Engineering Applications of Artificial Intelligence 160 (2025) 111901

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

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

Research paper

GHR-2D: Gaze and head redirection via disentanglement and diffusion for gaze estimation

Daosong Hu, Mingyue Cui *, Kai Huang

School of Computer Science and Engineering, Sun Yat-sen University, Guangdong, China

ARTICLE INFO

Keywords:
Gaze estimation
Gaze redirection
Eye tracking

ABSTRACT

The performance of appearance-based gaze estimation models significantly decreases during cross-domain evaluation due to domain gaps. Gaze redirection edits the gaze direction and head pose of existing faces to expand the dataset and bridge the inter-domain gap. However, the generated results are contaminated by gaze-irrelevant information. In addition, physiological structures can cause gaze shifts. We introduce a face image redirection method based on a diffusion model that interprets head pose and gaze redirection in the latent space. An encoder maps two-dimensional vectors to the latent space, aligning them with image features while preserving angle information. Relative gaze compensates for deviations in the visual and optical axes caused by physiological structures. We also propose a training pipeline for disentangling identity, gaze, and head pose embeddings. An encoder mines the relationship between gaze and head, mapping input pitch and yaw into conditions to guide the generation of gaze and head pose embeddings in a diffusion model. The decoder projects the connected identities and modified embeddings back into the feature space. Finally, the redirected samples expand the original dataset and facilitate cross-domain evaluation. Compared to state-of-the-art methods, our approach significantly improves gaze accuracy.

1. Introduction

Appearance-based gaze estimation is a task of great significance due to its potential applications in various scenarios, where facial images captured by cameras serve as indicators of individuals' attention. Applications include safe driving (Sharma and Chakraborty, 2024), disease diagnosis (Islam et al., 2025), and saliency detection (Sun et al., 2021). Compared to feature-based methods (Cheng et al., 2017), end-to-end models powered by deep learning have demonstrated remarkable estimation capabilities (Wu et al., 2024). From early convolutional neural network (CNN) models (Cheng et al., 2020) to various transformer-based frameworks (Wu et al., 2023), the ultimate goal is to achieve more robust and accurate predictions across diverse real-world scenarios. However, the challenges of comprehensively covering all possible scenarios during dataset collection have limited the practical application of these models in real-world contexts.

Recent research has provided quantitative findings in cross-dataset evaluations to substantiate the effectiveness of the proposed method (Cheng et al., 2021). The experimental findings reveal that prediction accuracy significantly declines when confronted with variations in head pose, facial features, and gaze direction. Notably, as the dataset size increases, the model's stability improves (Kellnhofer et al., 2019). Nevertheless, due to the inherent challenges of annotating gaze and head pose in the real world, data collection is often confined to controlled laboratory environments. Moreover, acquiring gaze labels is laborious and resource-intensive, requiring fixed cameras for subsequent calculations, which makes it difficult for the dataset to cover all conceivable conditions.

Numerous studies are exploring methods to tackle this challenge. Since gaze datasets are difficult to collect, the method of obtaining new gaze images to fill the label gap by modifying the gaze direction of existing face images is named gaze redirection. Existing method synthesizes eye images with different gaze directions to modify existing facial images and enhance the dataset. However, it does not address modifications in head pose (Fu et al., 2020). Another technique uses facial images to generate 3D graphics for head pose manipulation, which has gained popularity in recent research (Qin et al., 2023). However, it fails to account for changes in gaze direction. Facial redirection, powered by Generative Adversarial Networks (GAN), can alter head pose (Hsu and Chung, 2020; Jin et al., 2023), however it does not take into account physiological structural differences. Hence, there is a pressing need for a model that can freely and accurately adjust both head pose and gaze direction.

The diffusion model has gained popularity due to its excellent generation and extrapolation capabilities, as shown by recent studies (Ding

* Corresponding author.
E-mail addresses: huds@mail2.sysu.edu.cn (D. Hu), cuimy@mail2.sysu.edu.cn (M. Cui), huangk36@mail.sysu.edu.cn (K. Huang).

https://doi.org/10.1016/j.engappai.2025.111901
Received 10 March 2025; Received in revised form 26 June 2025; Accepted 27 July 2025
Available online 7 August 2025

et al., 2023). We propose leveraging the Diffusion model to facilitate versatile adjustments to facial images. However, due to disparities in gaze direction, head pose, and facial features, addressing alignment issues within the latent space is crucial. In addition, the differences between personal visual axis and optical axis need to be evaluated. By expanding the source domain data with generated data, our goal is to adapt to the target domain. We aim to redirect gaze and head posture in a way that is intuitive to humans, ultimately improving the model's ability to extend its applicability to uncharted domains.

We propose a method to redirect head posture and gaze in a single image using pitch and yaw. To improve generation quality, we adopted a new encoder structure and training pipeline. The angle encoder effectively aligns the input conditions with the image features in the latent space. To modify head posture and gaze, either simultaneously or separately, our training pipeline uses multiple loss functions to disentangle identity, gaze, and head posture in the feature space. Relative gaze addresses the inherent differences in physiological structures. To explore the connections between features, we designed a feature fusion network based on self-attention. The diffusion model is guided by encoded conditional embeddings to modify the gaze and head posture embeddings. The edited embeddings are connected, deprojected into the feature space, and the redirected face image is obtained through the decoder. In summary, our contributions are as follows:

- We introduce a novel framework that harnesses the latent diffusion model to modify facial image attributes, generating high-quality training data with extensive coverage and significant variations in angles.
- We introduce a innovative encoder structure that maps two-dimensional vectors to the latent space, aligning them with facial image encoding features while retaining their angular information. We define a feature fusion network that achieves projection and deprojection of embedding space and feature space by mining the connections between different features.
- For precise and effective image modifications, we employ a new framework that utilizes multiple loss functions to disentangle identity, gaze, and head posture features in the latent space.
- We propose to use relative gaze to address gaze errors caused by physiological construction and achieve gaze-invariant head rotation. Eliminating estimation errors caused by head posture by promoting face-forward gaze.

2. Related work

2.1. Cross-domain gaze estimation

Cross-domain evaluation remains a major challenge in appearance-based gaze estimation due to disparities between real-world conditions and controlled laboratory data (Lee et al., 2022; Kim et al., 2024). Previous efforts have addressed this challenge from different perspectives. One approach is unsupervised learning (Yu and Odobez, 2020), which can be more complex than annotating facial images for gaze. Liu et al. (2021) introduced a plug-and-play adaptive framework that operates without ground-truth labels, using adaptive training guided by outlier detection. Although these methods have shown promise in improving cross-domain prediction, they do not address the challenge of imbalanced datasets.

Some researchers have shifted focus from model enhancements to improving dataset quality (He et al., 2019; Zheng et al., 2020; Fu et al., 2020). Wood et al. (2015, 2016) pioneered the use of computer graphics to synthesize eye gaze samples. Other researchers, such as Yu et al. (2019), proposed a gaze redirection method that alters gaze direction by manipulating pupil position and opening. Wang et al. (2018) introduced a hierarchical eye data synthesis approach that generates eye shapes based on gaze directions and creates realistic eye images using Generative Adversarial Networks (GAN). He et al. (2019) proposed a GAN-based method to generate eye images with new gaze directions. ST-ED (Zheng et al., 2020) noticed that the encoded facial features contain gaze-irrelated features, so directly editing the features could lead to redirection errors. ST-ED further used GAN to extend feature editing to gaze-related features by disentangling the latent vector. However, the generated faces have low resolution and limited range. Jin et al. (2023) proposed high-resolution face redirection based on e4e-StyleGAN. However, differences in physiological structure were not fully considered, affecting the accuracy of the generated results. Qin et al. (2023) introduced a technique for mapping single or multi-view faces into 3D space to achieve 3D face reconstruction, where gaze data for various head poses can be acquired by rotating the 3D model. However, this method indirectly influences gaze through alterations in head pose. In summary, these methods mainly modify gaze or head pose to generate synthetic data. A more versatile model would be better suited for comprehensive editing of facial images.

2.2. Diffusion-based image synthesis

The early Diffusion model, initially developed for unconditional image generation (Song et al., 2020), primarily produced qualitative results. As demand for more controlled image generation grew, researchers began incorporating external information, such as text (Avrahami et al., 2022) or images (Benny and Wolf, 2022), to guide the synthesis process. Kim et al. (2022) introduced a method that uses a pre-trained CLIP model to encode text and fine-tunes the reverse diffusion process through CLIP loss. While text offers useful guidance, images convey richer information. Yang et al. (2023) proposed an example-based image editing technique with the Diffusion model, blending the encoded features of a reference image into the reverse diffusion process to guide image generation. Kawar et al. (2023) introduced an attribute binding method, optimizing text embeddings to generate images similar to the input and then fine-tuning to improve the reconstruction of target domain images. They also interpolate between optimized text embeddings and target embeddings to guide image editing. These methods control the diffusion process mainly through semantic information, which remains somewhat ambiguous. The gaze redirection task, however, requires more precise control, such as specific pitch and yaw angles.

3. Preliminaries

Diffusion model is widely applied in image generation tasks and has achieved satisfactory results (Zhu et al., 2023). In some downstream tasks, it is considered a better method, such as image denoising (Özdenizci and Legenstein, 2023; Kulikov et al., 2023; Chung et al., 2022), controllable image generation (Bar-Tal et al., 2023), and multimodal tasks (Ham et al., 2023).

The core of the diffusion model is to map any noise image $\epsilon_t \sim \mathcal{N}(0, I)$ to the realistic target image sample $\mathbf{X}_0$ after T successive denoising. Each intermediate sample $\mathbf{X}_t$ can be derived as follows:

$$\mathbf{X}_t = \sqrt{\alpha_t}\mathbf{X}_0 + \sqrt{1-\alpha_t}\epsilon_t \quad t \in \{0, \ldots, T\} \tag{1}$$

where, α_t is the hyperparameters of the diffusion model, satisfying $1 = \alpha_0 > \alpha_1 > \alpha_2 > \cdots > \alpha_{T-1} > \alpha_T = 0$. ϵ_t is the sampled random noise, and $\epsilon_t \sim \mathcal{N}(0, I)$. Song et al. (2020) proposed Denoising Diffusion Implicit Model (DDIM) that enjoys the following deterministic generative process:

$$\mathbf{X}_{t-1} = \sqrt{\alpha_{t-1}}\left(\frac{\mathbf{X}_t - \sqrt{1-\alpha_t}\epsilon_t\left(\mathbf{X}_t, t\right)}{\sqrt{\alpha_t}}\right) + \sqrt{1-\alpha_{t-1}}\epsilon_t\left(\mathbf{X}_t, t\right) \tag{2}$$

and the following inference process:

$$q\left(\mathbf{X}_{t-1} \mid \mathbf{X}_t, \mathbf{X}_0\right) = \mathcal{N}\left(\sqrt{\alpha_{t-1}}\mathbf{X}_0 + \sqrt{1-\alpha_{t-1}}\frac{\mathbf{X}_t - \sqrt{\alpha_t}\mathbf{X}_0}{\sqrt{1-\alpha_t}}, \mathbf{0}\right) \tag{3}$$

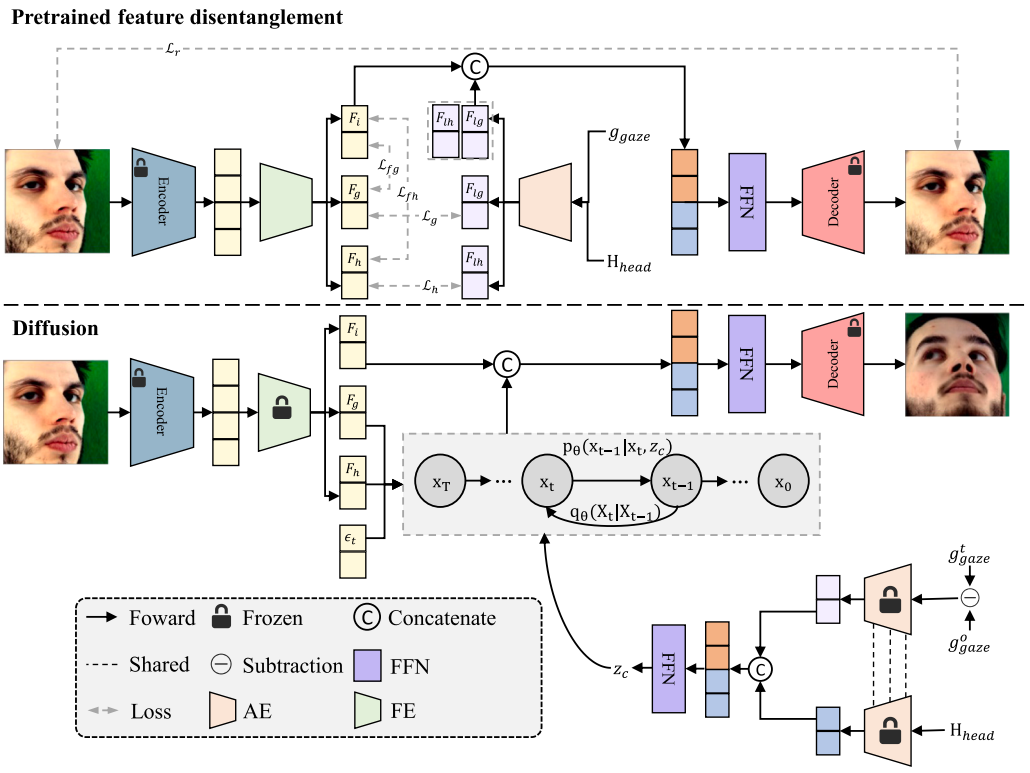

Fig. 1. Overview of the proposed method.

Ho et al. (2020) proposed UNet is used to learn a function $\epsilon_t(\mathbf{X}_t, t)$, that is, predicting the noise of a noisy image $\mathbf{X}_t$ to obtain X_{t-1}. The goal of the network is to make $\epsilon_t(\mathbf{X}_t, t) \approx \epsilon_t$, so the loss function is defined as $\|\epsilon_t(\mathbf{X}_t, t) - \epsilon_t\|$, and ϵ_t is the noise sampled from $\mathcal{N}(0, I)$, which is added to X_0 to get $\mathbf{X}_t$.

This method can be used to learn conditional distributions, that is, by inputting condition c into the network, $\epsilon_t(\mathbf{X}_t, t, c)$ can sample in the distribution of condition c, thereby achieving controllable generation of diffusion models. This condition can be text (Kawar et al., 2022), image (Yang et al., 2022), or others (Preechakul et al., 2022).

Early models typically operated directly in pixel space, resulting in optimization taking too long. Rombach et al. (2022) introduce the encoder–decoder into this task. By using pre-trained autoencoder, diffusion was innovated from pixel space to latent space, greatly reducing computational complexity. Simultaneously utilizing cross attention layers to flexibly control the generation process. The network is modified to $\epsilon_t(\mathcal{A}(\mathbf{X}_t), t, c)$, and $\mathcal{A}$ represents the encoder.

4. Proposed method

For gaze estimation tasks, our goal is to rotate the head and adjust the gaze direction within the camera coordinate system. However, the ambiguity between angles and image space complicates feature fusion. As shown in Fig. 1, we propose a gaze and head redirection model based on facial feature disentanglement. We extract three embeddings from the facial latent space: identity (F_i), gaze (F_g), and head pose (F_h). The head and gaze embeddings are input into the diffusion model, while the angle encoder (AE) maps the label embedding to angle space, serving as a control condition for editing the gaze and head embeddings. After connecting the identity and redirection features, the feature fusion network (FFN) projects them back into the initial latent space, and the redirected facial image is generated through the decoder. Our proposed framework incorporates pretrained feature disentanglement and diffusion processes.

4.1. Pretrained feature disentanglement

Our pretrained feature disentanglement module consists of two components: label alignment in the embedding space and identity information extraction through facial reconstruction. First, the input image is encoded into the feature space using CLIP to obtain its semantic representation. For redirection tasks, we need to modify the gaze and head pose information while preserving the identity. To achieve this, we define a feature extractor (FE) that maps the semantic information from the feature space to the embedding space, yielding gaze, head pose, and identity embeddings. Note that the structure of the FE is identical to that of the AE, with only the dimensions of the linear layer modified. This process can be formulated as follows:

$$F_i, F_g, F_h = \mathrm{FE}(\mathrm{CLIP}(I)) \tag{4}$$

In order to supervise the de entanglement of the embedding space, the basic facts are mapped by the angle encoder into the label embedding space, aligned with the gaze and head pose embeddings. Following the principle of non-intersection between gaze related and irrelevant features in Cheng et al. (2022), there is no intersection between identity features and gaze and head features. The gaze direction is influenced by the head posture, so there is a connection between the two. We use cosine distance to measure the similarity between features, thus defining four types of losses for supervised feature disentanglement. Specifically, as follows:

$$\mathcal{L}_f = \mathcal{L}_{fg} + \mathcal{L}_{fh} = 2 + \frac{F_i \cdot F_g}{|F_i| \cdot |F_g|} + \frac{F_i \cdot F_h}{|F_i| \cdot |F_h|} \tag{5}$$

$$\mathcal{L}_l = \mathcal{L}_g + \mathcal{L}_h = 2 - \frac{F_g \cdot F_{lg}}{|F_g| \cdot |F_{lg}|} - \frac{F_h \cdot F_{lh}}{|F_h| \cdot |F_{lh}|} \tag{6}$$

where $\mathcal{L}_{fg}$ represents the distance between F_i and F_g, while $\mathcal{L}_{fh}$ denotes the head pose. $\mathcal{L}_f$ is the distance between different attributes in the face embedding space, and as $\mathcal{L}_f$ decreases, F_i is gradually purified. $\mathcal{L}_l$ is the distance between the embedding of facial attributes and the label, and $\mathcal{L}_g$ is gaze, $\mathcal{L}_h$ is head posture. F_{lg} and F_{lh} are

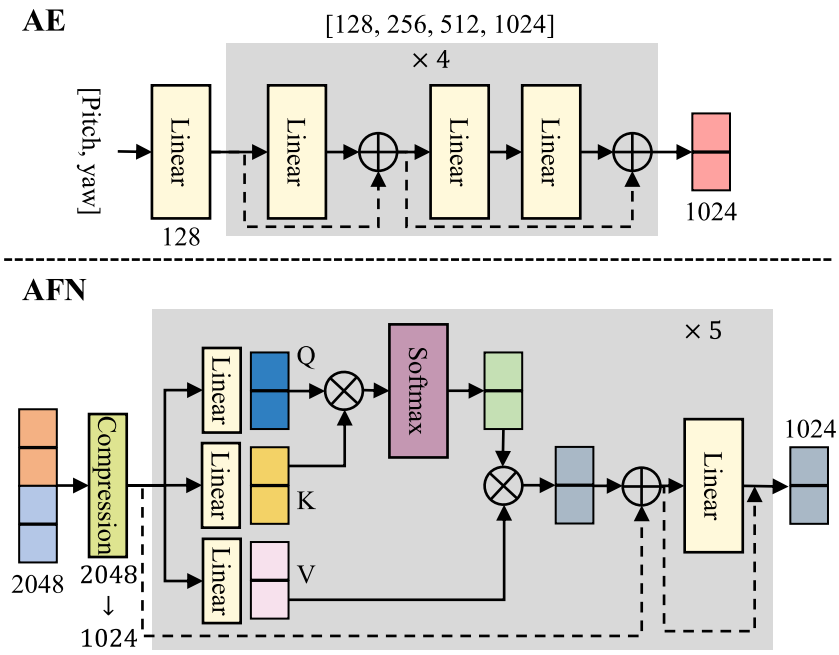

Fig. 2. The structures of AE and FFN.

label embeddings obtained by mapping the ground-truth of gaze (g_{gaze}) and head posture (H_{head}), respectively, such as $F_{lg} = \mathrm{AE}(g_{gaze})$ and $F_{lh} = \mathrm{AE}(\mathrm{H}_{head})$.

Assuming that the ground truth can be mapped to the embedding space, which, along with non changing factors such as appearance, contains all semantic information of the face. FFN is used to project embeddings back into the feature space. Finally, we decode the fused features to reconstruct the input image. We guide the reconstruction of images using a pixel-wise L2 reconstruction loss ($\mathcal{L}_r$) between the generated image (I_r) and the original image (I):

$$\mathcal{L}_r = \frac{1}{|I|}\|I_r - I\|_2 \tag{7}$$

The total loss function of the pre-training module is defined as $\mathcal{L}$ as follows:

$$\mathcal{L} = \mathcal{L}_r + \lambda\mathcal{L}_f + \lambda\mathcal{L}_l \tag{8}$$

where λ is a hyperparameter and is set to 0.5 to balance the weights of reconstruction and cosine losses.

4.1.1. AE and FFN

Pitch and yaw are descriptions in three-dimensional space used to represent angle information. MLP is a commonly used method that aligns low dimensional angle vectors with image features. However, although the number of mapping layers increases, the boundaries of pitch and yaw are blurred, resulting in overlapping features during information fusion, thereby affecting the accuracy of redirection. As shown in Fig. 2, we propose a skip connected MLP framework from coarse to fine. A fully connected layer is used to quickly locate low dimensional vectors into the feature space. Subsequently, residual connections are used to gradually refine the embedding of pitch and yaw in the feature space. The specific process is as follows:

$$F_c = L(A) \tag{9}$$

$$\mathrm{AE}(\cdot) = \sum_{j=1}^{n}((F_c + L(F_c)) + \sum_{i=1}^{2} L(F_c + L(F_c))) \tag{10}$$

where, A is the angle vector, represented by $[Pitch, Yaw]$. L is the linear layer, F_c denotes the extracted coarse embedding, and $\mathrm{AE}(\cdot)$ represents the angle encoder.

Gaze is affected by the movement of the pupils, and the head posture also changing the direction of gaze. Therefore, an angle fusion network based on self-attention is used to explore the correlation of concatenated embeddings, in order to obtain precise control conditions. Firstly, linear layers are used to compress the concatenated embeddings. Then, the Transformer Feed-Forward Network is used to mine the coupling relationship between gaze and head pose. Specifically, three parallel linear layers are used to obtain query (**Q**), key (**K**), and value (**V**) with overlapping attributes. By calculating the correlation between **Q** and **K** and applying the Softmax along the horizontal dimension, we can obtain an attention score $\mathbf{A}_{g,h\to j}$ that reflects the correlation between **Q** and **K**. Note that, $\mathbf{A}_{g,h\to j}$ also indicates the correlation between gaze and head posture. Following Scaled Dot-Product (Vaswani et al., 2017), and the scale coefficients $\sqrt{D}$ is applied in the calculation process. D is the dimension of the feature. Then, the $\mathbf{A}_{g,h\to j}$ is used to aggregate compressed feature representations in **V**, thereby obtaining joint feature representations. The residual connections are used to obtain the final embedding representation. The process is formulated as:

$$E_{con} = C(Concat(E_1, E_2, \ldots, E_n)) \tag{11}$$

$$\mathbf{A}_{g,h\to j} = \mathrm{Softmax}\left(\mathbf{Q}^T(\mathbf{K})/\sqrt{D}\right) \tag{12}$$

$$\mathrm{FFN}(\cdot) = E_{con} + \mathbf{A}_{g,h\to j}\mathbf{V} + L(E_{con} + \mathbf{A}_{g,h\to j}\mathbf{V}) \tag{13}$$

where $E_1, E_2, \ldots, E_n$ are different embeddings, E_{con} represents embedding compressed by linear layers (C).

4.2. Diffusion

We propose a face redirection framework based on diffusion models. The diffusion model is used for editing gaze and head pose embeddings, and guided by target label embeddings. Note that, due to physiological structure, there is a deviation between the visual and the optical axis, therefore gaze is defined as relative gaze (Δ_{gaze}). $z = (I, \mathbf{X}_T, \mathrm{H}_{head}, \Delta_{gaze})$ is a input, which consists of the input face image I, a completely Gaussian noise $\mathbf{X}_T$, target head pose H_{head}, and the relative gaze Δ_{gaze}.

Gaze direction is accurately labeled. Due to the deviation δ between the visual axis g_v and optical axis g_o of the human eyes, and the δ is not fixed. Hence, the real gaze direction g_{gaze} can be obtained by

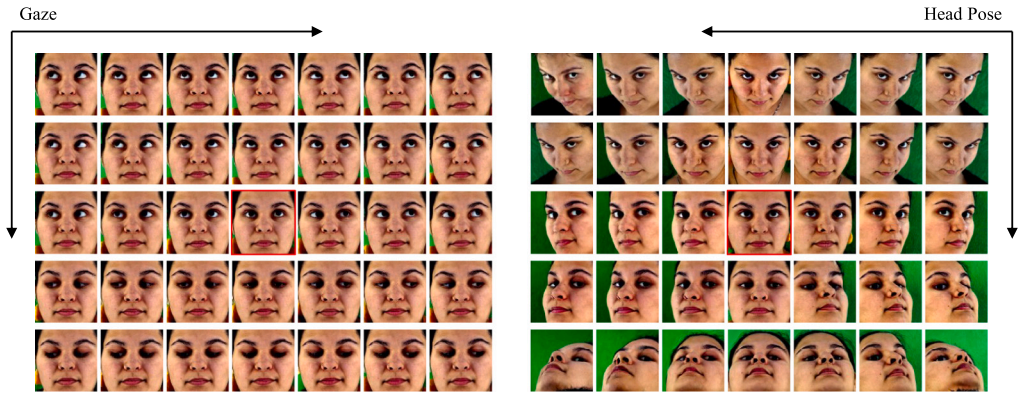

Fig. 3. The visual effects formed by individually modifying gaze or head pose. The red box represents the input image, and the sampling interval is 0.2 radians.

$g_{gaze} = g_o = g_v + \delta$. Therefore, this paper converts gaze conditions into relative gaze. that is,

$$\Delta_{gaze} = g^t_{gaze} - g^o_{gaze} \tag{14}$$

where g^t_{gaze} is the target gaze, and g^o_{gaze} denotes input image. By using the difference Δ_{gaze}, the impact of δ can be minimized.

The input conditions $(\mathrm{AE}(\mathrm{H}_{head}), \mathrm{AE}(\Delta_{gaze}))$ are integrated into a condition vector z_c. z_c contains all necessary information for adjusting the reverse diffusion process. that is,

$$z_c = \mathrm{FFN}(\mathrm{AE}(\mathrm{H}_{head}), \mathrm{AE}(\Delta_{gaze})) \tag{15}$$

Our diffusion model receives as input $z_d = (\mathbf{X}_0, \mathbf{X}_T, z_c)$ to produce the edited embedding. $\mathbf{X}_0$ is the embeddings of face image I. The DDIM is used to get $p\left(\mathbf{X}_{t-1} \mid \mathbf{X}_t, \mathbf{X}_0, z_c\right)$ by $q\left(\mathbf{X}_{t-1} \mid \mathbf{X}_t, \mathbf{X}_0\right)$, which defined in Eq. (3), with the following reverse process:

$$p\left(\mathbf{X}_{t-1} \mid \mathbf{X}_t, \mathbf{X}_0, z_c\right) = \begin{cases} \mathcal{N}\left(\mathbf{f}_\theta\left(\mathbf{X}_0, \mathbf{X}_1, 1, z_c\right), \mathbf{0}\right) & \text{if } t = 1 \\ q\left(\mathbf{X}_{t-1} \mid \mathbf{X}_t f\left(\mathbf{X}_0 \mathbf{X}_t t, z_c\right)\right) & \text{otherwise} \end{cases} \tag{16}$$

Unet is used to predict noise $\epsilon_t(\mathbf{X}_0, \mathbf{X}_t, t, z_c)$, and f can be restored using the predicted noise as follow:

$$f\left(\mathbf{X}_0, \mathbf{X}_t, t, z_c\right) = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \sqrt{1-\alpha_t}\epsilon_t\left(\mathbf{X}_0, \mathbf{X}_t, t, z_c\right)\right) \tag{17}$$

Specifically, the diffusion model generates by gradually reversing the Markov forward process. For the generative process, given the condition z_c, the diffusion model gradually denoises the noisy image from the last step by minimizing the following loss function:

$$\mathcal{L}_{simple} = \mathbb{E}_{\mathbf{X}_0, \epsilon_t}\left\|\epsilon_t(\mathbf{X}_0, \mathbf{X}_t, t, z_c) - \epsilon_t\right\|_2^2 \tag{18}$$

where ϵ_t is the random noise added during the forward process, that is, $\epsilon_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

F_g and F_h are input into the diffusion model to obtain redirected gaze $\hat{F}_g$ and head pose $\hat{F}_h$ embeddings. Similar to face reconstruction in pretrained feature disentanglement, F_i, $\hat{F}_g$ and $\hat{F}_h$ are connected and fed into the decoder to obtain the redirected image.

5. Experiments

5.1. Datasets

MPIIFaceGaze The MPIIFaceGaze (D_M) dataset is derived from the MPIIGaze dataset (Zhang et al., 2017). The method by Zhang et al. (2018) is employed for normalizing face and eye images.

EyeDiap EyeDiap (D_D) (Funes Mora et al., 2014) is categorized into continuous screen targets (CS) or 3D floating object targets (FT). The method by Cheng et al. (2021) is employed for preprocessing.

ETH-XGaze ETH-XGaze (D_E) (Zhang et al., 2020) comprises over one million images of 110 subjects with varying head poses. The method by Cheng et al. (2021) is employed for preprocessing.

GazeCapture GazeCapture (D_G) (Krafka et al., 2016) comprises approximately 2.4 million images featuring 1474 subjects with diverse head poses. The method by Cheng et al. (2021) is employed for preprocessing. The dataset is utilized for fine-tuning.

5.2. Implementation details

We implement proposed method by using PyTorch, initialize the face encoder and diffusion backbone of our model with weights pretrained using latent diffusion (Rombach et al., 2022), and the remaining layers with default values. a batch-size of 16 frames is set to train, and the learning rate is set as 0.0001. We train on $4 \times V100$ GPUs. Note that all our generate results in the main paper are for 256 image size. For our final approach, we use $T = 200$ timesteps.

To manipulate real-world images, we utilize a powerful generation model, Stable Diffusion (Rombach et al., 2022), as an initialization. The partial dataset D_E is selected for our pretraining phase. For fine-tuning, a subset of 1000 subjects from D_G is used, and the ground truth is not visible. The input condition z_c is the pseudo label generated by the VGG16-based gaze estimator, which is pretrained on D_E.

5.3. Generate results

ETH-XGaze covers a broader range of gaze and head postures and is used for pretraining. A subset of this dataset is employed to validate the pretraining results, with its ground truth hidden during training. To examine the impact of head pose ($\mathrm{H}_{head} = 0$) while varying the difference in gaze ($\Delta_{gaze} \in [-0.7, 0.7]$), a series of facial images is sampled. As shown on the left side of Fig. 3, the head pose remains constant while the pupil position changes, indicating a shift in gaze direction. Conversely, when Δ_{gaze} is set to zero, the head pose is varied to generate additional facial images, as shown on the right side of Fig. 3. The results demonstrate that GHR-2D is capable of achieving changes in both head pose and gaze direction. This approach allows the network to genuinely capture the intricate relationship between pupil positions and gaze directions.

5.4. Large head posture

D_G contains a larger number of subsets and is used for fine-tuning. However, due to limitations in the collection equipment, D_G has a restricted head posture range. As a result, redirecting the head pose in the samples of this dataset presents a challenge. As illustrated in Fig. 4, the design of feature disentanglement and relative gaze facilitates the model's ability to retain identity information while more freely redirecting head posture. Despite the larger range of head posture inputs, GHR-2D still delivers convincing results. When $\mathrm{H}_{head} = [0.01, -0.9]$, the generated image exhibits some distortion due to the limited number of relevant samples in the pretraining dataset. Modifying contrast and brightness also remains a limitation. Nevertheless, the quality of the redirected image is still sufficient to help expand the dataset.

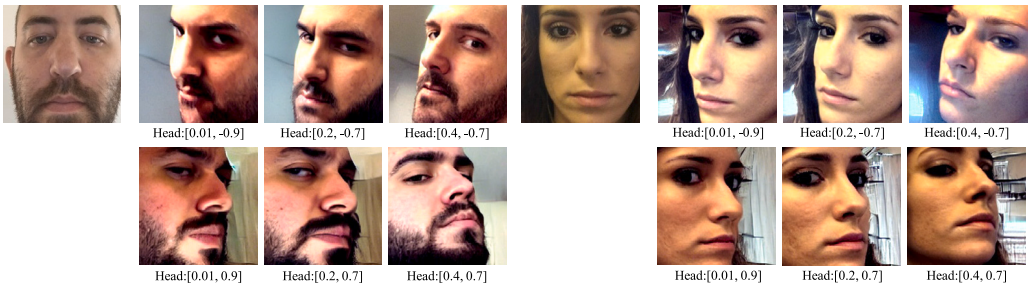

Fig. 4. The generation result of a large head posture.

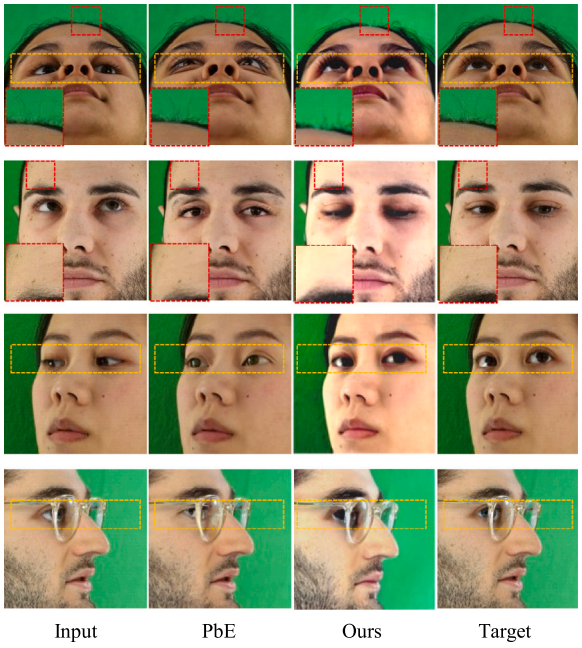

Fig. 5. Comparison of gaze redirection. The red dashed box denotes the blurring of gaze-irrelevant features, and the yellow dashed box is the redirection result of the pupil.

5.4.1. Comparative analysis

Most existing methods in gaze estimation rely on supervised learning, where the network learns to estimate gaze by minimizing the loss between predicted values and labels. In these approaches, head pose is typically estimated implicitly through separate network branches. Consequently, the accuracy of gaze labels is particularly critical. Compared to head pose, gaze is more prone to errors in generative models. We compared our method with the latest generative models, and the results are shown in Fig. 5. Our method's redirection results closely align with the target image, with the only limitation being the loss of some detail (highlighted in the red dashed box). This is because we edit the entire face, and due to incomplete disentanglement of features, some small gaze-irrelevant details may be blurred. In contrast, PbE (Yang et al., 2023) only masks the pupil area, preserving local details, but it struggles to accurately redirect the gaze of the reference image, often producing unrealistic results (yellow dashed box). For redirection tasks, the accuracy of gaze direction is the most critical indicator. Hence, our method is validated as effective for gaze redirection.

5.5. Comparison with adaptive methods

To enhance the model's cross-domain performance, domain adaptation is a key strategy. As shown in Table 3, we compare our method with existing domain adaptation approaches. PureGaze uses adversarial training to extract gaze-related features, while AGG replaces the MLP layer to analytically project high-dimensional features back into 3D space. The results indicate that domain adaptation strategies can improve cross-domain performance but often depend on unique training processes, network structures, or extra labels. Our method focuses on addressing data collection challenges. In the future, combining these two strategies could further promote the development of robust gaze models.

5.6. Comparison of methods

We conducted a quantitative performance comparison of various redirectors both within the dataset and across different datasets. Our experimental setup and quantitative standards align with those presented in Jin et al. (2023). Notably, as this paper applies the Diffusion model to gaze estimation, we also introduced the latest diffusion-based image editing model, which utilizes the target face image for redirection (Yang et al., 2022). We employed three evaluation metrics: Head (Gaze) Redir, Head (Gaze) Induce, and LPIPS (Jin et al., 2023) as indicators.

Head (Gaze) Redir: This metric measures the redirection accuracy in degrees between the redirected image and the target image when given fixed head orientations (gaze directions).

Head (Gaze) Induce: It quantifies the errors in degrees on gaze (head) when the head (gaze) is redirected.

LPIPS: This metric is used to assess the similarity and distortion of the face image in different gaze directions.

FID: This indicator is used to evaluate the similarity between the generated results and the real samples.

For our evaluation, we employed a pretrained ResNet-50-based head pose and gaze estimator. It is worth noting that the model is invisible during the redirector training process.

5.6.1. Performance of the redirectors

GHR-2D is fine-tuned and tested on D_G, with the subjects used for evaluation being unseen during training. When comparing the latest gaze and head pose redirection methods based on different frameworks for both within-dataset and cross-dataset performance, the results are summarized in Table 1 and Table 2, respectively.

For within-dataset performance, StarGAN produces unsatisfactory results. ReDirTrans (Jin et al., 2023), which also leverages a GAN framework, achieved state-of-the-art performance. However, due to the consideration of physiological structural differences, GHR-2D outperformed ReDirTrans in both gaze and head pose redirection, improving accuracy by 20.89% and 15.27%, respectively. Since head posture and gaze mutually influence each other, cross-validation between the two is essential. Our method consistently showed improvements over ReDirTrans, achieving a 10.15% reduction in LPIPS, which indicates that the redirected facial images align better with the source domain distribution. The PbE model (Yang et al., 2023), based on diffusion, demonstrates strong inference capabilities for more prominent features like head posture and facial attributes. However, its predictions for smaller features, such as pupil position, are less accurate.

Table 1
Within-dataset quantitative comparison (GazeCapture test subset) between different methods, all metrics are better when lower in value.

	Gaze Redir	Head ReDir	Gaze Induce	Head Induce	LPIPS	FID
StarGAN (Choi et al., 2018)	4.602	3.989	0.755	3.067	0.257	–
He et al. (2019)	4.617	1.392	0.560	3.925	0.223	–
VecGAN (Dalva et al., 2022)	2.282	0.824	0.401	2.205	0.197	–
ST-ED (Zheng et al., 2020)	2.385	0.800	0.384	2.187	0.208	4.854
ReDirTrans (Jin et al., 2023)	2.163	0.753	0.429	2.155	0.197	–
PbE (Yang et al., 2023)	2.554	0.948	0.493	2.301	0.213	**3.921**
GHR-2D	**1.711**	**0.638**	**0.343**	**1.817**	**0.177**	4.215

Table 2
Cross-dataset quantitative comparison (MPIIFaceGaze) between different methods, all metrics are better when lower in value.

	Gaze Redir	Head ReDir	Gaze Induce	Head Induce	LPIPS	FID
StarGAN (Choi et al., 2018)	4.488	3.013	0.786	2.783	0.260	–
He et al. (2019)	5.092	1.372	0.684	3.411	0.241	–
VecGAN (Dalva et al., 2022)	2.670	1.242	0.391	1.941	0.207	–
ST-ED (Zheng et al., 2020)	2.380	1.085	0.371	1.782	0.212	5.125
ReDirTrans (Jin et al., 2023)	2.380	0.985	0.391	1.782	0.202	–
PbE (Yang et al., 2023)	2.820	1.125	0.481	1.937	0.219	4.418
GHR-2D	**1.881**	**0.740**	**0.297**	**1.775**	**0.193**	**4.125**

Table 3
Comparison of cross-domain performance with adaptive methods.

Method	Year	$\mathcal{D}_E \rightarrow \mathcal{D}_M$	$\mathcal{D}_E \rightarrow \mathcal{D}_D$
PureGaze (Cheng et al., 2022)	AAAI 2022	7.08°	7.48°
AGG (Bao and Lu, 2024)	CVPR 2024	5.91°	6.75°
Ours		6.59°	8.05°

For cross-dataset performance, PbE exhibits fewer LPIPS fluctuations, suggesting that the generated faces are realistic. However, it still have larger redirection errors. In contrast, GHR-2D, by effectively disentangling gaze and head pose in the latent space, significantly outperforms PbE.

5.7. Proportion of extended data

We randomly selected facial images from the training set and sampled the conditions $[\mathrm{H}_{head}, \Delta_{gaze}]$ from a uniform distribution, i.e., $[\mathrm{H}_{head}, \Delta_{gaze}] \sim \mathcal{U}(-0.2\pi, 0.2\pi)$. A fixed proportion, $M\%$, of the redirected images is selected to expand the training set. The values for M were chosen as 0, 20, 40, 60, 80. When $M = 0$, this corresponds to using the original training set. The experimental results are summarized in Table 4. As the proportion of redirected data increased, errors across various mainstream frameworks gradually decreased. However, beyond a 60% expansion, the rate of error reduction slowed down. This is due to an increase in distorted samples. Considering the trade-off between accuracy improvement and resource consumption, it is recommended to set the data expansion ratio to 50%.

5.8. Gaze-invariant head rotation

Head posture can induce eye deformation, which affects feature extraction and results in poor gaze prediction accuracy. To achieve gaze-invariant head rotation, Δ_{gaze} is set to 0, as shown in Fig. 6. Gaze invariance does not mean fixing the position of the pupils; as the head rotates, the pupil position changes accordingly. In this experiment, ResNet50 is trained on the training set and used as the evaluation model. 1000 samples with larger head postures are selected in the test set (sorted by absolute pitch and yaw angle). The conditions $\Delta_{gaze} = 0$ and $\mathrm{H}_{head} = 0$ are applied to obtain redirected facial images. As shown in Table 5, the average error caused by head posture decreased by 46.54%. This result demonstrates that the redirector can effectively reduce the impact of head posture without the need for additional coordinate conversion. The main limitation, however, lies in the speed of model generation.

Fig. 6. The redirected images with parameters $\Delta_{gaze} = 0$ and $\mathrm{H}_{head} = 0$.

5.9. Ablation study

To evaluate the impact of the proposed structure on cross-domain performance, we conducted an ablation study with a data expansion ratio of 50%. The experimental results are summarized in Table 6, which includes the following variations: w.o. Ex (using raw data), w.o. Pfd (without pretrained feature disentanglement), w.o. AE (AE replaced by MLP), w.o. FFN (FFN replaced by MLP), w.o. RG (without relative gaze), w.o. F_i (without identity feature), and w. GHR-2D (using GHR-2D to extend the training set). From the experimental results, we can draw four key conclusions: (1) Pretraining on large-scale datasets better disentangles different attributes. (2) For features with a strong correlation between gaze and head posture, FFN effectively explores this correlation and produces well-fused features. (3) Relative gaze can help mitigate inherent differences caused by physiological structures. Specifically, removing relative gaze causes a performance loss of up to 6.23% on D_M and 9.89% on D_G. (4) identity information is beneficial for the decoupling of gaze-related and irrelevant features.

In order to analyze the impact of the output dimension of AE on the generation results, we set different output dimensions in a stepwise manner, and the results are shown in Table 7. Based on the evaluation of the generation results, we found that as the output dimension increases, the redirection results are more accurate, especially the

Table 4
Comparison of the number of redirected images. (unit: °).

Test set	Model	0%	20%	40%	60%	80%
$\mathcal{D}_M$	ResNet18 (He et al., 2016)	8.694	7.901	7.139	**6.615**	6.774
	ResNet50 (He et al., 2016)	8.109	7.887	6.913	6.822	**6.635**
	GazeTR (Cheng and Lu, 2022)	8.593	8.210	7.376	7.123	**6.985**
$\mathcal{D}_D$	ResNet18 (He et al., 2016)	23.153	13.956	10.349	**8.329**	8.391
	ResNet50 (He et al., 2016)	19.820	12.841	10.304	**8.158**	8.389
	GazeTR (Cheng and Lu, 2022)	17.128	11.031	9.461	**8.102**	8.152

Table 5
Quantitative analysis results. (unit: °).

Dataset	Original data	Redirected data	▽
$\mathcal{D}_M$	6.158	**3.104**	49.59%
$\mathcal{D}_D$	10.221	**6.155**	39.78%
$\mathcal{D}_E$	5.534	**2.387**	56.86%
$\mathcal{D}_G$	4.157	**2.496**	39.95%

Table 6
Cross-domain comparative ablation study. (unit: °).

Test	Model	w.o. Ex	w.o. Pfd	w.o. AE	w.o. RG	w.o. FFN	w.o. F_i	w. GHR-2D
$\mathcal{D}_M$	ResNet18	8.694	7.342	7.113	7.257	6.946	7.385	**6.831**
	ResNet50	8.109	7.291	6.813	6.741	6.795	6.811	**6.701**
	GazeTR	8.031	7.039	6.890	6.939	**6.418**	7.092	6.591
$\mathcal{D}_D$	ResNet18	23.153	12.731	8.739	9.133	8.516	9.352	**8.311**
	ResNet50	19.820	11.315	8.683	8.915	8.433	9.399	**8.193**
	GazeTR	17.128	10.973	8.229	8.775	8.573	9.083	**8.053**

Table 7
Ablation study of the output dimension of AE. (unit: °).

Dimension	Gaze Redir	Head Redir	Gaze Induce	Head Induce	LPIPS	FID
512	1.935	0.773	0.305	1.799	0.197	4.302
1024	1.881	0.740	0.297	**1.775**	0.193	4.125
2048	**1.812**	**0.738**	**0.291**	1.788	**0.190**	**4.015**

gaze direction. However, compared with 1024, the redirection results of 2048 are slightly improved. We speculate that this is due to the bottleneck caused by the limited gaze angles of the pre-training dataset. In addition, the increase in the output dimension of AE will bring additional computational overhead. Therefore, in this paper, the output dimension of AE is set to 1024.

5.10. Limitation

Gaze redirection is employed to modify the gaze direction of existing facial images, including pupil position and head posture. However, due to the challenges in obtaining large-scale and diverse gaze datasets, existing methods often rely on pseudo labels for fine-tuning, which can introduce annotation errors in the redirected results. The uneven distribution of training samples makes it prone to distortion when generating results with larger head poses. Under the default settings of this paper, the training time of the ETH dataset is about 12 h per epoch. The inference time is about 5.6 s per image. Therefore, the stable diffusion structure in the proposed method leads to relatively high computational cost. GHR-2D is more suitable for offline generation of extended datasets for subsequent gaze model training.

6. Conclusion

We propose a novel diffusion-based method for precise redirection of both gaze and head pose in the latent space. An encoder is introduced to map pitch and yaw into the latent space, aligning them with the encoded features of the face image, thereby generating embeddings that retain angle information. Relative gaze addresses inherent biases introduced by individual physiological structures. The pretraining pipeline maps semantic features to identity, gaze, and head pose embeddings. AE and FE are used for mutual projection across angle, feature, and embedding spaces. FFN ensures smooth fusion of correlated features, while the diffusion model only edits the target attributes, minimizing compression loss on identity. Through extensive experiments, we demonstrate that using synthetic data to extend the dataset significantly improves gaze estimation accuracy and reduces cross-domain evaluation errors. In future work, we aim to further explore face and gaze-related attributes, with the goal of synthesizing data tailored to specific scenarios.

CRediT authorship contribution statement

Daosong Hu: Writing – original draft, Validation, Software, Methodology, Investigation, Conceptualization. **Mingyue Cui:** Writing – review & editing, Validation, Software, Data curation. **Kai Huang:** Writing – review & editing, Supervision, Resources, Funding acquisition.

Declaration of competing interest

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

Acknowledgment

This work was supported by Shenzhen Medical Research Funds under Grant SZMRF-D2404009.

Data availability

Data will be made available on request.

References

Avrahami, O., Lischinski, D., Fried, O., 2022. Blended diffusion for text-driven editing of natural images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 18208–18218.

Bao, Y., Lu, F., 2024. From feature to gaze: A generalizable replacement of linear layer for gaze estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1409–1418.

Bar-Tal, O., Yariv, L., Lipman, Y., Dekel, T., 2023. Multidiffusion: Fusing diffusion paths for controlled image generation.

Benny, Y., Wolf, L., 2022. Dynamic dual-output diffusion models. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 11482–11491.

Cheng, Y., Bao, Y., Lu, F., 2022. Puregaze: Purifying gaze feature for generalizable gaze estimation. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, pp. 436–443.

Cheng, Y., Huang, S., Wang, F., Qian, C., Lu, F., 2020. A coarse-to-fine adaptive network for appearance-based gaze estimation. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 10623–10630.

Cheng, H., Liu, Y., Fu, W., Ji, Y., Yang, L., Zhao, Y., Yang, J., 2017. Gazing point dependent eye gaze estimation. Pattern Recognit. 71, 36–44.

Cheng, Y., Lu, F., 2022. Gaze estimation using transformer. In: Proceedings of the International Conference on Pattern Recognition. IEEE, pp. 3341–3347.

Cheng, Y., Wang, H., Bao, Y., Lu, F., 2021. Appearance-based gaze estimation with deep learning: A review and benchmark. arXiv preprint arXiv:2104.12668.

Choi, Y., Choi, M., Kim, M., Ha, J.-W., Kim, S., Choo, J., 2018. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8789–8797.

Chung, H., Lee, E.S., Ye, J.C., 2022. MR image denoising and super-resolution using regularized reverse diffusion. IEEE Trans. Med. Imaging 42 (4), 922–934.
Dalva, Y., Altındiş, S.F., Dundar, A., 2022. Vecgan: Image-to-image translation with interpretable latent directions. In: European Conference on Computer Vision. Springer, pp. 153–169.
Ding, Z., Zhang, X., Xia, Z., Jebe, L., Tu, Z., Zhang, X., 2023. DiffusionRig: Learning personalized priors for facial appearance editing. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 12736–12746.
Fu, X., Yan, Y., Yan, Y., Peng, J., Wang, H., 2020. Purifying real images with an attention-guided style transfer network for gaze estimation. Eng. Appl. Artif. Intell. 91, 103609.
Funes Mora, K.A., Monay, F., Odobez, J.-M., 2014. 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. pp. 255–258.
Ham, C., Hays, J., Lu, J., Singh, K.K., Zhang, Z., Hinz, T., 2023. Modulating pretrained diffusion models for multimodal image synthesis. arXiv preprint arXiv:2302.12764.
He, Z., Spurr, A., Zhang, X., Hilliges, O., 2019. Photo-realistic monocular gaze redirection using generative adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 6932–6941.
He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778.
Ho, J., Jain, A., Abbeel, P., 2020. Denoising diffusion probabilistic models, vol. 33, pp. 6840–6851.
Hsu, W.-Y., Chung, C.-J., 2020. A novel eye center localization method for head poses with large rotations. IEEE Trans. Image Process. 30, 1369–1381.
Islam, M.F., Manab, M.A., Mondal, J.J., Zabeen, S., Rahman, F.B., Hasan, M.Z., Sadeque, F., Noor, J., 2025. Involution fused convolution for classifying eye-tracking patterns of children with Autism Spectrum Disorder. Eng. Appl. Artif. Intell. 139, 109475.
Jin, S., Wang, Z., Wang, L., Bi, N., Nguyen, T., 2023. ReDirTrans: Latent-to-latent translation for gaze and head redirection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5547–5556.
Kawar, B., Zada, S., Lang, O., Tov, O., Chang, H., Dekel, T., Mosseri, I., Irani, M., 2022. Imagic: Text-based real image editing with diffusion models. arXiv preprint arXiv:2210.09276.
Kawar, B., Zada, S., Lang, O., Tov, O., Chang, H., Dekel, T., Mosseri, I., Irani, M., 2023. Imagic: Text-based real image editing with diffusion models. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6007–6017.
Kellnhofer, P., Recasens, A., Stent, S., Matusik, W., Torralba, A., 2019. Gaze360: Physically unconstrained gaze estimation in the wild. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 6912–6921.
Kim, G., Kwon, T., Ye, J.C., 2022. Diffusionclip: Text-guided diffusion models for robust image manipulation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2426–2435.
Kim, S., Nam, W.-J., Lee, S.-W., 2024. Appearance debiased gaze estimation via stochastic subject-wise adversarial learning. Pattern Recognit. 152, 110441.
Krafka, K., Khosla, A., Kellnhofer, P., Kannan, H., Bhandarkar, S., Matusik, W., Torralba, A., 2016. Eye tracking for everyone. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2176–2184.
Kulikov, V., Yadin, S., Kleiner, M., Michaeli, T., 2023. Sinddm: A single image denoising diffusion model. In: Proceedings of the International Conference on Machine Learning. PMLR, pp. 17920–17930.
Lee, I., Yun, J.-S., Kim, H.H., Na, Y., Yoo, S.B., 2022. LatentGaze: Cross-domain gaze estimation through gaze-aware analytic latent code manipulation. In: Proceedings of the Asian Conference on Computer Vision. pp. 3379–3395.
Liu, Y., Liu, R., Wang, H., Lu, F., 2021. Generalizing gaze estimation with outlier-guided collaborative adaptation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3835–3844.
Özdenizci, O., Legenstein, R., 2023. Restoring vision in adverse weather conditions with patch-based denoising diffusion models. IEEE Trans. Pattern Anal. Mach. Intell..
Preechakul, K., Chatthee, N., Wizadwongsa, S., Suwajanakorn, S., 2022. Diffusion autoencoders: Toward a meaningful and decodable representation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 10619–10629.
Qin, J., Shimoyama, T., Zhang, X., Sugano, Y., 2023. Domain-adaptive full-face gaze estimation via novel-view-synthesis and feature disentanglement. arXiv preprint arXiv:2305.16140.
Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B., 2022. High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 10684–10695.
Sharma, P.K., Chakraborty, P., 2024. A review of driver gaze estimation and application in gaze behavior understanding. Eng. Appl. Artif. Intell. 133, 108117.
Song, J., Meng, C., Ermon, S., 2020. Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502.
Sun, L., Chen, Z., Wu, Q.J., Zhao, H., He, W., Yan, X., 2021. AMPNet: Average-and max-pool networks for salient object detection. IEEE Trans. Circuits Syst. Video Technol. 31 (11), 4321–4333.
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. Adv. Neural Inf. Process. Syst. 30.
Wang, K., Zhao, R., Ji, Q., 2018. A hierarchical generative model for eye image synthesis and eye gaze estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 440–448.
Wood, E., Baltrušaitis, T., Morency, L.-P., Robinson, P., Bulling, A., 2016. Learning an appearance-based gaze estimator from one million synthesised images. In: Proceedings of the Symposium on Eye Tracking Research and Applications. pp. 131–138.
Wood, E., Baltrusaitis, T., Zhang, X., Sugano, Y., Robinson, P., Bulling, A., 2015. Rendering of eyes for eye-shape registration and gaze estimation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3756–3764.
Wu, C., Hu, H., Lin, K., Wang, Q., Liu, T., Chen, G., 2023. Attention-guided and fine-grained feature extraction from face images for gaze estimation. Eng. Appl. Artif. Intell. 126, 106994.
Wu, X., Li, L., Zhu, H., Zhou, G., Li, L., Su, F., He, S., Wang, Y., Long, X., 2024. EG-Net: Appearance-based eye gaze estimation using an efficient gaze network with attention mechanism. Expert Syst. Appl. 238, 122363.
Yang, B., Gu, S., Zhang, B., Zhang, T., Chen, X., Sun, X., Chen, D., Wen, F., 2022. Paint by example: Exemplar-based image editing with diffusion models. arXiv preprint arXiv:2211.13227.
Yang, B., Gu, S., Zhang, B., Zhang, T., Chen, X., Sun, X., Chen, D., Wen, F., 2023. Paint by example: Exemplar-based image editing with diffusion models. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 18381–18391.
Yu, Y., Liu, G., Odobez, J.-M., 2019. Improving few-shot user-specific gaze adaptation via gaze redirection synthesis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 11937–11946.
Yu, Y., Odobez, J.-M., 2020. Unsupervised representation learning for gaze estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7314–7324.
Zhang, X., Park, S., Beeler, T., Bradley, D., Tang, S., Hilliges, O., 2020. Eth-xgaze: A large scale dataset for gaze estimation under extreme head pose and gaze variation. In: Proceedings of the European Conference on Computer Vision. Springer, pp. 365–381.
Zhang, X., Sugano, Y., Bulling, A., 2018. Revisiting data normalization for appearance-based gaze estimation. In: Proceedings of the Symposium on Eye Tracking Research and Applications. pp. 1–9.
Zhang, X., Sugano, Y., Fritz, M., Bulling, A., 2017. Mpiigaze: Real-world dataset and deep appearance-based gaze estimation. IEEE Trans. Pattern Anal. Mach. Intell. 41 (1), 162–175.
Zheng, Y., Park, S., Zhang, X., De Mello, S., Hilliges, O., 2020. Self-learning transformations for improving gaze and head redirection. Proc. Adv. Neural Inf. Process. Syst. 33, 13127–13138.
Zhu, Y., Li, Z., Wang, T., He, M., Yao, C., 2023. Conditional text image generation with diffusion models. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 14235–14245.