in all metrics. For example, the proposed method achieved err_g of $1.884°$, which was improved by 14.2% compared to STED. Also, the proposed method showed $h \to g$ of $1.902°$, which is 6.7% better than STED. This shows that the consistency and disentanglement properties of latent features are important for learning auto-encoding of TA. Meanwhile, Table 1b shows the within-dataset evaluation results for the *MPIIGaze*, *Columbia*, and *EYEDIAP* datasets, respectively. Here, STED, which achieved the highest performance among existing methods, was compared with the proposed method. Comparison results with the other existing methods are reported in the **Appendix**. Note that the proposed method outper-

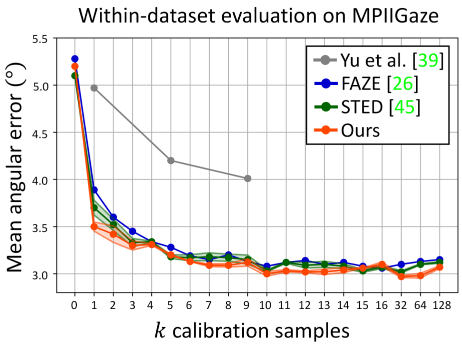


Figure 4: Performance of few-shot gaze estimation for the *MPIIGaze* dataset. We expressed the standard deviation of MAE as a shade overlaid on each curve. The curve of STED [45] was calculated by applying the learned representation of STED to the gaze estimator. For the results of [39, 26] were borrowed from the papers.

formed STED in most metrics. This means that the proposed method consistently contributes to the performance improvement regardless of the dataset.

Cross-Dataset Evaluation. Table 2 shows the strength of the proposed method through a cross-dataset evaluation protocol with different training and evaluation datasets. Similar to the within-dataset protocol in Table 1, the proposed method showed performance superiority over the baseline methods in the cross-dataset protocol. In particular, note that the proposed method achieved average 9.9% lower err_g than STED for the three datasets. Also, the proposed method generated redirected images of visually higher quality and showed slightly better performance even in terms of LPIPS metric. This is verified in the user study of Sec. 4.3.

Evaluation of learned representation. We evaluated the learned representation through a few-shot gaze estimation task. We trained the gaze estimator using only a few calibration samples. The gaze estimator is designed with a two-layer MLP, and it outputs a three-dimensional gaze direction vector by receiving the learned gaze representation. During the training of the gaze estimator, the encoder $\mathcal{E}$ is frozen. In the *MPIIGaze* dataset, 500 images per subject were used for evaluation. k calibration samples were randomly selected from the remaining samples and used as training data for the gaze estimator. Each experiment was repeated 10 times to calculate the mean and standard deviation. Fig. 4 shows the few-shot gaze estimation performance of several methods [39, 26, 45] in the *MPIIGaze* dataset. In most cases (for $k > 5$), the proposed method outperformed the previous works. This proves the superiority of the gaze representation learned by our model.

4.3. Qualitative Results

We used ContraCAM [24], an up-to-date visualization technique, to prove the effectiveness of the proposed discriminative learning. ContraCAM, which can utilize a continuous type of GT for calculating activation maps, is more suitable for the proposed method with continuous gaze or head pose as GT than a class probability score (cf. **Appendix** for implementation details).

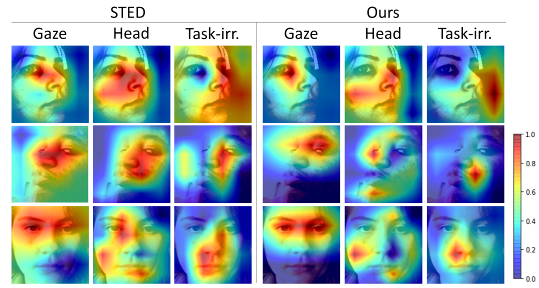


Figure 5: ContraCAM [24] visualization on the test split of *GazeCapture* dataset.

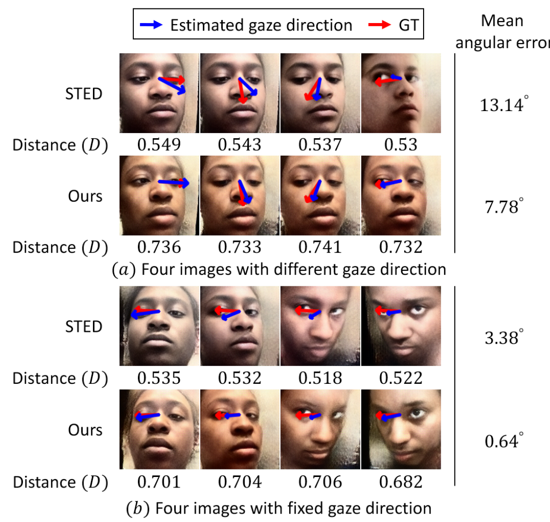


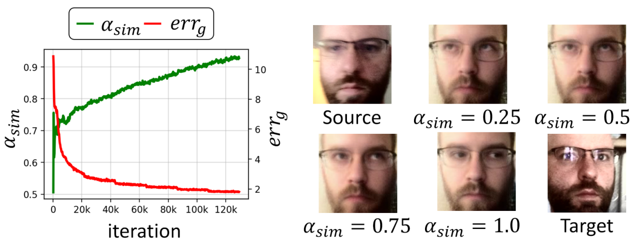
Figure 6: Experiments of latent space traversal on *GazeCapture* dataset. A sequence of facial images (a) with randomly selected four gaze directions and (b) with the same gaze direction while changing inessential features.

Figure 5 visualizes the feature maps of STED and the proposed method. The gaze features of STED pay attention to the non-eye regions that are little related to gaze. On the other hand, the gaze features of the proposed method focus only on the eye region, and the inessential features point to regions independent of gaze features.

Figure 6 analyzes the effect of gaze feature discrimination on gaze redirection through a qualitative comparison of the proposed method and STED. In Fig. 6(a), the proposed method tracks the direction change of GT well and shows a significantly lower MAE than STED. In addition, Euclidean distance (D) quantitatively measures how much the gaze features and the inessential features are disentangled from each other. In Fig. 6(b), the same tendency was observed even when the characteristics of the inessential

Table 3: Voting results of user study comparing STED with our method. Each column sums up to 100%. The degree indicates the projection of the gaze direction (pitch, yaw, roll) onto the image plane and increases clockwise. 0° is the left side from the center of the face.

Method	[0°,120°)	[120°,240°)	[240°,360°)	Mean
STED	14.6%	29%	20.2%	21.3%
Ours	85.4%	71%	79.8%	78.7%



(a) α_{sim} and err_g (b) Generated images using $\mathbf{z}^g_{tr}$

Figure 7: (a) Learning procedure of err_g and α_{sim}. (b) Some samples of generated images according to α_{sim}.

features were changed while the direction of the gaze feature was fixed.

In addition, we conducted a user study to evaluate the proposed method. We randomly chose 50 pairs of images generated by the proposed method and STED, with the same input image and gaze direction. For each image, 13 subjects were asked to select the redirected image that looks more similar with the GT. As in Table 3, the proposed method outperformed STED by up to 57%.

4.4. Ablation Study

This section regards an ablation study analyzing the effects of key components of the proposed method. First, Fig. 7(a) shows the transition of α_{sim} and err_g during training. We can observe that err_g decreases every 20K iterations thanks to $\mathbf{z}^g_{tr}$, which can alleviate the overfitting problem of the network at the later stage of learning. Fig. 7(b) shows the phenomenon that the subject's gaze moves from the direction of the source to that of the target according to α_{sim}, which adjusts the ratio of the target's gaze direction.

Next, we analyzed the proportion of the proposed $\mathcal{L}_{cns}$ and $\mathcal{L}_{sg}$ in performance improvement. As shown in Table 4, the contribution of $\mathcal{L}_{cns}$ to the performance improvement was slightly greater than that of $\mathcal{L}_{sg}$. Case (d) shows the effect of SG loss without using $\mathbf{z}^-_s$ and $\mathbf{z}^+_s$ (cf. Section 3.3), i.e., $\mathcal{L}^{\text{wo-ft}}_{sg}$ on performance. Compared to case (c) which showed only marginal improvement, case (d) showed a significant performance increase in all metrics. This proves the effect of feature transformation to generate hard negative and positive samples for SG loss.

Table 4: Effect of gaze consistency loss ($\mathcal{L}_{cns}$), SG loss without feature transformation ($\mathcal{L}^{\text{wo-ft}}_{sg}$) and full SG loss ($\mathcal{L}_{sg}$) on the entire performance. *GazeCapture* dataset was used for this experiment.

Case	$\mathcal{L}_{cns}$	$\mathcal{L}^{\text{wo-ft}}_{sg}$	$\mathcal{L}_{sg}$	err_g	$h \rightarrow g$	LPIPS
(a)				2.334	2.414	0.237
(b)	✓			2.100	2.339	0.211
(c)		✓		2.221	2.329	0.233
(d)			✓	2.134	2.018	0.219
(e)	✓		✓	**1.884**	**1.902**	**0.199**

Table 5: Performance of the proposed method according to the number of $\mathbf{z}^g_{tr}$ on *Columbia* dataset.

# of $\mathbf{z}^g_{tr}$	$1N$	$10N$	$20N$	$50N$
err_g	2.872	2.714	2.364	**2.112**

Also, case (e) shows that the two loss functions cause a synergistic effect with each other. Finally, Table 5 shows the performance of the proposed method according to the number of $\mathbf{z}^g_{tr}$. As the number of $\mathbf{z}^g_{tr}$ increases, err_g becomes lower because our model can learn fine-grained gaze directions between source and target. We reported the additional results of the ablation study in **Appendix**. They include the influence of M^h (or M^u), $\mathbf{z}^e_s$ and batch-size. Finally, the result when the other metric loss (margin loss [35] and signal-to-noise (SNR) loss [41]) is reported as well.

5. Conclusion

We succeeded in augmenting and manipulating gaze features including various gaze directions through GHT. The generated gaze features serve as additional supervision, improving the generalization performance of gaze redirection. In the future, GHT will be used for various purposes in gaze representation learning requiring heavy annotation costs. Also, the SG loss function for discriminative learning of features can be extended to other computer vision tasks such as recognition of facial emotions or gestures.

Acknowledgement

This work was supported by IITP grants funded by the Korea government (MSIT) (No. 2021-0-02068, AI Innovation Hub and RS-2022-00155915, Artificial Intelligence Convergence Research Center (Inha University)), and was supported by the NRF grant funded by the Korea government (MSIT) (No. 2022R1A2C2010095 and No. 2022R1A4A1033549).

References

[1] Jingjing Chen, Jichao Zhang, Enver Sangineto, Tao Chen, Jiayuan Fan, and Nicu Sebe. Coarse-to-fine gaze redirection with numerical and pictorial guidance. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3665–3674, 2021.

[2] Weihua Chen, Xiaotang Chen, Jianguo Zhang, and Kaiqi Huang. Beyond triplet loss: a deep quadruplet network for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 403–412, 2017.

[3] De Cheng, Yihong Gong, Sanping Zhou, Jinjun Wang, and Nanning Zheng. Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In *Proceedings of the iEEE conference on computer vision and pattern recognition*, pages 1335–1344, 2016.

[4] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8789–8797, 2018.

[5] Yueqi Duan, Wenzhao Zheng, Xudong Lin, Jiwen Lu, and Jie Zhou. Deep adversarial metric learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2780–2789, 2018.

[6] Yi Fang, Jiapeng Tang, Wang Shen, Wei Shen, Xiao Gu, Li Song, and Guangtao Zhai. Dual attention guided gaze target detection in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11390–11399, 2021.

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

[8] Yaroslav Ganin, Daniil Kononenko, Diana Sungatullina, and Victor Lempitsky. Deepwarp: Photorealistic image resynthesis for gaze manipulation. In *European conference on computer vision*, pages 311–326. Springer, 2016.

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

[10] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020.

[11] Zhe He, Adrian Spurr, Xucong Zhang, and Otmar Hilliges. Photo-realistic monocular gaze redirection using generative adversarial networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6932–6941, 2019.

[12] Geoffrey E Hinton, Alex Krizhevsky, and Sida D Wang. Transforming auto-encoders. In *International conference on artificial neural networks*, pages 44–51. Springer, 2011.

[13] Junlin Hu, Jiwen Lu, and Yap-Peng Tan. Discriminative deep metric learning for face verification in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1875–1882, 2014.

[14] Chen Huang, Chen Change Loy, and Xiaoou Tang. Local similarity-aware deep feature embedding. *Advances in neural information processing systems*, 29:1262–1270, 2016.

[15] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016.

[16] Harsimran Kaur and Roberto Manduchi. Subject guided eye image synthesis with application to gaze redirection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 11–20, 2021.

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

[18] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[19] Nathan L Kluttz, Brandon R Mayes, Roger W West, and Dave S Kerby. The effect of head turn on the perception of gaze. *Vision research*, 49(15):1979–1993, 2009.

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

[21] Eui Chul Lee, Kang Ryoung Park, Min Cheol Whang, and Junseok Park. Robust gaze tracking method for stereoscopic virtual reality systems. In *International Conference on Human-Computer Interaction*, pages 700–709. Springer, 2007.

[22] Gang Liu, Yu Yu, Kenneth Alberto Funes Mora, and Jean-Marc Odobez. A differential approach for gaze estimation. *IEEE transactions on pattern analysis and machine intelligence*, 2019.

[23] Yoshio Matsumoto and Alexander Zelinsky. An algorithm for real-time stereo vision implementation of head pose and gaze direction measurement. In *Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580)*, pages 499–504. IEEE, 2000.

[24] Sangwoo Mo, Hyunwoo Kang, Kihyuk Sohn, Chun-Liang Li, and Jinwoo Shin. Object-aware contrastive learning for debiased scene representation. *arXiv preprint arXiv:2108.00049*, 2021.

[25] Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. Deep metric learning via lifted structured feature embedding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4004–4012, 2016.

[26] Seonwook Park, Shalini De Mello, Pavlo Molchanov, Umar Iqbal, Otmar Hilliges, and Jan Kautz. Few-shot adaptive gaze estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9368–9377, 2019.

[27] Seonwook Park, Adrian Spurr, and Otmar Hilliges. Deep pictorial gaze estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 721–738, 2018.

[28] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32:8026–8037, 2019.

[29] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

[30] Ali Shafti, Pavel Orlov, and A Aldo Faisal. Gaze-based, context-aware robotic system for assisted reaching and grasping. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 863–869. IEEE, 2019.

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

[32] Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. In *Advances in neural information processing systems*, pages 1857–1865, 2016.

[33] Roberto Valenti, Nicu Sebe, and Theo Gevers. Combining head pose and eye location information for gaze estimation. *IEEE Transactions on Image Processing*, 21(2):802–815, 2011.

[34] William Hyde Wollaston. Xiii. on the apparent direction of eyes in a portrait. *Philosophical Transactions of the Royal Society of London*, 114:247–256, 1824.

[35] Chao-Yuan Wu, R Manmatha, Alexander J Smola, and Philipp Krahenbuhl. Sampling matters in deep embedding learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2840–2848, 2017.

[36] Min Wu, Tyron Louw, Morteza Lahijanian, Wenjie Ruan, Xiaowei Huang, Natasha Merat, and Marta Kwiatkowska. Gaze-based intention anticipation over driving manoeuvres in semi-autonomous vehicles. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 6210–6216. IEEE, 2019.

[37] Yong Wu, Hanbang Liang, Xianxu Hou, and Linlin Shen. Gazeflow: Gaze redirection with normalizing flows. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2021.

[38] Weihao Xia, Yujiu Yang, Jing-Hao Xue, and Wensen Feng. Controllable continuous gaze redirection. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1782–1790, 2020.

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

[40] Yu Yu and Jean-Marc Odobez. Unsupervised representation learning for gaze estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7314–7324, 2020.

[41] Tongtong Yuan, Weihong Deng, Jian Tang, Yinan Tang, and Binghui Chen. Signal-to-noise ratio: A robust distance metric for deep metric learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4815–4824, 2019.

[42] Xucong Zhang, Yusuke Sugano, and Andreas Bulling. Revisiting data normalization for appearance-based gaze estimation. In *Proceedings of the 2018 ACM symposium on eye tracking research & applications*, pages 1–9, 2018.

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

[44] Wenzhao Zheng, Zhaodong Chen, Jiwen Lu, and Jie Zhou. Hardness-aware deep metric learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 72–81, 2019.

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

[46] Rui Zhu, Bingchen Zhao, Jingen Liu, Zhenglong Sun, and Chang Wen Chen. Improving contrastive learning by visualizing feature transformation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10306–10315, 2021.