Denoising Diffusion Models are Good General Gaze Feature Learners

Guanzhong Zeng[1] , **Jingjing Wang**[1] , **Pengwei Yin**[1] , **Zefu Xu**[1] , **Mingyang Zhou**[2]
[1]Hikvision Research Institute
[2]Shenzhen University
{zengguanzhong, wangjingjing9, yinpengwei, xuzefu}@hikvision.com, zmy@szu.edu.cn

Abstract

Since the collection of labeled gaze data is laborious and time-consuming, methods which can learn generalizable features by leveraging large-scale available unlabeled data are desirable. In recent years, we have witnessed the tremendous capabilities of diffusion models in generating images as well as their potential in feature representation learning. In this paper, we investigate whether they can acquire discriminative representations for gaze estimation via generative pre-training. To achieve this goal, we propose a self-supervised learning framework with diffusion models for gaze estimation, called GazeDiff. Specifically, we utilize a conditional diffusion model to generate target image with gaze direction specified by the reference image as the pre-training task. To facilitate the diffusion model to learn gaze related features as condition, we propose a disentangling feature learning strategy, which first learns appearance feature, head pose feature, and eye direction feature respectively, and then combines them as the conditional features. Extensive experiments demonstrate denoising diffusion models are also good general gaze feature learners.

1 Introduction

Gaze estimation is the task of measuring the gaze direction of a human in an image. It has been widely applied in human-computer interaction [Zhang *et al.*, 2017a; Sugano *et al.*, 2016; Park *et al.*, 2021], augmented reality [Padmanaban *et al.*, 2017] and driver monitoring system [Mavely *et al.*, 2017]. With the rapid development of deep learning technologies, appearance-based gaze estimations have achieved remarkable breakthroughs and promising performances.

Although deep learning methods have shown promising results for gaze estimation. Training a well-performed model needs sufficiently large and diverse labeled data, covering a wide range of gaze directions, appearances, and head poses, which is very laborious and time-consuming. The limited labeled data hinders the development of gaze estimation methods. When trained on only a small amount of annotated samples, supervised learning methods are easy to overfit the training data, and their performance significantly degrades when encounter a new scenario with different data distribution. Therefore, methods which facilitate training with limited gaze annotations are highly desirable.

Self-supervised learning (SSL) has proven successful at learning generalizable features by leveraging large-scale available unlabeled data. Some researchers use the SSL techniques (e.g. contrastive learning [Du *et al.*, 2023; Jindal and Manduchi, 2023] and cross-encoder disentangling [Sun *et al.*, 2021]) to tackle the issue of data scarcity in the field of gaze estimation, and have achieved promising results. However, most of them are not generative pre-training methods. Recently, we have witnessed remarkable progress at the generative domain, especially with diffusion modeling proving to be a powerful technique which can create vivid imagery of astonishing realism. Following the idea, which considers the ability to create is among the highest manifestations of learning, more recently, a few researchers [Chen *et al.*, 2024; Xiang *et al.*, 2023] start to explore the representational capacity of diffusion models, and show promising results. However, its representational capacity for gaze estimation is still unexplored.

To verify whether denoising diffusion models are good general gaze feature learners, in this paper, we propose an unsupervised gaze representation learning framework based on diffusion modeling as shown in Figure 1. Our intuition is that if the network learns how to generate target images with gaze direction corresponding to the reference images, then it has learned to extract gaze-related features from given reference images. Specifically, we utilize a conditional diffusion model to generate target image with gaze direction specified by the reference image as the pre-training task. The reference image goes through a feature extractor and gets a latent feature as the condition. If we utilize the target image as the reference image, then it behaves more like an autoencoder, without focusing on extracting gaze-related features, which goes against our original intention. To make the extractor learn meaningful features for gaze direction control, we propose a disentangling feature learning strategy. Considering gaze direction is mainly influenced by head pose and eye direction (as illustrated in supplementary material Figure 1), a image with same identity as the target image but with different gaze directions is used as input to extract the appearance

feature. The eye region of the target image is used as input to extract the eye direction feature, and the target image with Gaussian blur augmentation is used as the input to extract the head pose feature. Finally these features are combined as the conditional features to guide the target image generation.

In summary, the contributions of our work are as follows:

- We are the first to propose a self-supervised learning framework based on diffusion modeling for gaze estimation. It leverages the powerful generative ability of diffusion models to enhance the gaze representation learning.
- To facilitate the diffusion model to learn gaze related features as condition to control the gaze direction of generated image accurately, we design a disentangling feature learning strategy. It can learn appearance feature, head pose feature, and eye direction feature respectively, and combine them as the conditional features.
- Our method achieves the state-of-the-art performance in extensive evaluation settings, which demonstrates that, denoising diffusion models are also good general gaze feature learners.

2 Related Work

Appearance-based Gaze Estimation

Result from the availability of large-scale datasets [Zhang *et al.*, 2020; Park *et al.*, 2020; Krafka *et al.*, 2016] and dramatic improvements in computing power, appearance-based gaze estimation methods have received widespread attention and achieved remarkable performance. However, these methods still suffer from obvious performance deterioration on cross-dataset evaluation, owing to the data distribution shifts. The common approach is to learn a more generalizable gaze representation. Such as, PureGaze [Cheng and Bao, 2022] and [Xu *et al.*, 2023] improve robustness to unfavorable interference through adversarial learning. CLIP-Gaze [Yin *et al.*, 2024c] and LG-Gaze [Yin *et al.*, 2025] exploit visual language models to strengthen feature representation. GLA [Zeng *et al.*, 2025] further eliminates label distribution shifts. And [Wang and Yin, 2025; Yin *et al.*, 2024b] aligns feature distribution at test time to improve target domain performance. But these methods are still limited by the diversity of training datasets. In addition, [Ruzzi *et al.*, 2023; Yin *et al.*, 2024a] generates data with gaze annotations to improve sample diversity, but the realness of generated data cannot be guaranteed. There are still certain risk in transfering to unseed domains. So that, fully utilizing massive amounts of unlabeled data for self-supervised learning has gradually become popular.

Gaze Self-supervised Learning

Contrastive learning is widely popular due to its superior performance in self-supervised learning. It maximizes mutual information between latent representations to promote the discrimination ability of feature extractor [Bachman *et al.*, 2019]. Inspired by this idea, GazeCLR [Jindal and Manduchi, 2023] and ConGaze [Du *et al.*, 2023] construct various contrastive tasks to learn effective gaze representation. On the other hand, [Yu and Odobez, 2020] learns low-dimensional gaze representation by gaze redirection, [Sun *et al.*, 2021] proposes latent-code-swapping mechanism to decouple eye features and gaze features. Furthermore, MV-DE [Bao and Lu, 2024] draws upon multi-view constraints and designs a multi-view gaze representation swapping strategy to derive gaze-related feature. Nevertheless, most of them are not generative pre-training methods. Following the idea, which considers the ability to create is among the highest manifestations of learning. A few researchers start to explore the representational capacity of diffusion models, and show promising results. However, its representational capacity for gaze estimation is still unexplored.

Denoising Diffusion Models as Feature Learners

Since the advent of denoising diffusion model, marked breakthroughs have been made in numerous tasks and modalities. Based on its powerful generative capacity, some works [Clark and Jaini, 2024] explored the representation ability of diffusion model. l-DAE [Chen *et al.*, 2024], DDAE [Xiang *et al.*, 2023] and DiffMAE [Wei *et al.*, 2023] auto encode the same image to derive visual representation capability, the final feature learner is detached from a part of the diffusion model. Meanwhile, SODA [Hudson *et al.*, 2024] is proposed to generate novel images from another image and consists of an encoder and denoiser. It considers images as different views that relate visually or semantically, which significantly enhances the encoder's representation skills. Inspired by its success, we explore the potential of diffusion-based representation learning for gaze estimation, and propose an efficient self-supervised framework to learn general gaze representations.

3 Prilimilary

Our research is based on Denoising Diffusion Model (DDM) [Sohl-Dickstein *et al.*, 2015; Ho *et al.*, 2020]. DDM can be briefly considered to contain a pair of forward and backward Markov chains. The forward process starts from a clean data x_0 and sequentially adds noise to it. At a specific time step t, the noised data x_t can be formulated by:

$$\mathrm{x}_t = \sqrt{\bar{\alpha}_t}\mathrm{x}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, \quad t \in \{0, ..., T\} \tag{1}$$

where $\bar{\alpha}_t$ defines the scaling factors of the signal and noise [Ho *et al.*, 2020], $\epsilon \in \mathcal{N}(\mathbf{0}, \mathbf{I})$ is the sampled Gaussian noise.

The backward process performs image denoising in order to recover the clean data. Specifically, DDM employs trainable network to estimate ϵ_t by minimizing:

$$\mathcal{L}_{denoise} = \mathbb{E}_{\mathrm{x}_t,\epsilon,t}\left[\|\epsilon - \epsilon_\theta(\mathrm{x}_t, t)\|^2\right] \tag{2}$$

where $\epsilon_\theta(\cdot)$ is a function approximator intended to predict ϵ from x_t. Therefore, DDM can generate realistic images from noise signal as $\epsilon_\theta(\cdot)$ be iteratively applied. Similar to other types of generative models [Mirza and Osindero, 2014], DDM also are capable of modeling conditional distributions of the form $p(\mathrm{x}|y)$ [Rombach *et al.*, 2022]. This can be implemented by adding condition c into $\epsilon_\theta(\mathrm{x}_t, t, c)$. Based on the additional condition c, DDM can generate the images we need instead of the images with random content.

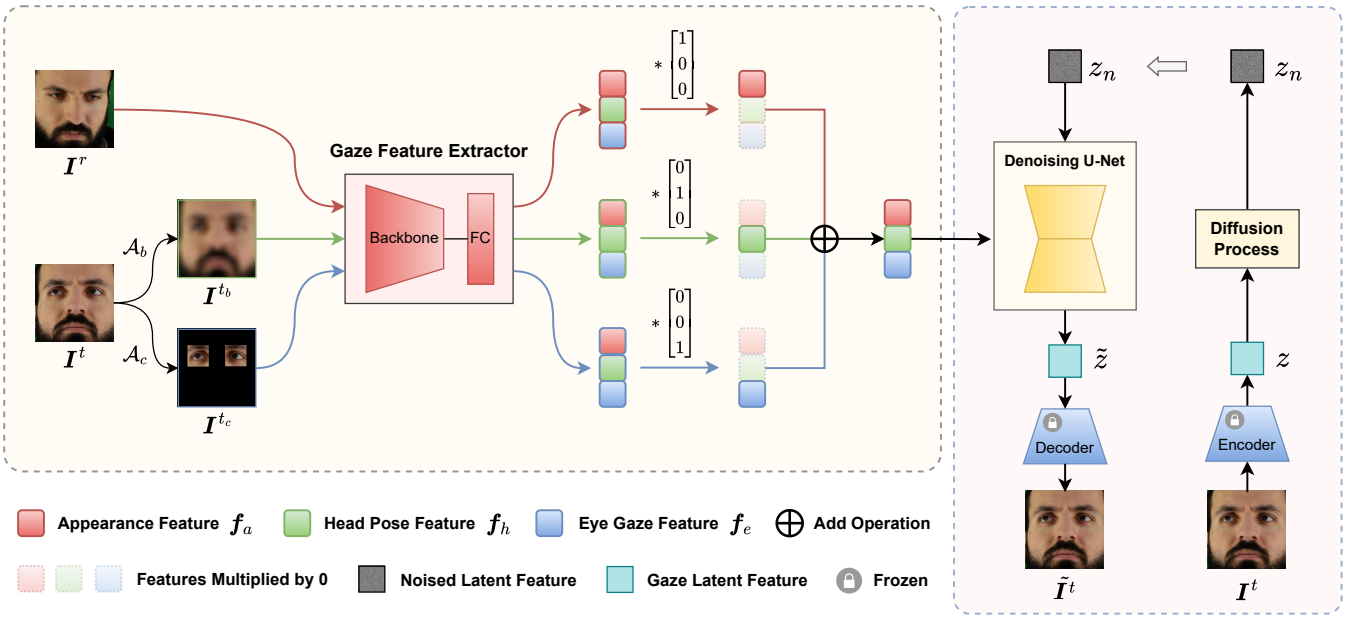

Figure 1: Overview of GazeDiff framework. It consists of two major modules, namely, a gaze feature extractor $\mathcal{G}$ and a condition diffusion model $\mathcal{C}$. $\mathcal{G}$ extracts representations from different images. $\mathcal{C}$ generates target gaze images based on constructed specific representations.

4 Method

4.1 Self-supervised Gaze Representation Learning Framework

GazeDiff is a self-supervised representation learning framework for gaze feature extractor training. As shown in Figure 1, it consists of two parts, a model agnostic gaze feature extractor $\mathcal{G}$ and a conditional diffusion model $\mathcal{C}$. Specifically, $\mathcal{G}$ includes a general network and a Fully Connected (FC) layer, the FC layer maps image features to a specified dimension. Such an architecture has been widely used in gaze estimation [Cheng and Bao, 2022; Wang *et al.*, 2022; Bao and Lu, 2023]. As a result, different gaze estimation models can be easily inserted into the GazeDiff framework. Additionally, in order to reduce computational complexity and accelerate training speed, we use the popular latent diffusion model [Rombach *et al.*, 2022] as $\mathcal{C}$.

Since gaze direction is mainly influenced by head pose and eye direction, we hope gaze feature extractor can pay attention to the head pose and eye areas information from the whole image. Thus, we extract multiple images features with $\mathcal{G}$, and explicitly disentangle the image representations into appearance representation, head pose representation and eye direction representation. After that, we manipulate the disentangling representations to construct a gaze-specific representation and send it to $\mathcal{C}$ as a condition to generate a target image with specific appearance, head pose and eye direction. The operational details will be explained later.

4.2 Conditional Diffusion Model

We generate gaze images to enhance the capability of gaze representations extracted by $\mathcal{G}$. In detail, we freeze the encoder and encoder, and fine-tune the denoising U-Net only as described in Figure 1. First, the encoder $\mathcal{E}$ encodes face image $\boldsymbol{I}^t$ into a latent representation $z = \mathcal{E}(\boldsymbol{I}^t)$. After the diffusion process, we got a noised latent representation z_t. Second, we condition the denoising process through cross-attention (similar to txt-to-image diffusion models [Rombach *et al.*, 2022]). To be more specific, we project the conditional recombined feature $\tilde{\boldsymbol{f}}^t$ to an intermediate representation $\tau_\theta(\tilde{\boldsymbol{f}}^t)$ with a gaze specific encoder τ_θ. Then, we map it to the intermediate layers of U-Net via $\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d}}) \cdot V$:

$$Q = W_Q^{(i)} \cdot \varphi_i\left(z_t\right), K = W_K^{(i)} \cdot \tau_\theta(\tilde{\boldsymbol{f}}^t), V = W_V^{(i)} \cdot \tau_\theta(\tilde{\boldsymbol{f}}^t) \tag{3}$$

Here, $\varphi_i\left(z_t\right)$ indicates a flattened intermediate representation of U-Net, $W_Q^{(i)}$, $W_K^{(i)}$ and $W_V^{(i)}$ are trainable projection matrices. Next, the U-Net gradually denoises z_n to restore the target latent representation $\tilde{z}$. Finally, the decoder $\mathcal{D}$ decodes $\tilde{z}$ to a realistic face image $\tilde{\boldsymbol{I}}^t = \mathcal{D}(\tilde{z})$. We fine-tune $\mathcal{C}$ by minimizing the loss function:

$$\mathcal{L}_\mathcal{C} = \mathbb{E}_{\mathrm{x}_t, \epsilon, t, \tilde{\boldsymbol{f}}^t}\left[\left\|\epsilon - \epsilon_\theta(\mathrm{x}_t, t, \tau_\theta(\tilde{\boldsymbol{f}}^t))\right\|^2\right] \tag{4}$$

4.3 Disentangling Feature Learning Strategy

In order to control the appearance, head pose and eye direction of generated image, we need to ensure that the condition of $\mathcal{C}$ contains these corresponding information. To achieve this goal, we input various data pairs and introduce a disentangling feature learning strategy.

A General Manner

In addition to inputting a single face image (we call it *base* input data), just like Figure 2 - (a). Easy-to-obtain unlabeled

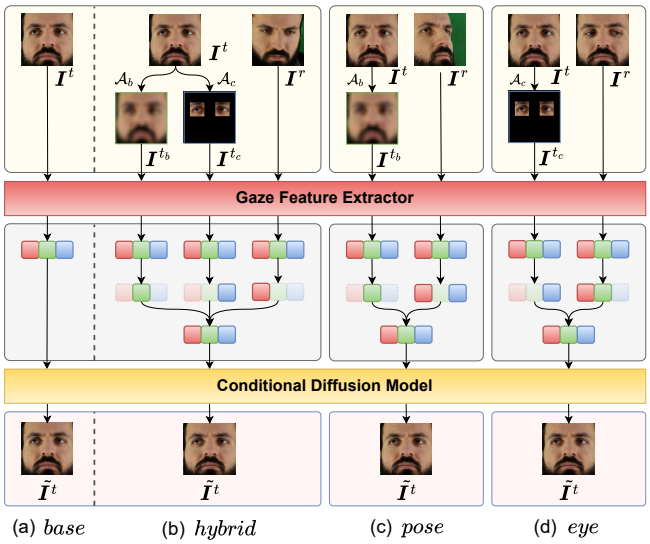

Figure 2: Input data pairs and disentangling feature learning strategy. (a) and (b) are the general input data that are applicable to the vast majority of facial datasets. (c) and (d) are enhanced data pairs for the facial datasets with specific collection settings.

facial datasets can usually construct such a *hybrid* data pair $(\boldsymbol{I}^r, \boldsymbol{I}^t)$, which consists of two images from the same identity with different head poses and eye directions, as shown in Figure 2 - (b). Wherein, $\boldsymbol{I}^r$ is a reference image, and $\boldsymbol{I}^t$ is the target image that is the generate target of conditional diffusion model $\mathcal{C}$. We respectively perform a Gaussian blur augmentation $\mathcal{A}_b$ and eye region cropping augmentation $\mathcal{A}_c$ on $\boldsymbol{I}^t$, then we obtained two augmented images $\boldsymbol{I}^{t_b}$ and $\boldsymbol{I}^{t_c}$. $\mathcal{A}_b$ removes appearance and eye direction information, while retaining the head contour information representing the head pose. And $\mathcal{A}_c$ only retains the eye region image including the information of eye direction, and losing appearance information and head pose information.

As shown in Figure 1, the *hybrid* data pair is inputted to GazeDiff. $\mathcal{G}$ extracts image features from $\boldsymbol{I}^r$, $\boldsymbol{I}^{t_b}$ and $\boldsymbol{I}^{t_c}$, and combines the appearance feature $\boldsymbol{f}_a^r$ from $\boldsymbol{I}^r$, the head pose feature $\boldsymbol{f}_h^{t_b}$ from $\boldsymbol{I}^{t_b}$ and the eye direction feature $\boldsymbol{f}_e^{t_c}$ from $\boldsymbol{I}^{t_c}$ into a new image features $\tilde{\boldsymbol{f}}^t$. Since $\boldsymbol{I}^r$ and $\boldsymbol{I}^t$ have the same identity, the generated image based on feature $\tilde{\boldsymbol{f}}^t$ should have the same appearance, head pose and eye direction as $\boldsymbol{I}^t$. Finally, for most facial datasets, We feed *base* and *hybrid* data pairs together into GazeDiff for self-supervised training.

Enhanced Manners

For some special datasets, we can construct enhanced data pairs to further promote the extraction and disentanglement of gaze features. Specifically, some facial datasets are collected synchronously by multiple cameras, such as ETH-XGaze [Zhang *et al.*, 2020], EVE [Park *et al.*, 2020] and ColumbiaGaze [Smith *et al.*, 2013]. At the same time, face images captured by different cameras have different head pose and the same eye direction. So we can also use the images from different cameras and a same timestamp to construct additional data pair to further enhance the ability to extract head pose information, as shown in Figure 2 - (c). We call this pair as *pose*, because *pose* is used to enhance the disentangling of head pose. Moreover, there are some datasets collected with fixed head pose, such as ETH-XGaze and ColumbiaGaze. Multiple images captured by the same camera have different eye directions and similar head poses. So we construct a new data pair as shown in Figure 2 - (d) and call this pair as *eye*.

4.4 Supervised Gaze Estimation

After the self-supervised gaze representation learning, the gaze feature learner $\mathcal{G}$ pre-trained by GazeDiff can be used for the task of gaze estimation. When fine-tuning on labeled gaze datasets, we add a Multi Layer Perceptrons (MLP) as gaze regressor $\mathcal{M}$ to predict 3D gaze direction. We train $\mathcal{M}$ by the loss function:

$$\mathcal{L}_{gaze}\left(\hat{\boldsymbol{g}}, \boldsymbol{g}\right) = \arccos\left(\frac{\hat{\boldsymbol{g}} \cdot \boldsymbol{g}}{\|\hat{\boldsymbol{g}}\|\|\boldsymbol{g}\|}\right) \tag{5}$$

where $\hat{\boldsymbol{g}}$ is model predicted gaze direction and $\boldsymbol{g}$ is the ground truth label.

5 Experiments

5.1 Dataset

GazeDiff is a general self-supervised representation learning framework. Therefore, we use the following datasets in our experiments as previous methods do: ETH-XGaze [Zhang *et al.*, 2020], EVE [Park *et al.*, 2020], Gaze360 [Kellnhofer *et al.*, 2019], GazeCapture [Krafka *et al.*, 2016], ColumbiaGaze [Smith *et al.*, 2013], MPIIFaceGaze [Zhang *et al.*, 2017b], EyeDiap [Funes Mora *et al.*, 2014] and VGG-Face2 [Cao *et al.*, 2018]. More details about above datasets can be found in the supplementary material.

Respectively, we denote them as $\mathcal{D}_{\mathrm{X}}$ (ETH-XGaze), $\mathcal{D}_{\mathrm{E}}$ (EVE), $\mathcal{D}_{\mathrm{G}}$ (Gaze360), $\mathcal{D}_{\mathrm{C}}$ (GazeCapture), $\mathcal{D}_{\mathrm{O}}$ (ColumbiaGaze), $\mathcal{D}_{\mathrm{M}}$ (MPIIFaceGaze), $\mathcal{D}_{\mathrm{D}}$ (EyeDiap) and $\mathcal{D}_{\mathrm{V}}$ (VGG-Face2).

5.2 Baseline Methods

We compare our approach with six following baselines: (i) **SimCLR** [Chen *et al.*, 2020], (ii) **MoCo v3** [Chen *et al.*, 2021], (iii) **DINO v2** [Oquab *et al.*, 2023], (iv) **MAE** [He *et al.*, 2022], (v) **DDAE** [Xiang *et al.*, 2023], (vi) **GazeDiff**$^-$, (vii) **GazeDiff** and (viii) **Supervised**. All methods except GazeDiff$^-$ load the released pre-trained weights on ImageNet, which means only the parameters of GazeDiff$^-$ are initialized randomly. Supervised method is trained on the annotated dataset and serves as the possible performance upper bound of the self-supervised gaze representation learning, the other methods are trained according to the self-supervised paradigm without labels. More details can be found in the supplementary materials.

5.3 Linear Probe Analysis

We perform linear probe analysis through within-dataset evaluation and cross-dataset evaluation, respectively.

Within-dataset Evaluation For the self-supervised methods, we pre-train in $\mathcal{D}_{\mathrm{X}}$, then freeze the pre-trained learner $\mathcal{G}$ and add a FC layer as gaze regressor $\mathcal{M}$ for linear-probe analysis

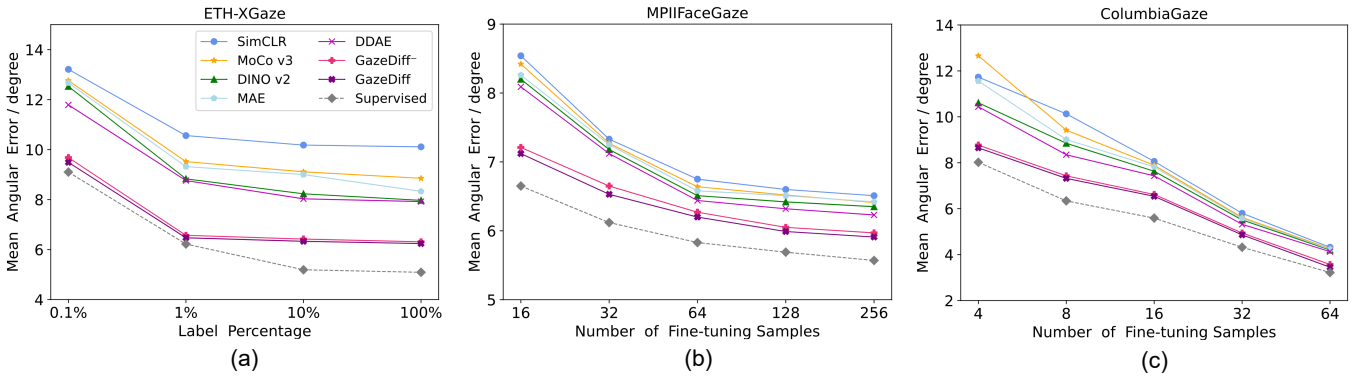

Figure 3: The experiment results of within-dataset and cross-dataset evaluation. All models are pre-trained in $\mathcal{D}_{\mathrm{X}}$.

following [Chen *et al.*, 2020]. We use different proportions of annotated training data to adapt $\mathcal{M}$ and evaluate on the separate validation data. As for the Supervised method, we finetune the whole gaze model on different proportions of annotated data. We report the experiments results in Figure 3 - (a). As we can see, regardless of whether the pre-trained weights are loaded, our proposed method GazeDiff outperforms other pre-training methods at different percentage of fine-tuning data. And when fine-tuning on a small subset of labeled data (such as 0.1% or 10% labels), GazeDiff achieves the similar performance to Supervised method, which suggesting that the learned representations are very effective for gaze estimation task.

Cross-dataset Evaluation We perform the cross-dataset evaluation following [Jindal and Manduchi, 2023]. The gaze feature learner $\mathcal{G}$ is pre-trained in $\mathcal{D}_{\mathrm{X}}$, then we finetune the gaze regressor $\mathcal{M}$ on $\mathcal{D}_{\mathrm{M}}$ and $\mathcal{D}_{\mathrm{O}}$ separately. In order to verify the generality of $\mathcal{G}$, we sample a few samples from each test subject for fine-tuning and evaluate on the remaining samples of the same subject. We conduct 10 times for each subject on both datasets and report the mean angular errors in Figure 3 - (b) and (c).

We can observe that our GazeDiff is still obviously superior to other self-supervised pre-training methods. The angular errors decrease as the number of fine-tuning samples increasing for all methods. Meanwhile, the performance gap between GazeDiff and Supervised reduces.

5.4 Comparsion with SOTA Gaze Representation Learning Methods

We compare GazeDiff with existing SOTA supervised [Cheng and Bao, 2022; Lee *et al.*, 2022; Wang *et al.*, 2022; Xu *et al.*, 2023; Bao and Lu, 2023; Yin *et al.*, 2024c; Yin *et al.*, 2025; Zeng *et al.*, 2025] and unsupervised [Yu and Odobez, 2020; Sun *et al.*, 2021; Du *et al.*, 2023; Jindal and Manduchi, 2023] gaze representation learning methods. Since [Bao and Lu, 2024] is limited to multi-view datasets and its results are not available in $\mathcal{D}_{\mathrm{M}}$ and $\mathcal{D}_{\mathrm{O}}$, we exclude their method in comparison.

Comparison with Supervised Methods

To compare the quality of learned representations with gaze labels, we fine-tune the entire gaze estimation network in a supervised manner based on unsupervised pre-training weights of GazeDiff. Besides, we add two baselines with random initialized and ImageNet pre-trained weights as a reference. Other SOTA methods trained on $\mathcal{D}_{\mathrm{x}}$, then evaluated on $\mathcal{D}_{\mathrm{M}}$ and $\mathcal{D}_{\mathrm{D}}$, the experimental results on are shown in Table 1. We can observe that the baseline with ImageNet pre-training weights shows a performance improvement compared to the baseline with random initialized weights. When we load the GazeDiff pre-trained weights, the cross-dataset errors further reduce significantly. Compared with other SOTA methods, GazeDiff achieves the best performance on $\mathcal{D}_{\mathrm{D}}$ and gains the second-best performance on $\mathcal{D}_{\mathrm{M}}$. In summary, GazeDiff derives the best average performance on two evaluation datasets. This indicates that GazeDiff provides a excellent initialization parameters for gaze estimation task.

Methods	Train Data	$\mathcal{D}_{\mathrm{M}}$	$\mathcal{D}_{\mathrm{D}}$	Avg.
Baseline	Random Init.	7.94	8.78	8.36
	ImageNet	7.40	8.22	7.81
PureGaze	$\mathcal{D}_{\mathrm{X}}$	7.08	7.48	7.28
LatentGaze		7.98	9.81	8.90
CDG†		6.73	7.95	7.34
Xu *et al.*		6.50	7.44	6.97
PCFGaze		7.40	7.30	7.35
CLIP-Gaze		**6.41**	7.51	6.96
LG-Gaze		<u>6.45</u>	<u>7.22</u>	<u>6.84</u>
GLA		6.83	7.38	7.11
GazeDiff		6.57	**6.95**	**6.76**

Table 1: Comparison of supervised gaze representation learning methods. The Avg. in the last column means the average error on both $\mathcal{D}_{\mathrm{M}}$ and $\mathcal{D}_{\mathrm{D}}$. Bold and underline denote the best and the second-best result among each column. † expresses the model employs ResNet-50 as backbone.

Comparison with Unsupervised Methods

For a fair comparison, we follow the experimental settings of [Yu and Odobez, 2020; Sun *et al.*, 2021]. Specifically, we use the leave-one-out (15-fold) and 5-fold cross-validation for $\mathcal{D}_{\mathrm{D}}$ and $\mathcal{D}_{\mathrm{O}}$. In each fold, we randomly select 50 samples with labels to fine-tune $\mathcal{M}$ and repeat each experiment for 10 times. Finally, we report the average error in Table 2.

Methods	Pre-train Data	$\mathcal{D}_{\mathrm{M}}$	$\mathcal{D}_{\mathrm{O}}$	Avg.
Yu *et al.*	$\mathcal{D}_{\mathrm{O}}$	-	8.9	8.9
Sun *et al.*	$\mathcal{D}_{\mathrm{M}}$	8.5	-	7.8
	$\mathcal{D}_{\mathrm{O}}$	-	7.0	
ConGaze*	$\mathcal{D}_{\mathrm{M}}$	<u>7.0</u>	-	<u>6.3</u>
	$\mathcal{D}_{\mathrm{O}}$	-	<u>5.5</u>	
GazeDiff	$\mathcal{D}_{\mathrm{M}}$	**6.1**	-	**5.7**
	$\mathcal{D}_{\mathrm{O}}$	-	**5.2**	
GazeCLR(Equiv)	$\mathcal{D}_{\mathrm{E}}$	7.0	<u>6.1</u>	6.6
GazeCLR(Inv+Equiv)		<u>6.5</u>	6.6	6.6
GazeDiff		**6.4**	**5.5**	**6.0**

Table 2: Comparison of unsupervised gaze representation learning methods. * means the method select 100 annotated samples for fine-tuning.

Note that Yu *et al.* [Yu and Odobez, 2020], Sun *et al.* [Sun *et al.*, 2021] and ConGaze [Du *et al.*, 2023] pre-trained on the evaluation dataset, [Jindal and Manduchi, 2023] uses $\mathcal{D}_{\mathrm{E}}$ as pre-training dataset, so we provide results with the same experimental setup for each method. The results show that regardless of which dataset is selected for pre-training, GazeDiff always achieves the best performance.

5.5 Ablation Study

Next, we conduct detailed ablation studies to explore the effectiveness of the proposed disentangling feature learning strategy and the stable performance improvement of GazeDiff in different setting.

Ablation Study on Disentangling Feature Learning Strategy

We proposed disentangling feature learning strategy to focus on extracting gaze-related features, and mentioned earlier that datasets with different collection settings can construct different input data pairs. For example, we can construct the four input pairs in Figure 2 on $\mathcal{D}_{\mathrm{X}}$. So we pre-train GazeDiff on $\mathcal{D}_{\mathrm{X}}$, and fine-tune gaze model in a supervised manner to analyze the contribution of our learning strategy. Hence, we report the cross-dataset performance of different pre-training input manners. Additionally, we also provide the results of Supervised baseline with ImageNet pre-training weights as a reference. The experimental results are reported in Table 3. Compared with the Supervised baseline, we can see that even with the powerful generation capability of diffusion model, GazeDiff with *base* input data (equivalent to autoencoder pre-training) only achieve a slight improvement of 0.11° on the average error. With regard to our learning strategy, after adding the *hybrid* data pair as input, the model gains an obvious improvement of 0.58° compared to the GazeDiff with *base* input data. Next, we add *pose* and *eye* pairs as input, the cross-dataset performances improve again. The absolute performance gains brought by *pose* and *eye* pairs are 0.25° and 0.11°. Overall, the *hybrid* data pair contributes the most to performance enhancement, and this data pair can be constructed from a vast majority of facial datasets. Above results prove that our GazeDiff is a general and effective framework.

Methods	Input Data *base*	*hybrid*	*pose*	*eye*	$\mathcal{D}_{\mathrm{M}}$	$\mathcal{D}_{\mathrm{D}}$	Avg.
Supervised					7.40	8.22	7.81
GazeDiff	✓				7.25	8.15	7.70
	✓	✓			7.10	7.13	7.12
	✓	✓	✓		<u>6.83</u>	**6.90**	<u>6.87</u>
	✓	✓	✓	✓	**6.57**	<u>6.95</u>	**6.76**

Table 3: Ablation study results on disentangling feature learning strategy. There are four input data pairs, *base*, *hybrid*, *pose* and *eye* correspond to the Figure 2 - (a) to (d) respectively.

Estimation Task	Input type	Mean error
head pose	blurred image ($\mathcal{A}_b$)	10.30
	eye cropped image ($\mathcal{A}_c$)	24.03
eye gaze	blurred image ($\mathcal{A}_b$)	12.20
	eye cropped image ($\mathcal{A}_c$)	7.53

Table 4: Ablation study results on disentangling augmentation.

Ablation Study on Disentangling Augmentation

We apply Gaussian blur augmentation and eye region cropping to reduce the impact of undesirable information in the face image. Nevertheless, information remains in blurred and cropped images, for example, there may be residual head posture information in the eye region cropped image. We train models on $\mathcal{D}_{\mathrm{X}}$ to estimate head pose and eye gaze respectively, then test on $\mathcal{D}_{\mathrm{M}}$. The results are reported in Table 4.$\mathcal{A}_b$ indeed discards most of eye direction information, and $\mathcal{A}_c$ removes most of head pose information. While it cannot be entirely eliminated, the network is forced to extract more accurate poses from blurred images and eye directions from eye images in order to generate high-precision images.

Ablation Study on Pre-training Datasets

In the previous experiments, in order to maintain consistent experimental settings with other methods, we only reported the results of pre-training on the gaze dataset. In this section, we pre-train on $\mathcal{D}_{\mathrm{V}}$, which is a large facial recognition dataset without gaze annotation, then fine-tune the model on the gaze dataset to demonstrate the performance in actual applications. The results are reported in Table 5. We can observe that, the models with pre-training weights perform consistent better than the models without pre-training weights on all training datasets. Then, the models pre-trained on $\mathcal{D}_{\mathrm{V}}$ have a better performance than the models pre-trained on $\mathcal{D}_{\mathrm{X}}$. This demonstrates that a more diversity and larger pre-training dataset can further improve the performance of GazeDiff. It is very practical and valuable to perform unsupervised gaze pre-training on a face dataset even without gaze labels.

For more ablation studies on training datasets and network architectures, please refer to the supplementary materials.

5.6 Visualization Analysis

Visualization of learned feature distributions

Following [Du *et al.*, 2023], We use t-SNE [Van der Maaten and Hinton, 2008] to visualize the distributions of gaze repre-

Pre-train Data	Train Data	$\mathcal{D}_{\mathrm{M}}$	$\mathcal{D}_{\mathrm{D}}$	Avg.
-	$\mathcal{D}_{\mathrm{G}}$	8.32	7.72	8.02
	$\mathcal{D}_{\mathrm{C}}$	6.35	6.61	6.48
$\mathcal{D}_{\mathrm{X}}$	$\mathcal{D}_{\mathrm{G}}$	7.30	7.59	7.45
	$\mathcal{D}_{\mathrm{C}}$	5.23	**5.97**	5.60
$\mathcal{D}_{\mathrm{V}}$	$\mathcal{D}_{\mathrm{G}}$	**7.06**	**7.10**	**7.08**
	$\mathcal{D}_{\mathrm{C}}$	**4.96**	6.12	**5.54**

Table 5: Ablation study results on pre-training datasets.

sentation on $\mathcal{D}_{\mathrm{M}}$ and $\mathcal{D}_{\mathrm{D}}$. Models are pre-trained on $\mathcal{D}_{\mathrm{X}}$. Different colors indicate different subjects, and each point corresponding to one image. We visualize the feature distributions of appearance, head pose and eye direction separately. The visualization results are shown in Figure 4. For the feature distributions of GazeDiff on evaluation datasets, it can be seen that features of the same identity clustered together and far away from the features of other people. As for head pose features and eye direction features, they are continuously distributed and invariant to subject identity. These feature distributions indicate the image features have been successfully disentangled into appearance, head pose and eye direction. Overall, these results show that the representations learned by GazeDiff are disentangled and can be easily transferred to unseen datasets.

Latent Interpolations

As shown in Figure 5, we explore $\mathcal{G}$'s feature space and discover directions that corresponding to semantic attributes of appearance, head pose and eye direction. By traversing its feature space, we can interpolate between images, morphing from one head pose to another head pose smoothly, and the same goes for appearance and eye direction contributes. For example, we disentangle the image features into appearance feature, head pose feature and eye direction feature. In order to control the head pose, we linearly interpolate from one head pose feature $\boldsymbol{f}_{h_1}$ to another $\boldsymbol{f}_{h_2}$, and keep the appearance feature and eye feature unchanged. The visualization demonstrates that, GazeDiff is trained fully unsupervised over facial images only, but successfully disentangles the image features into different parts of gaze-related representations, which proves that the effectiveness of our framework and proposed disentangling feature learning strategy. more visualization analysis can be found in the supplementary material.

6 Conclusion

We propose a diffusion-based framework (GazeDiff) for self-supervised gaze representation learning. GazeDiff exploits the generative power of diffusion models and enhances the quality, informativeness and interpretability of gaze representations. In order to guarantee the important head pose and eye information are learned from face images, we design a disentangling feature learning strategy. Specifically, we use Gaussian blur augmentation to keep head pose information and extract head pose feature. Then, we crop eyes region to extract eye direction feature. Finally, we take appearance feature from a reference image and combine these features as a condition to guide the generation. Our method achieves the SOTA performance in extensive evaluation settings, and the detailed ablation study and visualization analysis have demonstrate the generality and effectiveness of our work.

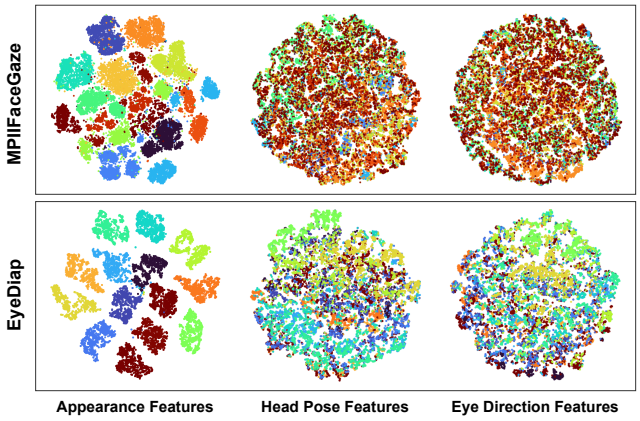

Figure 4: Visualization of the representations learned by GazeDiff. Different colors indicate different subjects, and one scatter corresponding to one image. The first row is the feature distribution of $\mathcal{D}_{\mathrm{M}}$, the second row corresponds to $\mathcal{D}_{\mathrm{D}}$.

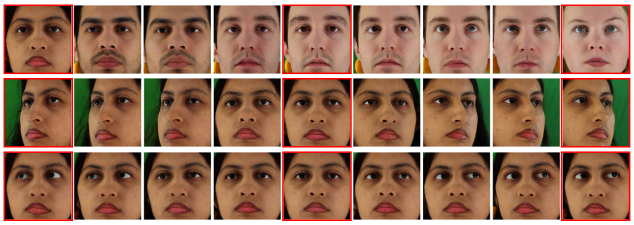

Figure 5: Latent controllability of appearance, head pose and eye direction. The red box represents the input images. The first row represents the results of appearance feature interpolations, the second row denotes the results of head pose feature interpolations and the last row indicates the results of eye direction feature interpolations.

Acknowledgements

This work was sponsored by National Natural Science Foundation of China (62476173), Shenzhen Fundamental Research Fundation(JCYJ20240813142610014) and National Key R&D Program of China (2023YFE0204200).

Contribution Statement

This work was a collaborative effort by all contributing authors. Guanzhong Zeng and Jingjing Wang made equal contributions to this study and are designated as co-first authors. Mingyang Zhou serving as the corresponding author, is responsible for all communications related to this manuscript.

References

[Bachman *et al.*, 2019] Philip Bachman, R Devon Hjelm, and William Buchwalter. Learning representations by

maximizing mutual information across views. *Advances in neural information processing systems*, 32, 2019.

[Bao and Lu, 2023] Yiwei Bao and Feng Lu. Pcfgaze: Physics-consistent feature for appearance-based gaze estimation. *arXiv preprint arXiv:2309.02165*, 2023.

[Bao and Lu, 2024] Yiwei Bao and Feng Lu. Unsupervised gaze representation learning from multi-view face images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1419–1428, 2024.

[Cao *et al.*, 2018] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 67–74. IEEE, 2018.

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

[Chen *et al.*, 2021] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9640–9649, 2021.

[Chen *et al.*, 2024] Xinlei Chen, Zhuang Liu, Saining Xie, and Kaiming He. Deconstructing denoising diffusion models for self-supervised learning. *arXiv preprint arXiv:2401.14404*, 2024.

[Cheng and Bao, 2022] Yihua Cheng and Yiwei Bao. Puregaze: Purifying gaze feature for generalizable gaze estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 436–443, 2022.

[Clark and Jaini, 2024] Kevin Clark and Priyank Jaini. Text-to-image diffusion models are zero shot classifiers. *Advances in Neural Information Processing Systems*, 36, 2024.

[Du *et al.*, 2023] Lingyu Du, Xucong Zhang, and Guohao Lan. Unsupervised gaze-aware contrastive learning with subject-specific condition. *arXiv preprint arXiv:2309.04506*, 2023.

[Funes Mora *et al.*, 2014] 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.

[He *et al.*, 2022] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.

[Ho *et al.*, 2020] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[Hudson *et al.*, 2024] Drew A Hudson, Daniel Zoran, Mateusz Malinowski, Andrew K Lampinen, Andrew Jaegle, James L McClelland, Loic Matthey, Felix Hill, and Alexander Lerchner. Soda: Bottleneck diffusion models for representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23115–23127, 2024.

[Jindal and Manduchi, 2023] Swati Jindal and Roberto Manduchi. Contrastive representation learning for gaze estimation. In *Annual Conference on Neural Information Processing Systems*, pages 37–49. PMLR, 2023.

[Kellnhofer *et al.*, 2019] Petr Kellnhofer, Adria Recasens, Simon Stent, Wojciech Matusik, and Antonio Torralba. Gaze360: Physically unconstrained gaze estimation in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6912–6921, 2019.

[Krafka *et al.*, 2016] 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.

[Lee *et al.*, 2022] Isack Lee, Jun-Seok Yun, Hee Hyeon Kim, Youngju Na, and Seok Bong Yoo. Latentgaze: Cross-domain gaze estimation through gaze-aware analytic latent code manipulation. In *Proceedings of the Asian Conference on Computer Vision*, pages 3379–3395, 2022.

[Mavely *et al.*, 2017] Annu George Mavely, JE Judith, PA Sahal, and Steffy Ann Kuruvilla. Eye gaze tracking based driver monitoring system. In *2017 IEEE international conference on circuits and systems (ICCS)*, pages 364–367. IEEE, 2017.

[Mirza and Osindero, 2014] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[Oquab *et al.*, 2023] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[Padmanaban *et al.*, 2017] Nitish Padmanaban, Robert Konrad, Emily A Cooper, and Gordon Wetzstein. Optimizing vr for all users through adaptive focus displays. In *ACM SIGGRAPH 2017 Talks*, pages 1–2. 2017.

[Park *et al.*, 2020] Seonwook Park, Emre Aksan, Xucong Zhang, and Otmar Hilliges. Towards end-to-end video-based eye-tracking. In *European Conference on Computer Vision (ECCV)*, 2020.

[Park *et al.*, 2021] Wooyeong Park, Jeongyun Heo, and Jiyoon Lee. Talking through the eyes: User experience design for eye gaze redirection in live video conferencing. In *Human-Computer Interaction. Interaction Techniques and Novel Applications: Thematic Area, HCI 2021, Held as Part of the 23rd HCI International Conference, HCII*

2021, Virtual Event, July 24–29, 2021, Proceedings, Part II 23, pages 75–88. Springer, 2021.

[Rombach *et al.*, 2022] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bj¸orn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[Ruzzi *et al.*, 2023] Alessandro Ruzzi, Xiangwei Shi, Xi Wang, Gengyan Li, Shalini De Mello, Hyung Jin Chang, Xucong Zhang, and Otmar Hilliges. Gazenerf: 3d-aware gaze redirection with neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9676–9685, 2023.

[Smith *et al.*, 2013] 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.

[Sohl-Dickstein *et al.*, 2015] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015.

[Sugano *et al.*, 2016] Yusuke Sugano, Xucong Zhang, and Andreas Bulling. Aggregaze: Collective estimation of audience attention on public displays. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*, pages 821–831, 2016.

[Sun *et al.*, 2021] Yunjia Sun, Jiabei Zeng, Shiguang Shan, and Xilin Chen. Cross-encoder for unsupervised gaze representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3702–3711, 2021.

[Van der Maaten and Hinton, 2008] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[Wang and Yin, 2025] Jingjing Wang and Pengwei Yin. Test time prompt tuning for domain adaptive gaze estimation. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025.

[Wang *et al.*, 2022] Yaoming Wang, Yangzhou Jiang, Jin Li, Bingbing Ni, Wenrui Dai, Chenglin Li, Hongkai Xiong, and Teng Li. Contrastive regression for domain adaptation on gaze estimation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19354–19363, 2022.

[Wei *et al.*, 2023] Chen Wei, Karttikeya Mangalam, Po-Yao Huang, Yanghao Li, Haoqi Fan, Hu Xu, Huiyu Wang, Cihang Xie, Alan Yuille, and Christoph Feichtenhofer. Diffusion models as masked autoencoders. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16284–16294, 2023.

[Xiang *et al.*, 2023] Weilai Xiang, Hongyu Yang, Di Huang, and Yunhong Wang. Denoising diffusion autoencoders are unified self-supervised learners. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15802–15812, 2023.

[Xu *et al.*, 2023] Mingjie Xu, Haofei Wang, and Feng Lu. Learning a generalized gaze estimator from gaze-consistent feature. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(3):3027–3035, Jun. 2023.

[Yin *et al.*, 2024a] Pengwei Yin, Jingjing Wang, Jiawu Dai, and Xiaojun Wu. Nerf-gaze: A head-eye redirection parametric model for gaze estimation. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2760–2764. IEEE, 2024.

[Yin *et al.*, 2024b] Pengwei Yin, Jingjing Wang, and Xiaojun Wu. Test-time adaptation with self-supervised learning for gaze estimation. *IEEE Transactions on Consumer Electronics*, 2024.

[Yin *et al.*, 2024c] Pengwei Yin, Guanzhong Zeng, Jingjing Wang, and Di Xie. Clip-gaze: Towards general gaze estimation via visual-linguistic model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 6729–6737, 2024.

[Yin *et al.*, 2025] Pengwei Yin, Jingjing Wang, Guanzhong Zeng, Di Xie, and Jiang Zhu. Lg-gaze: Learning geometry-aware continuous prompts for language-guided gaze estimation. In *European Conference on Computer Vision*, pages 1–17. Springer, 2025.

[Yu and Odobez, 2020] 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.

[Zeng *et al.*, 2025] Guanzhong Zeng, Jingjing Wang, Zefu Xu, Pengwei Yin, Wenqi Ren, Di Xie, and Jiang Zhu. Gaze label alignment: Alleviating domain shift for gaze estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 9780–9788, 2025.

[Zhang *et al.*, 2017a] Xucong Zhang, Yusuke Sugano, and Andreas Bulling. Everyday eye contact detection using unsupervised gaze target discovery. In *Proceedings of the 30th annual ACM symposium on user interface software and technology*, pages 193–203, 2017.

[Zhang *et al.*, 2017b] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. It's written all over your face: Full-face appearance-based gaze estimation. *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2299–2308, 2017.

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