Institute for Visualization and Interactive Systems

University of Stuttgart
Universitätsstraße 38
D–70569 Stuttgart

Masterarbeit Nr. 3578461

A 3D-Aware Conditional Diffusion Model for Gaze Redirection

YeonJoo Cho

Course of Study:	Computer Science
Examiner:	Prof. Dr. Andreas Bulling
Supervisor:	Chuhan Jiao, M.sc.
Commenced:	December 4, 2023
Completed:	June 4, 2024

Abstract

Gaze redirection refers to the task of modifying the direction of eye gaze and its corresponding facial counterparts to a targeted direction, while preserving the original identity of the subject. An effective gaze redirection approach must (i) be aware of the 3D nature of the task, (ii) accurately redirect the gaze into any specified direction, and (iii) generate photorealistic output images that preserve the shape and texture details from the input images. In response to these requirements, this thesis presents a novel approach to gaze redirection using a 3D-aware conditional diffusion model that leverages the intrinsic geometric properties of human faces. This approach effectively transforms the task into a conditional image-to-image translation. To embed 3D awareness comprehensively, we adopt a viewpoint-conditioned diffusion model, that can learn the 3D context of the facial geometry. Then, the conditions to this model are unique gaze rotations and latent facial parameters from the face images. These strategies are further reinforced by a novel loss function focused on gaze direction and head orientation, which enhances the model's ability to learn and apply accurate gaze and head adjustments effectively. Together, these elements underscore the potential of our approach to produce high-quality, accurate gaze redirection, fulfilling the complex demands of this sophisticated visual task.

Contents

List of Figures

List of Tables

List of Algorithms

1 Introduction

1.1 Motivation

As individuals, humans possess a myriad of distinctive attributes that define our human nature. This includes vocal characteristics, unique fingerprints, and the capacity for logical thinking. Along with these distinguishing traits, human gaze emerges as a significant non-verbal cue that serves as a rich source of information. This subtle yet powerful aspect of human behavior is capable of conveying a spectrum of emotional states, levels of visual attention, underlying intentions, and even physical well-being. As a result, it has found applications across diverse domains.

- Human-Computer Interaction

  In the field of human-computer interaction, gaze data bridges the communication between human and non-human or human-like devices, to eventually enhance the user's interaction experiences. The system infers the user's attentional state and adapts its behavior accordingly, by simply tracking the data. This enables hands-free interaction between the user and the system. For example, an intelligent tutoring system can analyze collected gaze data and detect when a student is struggling or disengaged and provide targeted assistance or feedback [RJ21]. Furthermore, in immersive virtual reality (VR) and augmented reality (AR) experiences, gaze-based interaction techniques enable a more natural interaction with virtual objects and environments. Gaze-based interfaces in these systems can support disabled users in accessing computers, communicating, and controlling electronic devices solely relying on their gaze.

- Human-Robot Interaction

  Gaze data also proves its competency in human-robot interaction (HRI). Studies [Kle86; Mut09; SCV21] have verified that human gaze enhances the ability of robots in HRI tasks. Precisely, it empowers robots to perceive and respond to human social cues, visual intentions, and communicate attentional states. For instance, using human-like gaze cues during human-robot handover events, i.e., for a robot

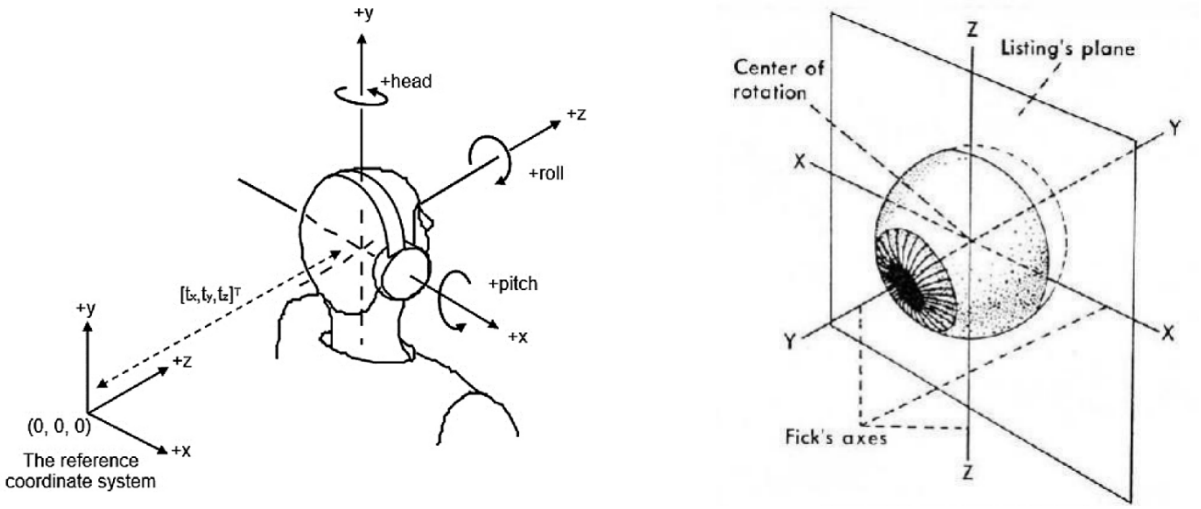
to fetch and handover an object to a person, can improve the synchronization and the perceived quality of the handover event [MTG+14].

- Medical Application

  In psychological research [ACC+21; BSP+15; CST+22; LSP+23] and clinical settings, gaze data is utilized to study cognitive processes and social interactions. It provides researchers with valuable information about how individuals perceive and respond to stimuli. By tracking these data, it is possible to assess neurological disorders, such as autism spectrum disorder (ASD), attention deficit hyperactivity disorder (ADHD), and Parkinson's disease. It can also assist in rehabilitation programs for patients with visual impairments or motor disabilities.

The extensive use-cases of human gaze demonstrate its potential to substantial advancements in many research areas, contingent upon the utilization of gaze data. Thus, gaze data of both high quality and quantity is an essential prerequisite for any application that relies on this information. Particularly in systems powered by artificial intelligence, it provides richer contextual information to understand the user's intention and engagement, thereby enhancing the system's performance. To increase the robustness of these systems under broader range of environmental conditions, gaze dataset should encompass a variety of viewpoints, extreme gaze angles, lighting variation, and image resolutions. Unfortunately, existing gaze datasets are quantitatively insufficient. They are mostly constrained to frontal, stationary settings, covering a relatively narrow range of gaze directions or obtained through well-defined lab environments, which lacks generalization in real-world scenarios. To circumvent the data need, gaze redirection technique can serve as a powerful tool to enhance the quantity and variability of gaze data.

Gaze redirection is the process of modifying the eye gaze and its correlated facial regions within an input image to a desired direction, while preserving the original identity of the subject. This technique can synthesize gaze samples by artificially modifying the gaze direction in existing images or video frames. As a result, it can generate varied datasets that simulate how people look at different view points or objects under various 3D conditions, eliminating the need to collect real-world data for every new scenario. Gaze redirection can be further extended to a wider range of applications. In video conferencing, it can simulate eye contact between the participants, creating a more natural and engaging interaction. It also proves useful in photography and film production. For instance, during group photo sessions, participants often do not look at the camera simultaneously, and in film, the relocation of CGI characters might necessitate changes in an actor's gaze direction. Gaze redirection can resolve the common challenges in these scenarios by efficiently adjusting the eye gaze toward a target direction.



(a) Head Rotation [DE16] **(b)** Gaze Rotation [KAY08]

Figure 1.1: 3D nature of the gaze redirection task. Both eyes and the head respectively have 3D coordinate system. (a) shows the pitch, yaw, and roll of the head's coordinate system. (b) describes the three-dimensional orientation of the eye during rotation, typically within the concept of Listing's Plane, which is commonly used in theoretical studies of vision science.

Most of the existing approaches in gaze redirection deploy this as a 2D problem and are not explicitly 3D aware. In fact, gaze and head redirection is inherently a 3D problem, as both head and eyes move in three-dimensional space. The head can rotate around three axes (yaw, pitch, and roll), and the eyes also have three degrees of freedom (vertical, horizontal, and torsional rotations), illustrated in Figure 1.1. Accurately estimating and redirecting these movements necessitates a thorough understanding and representation in 3D. Furthermore, applications such as virtual environments, films, and video games that utilize gaze and head redirection operate within a 3D space [DE16]. Thus, the task of redirection is naturally 3D aware.

Being conscious of the 3D nature of the task, a recent study [RSW+23] proposed a novel 3D-aware method. While it outperforms the existing 2D approaches, the performance is constrained to the images that were seen during the training process. That is, to perform the redirection for a person who was not shown during the training, the model has to be fine-tuned on one or more images of this person.

Therefore, this work aims to develop a method that can operate directly on any input image while incorporating high degree of 3D awareness. We achieve this by redefining the task as conditional image generation. Our approach introduces three significant contributions to enhance 3D awareness:

1. Adoption of a Viewpoint-Conditioned Diffusion Model: Originally used for novel view synthesis, this model has proven effective at learning the transformations within three-dimensional spaces, capturing extensive spatio-temporal information.
2. Explicit 3D Gaze Rotation: We calculate gaze rotations directly from the gaze labels associated with each image.
3. Integration of 3D Facial Geometry Priors: To deepen the model's understanding of human facial structures, we incorporate 3D priors from latent facial parameters into the training process.

Additionally, we introduce a novel loss function that targets gaze direction and head orientation, designed to enhance the model's task-specific learning capabilities. Through these contributions, this thesis demonstrates the potential of our proposed method to effectively address the gaze redirection task using an image generative approach.

1.2 Outline

This thesis is structured as follows:

Chapter 2 – Related Works introduces two streams of work, gaze redirection and novel view synthesis, that are required for the task-driven formulation of this thesis.

Chapter 3 – Ground Works provides background knowledge and detailed formulas of diffusion models, extending to conditional diffusion models, which are necessary for the methodological formation of this thesis.

Chapter 4 – Method presents the proposed methodology of this work, building upon the understanding of conditional diffusion models. Three distinct strategies and a task-specific loss function are demonstrated to infuse the 3D-awareness, supported by detailed algorithms for training and generating images.

Chapter 5 – Experiments accounts for an in-depth explanation of the experimental setting, including the dataset, implementation details, and evaluation metrics. This chapter further presents the results obtained from extensive experiments and evaluate the effectiveness and performance of our proposed method.

Chapter 6 – Discussion&Limitations analyze and discuss our approach, examining the three distinct methods we have proposed. It also discusses the hypothesized limitations and outlines potential avenues for further improvements in future work.

2 Related Works

Our work integrates two distinct lines of research, addressing the task of gaze redirection as an approach of novel view synthesis. By discussing these two lines of work in depth, this chapter aims to establish a solid foundation and motivation for our proposed method.

Section 2.1 provides a comprehensive overview of the existing research in gaze redirection, including its background, key features, and limitations. It highlights the gaps in the recent methods and establishes the basis of our proposed methodology. Section 2.2 delves into the concept of novel view synthesis, which our work adopts as a solution to enhance the gaze redirection task. This section outlines how synthesizing new viewpoints can be applied to alter gaze direction in images.

2.1 Gaze Redirection

Conventional approaches to gaze redirection are formulated on a 3D graphics model, which involves re-rendering the entire input region. Consequently, the output quality is limited to its 3D modeling capabilities and requires expensive computation. For example, GazeDirector [WBM+17] fits a 3D morphable model to restore the shape, pose, and appearance of the eye region in 3D. A dense flow field that matches the eyelid motion between the original and the target gaze directions is then derived from the fitted model. Finally, the redirection is performed by warping the eyelids using the computed flow field and compositing the newly redirected eyeballs into the output image. While the results are aesthetically highly plausible, some components that are difficult to render or induce distortions to the model are occluded, such as eyelashes, eyeglasses, or shadows from hooded eyes.

Recent advancements in redirecting gaze and head orientation increasingly rely on deep learning-based approaches. DeepWarp [GKSL16] utilizes a deep convolutional network to predict a warping field, which is directly applied to the input image to produce a redirected eye image. This method is simple but has several drawbacks. First, it often struggles when the redirection angle between the source and target image is beyond a specific range. Second, the pixel-wise minimization used during the training process

does not accurately reflect the perceptual quality of the outputs. Lastly, the fundamental weakness of warping-based methods is that the resulting image is merely a modified version of the original input, unable to generate entirely new pixels out of the input image. Therefore, these methods cannot synthesize extreme changes in gaze and head directions or variations in lighting conditions.

To introduce generative powers to the gaze redirection task, He et al. [HSZH19] proposed a framework based on generative adversarial networks (GANs). They integrate a gaze estimator into the discriminator network to ensure photorealistic outputs and high redirection accuracy. To maintain the perceptual consistency of the results, a perceptual loss and cycle-consistency loss are also employed. Despite the efforts, the work is limited to high-resolution images under well-paired head and gaze orientations and does not generalize effectively to in-the-wild images.

These previous methods, unfortunately, have been restricted to using inputs from the eye region, requiring high-quality images for training, but often fail to preserve gaze accurately. The following two works, which will be introduced in more detail, generate full-face images instead of eye patches alone. A notable difference from previous papers is that they address the problem in a disentangled manner and explicitly simulate redirection by exploiting a 3D rotation matrix.

2.1.1 Self-Transforming Encoder-Decoder

Zheng et al. [ZPZ+20] propose a self-transforming encoder-decoder architecture that imposes a disentanglement between the task-relevant (gaze, head) and the task-irrelevant factors (e.g., lighting, hue, blurriness, etc.). They typically follow the initial transforming encoder-decoder architecture from Hinton et al. [HKW11], where the encoder predicts an embedding, and this embedding is transformed by pre-defined steps of transformation. STED [ZPZ+20] defines this transformation as rotations, which is the transition from the canonical representation of the gaze and head system to the general world representation. Given a pair of source and target images, the encoder estimates the distinct pseudo-label conditions of gaze and head orientation along with the personal embeddings. The embeddings from the source image are reverted to the canonical representation by an inverse rotation matrix derived from the pseudo-labels. Then, the embeddings are transformed back to the world representation through a rotation matrix based on the target image's pseudo-conditions. Finally, the discriminator collects the embeddings and predicts a redirected output image. They demonstrate that this latent space transformation with pseudo-condition labels allows the model to be capable of learning the unknown peripheral factors, such as lighting, hue, shadow, and camera distance. Additionally, a new functional loss is introduced to prioritize the minimization

of task-relevant discrepancies between the generated and target images. At inference time, the pseudo-target conditions are replaced with the target's actual ground-truth conditions to generate the final redirected face image.

STED contributes to great advances in gaze redirection tasks, surpassing earlier models in terms of achieving high-fidelity in gaze redirection. Nevertheless, it suffers from the problem of maintaining the identity of the subjects. This includes difficulties in retaining person-specific features like face shape or unusual facial expressions, as well as finer facial details such as moles and freckles. Above all, it lacks 3D-awareness. The rotation matrix, which represents the 3D transformation, is applied to the 2D latent embeddings, mixed with the eyes and the rest of the face. Such operation does not take the inherent characteristics of the non-flexible eyeball rotation and the flexible deformation of the residual face regions into consideration.

2.1.2 GazeNeRF

GazeNeRF [RSW+23] proposes a novel 3D-aware gaze redirection method to realize the actual 3D eyeball rotation. The underlying idea is that the physical face and eyes are separate 3D structures: the deformable face without eyes and the eyeballs that solely rotate during eye movement. To disentangle the eyes and the rest of the face while incorporating a sense of three-dimensionality, it typically takes advantage of the NeRF [MST+20] architecture. NeRF, an approach to novel view synthesis tasks, deploys a single multilayer perceptron (MLP) network. The network is optimized to learn the mapping of a 3D spatial point and view direction to an RGB color and volume density. Eventually, the outputs from the network are used for volume rendering which naturally enables NeRF-based architectures to learn the 3D volumetric information of objects. Similarly, GazeNeRF employs two separate MLP streams for the eyes and face, respectively. This ensures that the feature maps from the eyes stream already include the eye information in a 3D manner, and directly applying the rotation matrix tackles the problem of rigid 3D eyeball rotation. To render the redirected output image, the respective feature maps of the two streams are merged in the end.

GazeNeRF outperforms STED in preserving the identity of the input image after redirection. Yet, it struggles with generalizing to unseen or in-the-wild images that were not present during training. Redirecting a new image of a person requires additional fine-tuning, which is not practical for real-world applications and adds significant computational load.

Our work aims to preserve the advantages of the 2D approach, where arbitrary directions can be applied to any input image. At the same time, to infuse 3D awareness in a different

way from the aforementioned approach, an alternative line of work is explored from the novel view synthesis task.

2.2 Novel View Synthesis

Novel view synthesis (NVS) is the task of predicting the physical appearance of an object in a 3D scene from new, unexplored viewpoints. The goal is to create realistic and accurate 3D-consistent views, using the limited contextual information available from 2D images to facilitate a natural transition from 2D to 3D mapping. Humans are naturally able to infer and imagine how an object would be depicted from different perspectives. We can even envision the 3D shape of objects that do not or cannot exist in reality. This level of generalization is made possible by the extensive visual knowledge that is accumulated over a lifetime. Approaches to NVS aim to encapsulate this visual knowledge as known-priors in various forms, such as geometry priors, generative priors, and language-guided priors.

2.2.1 Geometry-Prior

Most recently, the Neural Radiance Fields (NeRF) [MST+20] has achieved great advancements in geometry-prior-based methods. Given a set of multi-view images along with their corresponding camera poses, it recovers the underlying 3D scene as a radiance field parametrized by a neural network. To render a novel view from a particular viewpoint, a series of 3D points are sampled along a ray that passes through the scene. These spatial points, coupled with their viewing directions, are inputs to the model, which then produces the corresponding colors and densities. Finally, these colors and densities are compiled into a 2D image by classical volume rendering techniques. NeRF-class models inherently guarantee 3D consistency due to their structural design. Yet, the process itself is computationally intensive and requires precise camera calibration and poses for training. Misestimation in camera poses can lead to significant inaccuracies in the rendered images, making it vulnerable to errors in the input data. To reduce NeRF inputs, follow-up works [JTA21; MCL+21; NBM+21] have focused on probing less informative data, such as unposed images or sparse views, with various regularization losses.

2.2.2 Generative-Prior

As an alternative strategy to achieve 3D consistency, generative-based approaches rely on diffusion models. It is inspired by the remarkable success of 2D diffusion models in generative tasks, which have demonstrated the potential for establishing well-founded priors of the physical world. Especially, so-called large-scale text-to-image diffusion models like Imagen [SCS+22], Dall-E [RPG+21], and StableDiffusion [RBL+22] have shown that stronger semantic and geometric priors can be learned, even though they were purely trained on 2D images. Building upon this idea, it seems feasible to realize 3D diffusion models by training them on 3D data and capture robust priors of the real 3D world. However, transforming or synthesizing vast amount of 3D structures into usable data is yet another significant challenge and is non-trivial under current conditions [LLW+23].

Recent works [CNC+23; LWH+23; WCM+22] integrate the 3D capabilities using a conditional diffusion model and redefine the novel view synthesis task as a conditional image-to-image translation task. These models are conditioned by an input image and a particular viewpoint (R,T), i.e., rotation matrix R and a translation matrix T. They demonstrate that their viewpoint-conditioned diffusion models learn the rotations and translations of three-dimensional space and can have rich geometric information. These lines of work also perform 3D reconstruction using the predicted images from novel views. This once again verifies that viewpoint-conditioned diffusion models can learn the 3D priors of the physical world.

Building on this approach, our work addresses the gaze redirection task as a novel view synthesis task using a generative model. Precisely, a conditional diffusion model is adopted, where the conditions include a reference image and specific viewpoints to guide the diffusion process. The core idea is that generative novel view synthesis is analogous to conditional image generation task. Essentially, all that is required is to condition a 2D image diffusion model on the input image along with the corresponding conditions.

3 Ground Works

Building on the theoretical context provided in Chapter 2, this section develops the mathematical and systematic foundations of our work. First, the origins to the diffusion models are explored, introducing the scientific motivation behind their development and the detailed mechanisms in terms of mathematical equations and a specific model architecture integrating the diffusion process. Then, it is further extended to diffusion models under specific conditions, known as conditional diffusion models. This concept represents the core mathematical framework of our approach, enabling targeted generation and manipulation within the model.

3.1 Diffusion Probabilistic Models

Diffusion Probabilistic Models (DPMs), commonly referred to as diffusion models, was first introduced by Sohl-Dickstein et al. [SWMG15]. It uses a Markov chain to gradually convert one distribution into another. The foundational concept comes from the "diffusion process" within the context of non-equilibrium statistical physics. Diffusion process describes the random movement of particles (such as molecules, atoms, or photons) that change over time from areas of higher concentration to areas of lower concentration, driven by the laws of statistical thermodynamics. In the absence of external forces, this process continues until a state of equilibrium is reached, where the particles no longer change in time and become uniform across the system.

Following this law of physics, the diffusion model consists of a forward process that gradually transforms data distribution into Gaussian distribution (random noise) and a reverse process that recovers the data from the noise. The reverse process is the critical mechanism that empowers diffusion models with their generative capabilities, allowing them to transform completely random noise into specific desired outputs, such as images, speech, or text. This diffusion process formulated based on the methodologies by Sohl-Dickstein et al. and He et al. [HJA20; SWMG15], forms the backbone of subsequent research that broadens this concept.

In a nutshell, a diffusion model is a noise predictor parameterized as $\boldsymbol{\epsilon}_\theta\left(\boldsymbol{x}_t, t\right)$. For a diffusion model of total T diffusion steps, t is the timestep index for $t = 0, 1, 2, \cdots, T$.

It starts with a fully random noise $\boldsymbol{x}_T$ and aims to produce a slightly more "denoised" sample $\boldsymbol{x}_{t-1}$ from the previous sample $\boldsymbol{x}_t$, until reaching the final image sample $\boldsymbol{x}_0$. The noisy sample $\boldsymbol{x}_t$ can be thought of as a linear combination of the random noise $\boldsymbol{x}_T$ with a predicted noise level $\boldsymbol{\epsilon}$, drawn from a Gaussian distribution at timestep t.

The forward (diffusion) process q operates as a fixed Markov chain that gradually adds noise into the initial data distribution $\boldsymbol{x}_0 \sim q(\boldsymbol{x}_0)$ to produce increasingly noisy samples from $\boldsymbol{x}_1$ to $\boldsymbol{x}_T$. A certain variance schedule, given by $\beta_{1:T}$, defines how much noise is added at each time step:

$$q\left(\boldsymbol{x}_{1:T} \mid \boldsymbol{x}_0\right) := \prod_{t=1}^{T} q\left(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}\right) \tag{3.1}$$

$$q\left(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}\right) := \mathcal{N}\left(\boldsymbol{x}_t; \sqrt{1-\beta_t}\boldsymbol{x}_{t-1}, \beta_t \mathbf{I}\right) \tag{3.2}$$

Eventually when $T \to \infty$, $\boldsymbol{x}_T$ becomes equivalent to an isotropic Gaussian distribution. Under this property, sampling from an arbitrary time step $\boldsymbol{x}_t \sim q\left(\boldsymbol{x}_t \mid \boldsymbol{x}_0\right)$ can be reduced to a closed form. Using the notation $\alpha_t := 1 - \beta_t$, $\bar{\alpha}_t := \prod_{s=0}^{t} \alpha_s$, and $1 - \bar{\alpha}_t$ denoting the variance of the noise for an arbitrary timestep:

$$q\left(\boldsymbol{x}_t \mid \boldsymbol{x}_0\right) = \mathcal{N}\left(\boldsymbol{x}_t; \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0, \left(1-\bar{\alpha}_t\right)\mathbf{I}\right) \tag{3.3}$$

$$\boldsymbol{x}_t = \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \tag{3.4}$$

By reversing the above forward process from $\boldsymbol{x}_T$ to $\boldsymbol{x}_0$, the noise can be gradually subtracted by sampling from $q\left(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t\right)$. This leads to regenerating the true data sample from a Gaussian noise input $\boldsymbol{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Using Bayes theorem, the posterior $q\left(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t, \boldsymbol{x}_0\right)$ is also a Gaussian with mean $\tilde{\mu}_t\left(\boldsymbol{x}_t, \boldsymbol{x}_0\right)$ and variance $\tilde{\beta}_t$ defined as follows:

$$q\left(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t, \boldsymbol{x}_0\right) = \mathcal{N}\left(\boldsymbol{x}_{t-1}; \tilde{\mu}\left(\boldsymbol{x}_t, \boldsymbol{x}_0\right), \tilde{\beta}_t \mathbf{I}\right) \tag{3.5}$$

$$\tilde{\beta}_t := \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t \tag{3.6}$$

$$\tilde{\mu}_t\left(\boldsymbol{x}_t, \boldsymbol{x}_0\right) := \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\boldsymbol{x}_0 + \frac{\sqrt{\alpha_t}\left(1-\bar{\alpha}_{t-1}\right)}{1-\bar{\alpha}_t}\boldsymbol{x}_t \tag{3.7}$$

However, it is not trivial to compute $q\left(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t\right)$, especially when the data distribution is unknown. As an alternative, a neural network can be utilized to approximate $q\left(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t\right)$ as:

$$p_\theta\left(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t\right) := \mathcal{N}\left(\boldsymbol{x}_{t-1}; \mu_\theta\left(\boldsymbol{x}_t, t\right), \sigma_\theta\left(\boldsymbol{x}_t, t\right)\right) \tag{3.8}$$

In the initial approach of Sohl-Dickstein et al. [SWMG15], the network is trained in a way that predicts the mean μ_θ and the variance σ_θ. Ho et al. [HJA20] optimize this by

training the model $\boldsymbol{\epsilon}_\theta\left(\boldsymbol{x}_t, t\right)$ to predict the noise $\boldsymbol{\epsilon}$ in Equation 3.4. They ignore the fact that the variances are learnable and instead fix them to constants $\beta_t\mathbf{I}$ or $\tilde{\beta}_t\mathbf{I}$, for every step t. This leads to a simplified training objective as a mean-squared error loss between the true noise and the predicted noise:

$$L_{\text{simple}} := E_{t\sim[1,T],x_0\sim q(x_0),\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[\left\|\boldsymbol{\epsilon}-\boldsymbol{\epsilon}_\theta\left(\boldsymbol{x}_t, t\right)\right\|^2\right] \tag{3.9}$$

During sampling, the mean $\mu_\theta\left(\boldsymbol{x}_t, t\right)$ can be calculated from $\boldsymbol{\epsilon}_\theta\left(\boldsymbol{x}_t, t\right)$ as:

$$\mu_\theta\left(\boldsymbol{x}_t, t\right)=\frac{1}{\sqrt{\alpha_t}}\left(\boldsymbol{x}_t-\frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta\left(\boldsymbol{x}_t, t\right)\right) \tag{3.10}$$

3.1.1 Diffusion process in continuous time

Since the foundational works on diffusion model was established by Sohl-Dickstein et al. and He et al. [HJA20; SWMG15], there have been several significant follow-up works that build upon and extend this original concept. These enhancements include a variety of improvements such as adjusting the variance schedules, reducing the number of sampling steps, refining the denoising function, modifying network architectures, enabling conditional generation, and expanding to cross-modal applications. Specifically, works from Kingma et al., Saliman et al, Chen et al., Song et al. [CZZ+20; KW13; SH22; SSK+21] consider the diffusion process to be performed over a continuous time variable, in contrast to the discrete Markov chain approach described in the original works [HJA20; SWMG15]. This thesis follows the parametrization of the diffusion process in these works.

A diffusion model has latent variables $\boldsymbol{z}=\{\boldsymbol{z}_t \mid t\in[0,1]\}$, which represent noisy images. These are generated by adding a specific amount of noise to images from the training dataset $\boldsymbol{x}\sim p(\boldsymbol{x})$. A noise scheduler consisting of differentiable functions α_t and σ_t is defined in a way that the log signal-to-noise ratio, $\lambda_t=\log[\alpha_t^2/\sigma_t^2]$, decreases monotonically over time t. These components define the forward process $q(\boldsymbol{z}\mid\boldsymbol{x})$ as a Gaussian process satisfying the following Markovian structure:

$$q\left(\boldsymbol{z}_t \mid \boldsymbol{x}\right)=\mathcal{N}\left(\boldsymbol{z}_t;\alpha_t\boldsymbol{x},\sigma_t^2\boldsymbol{I}\right) \tag{3.11}$$

In the diffusion model, the purpose of function approximation is to denoise the latent variable $\boldsymbol{z}_t$, sampled from the conditional distribution $q(\boldsymbol{z}_t \mid \boldsymbol{x})$, with an estimated noise value $\boldsymbol{\epsilon}_\theta(\boldsymbol{z}_t,\lambda_t)$ so that it closely approximates the original clean data $\boldsymbol{x}$. It is also possible to express the variable $\boldsymbol{z}$ as a deterministic variable $\boldsymbol{z}=f_\theta(\boldsymbol{\epsilon},\boldsymbol{x})$, where $\boldsymbol{\epsilon}$ is an auxiliary variable with independent marginal $p(\boldsymbol{\epsilon})$, and $f_\theta(\cdot)$ is some vector-valued function parameterized by θ, i.e, the neural network in our case.

During sampling, the mean from Equation 3.10 and the variance estimate each denoising step as:

$$\boldsymbol{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(\boldsymbol{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \boldsymbol{\epsilon}_\theta \left(\boldsymbol{x}_t, \lambda_t \right) \right) + \sigma_t \boldsymbol{z} \tag{3.12}$$

3.1.2 UNet Architecture

UNet [HJA20; RFB15] is the standard architecture exploited in diffusion models, primarily due to its structural features that are suitable for iterative image denoising, which is essential for the diffusion process. The "U" shaped architecture comprises a series of convolutional layers in two networks: the encoder and the decoder. The encoder encodes an input image into a compressed representation by reducing the spatial dimensions while increasing the depth of the feature maps. The decoder then reverses this process to transform the compressed information back to an image. This progressive downsampling in the encoder and upsampling in the decoder makes it particularly efficient in reconstructing images from noisy data.

A pivotal feature of the UNet is the skip connections that directly connect corresponding layers between the encoder and decoder. At each level of the encoder, feature maps are generated from the convolutional and pooling operations. These feature maps are stored and subsequently concatenated to the matching layers in the decoder of same spatial levels. The concatenated data is then jointly processed in the decoder. Skip connections facilitate a smoother flow of gradients through the model, effectively addressing the issue of vanishing gradients. They provide direct access to detailed information from the encoder side to be connected to the decoder side of the network, adding crucial details or high-resolution information that might be lost during the downsampling. Skip connections greatly improve the quality of the reconstructed image, ensuring that fine details crucial for precise image denoising and reconstruction are preserved.

3.2 Conditional Diffusion Models

Endowed with strong generative powers, diffusion models have achieved significant advancements in the field of image synthesis. The development of these models has been particularly accelerated by the introduction of techniques that can incorporate specific conditions into traditional noise-to-image models. Specifically referred to as conditional diffusion models, they are designed to integrate targeted guidance into their

generative process. With the conditions denoted as $\mathbf{c}$, the task then becomes training a neural network to learn:

$$p_\theta\left(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t, \mathbf{c}\right) := \mathcal{N}\left(\boldsymbol{x}_{t-1}; \mu_\theta\left(\boldsymbol{x}_t, t, \mathbf{c}\right), \sigma_\theta\left(\boldsymbol{x}_t, t, \mathbf{c}\right)\right) \tag{3.13}$$

The first approach to gaining control over the diffusion model was using class-labels to produce images of a specific class [DN21]. It trains a separate classifier to distinguish between different classes or features in the data. Then, the trained classifier fits into the model during the reverse diffusion process. The gradients from the classifier penalize the deviations from undesired attributes or directly guide the generation process towards higher probabilities of desired outputs. This can be expressed as:

$$p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, y) \propto p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) \cdot p_\phi(y|\boldsymbol{x}_t) \tag{3.14}$$

where $p_\phi(y|\boldsymbol{x}_t)$ is the probability of class y given the noisy data $\boldsymbol{x}_t$, from the classifier. For example, if generating images of dogs, the classifier guides the diffusion steps to ensure that the characteristics of the generating image match those of dogs. While this method can gain control over the outputs, a severe drawback is the need to train a separate classifier.

Classifier-free guidance[HS22] enhances this approach by directly embedding the classifier into the model. By applying the Bayes rule to the gradient of the separate classifier $\nabla \log p(y|\boldsymbol{x}_t)$, it is divided into two score functions:

$$\nabla \log p(y|\boldsymbol{x}_t) = \nabla \log p(\boldsymbol{x}_t|y) - \nabla \log p(\boldsymbol{x}_t) \tag{3.15}$$

$\nabla \log p(\boldsymbol{x}_t|y)$ is the score of the data $\boldsymbol{x}_t$ conditioned on class y and $\nabla \boldsymbol{x}_t \log p(\boldsymbol{x}_t)$ is the score of all the data $\boldsymbol{x}_t$. As this divided formation can leverage both the conditional and unconditional information, it alleviates the need for a separate, pre-trained classifier. Specifically, in training time, the model learns a shared representation of both conditioned and unconditioned data. It also zeros out the conditioning information, teaching the model to generate high-quality unconditioned samples as well. At the inference stage, the model can be used either conditionally or unconditionally. The strength of conditioning can be adjusted by a hyperparameter that scales the influence of the condition during the generation process. This flexibility allows the model to generate a wide range of outputs. A higher guidance scale adheres more to the condition with reduced diversity, whereas a lower guidance scale allows more varied and rich samples to be generated.

State-of-the-art diffusion models build on this approach and extend the conditioning control to include text, additional image, or latent information. They have demonstrated remarkable success across a diverse range of tasks including text-to-video, text-to-image,

image-to-image translation, super-resolution, inpainting, colorization, uncropping, and artifact removal. For example, the recent explosion of language-guided 2D image generators like DALL-E[RPG+21], Imagen[SCS+22], and Stable Diffusion[RBL+22] can solve highly-ambiguous generation tasks, obtaining photo-realistic 2D images from strong semantic correlation with the given text-prompt inputs, textual descriptions, semantic maps, partially-complete images, or simply unconditionally from random noise.

4 Method

4.1 Problem Setting

The main idea of our approach is to develop a method where the 3D-aware model generates a 2D face image, with the gaze adjusted to any direction of our desire. To this end, we redefine the task of gaze redirection as a novel view synthesis task, by using a conditional diffusion model that can learn the 3D priors of the facial geometry. The intuition is that this is similar to any other conditional image generation task.

Given a pair of images $\boldsymbol{x}_{\text{input}}$ and $\boldsymbol{x}_{\text{target}}$, along with the relevant conditions $\boldsymbol{c}$ that we propose to incorporate the 3D awareness, our goal is to sample the redirected target image from the conditional distribution:

$$p\left(\boldsymbol{x}_{\text{target}} \mid \boldsymbol{x}_{\text{input}}, \boldsymbol{c}\right) \tag{4.1}$$

4.2 Incorporating 3D-Awareness

To achieve 3D awareness, we integrate this concept through three distinct approaches:

1. First, we employ a model that is inherently 3D-aware in its structure, enabling it to intuitively understand and interpret 2D data within the 3D context.
2. Second, we apply 3D transformations, which adapt the spatial orientation of the eye gaze, allowing the model to handle and analyze input from multiple perspectives accurately.
3. Third, we capitalize on the 3D priors of the facial geometry. By leveraging an existing model that already has advanced priors of the human face topology, we aim to extract relevant information that will enhance our model's capability to generate and manipulate faces images in three-dimensional space.

4.2.1 3D-aware diffusion model architecture

This thesis builds on the works of novel view synthesis that introduce 3D generative models. These models are typically conditional diffusion models, conditioned on images and their corresponding poses. A significant advancement in this field is the X-UNet architecture introduced in 3DiM [WCM+22]. It has demonstrated remarkable effectiveness in generating 3D-consistent frames for novel view synthesis using only 2D images for training. Unlike other state-of-the-art methods [LWH+23; MRLV23], it doesn't rely on any pre-existing knowledge from large-scale diffusion models, and the architecture is claimed to be naturally 3D-aware. Inspired by these capabilities, we have adopted the X-UNet architecture to integrate 3D awareness into our gaze redirection task.

The architecture of X-UNet differs from the original UNet [HJA20; SWMG15] in several aspects. First, unlike DDPM[HJA20] which denoises multiple frames simultaneously at a single noise level, X-UNet assigns a unique noise level to each frame. Second, the poses conditions are encoded as pose embeddings, which are formatted to match the dimensionality of the images. These pose embeddings are subsequently combined with noise-level positional encodings to produce the final format of the embeddings. This contrasts the DDPM approach, which uses noise-level encodings alone. Lastly, X-UNet defines a cross-attention layer, allowing the feature maps of each frame to be interchangeable with other frames. This cross-attention mechanism, which shares parameters between the two views, enhances the learning of complex, nonlinear image transformations. Conversely, DDPM uses successive self-attention layers, processing sequences over time. These modifications collectively improve 3D consistency and better alignment with the conditioning image.

4.2.2 Pose embeddings from 3D gaze rotations

One of the core elements that outline 3DiM [WCM+22] as a 3D-aware method is the use of 3D transformations as conditions within its conditional diffusion model. Specifically, the 3D pose information is encoded as pose embeddings for the images. To incorporate this approach, we utilize the gaze label $\boldsymbol{g} = (\theta, \phi)$ from the images. The gaze labels, pitch θ and yaw ϕ, are angles of the spherical coordinate systems of head orientation and gaze direction [ZSB18]. A rotation matrix can be derived from these gaze labels as:

$$\boldsymbol{R} = \begin{pmatrix} \cos\phi & 0 & \sin\phi \\ 0 & 1 & 0 \\ -\sin\phi & 0 & \cos\phi \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\theta & -\sin\theta \\ 0 & \sin\theta & \cos\theta, \end{pmatrix} \tag{4.2}$$

This rotation formulates the transformation from the canonical space to the world(target) space [RSW+23; ZPZ+20]. The zero-representation indicates a frontal direction, where the face is oriented directly to the camera. In the original work, 3D transformations are utilized within their ray construction framework to create pose embeddings. However, due to the incompatibility of this approach with our setup, we replace it with a fully-connected layer. The rotation matrix $\boldsymbol{R} \in \mathbb{R}^{3\times3}$ is encoded to rotation embedding $\boldsymbol{R} \in \mathbb{R}^{128\times128\times3}$, that match the dimensionality of the images. Then, we stack the rotation embedding from the input and target images along a new dimension, resulting in $\boldsymbol{R} \in \mathbb{R}^{2\times128\times128\times3}$. This approach is in parallel with 3DiM, which stacks input and target images along a new dimension to facilitate its cross-attention mechanism that enables weight sharing across frames.

Rotation embeddings are further augmented by its positional encodings, as detailed in [MST+20; WCM+22]. For a given number of frequency levels L, the positional encoding function $PE(\cdot)$ maps each coordinate values x to $2L$ features and applies a series of sinusoidal function, represented as follows:

$$PE(x) = \Big(\sin(2^0\pi x), \cos(2^0\pi x) \ldots, \sin(2^{L-1}\pi x), \cos(2^{L-1}\pi x)\Big) \tag{4.3}$$

The intuition behind this positional encoding is to transform each input coordinate into a higher-dimensional space, thereby enhancing the model's ability to capture the details and subtle complexities in the data. As this encoding strategy originates from NeRF [MST+20], it significantly aids the reconstruction and rendering of 3D objects by providing richer spatial information.

4.2.3 3D Priors of the facial geometry

As we adapt a model initially designed to predict novel views of objects, there exists a fundamental difference between the goal of this task and ours of generating novel gaze directions of humans. Compared to standard 3D objects in the world, modeling human facial geometry is significantly more complex and demands a higher level of understanding from the model. For instance, human faces exhibit a vast range of expressions and subtle movements from facial muscles, offering nearly limitless degrees of freedom in manipulating the details. This complexity is greater than that found in static objects, where the range of potential variations is typically more constrained and predictable. Although 3DiM [WCM+22] is capable of performing 3D object reconstruction based on their predicted views, demonstrating their ability of learning 3D priors, 3D face reconstruction from 2D image is yet another specialized field of research [ZTB+18].

To bridge this gap between the task, the model should be provided with additional information of the face. Thus, we utilize established models known for their ability

to conduct 3D face modeling, where detailed facial parameters can be extracted. One such model is the Detailed Expression Capture and Animation (DECA) [FFBB21], an animatable displacement model capable of synthesizing realistic geometric details through varying expression parameters. DECA takes a single two-dimensional image as input and outputs person-specific details that can be used for realistic animation. It demonstrates exceptional robustness in reconstructing extreme geometric details, including common occlusions, wide pose variations, and significant illumination changes.

In our work, we specifically use the encoder architecture of DECA, which generates a latent code representing shape, expression, pose, texture, and lighting parameters from a 2D face image. This results in a 233dimensional latent vector. We transform this latent code into a latent matrix and integrate it with the input image $\boldsymbol{x} \in \mathbb{R}^{C \times H \times W}$, as additional information. H and W denote the height and width of the input image, respectively, both set at 128, while C represents the channel dimension of 3, corresponding to the RGB color space. We have developed two methods for this integration:

- First, we expand the latent code $\boldsymbol{l} \in \mathbb{R}^{236}$ to $\boldsymbol{l} \in \mathbb{R}^{1 \times 128 \times 128}$, and concatenate it along the channel dimension, resulting in $\boldsymbol{x} \in \mathbb{R}^{4 \times 128 \times 128}$.
- Second, we extend the latent code $\boldsymbol{l} \in \mathbb{R}^{236}$ to $\boldsymbol{l} \in \mathbb{R}^{3 \times 128 \times 128}$, and stack it to a new dimension, resulting in $\boldsymbol{x} \in \mathbb{R}^{2 \times 3 \times 128 \times 128}$.

The first method is inspired by the Concat-UNet [SHC+21] architecture, where it concatenates the input and target image along the channel axis, as it does not share weights across frames. In constrast, 3DiM stacks these images along a new dimension, as it uses a cross-attention mechanism that shares weights across frames. The second method draws inspiration from this approach. As the initial operation in our model is a convolutional layer that expands the channel dimensions from 3 to a feature dimension of 128, these variations in image size are straightforwardly manageable.

By integrating these strategic approaches, which differ in three distinct ways, we aim to infuse 3D awareness into our methodology. This enhancement is expected to significantly improve the fidelity and accuracy of our generative outputs of human face. As a result, our conditional diffusion model is designed to process a noisy target image $\boldsymbol{z}$, using a set of complex conditioning information. These include a conditioning input image $\boldsymbol{x}_{\text{input}}$, rotational data from both the input and target $\boldsymbol{R}_{\text{input}}, \boldsymbol{R}_{\text{target}}$, and latent codes for the input image $\boldsymbol{l}_{\text{input}}$, as well as a specified noise level λ_t.

4.3 Objectives

Given a data distribution $q(\boldsymbol{x}_1, \boldsymbol{x}_2)$ of image pairs from a common person and the relevant conditions $\boldsymbol{c}$ from the images, we define an isotropic Gaussian process. It incrementally adds noise to data samples as the signal-to-noise-ratio λ_t decreases.

Following the methodology described by Saliman et al.[SH22], we employ a cosine scheduler where $\alpha_t = \cos(0.5\pi t)$. This is based on a standard variance-preserving diffusion process, where $\alpha_t^2 + \sigma_t^2 = 1$. Under this configuration, the log signal-to-noise ratio $\lambda_t = \log[\alpha_t^2/\sigma_t^2]$ breaks down to $\alpha_t = s(\lambda_t)^{\frac{1}{2}}$ and $\sigma_t = s(-\lambda_t)$, with $s(\cdot)$ denoting the sigmoid function $s(x) = 1/(e^{-x} + 1)$. We omit the t in λ_t for simpler notation.

The forward process from 3.11 comes down to:

$$q\left(\boldsymbol{z}_k^{(\lambda)} \mid \boldsymbol{x}_k\right) := \mathcal{N}\left(\boldsymbol{z}_k^{(\lambda)}; \sigma(\lambda)^{\frac{1}{2}}\boldsymbol{x}_k, \sigma(-\lambda)\mathbf{I}\right) \tag{4.4}$$

where $\sigma(\cdot)$ is the sigmoid function. By applying the reparametrization trick from [KW13], $\boldsymbol{z}$ can be further expressed as a deterministic variable sampled from these marginal distributions as:

$$\boldsymbol{z}_k^{(\lambda)} = \sigma(\lambda)^{\frac{1}{2}}\boldsymbol{x}_k + \sigma(-\lambda)^{\frac{1}{2}}\boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \tag{4.5}$$

Then, the task of our neural network is to reverse this process in one of the two frames. That is, the network learns the conditional distribution outlined in 4.1 to approximate the correct noise that was injected in the forward process.

4.3.1 Noise Loss

Following the simplified training objective of Ho et al. [HJA20], we minimize the loss between the predicted noise and true noise to recover the data distribution of the target image. Here, the l2 loss is used:

$$\mathcal{L}_{\text{noise}} = \mathbb{E}_{q(\boldsymbol{x}_1, \boldsymbol{x}_2)}\mathbb{E}_{\lambda, \boldsymbol{\epsilon}} \left\| \boldsymbol{\epsilon}_\theta\left(\boldsymbol{z}_2^{(\lambda)}, \boldsymbol{x}_1, \boldsymbol{c}, \lambda\right) - \boldsymbol{\epsilon} \right\|_2^2 \tag{4.6}$$

The neural network ϵ_θ is tasked with denoising the noisy frame $\boldsymbol{z}_2^{(\lambda)}$, using the clean conditioning frame $\boldsymbol{x}_1$ and the feature-specific conditions $\boldsymbol{c}$ from the frames, along with the logarithmic signal-to-noise ratio λ as guidance. To enhance the clarity of our notation, we will simply use $\boldsymbol{\epsilon}_\theta\left(\boldsymbol{z}_2^{(\lambda)}, \boldsymbol{x}_1\right)$ to denote the noise predicted by the model.

4.3.2 Gaze and Head loss

While L_{noise} leads the overall diffusion process for image generation, a more task-specific guidance is required to ensure the accuracy of gaze and head redirection. A novel functional loss is proposed in STED [ZPZ+20] which prioritizes the minimization of task-relevant inconsistencies between the generated and target images. Inspired by this work, we adopt gaze and head loss to gain control over the redirection task.

To accomplish this, it is necessary to first extract the gaze direction and head orientation from the target images. However, direct access to the generated target image is unavailable during the training phase. As an alternative, we reverse the the sampling process in Equation 4.5 to simulate the target image. During the sampling process, a certain amount of noise is added to the target image, resulting in a noisy target image. The model then predicts this injected noise. By reverting this process, it is possible to approximate the ground truth target image $\hat{\boldsymbol{x}}_k$ using the predicted noise $\tilde{\boldsymbol{\epsilon}}$ and the noisy target image $\boldsymbol{z}_k^{(\lambda)}$:

$$\hat{\boldsymbol{x}}_k = \frac{1}{\sigma(\lambda)^{\frac{1}{2}}} \left(\boldsymbol{z}_k^{(\lambda)} - \sigma(-\lambda)^{\frac{1}{2}} \tilde{\boldsymbol{\epsilon}} \right) \tag{4.7}$$

Gaze direction and head orientation can be determined from the recovered target image using a external estimator, denoted as $\boldsymbol{F}$. As these measurements are initially extracted as pitch (θ) and yaw (ϕ) angles of the head's coordinate system, converting these angles into a 3D vector simplifies the computation of loss functions. With rotation matrix $\boldsymbol{R}$ from Equation 4.2 and a unit vector $(0, 0, 1)$ denoting the frontal direction of the head's coordinate system, the vector can be computed as:

$$\mathbf{v} = \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \boldsymbol{R} \cdot \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} \sin\phi\cos\theta \\ -\sin\theta \\ \cos\phi\cos\theta \end{pmatrix} \tag{4.8}$$

A useful method to measure the similarity between two vectors is cosine similarity. This method evaluates the cosine of the angle between the vectors, indicating whether they are pointing in the same direction. By simply calculating the inverse cosine similarity, the angle between the vectors can be derived.

$$\mathcal{E}_{\text{ang}}(\mathbf{v}, \hat{\mathbf{v}}) = \arccos \frac{\mathbf{v} \cdot \hat{\mathbf{v}}}{\|\mathbf{v}\| \|\hat{\mathbf{v}}\|} \tag{4.9}$$

Gaze and head loss is formulated as the angular error between the gaze direction and head orientation estimated from the reconstructed $\hat{\boldsymbol{x}}_k$ and ground truth $\boldsymbol{x}_k$.

$$\mathcal{L}_{\text{gaze}} = \mathcal{E}_{\text{ang}} \left(\boldsymbol{F}^g \left(\hat{\boldsymbol{x}}_k \right), \boldsymbol{F}^g \left(\boldsymbol{x}_k \right) \right) \tag{4.10}$$

$$\mathcal{L}_{\text{head}} = \mathcal{E}_{\text{ang}}\left(\boldsymbol{F}^{h}\left(\hat{\boldsymbol{x}}_{k}\right), \boldsymbol{F}^{h}\left(\boldsymbol{x}_{k}\right)\right) \tag{4.11}$$

$\boldsymbol{F}^{g}$ and $\boldsymbol{F}^{h}$ denote the gaze and head vector from the estimator, respectively. Finally, the training objective becomes:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{F}_{\text{noise}}} + \lambda_{F_{\text{gaze}}}\mathcal{L}_{\text{F}_{\text{gaze}}} + \lambda_{F_{\text{head}}}\mathcal{L}_{\text{F}_{\text{head}}}. \tag{4.12}$$

Algorithm 4.1 summarizes the training phase.

Algorithmus 4.1 Training

Require: Gaze estimator $\boldsymbol{F}_t$ for training phase
Require: Model $\boldsymbol{\epsilon}_\theta(\cdot)$ to be trained
Require: Noise scheduler $logsnr\_cosine(\cdot)$
Require: Conditioning mask m_{cond} for classifier-free guidance approach
while not converged **do**
$(x_1, \boldsymbol{c}_1), (x_2, \boldsymbol{c}_2, g, h) \sim \mathcal{D}$ ▷ Query data and relevant conditions
$\epsilon \sim N(0, I)$ ▷ Sample random noise
$t \sim U[0, 1]$ ▷ Sample time
$\lambda_t = logsnr\_cosine(t)$ ▷ log-SNR
$\tilde{\boldsymbol{x}} = x_1 + m_{\text{cond}}$ ▷ Randomly discard conditioning to train unconditioned
$\boldsymbol{z} = \sigma(\lambda_t)^{\frac{1}{2}} x_2 + \sigma(-\lambda_t)^{\frac{1}{2}}\boldsymbol{\epsilon}$ ▷ Add noise to target image
$\tilde{\boldsymbol{\epsilon}} = \boldsymbol{\epsilon}_\theta\left(\tilde{\boldsymbol{x}}, \boldsymbol{z}, \boldsymbol{c}_1, \boldsymbol{c}_2, \lambda_t, m_{\text{cond}}\right)$ ▷ Predict noise
$\hat{x}_2 = (\boldsymbol{z} - \sigma(-\lambda_t)^{\frac{1}{2}}\tilde{\boldsymbol{\epsilon}})/\sigma(\lambda_t)^{\frac{1}{2}}$ ▷ Recover target image with predicted noise
$\tilde{g}, \tilde{h} = \boldsymbol{F}_t(\hat{x}_2)$ ▷ Predict gaze and head vector from the gaze estimator
$L_{\text{noise}} = \|\tilde{\boldsymbol{\epsilon}} - \boldsymbol{\epsilon}\|_2^2$ ▷ Noise loss
$\mathcal{L}_{\text{gaze}} = \mathcal{E}_{ang}(\tilde{g}, g)$ ▷ Gaze angular loss
$\mathcal{L}_{\text{head}} = \mathcal{E}_{ang}(\tilde{h}, h)$ ▷ Head angular loss
$\mathcal{L}_\theta = \mathcal{L}_{\text{noise}} + \lambda_{F_{\text{gaze}}}\mathcal{L}_{\text{gaze}} + \lambda_{F_{\text{head}}}\mathcal{L}_{\text{head}}$ ▷ Total loss
$\theta \leftarrow \theta - \gamma\nabla_\theta\mathcal{L}_\theta$ ▷ Optimization
end while

4.4 Generating Redirected Image at Inference

The power of diffusion models becomes evident during the inference phase. Once the model has been effectively trained, it is capable of generating a redirected face image starting from completely random noise. During the sequence of time steps in $\lambda_{\min} = \lambda_T < \lambda_{T-1} < \ldots < \lambda_0 = \lambda_{\max}$, the random noise is gradually removed until it reaches a complete clean image. Detailed notation is provided in Algorithm 4.2.

The model predicts the conditional noise $\boldsymbol{\epsilon}_\theta(\boldsymbol{z}^{(\lambda_t)}, \boldsymbol{x})$ at timestep t, where $\boldsymbol{x}$ is the conditioning image. Incorporating the classifier-free guidance [HS22], the model also predicts the unconditional noise $\boldsymbol{\epsilon}_\theta(\boldsymbol{z}^{(\lambda_t)}, \tilde{\boldsymbol{x}})$. In this case, the positionally encoded pose is zeroed out and the conditioning image is replaced with randomly sampled image $\tilde{\boldsymbol{x}}$ from the standard Gaussian noise. Then, the final noise level is determined by a linear combination of both conditional and unconditional noise, where the parameter w controls the strength of the conditioning effect:

$$\hat{\boldsymbol{\epsilon}} = (w+1)\boldsymbol{\epsilon}_\theta(\boldsymbol{z}^{(\lambda_t)}, \boldsymbol{x}) - w\boldsymbol{\epsilon}_\theta(\boldsymbol{z}^{(\lambda_t)}, \tilde{\boldsymbol{x}}) \tag{4.13}$$

This integration allows for a precise manipulation of how much the conditioning image will influence the final output. It is further possible to denoise $\boldsymbol{z}^{(\lambda_t)}$ at timestep t using the final predicted noise $\hat{\boldsymbol{\epsilon}}$:

$$\hat{\boldsymbol{x}} = \frac{1}{\sigma\left(\lambda_t\right)^{\frac{1}{2}}}\left(\boldsymbol{z}^{(\lambda_t)} - \sigma\left(-\lambda_t\right)^{\frac{1}{2}}\hat{\boldsymbol{\epsilon}}\right) \tag{4.14}$$

This equation is analogous to Equation 4.7, where we reverse the sampling process to recover the ground truth target image. The noisy target image for the next timestep $\boldsymbol{z}^{(\lambda_{t-1})}$ can be sampled from the following distribution, until it becomes completely noiseless.

$$\boldsymbol{z}^{(\lambda_{t-1})} \sim q\left(\boldsymbol{z}^{(\lambda_{t-1})} \mid \boldsymbol{z}^{(\lambda_t)}, \hat{\boldsymbol{x}}\right) \tag{4.15}$$

In practice, $\boldsymbol{z}^{(\lambda_{t-1})}$ is computed as a linear operation of the computed mean from the predicted noise and the variance value from the noise schedule, with the fully noisy $\boldsymbol{z}^{(\lambda_T)}$ from the random Gaussian distribution. The variance $\sigma_\theta\left(\boldsymbol{x}_t, \lambda_t\right)$ is determined using a fixed value from the noise schedule for the current timestep λ_t and the next, less noisy timestep λ_{t-1}. The mean $\mu_\theta\left(\boldsymbol{x}_t, \lambda_t\right)$ is derived using the denoised $\hat{\boldsymbol{x}}$ and noisy $\boldsymbol{z}^{(\lambda_t)}$ at timestep t.

$$\boldsymbol{z}^{(\lambda_{t-1})} = \mu_\theta\left(\boldsymbol{x}_t, \lambda_t\right) + \sigma_\theta\left(\boldsymbol{x}_t, \lambda_t\right) \cdot \boldsymbol{z}^{(\lambda_T)} \tag{4.16}$$

$$\mu_\theta\left(\boldsymbol{x}_t, \lambda_t\right) = \sigma\left(-\lambda_t\right)^{\frac{1}{2}}\left(\frac{\boldsymbol{z}^{(\lambda_t)}\left(1-c\right)}{\sigma\left(\lambda_t\right)^{\frac{1}{2}}} + c \cdot \hat{\boldsymbol{x}}\right) \tag{4.17}$$

$$\sigma_\theta\left(\boldsymbol{x}_t, \lambda_t\right) = \sigma(-\lambda) \cdot c \tag{4.18}$$

$$\textbf{where} \quad c = 1 - \exp\left(\lambda_t - \lambda_{t-1}\right) \tag{4.19}$$

Algorithmus 4.2 Inference

Require: Conditioning mask m_{cond} for classifier-free guidance approach
$\boldsymbol{z} \sim N(0, I)$ ▷ Sample random noise
$\boldsymbol{x}, \boldsymbol{c}_x, \boldsymbol{c}_z \sim \mathcal{D}$ ▷ Query data
for $t = T, ..., 1$ **do**
$\tilde{\boldsymbol{x}} \sim N(0, I)$ ▷ Random sample conditioning image
$\tilde{\boldsymbol{\epsilon}}_{\text{cond}} = \boldsymbol{\epsilon}_\theta(\boldsymbol{x}, \boldsymbol{z}_t, \boldsymbol{c}_x, \boldsymbol{c}_z, \lambda_t, m_{\text{cond}})$ ▷ Predict conditioned noise
$\tilde{\boldsymbol{\epsilon}}_{\text{uncond}} = \boldsymbol{\epsilon}_\theta(\tilde{\boldsymbol{x}}, \boldsymbol{z}_t, \boldsymbol{c}_x, \boldsymbol{c}_z, \lambda_t, m_{\text{cond}})$ ▷ Predict unconditioned noise
$\hat{\boldsymbol{\epsilon}} = (w+1)\tilde{\boldsymbol{\epsilon}}_{\text{cond}} - w\tilde{\boldsymbol{\epsilon}}_{\text{uncond}}$ ▷ Final noise
$\hat{\boldsymbol{x}}_t = (\boldsymbol{z}_t - \sigma(-\lambda_t)^{\frac{1}{2}}\hat{\boldsymbol{\epsilon}})/\sigma(\lambda_t)^{\frac{1}{2}}$ ▷ Recover z at timestep t
if $t > 1$ **then**
$\boldsymbol{z}_{t-1} \sim q(\boldsymbol{z}_{t-1} \mid \boldsymbol{z}_t, \hat{\boldsymbol{x}}_t)$ ▷ Sample z at timestep t+1
else
$\boldsymbol{z}_{t-1} = \hat{\boldsymbol{x}}_t$ ▷ The final denoised z
end if
end for
return $\hat{\boldsymbol{x}}_0$

5 Experiments

This chapter details the experiments conducted to evaluate the effectiveness of our approach. We focus on two primary objectives:

1. Demonstrate that using pose embeddings as conditions for the diffusion model can successfully perform gaze redirection.
2. Incorporating latent parameters of the face can provide additional geometric priors of the facial structures, ultimately improving the results.

Initially, our experiments utilized the ETH-XGaze dataset. However, subsequent tests led us to conclude that this dataset was not compatible with our experimental framework. Section 5.1.1 provides a brief overview of these extensive experiments, discussing the points of failure and the reasons for shifting to an alternative dataset. The remaining sections focus on the experiments conducted using the GazeCapture dataset, which we selected as a more suitable alternative. These sections include details for the implementation, along with both qualitative and quantitative evaluations of the results, providing a comprehensive overview of the performance and applicability of our approach using this dataset.

5.1 Datasets

5.1.1 ETH-XGaze

ETH-XGaze [ZPB+20] is a large-scale gaze estimation dataset featuring high-resolution images that capture wide range of gaze variation under extreme head poses. This dataset was collected from 110 subjects using a multi-view setup with 18 cameras and varying lighting conditions.

Recently, GazeNeRF[RSW+23] achieved state-of-the-art results in novel view synthesis with this dataset. Their preprocessed format contains the rotation (R), translation (T), and camera intrinsic (K) matrices, which can be used for the ray construction framework in 3DiM, to embed the pose conditions. It also includes the latent facial

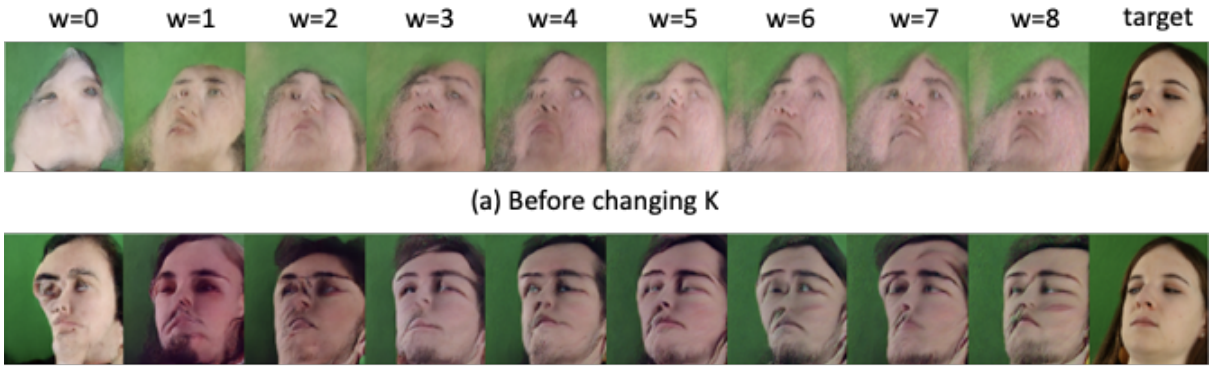


Figure 5.1: Comparison of the results in accordance with the change in intrinsic matrix. In both cases, the conditions are (R,T,K). (a) shows the results using the original intrinsic matrix from the preprocessed dataset, i.e., image size of 512. (b) shows the results after changing the intrinsic matrix to our image size setting of 128. The labels denoted as w is the hyperparameter that controls the strength of the conditioning information.

parameters achieved from the 3D morphable model (3DMM), such as shape, expression, and texture of the face and the illumination of the image. 3DMM is a pre-trained parametric model that transforms the face image into vector space representations. In the original implementation of GazeNeRF, these latent parameters are the ingredients for reconstructing a face mesh in their post-processing steps. Adapting to our settings, we use them as additional feature information of the images to provide the model with deeper insights into complex facial features and three-dimensional structures. Although the raw dataset is not publicly available, the preprocessed formats are readily accessible from our GPU servers. For these reasons, we initially utilized the preprocessed version of the ETH-XGaze dataset for training. Our model, containing approximately 141 million parameters, could only support an image resolution of 128 due to GPU memory constraints, despite employing data parallelism. Consequently, we resized the images from 512×512 pixels to 128×128 pixels and conducted various experiments. At this stage, we are yet to adopt the task-specific head and gaze loss discussed in Section 4.3.2 and solely rely on the noise loss. A timestep of 1000 is used during training, and 256 sampling steps are used during inference.

In our experiments, we made two main observations. Here we only provide qualitative evaluation of the results, without delving into detailed error metrics. To enhance diversity in the visualizations, different samples are presented for each comparison. First, acknowledging the discrepancy in image size between the preprocessed dataset and our experimental settings, we made necessary adjustments to the structures where the

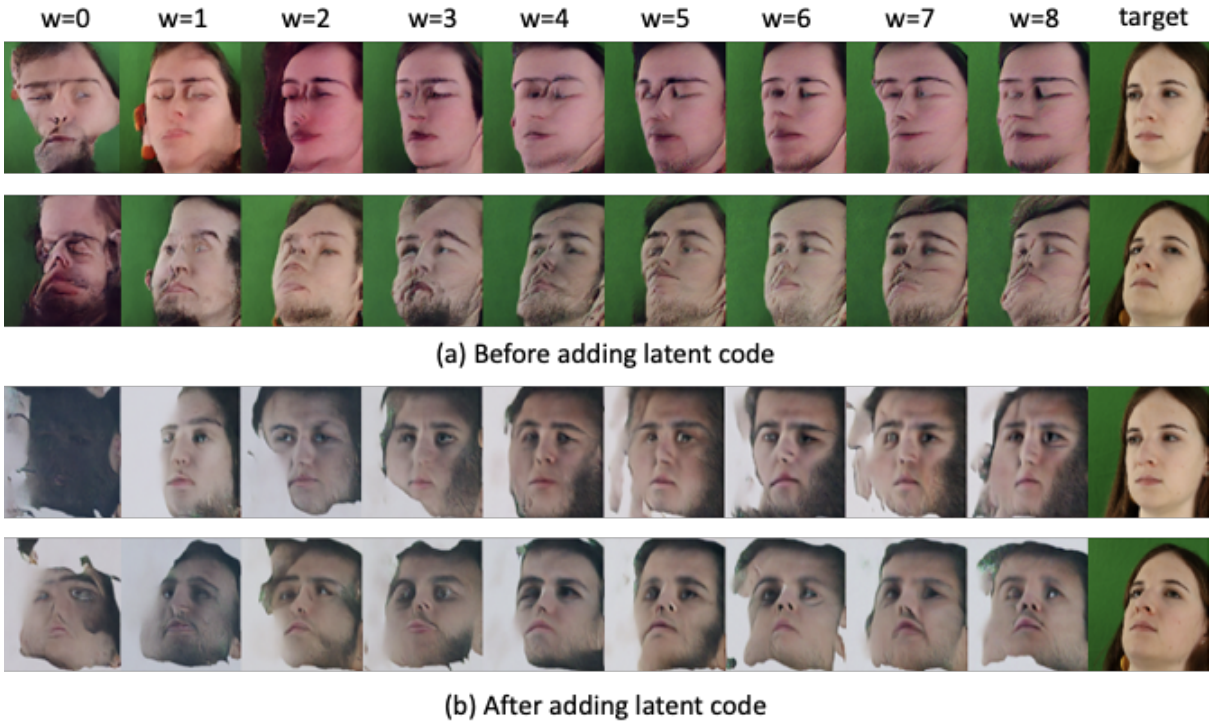


Figure 5.2: Comparison of the results within the incorporation of latent parameters. Respectively, conditions are (R,T, fixed K) in (a) and (R,T, fixed K, latent code) in (b). The labels denoted as w is the hyperparameter that controls the strength of the classifier-free guidance.

resolution of 512 is reflected. One such adjustment is the intrinsic matrix, which was initially defined based on an image size of 512 and distinct camera calibration values. Upon manually modifying the intrinsic matrix to our resized image resolution of 128, we observed improvements in facial structures compared to results before changing the intrinsic matrix. Figure 5.1 shows the comparison of the sampled results.

Second, incorporating latent information about the face images yielded improvements. Figure 5.2 shows the enhancements achieved with the latent codes. A particularly effective method was extending the latent codes into a matrix matching the height and width of the images and appending these to the channel dimension of the input image, which is visualized in Figure 5.3. Additional clamping of the latent codes to the range [-1, 1], consistent with the normalized image values, produced more plausible results than when clamping was not applied. In fact, this concatenation expands the image channels from RGB to RGBA. The alpha channel signifies the opacity level of each pixel, allowing images to be layered over each other through alpha compositing. Thus, we hypothesize that the latent codes embedded in the alpha layer can provide more precise

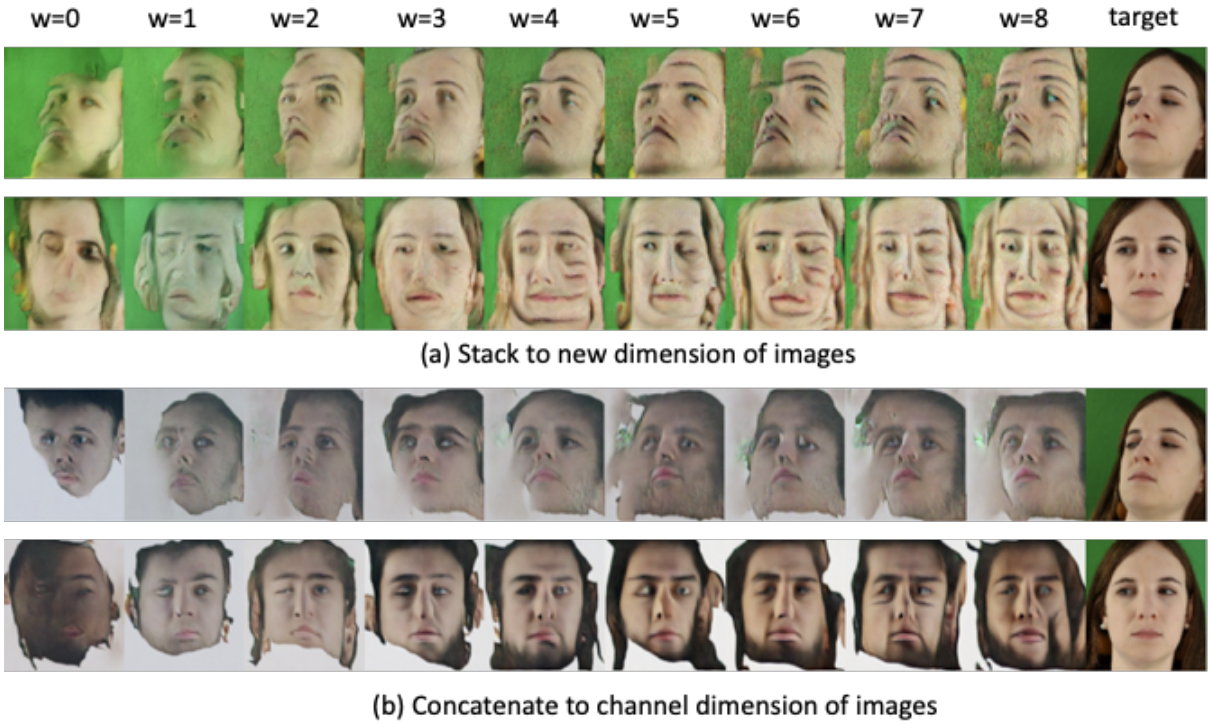


Figure 5.3: Comparison of the results with different methods to integrate the latent code. In both cases, the conditions are (R,T, fixed K, latent code). Results in (a) expands the latent vector $\boldsymbol{l}$ to $\boldsymbol{l} \in \mathbb{R}^{128\times128\times3}$ and stacks to a new dimension. Results in (b) expands the latent vector $\boldsymbol{l}$ to $\boldsymbol{l} \in \mathbb{R}^{128\times128}$ and concatenates to the channel dimension. The labels denoted as w is the hyperparameter that controls the strength of the classifier-free guidance.

control over the representation of complex facial features. The disappearance of the green background in (b) of Figure 5.2 and 5.3 illustrates the effect of integrating the latent codes to the alpha channel. A fully-connected layer was used to map the latent codes into a matrix of the image size. Since the static green background is common to every input image from the dataset and has no feature related to the face, we can assume that it resulted in zero gradients. Consequently, this led to zero opacity, moving out the background color to be fully transparent.

Despite making various adjustments, the outcomes remained unsatisfactory and potentially misleading. Upon deeper examination and empirical observations, we conclude that the main reason for these shortcomings is the mismatch in image resolution settings between the preprocessed dataset and our experimental framework. The preprocessed images were initially normalized for a setting of 512×512 pixels. This normalization process corrects the distortion caused by minor angular disparities during photography,

aligning the images to a common perspective and size for training purposes. The computation involves using 3D landmarks from the images, camera calibration settings (such as focal length and distance to the camera), and the image size. However, some image-specific information is stored in the individual annotation files from the raw ETH-XGaze dataset, making it impracticable to re-normalize the images to fit our adjusted resolution of 128. Thus, this necessitates a change to a dataset where direct access to raw data is possible and/or the image resolution is consistent with our setup.

5.1.2 Gaze Capture

GazeCapture [KKK+16] is selected as the primary dataset for our training and evaluation, replacing ETH-XGaze. It is a large-scale eye-tracking dataset used in wide range of gaze related research, featuring utmost 2.5 million images from more than 1,450 participants. Data was collected in unconstrained settings, through a mobile app that utilizes the front-facing camera to take selfies, resulting in significant variations in illumination, head pose, appearance, and background. This diversity in participants and environmental conditions make it highly applicable to real-world scenarios, allowing the learning of robust models that generalize well to novel faces.

We preprocess the GazeCapture dataset analogously to STED [ZPZ+20]. That is, we use the code provided by Park et al. [PMM+19] with additional changes made by Zheng et al. [ZPZ+20]. Here, the face images are normalized in 128×128 pixels, along with camera callibration settings, which fits our image resolution settings to accommodate GPU capacity. One preprocessed sample includes a normalized image, normalized gaze direction, normalized head pose, and normalized gaze directions for both the left and right eyes. Additionally, we extract latent facial codes from the images, along with gaze direction and head orientation, using an external estimator.

5.2 Implementation Details

5.2.1 Training

We adhere our training configuration to that of 3DiM [WCM+22], as their proposed method in combination with their novel architecture have demonstrated high 3D consistency and scalability across numerous scenes, without relying on hyper-networks or test-time optimization. The training subset of GazeCapture dataset is used for the training phase, which has 1,379,083 samples. To enable classifier-free guidance [HS22], each batch element is trained with 10% of unconditional example. This is done by

defining a boolean matrix as a conditioning mask and overriding the conditioning frame with a random Gaussian noise (maximum noise level) at 10% of time. The model is optimized using the Adam optimizer [KB17] with β_1 = 0.9 and β_2 = 0.99. For the noise schedule λ_t, a cosine-shaped log signal to noise ratio is used, monotonically decreasing from λ_{max} = 20 to λ_{min} = -20.

$$\theta_{\text{max}} = \arctan\left(\exp\left(-0.5 * \lambda_{\text{max}}\right)\right) \tag{5.1}$$

$$\theta_{\text{min}} = \arctan\left(\exp\left(-0.5 * \lambda_{\text{min}}\right)\right) - \theta_{\text{max}} \tag{5.2}$$

$$\lambda_t = -2\log\left(\tan\left(t * \theta_{\text{max}} + \theta_{\text{min}}\right)\right) \tag{5.3}$$

A batch size of maximum 4 fits within a single Tesla V100-SXM2-32GB GPU. Using distributed data parallelism with batch size of 16 (4 GPUs), it takes around 1 full day with timesteps of 500, and 1.5 days with timesteps of 1000 to train a single epoch. To generate samples that are at least qualitatively plausible, we found that a minimum of 7 epochs are needed in both cases. In this limited situation, conducting extensive experiments to empirically determine the optimal coefficients was impractical. As an alternative, we focused on monitoring individual losses while keeping a fixed total loss. We prioritized penalizing the noise loss most heavily, as it serves as the primary guide for model learning, and treated the remaining losses equally:

$$\mathcal{L} = \frac{\mathcal{L}_{\text{noise}}}{\mathcal{L}_{\text{noise}}.detach()} + 0.5\frac{\mathcal{L}_{\text{gaze}}}{\mathcal{L}_{\text{gaze}}.detach()} + 0.5\frac{\mathcal{L}_{\text{head}}}{\mathcal{L}_{\text{head}}.detatch()} \tag{5.4}$$

5.2.2 Sampling

As the input images are normalized to the range $[-1, 1]$, we clip each predicted $\hat{x}$ at each denoising step to the this range. We have incorporated classifier-free guidance [HS22] and varied the hyperparameter w in our sampling process to control the strength of the conditioning. The predicted noise for the sampling phase is computed as a combination of conditional and unconditional noise, as outlined in Equation 4.13. Empirical testing revealed that values of w above 6 tend to introduce more noisy artifacts. Consequently, we have restricted the range of w to $[0, 1, 2, 3, 4, 5, 6]$ to maintain the quality of the generated outputs. While 3DiM deploys denoising steps of 256 under the training timesteps of 1000, we use 128 steps to generate the samples, in our reduced training steps of 500. Under this situation, we empirically found that the qualitative results between 128 and 256 steps does not differ significantly.

5.2.3 Gaze direction and Head orientation estimator

To enforce a more task-relevant loss for training, we adopt an external gaze direction and head orientation estimator from STED [ZPZ+20]. When provided with a full-face image of 128×128 pixels, this estimator outputs a 4-dimensional vector, representing pitch and yaw values in spherical coordinates of the head system. During the training phase, we use an estimator based on the VGG-16 architecture, while another estimator network based on ResNet50 is employed for evaluation. They are both ImageNet pre-trained models, fine-tuned on gaze and head orientation estimation tasks. STED fine-tunes both of the estimator on the same training subset of the GazeCapture dataset for 100k iterations with a batch size of 64, using the Adam optimizer with momentum values β_1 = 0.9, β_2 = 0.95. The initial learning rate is 10^{-4} and is decayed by a factor of 0.5 after 50k iterations. Table 5.1 and 5.2 shows the architecture of the estimators.

Table 5.1: Architecture of the external gaze direction and head orientation estimation network based on VGG-16 used during training phase.

Nr.	layers / blocks
0	VGG-16 convolutional layers
1	FC(512, 64, w/bias), LeakyReLU()
2	FC(64, 64, w/bias), LeakyReLU()
3	FC(64,4, w/bias), 0.5 $\pi\cdot$ Tanh()

Table 5.2: Architecture of the external gaze direction and head orientation estimation network based on ResNet50 used during evaluation phase.

Nr.	layers / blocks
0	ResNet convolutional layers, stride of MaxPool 2d=1
1	FC(2048, 4, w/bias)

5.2.4 Evaluation metrics

We evaluate our model with 4 different metrics. The redirection error specifically measures the accuracy of gaze and head direction, targeting the model's effectiveness in this particular task. The other three metrics—LPIPS, SSIM, and PSNR—evaluate the image quality between the ground truth and generated images by the model.

Redirection Error

To verify the task-explicit performance of our approach, we report the gaze and head redirection error as angular errors using the external estimator outlined in section 5.2.3. As the estimator used for evaluation differs from the one employed during training, it ensures an unbiased means of assessment. The computation is analogous to computing $\mathcal{L}_{\text{gaze}}$ and $\mathcal{L}_{\text{head}}$ during the training phase, in Equations 4.10 and 4.11. We estimate the reverse cosine similarity between the vectors from the predicted target sample and the ground truth target sample.

LPIPS

Learned Perceptual Image Patch Similarity (LPIPS) [ZIE+18] is a commonly used metric in image generation models, as it measures the perceptual similarity in a way that human vision would interpret them. It uses high-level features that are extracted from pre-trained convolutional neural networks (CNNs), to compare images at a feature level rather than just pixel by pixel. This allows LPIPS to capture complex patterns and textures, closely mirroring human visual perception. Lower values signify that generated image and ground truth image are perceptualy similar. Given two images I_1 and I_2, and a set of layers L from a pretrained network, with the learned weights w_l it is computed as the squared Euclidean distance between the normalized feature maps $\hat{\phi}_l(I)$ from the corresponding layers of each image:

$$\text{LPIPS}(I_1, I_2) = \sum_{l \in L} w_l \cdot \|\hat{\phi}_l(I_1) - \hat{\phi}_l(I_2)\|^2 \tag{5.5}$$

SSIM

Structural Similarity Index Measure (SSIM) [WBSS04] evaluates image similarity by a comprehensive comparison of the brightness, contrast, and structural information between the images. It is an approach that aims to mimic human visual perception, similar to LPIPS. Higher SSIM values denote greater accuracy in maintaining the structural integrity of the original image. SSIM is defined as:

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{5.6}$$

μ_x and μ_y are the mean intensities, σ_x^2 and σ_y^2 are the variances of images x and y, and σ_{xy} is the covariance of x and y. C_1 and C_2 are constants to stabilize the division with weak denominator, to avoid the case when the denominator becomes 0.

PSNR

Peak Signal-to-Noise Ratio (PSNR) is a traditional metric used to assess the pixel-wise quality of reconstructed images. It is commonly used in image compression, video compression, and other fields where maintaining image quality is essential. Although

PSNR may not directly correlate with human visual perception, it can serve as an effective tool for monitoring the quality of image reconstruction from our generative model across different timesteps. Higher PSNR values indicate a higher quality of reconstruction, showing that the generated image retains the details from the original. The computation follows:

$$\text{PSNR} = 10 \log_{10} \left(\frac{\text{MAX}^2}{\text{MSE}} \right) \tag{5.7}$$

$$\text{MSE} = \frac{1}{mn} \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} [I(i,j) - K(i,j)]^2 \tag{5.8}$$

MAX is the maximum possible pixel value of the image. MSE is the Mean Squared Error between the images where m and n are the dimensions of the images, and $I(i,j)$ and $K(i,j)$ are the pixel values at position (i,j), respectively.

5.3 Results

During the evaluation phase, we utilize the test subset of the GazeCapture dataset and conduct both quantitative and qualitative analyses to provide a comprehensive overview of our proposed approach. We present results where the conditions are solely pose embeddings from the rotation matrix, and shift the discussion of using the latent facial parameters to Chapter 6. These results were obtained after training the model for 9 epochs with a reduced timestep of 500. For the quantitative evaluation, we employ the STED [ZPZ+20] model as our baseline to compare the performance of our method against a recognized benchmark in the field, shown in Table 5.3.

	Gaze Redir.	Head Redir.	LPIPS	PSNR	SSIM
STED	4.242	1.466	0.260	12.942	0.422
Ours	**7.717**	**14.792**	**0.543**	**9.882**	**0.129**

Table 5.3: Evaluation of our results with STED [ZPZ+20]. Our results are taken from the guidance weight $w = 4$ for the classifier-free guidance approach, as it has the lowest gaze redirection error. Redirection error and LPIPS are better when lower in value, whereas SSIM and PSNR are better when values are higher.

As we adopt the classifier-free guidance to our approach, we provide a comparison between different w values, the weight that controls the influence of conditioning

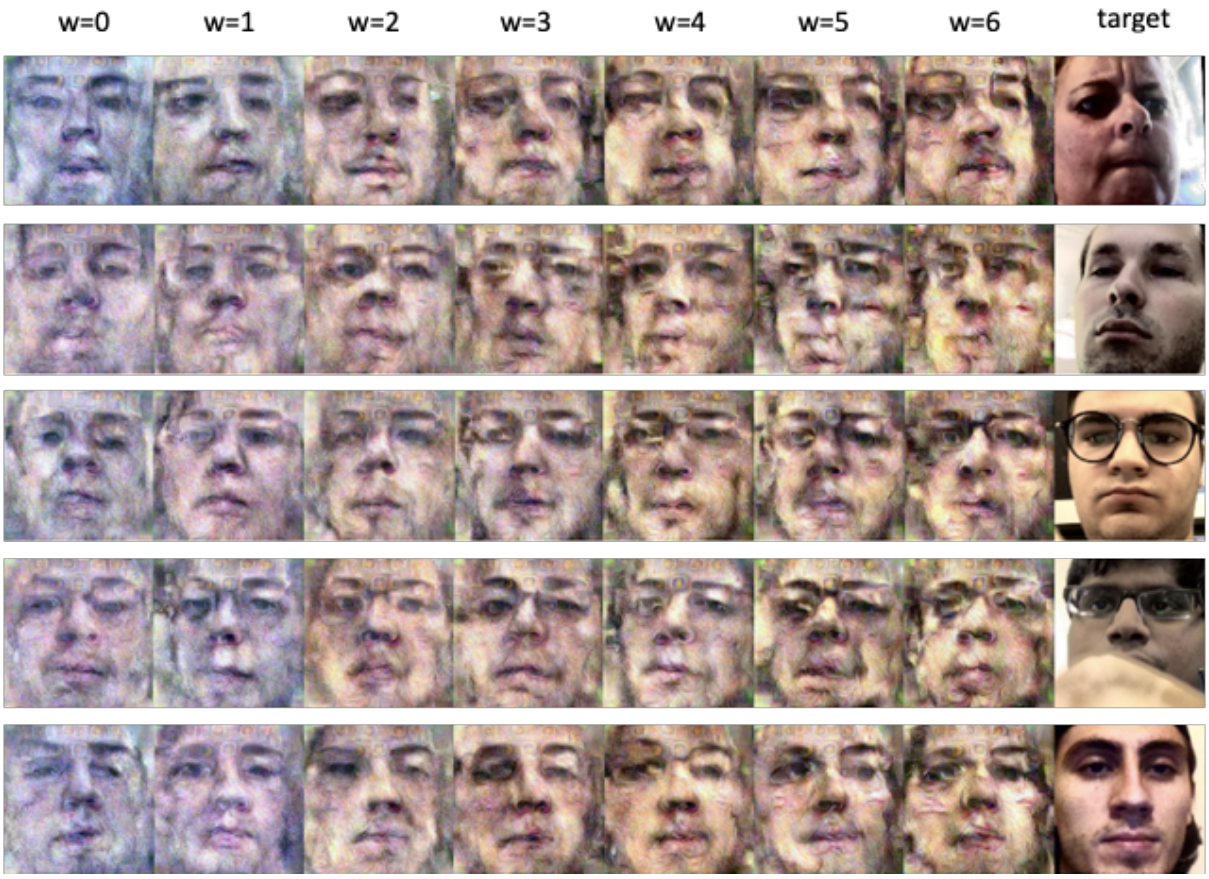


Figure 5.4: Qualitative comparison of different w values that control the strength of the conditioning information. As the weight w increases, the target image is generated more conditionally.

	w=0	w=1	w=2	w=3	w=4	w=5	w=6
Head Redir.	10.979	20.817	17.842	16.923	**12.202**	**12.958**	15.228
Gaze Redir	13.307	18.687	14.869	**7.963**	**8.344**	10.783	10.822

Table 5.4: Quantitative results corresponding to the sample from the **second row** in Figure 5.4. The head redirection direction error is at lowest in weights 4 or 5, while gaze redirection direction error is lowest in weights 3 or 4. This clarifies our selection of the weight $w = 4$ to be the best sample results.

information during the generation process. When w is set to a higher value, the model emphasizes the conditioning information more strongly, making the output more closely aligned with the specific conditions. Conversely, a lower w value makes the model behave more like an unconditioned generative model, where the output is less dependent on

the conditions. These effects are illustrated in Figure 5.4. It is noticeable that from the weight $w = 2$, the generated sample begins to adopt the color range of the target image. Subsequently, when the weight reaches $w = 3$, the sample starts to more closely match the head pose alignment of the target image, evidenced quantitatively in Table 5.4. With increasing weight, the adaptation to the target's characteristics becomes more apparent. Nonetheless, excessively high weights tend to introduce noisy artifacts, suggesting that there is a threshold beyond which additional increases in weight can negatively impact image quality.

The sample results from the third and fourth rows indicate that the model struggles to capture task-irrelevant details, such as eyeglasses. However, it is also noteworthy that in the fourth row, despite the presence of obstacles that partially cover the face, the model is able to predict the covered parts of the face using its learned knowledge. This demonstrates the model's capability to infer and reconstruct occluded facial features.

We hypothesize that the presence of noisy artifacts in the results may be attributed to the reduced number of timesteps during training. Initially, the standard DDPM model recommends 1000 timesteps, which we reduced to 500 in an effort to accelerate the training process. This adjustment likely compromised the model's ability to fully learn the data's complexities, resulting in less refined outputs.

Table 5.5 presents the quantitative results with varying weights for w. We selected results from the guidance weight $w = 4$ for detailed comparison with STED, as this yielded the lowest gaze redirection error, aligning closely with the main objective of

	Gaze Redir.	Head Redir.	LPIPS	PSNR	SSIM
w=0	9.370	16.413	0.692	9.146	0.111
w=1	8.537	15.143	0.654	9.457	0.123
w=2	7.933	14.709	0.603	9.740	0.128
w=3	7.850	15.549	0.567	9.877	0.128
w=4	**7.717**	**14.792**	**0.543**	**9.882**	**0.129**
w=5	8.224	15.687	0.541	9.826	0.122
w=6	8.351	15.888	0.535	9.757	0.118

Table 5.5: Evaluation with different w values that control the strength of the conditioning information. The respective metrics in each row is the averaged value across the entire test samples. Redirection error and LPIPS are better when lower in value, whereas SSIM and PSNR are better when values are higher.

our task. We observe that the LPIPS score decreases as w increases, reinforcing that integrating more conditioning information into the generation process tends to produce samples that resemble more closely to the target image, in a way that human perceives the image. Conversely, PSNR and SSIM scores hover at similar range across different w values. When comparing these metrics with those from STED, it shows that our model tends to produce images of fundamentally lower quality in terms of standard image quality metrics.

6 Discussion&Limitations

To integrate 3D awareness into the gaze redirection task, we developed three distinct approaches focusing on model architecture, 3D gaze rotation, and latent facial parameters. In this chapter, we will thoroughly analyze and discuss the effectiveness and limitations of each proposed method, providing a comprehensive overview of their respective contributions to the task. Moving forward, we will outline future research directions to address the limitations and demonstrate the significant potential of our theoretic approach for performing 3D-aware gaze redirection as a conditional image generation task.

6.1 Latent Facial Parameters

In Chapter 5, we conducted a series of experiments to extensively explore and demonstrate the two specific goals we aim to achieve. Initially, we investigated the effectiveness of pose embeddings derived from gaze rotation matrices. However, we have yet to address the impact of latent facial parameters within the GazeCapture [KKK+16] dataset. In this section, we will provide a detailed summary of these findings and engage in a more in-depth discussion about the role and influence of latent parameters within our study.

In our experiments with the ETH-XGaze [ZPB+20] dataset, we found that the incorporation of latent codes significantly enhanced outcomes, particularly in terms of achieving more precise facial structures, as illustrated in Figure 5.2. Concatenating the latent code to the channel dimension produced better results compared to stacking it into a new dimension. Based on these empirical findings, we apply this approach to the GazeCapture dataset. We transformed the latent code $\boldsymbol{l} \in \mathbb{R}^{233}$ from the DECA encoder into $\boldsymbol{l} \in \mathbb{R}^{1\times128\times128}$, and concatenated it along the channel dimension of the input image. The quantitative results, presented in Table 6.1, show a notable decrease in gaze redirection error compared to the results where only pose embeddings were used as conditions. Remarkably, the gaze redirection error was comparable to the STED model at weight $w = 4$.

	Gaze Redir.	Head Redir.	LPIPS	PSNR	SSIM
pose embeddings	7.717	14.792	0.543	9.882	0.129
pose+latent code	**4.250**	**16.034**	**0.824**	**9.885**	**0.081**
STED	4.242	1.466	0.260	12.942	0.422

Table 6.1: Comparison of results using pose embeddings alone, pose embeddings with latent code, and STED model. As the weight $w = 4$ resulted in the lowest gaze redirection error in both cases, we present this specific result for more practical comparison.

Figure 6.1: Results shown for conditions that include pose embeddings and latent codes. Alongside the quantitative results, we specifically present cases where the weight is set at $w = 4$.

With the latent vector encompassing information on shape, expression, texture, pose, and illumination, improvements in facial expression, apparent colors, and distinct facial features are evident in Figure 6.1. This contrasts with the results shown in Figure 5.4, which relied solely on pose embeddings. These outcomes confirm that incorporating latent parameters enhances the model's capacity to interpret and reproduce facial features, compared to using pose embeddings alone.

Despite these partial improvements, the generated images were qualitatively misleading and diverged significantly from the ground truth images, contradicting our initial expectations. This is reflected in the increased LPIPS and the decreased SSIM scores, as both metrics assess perceptual similarity in a manner consistent with human visual interpretation. We had hypothesized that embedding the latent codes into the alpha channel of

the input mage would allow for more sophisticated control over the representation of complex facial features through adjustments in opacity. Unfortunately, this approach did not yield the anticipated subtle controls, leading to outcomes that did not align with our hypothesized effectiveness of this technique.

We attribute the qualitative discrepancies in the generated images to significant differences in image quality between the two datasets. The ETH-XGaze dataset was collected in a controlled lab environment, resulting in very high-quality images with uniform backgrounds and clear facial outlines, which make it easy to distinguish human subjects from the background. Additionally, the dataset lacks complicating factors that could obscure facial features, such as obstacles, blurry images, or extreme lighting conditions. This allows the model to focus on learning the distribution of the human face.

In contrast, the GazeCapture dataset consists of selfies taken in uncontrolled environments, exhibiting severe challenges. These include obstacles partially covering faces, extreme illumination conditions (e.g., colored lights or high contrast from flashes), blurriness, and some images being cropped such that edges are depicted as black. These factors lead to significant degradation in image quality and considerable variation in the backgrounds. Such variability complicates the model's learning of the data distribution of the human figure, as it also needs to discern between the main subject and a noisy, inconsistent background. Furthermore, the original 3DiM model, which we adapted for our task, was trained on the SRN ShapeNet[SZW19] dataset, characterized by its clean white backgrounds and centrally positioned objects. This consistency in background settings supports our hypothesis that the variability in the GazeCapture dataset introduces significant challenges to the learning process, impacting the performance of the 3DiM model.

Due to the constrained timeframe and the restricted number of GPUs available, we were unable to conduct additional experiments involving the latent code, as further training for a single adjustment would require at least a week. Nevertheless, our extensive experiments with the ETH-XGaze dataset, supplemented by partial experiments with the GazeCapture dataset, provided valuable insights. Specifically, we verified that incorporating latent facial parameters enhances the model's structural understanding of the human face. These latent parameters help capture the intricate geometric and texture details, thereby improving the model's ability to generate more accurate face images with apparent gaze direction. While our findings indicate a positive impact, we recognize the need for further experiments to fully exploit the potential of latent code integration in enhancing model performance.

6.2 3D Gaze Rotation

The viewpoint-conditioned diffusion model we have adopted from the novel view synthesis task relies on viewpoints, specifically the 3D transformations for creating the pose embeddings. In the original 3DiM approach, pose embeddings are created within their ray construction methodology. Specifically, the global 3D rotation and translation matrices, which define the object's offset relative to the camera, along with the intrinsic parameters of the camera setup, are utilized to construct a ray. This concept was adopted from a previous study by Sajjadi et al. [SMP+22], which is a geometry-free approach (i.e., methods without explicit geometric assumptions compared to those used in volume rendering) that achieved competitive results in novel view synthesis. The ray construction process takes the relative rotation and translation between two views, generating a camera ray that captures the 3D scene from these perspectives. The ray's origin and direction vectors are then aggregated and transformed into positional-pose encodings, which serve as the final pose embeddings.

To leverage their structured approach, we initially utilized the preprocessed ETH-XGaze dataset [ZPB+20], where the rotation matrix (R), translation vector (T), and camera intrinsic matrix (K) were readily available. However, the 512-pixel settings in the preprocessed dataset didn't match our GPU capacity and simply resizing the images to our setting of 128 pixels resulted in misleading outcomes. It necessitated access to the raw ETH-XGaze dataset for further adjustments, which was unavailable. Consequently, we switched to using the GazeCapture dataset [KKK+16], which has images naturally sized at 128.

Since the GazeCapture dataset does not include the T and K parameters, we have manually set T to a zero vector to denote no translation, and K to an identity matrix, following the original authors [SMP+22] setting the intrinsic matrix to identity when the information is not available. However, our empirical findings indicate that this approach introduced irrelevant information, adding unnecessary noise during training. This was evident as the noise loss failed to converge below 1.3721 after 7 epochs for one week of training, whereas ideally, it should fall below 0.1 to generate outputs that are distinctly non-noisy. Consequently, we replaced the ray constructing framework with a fully connected layer. To further leverage the cross-attention mechanism suggested by the 3DiM model, we stacked the rotation embeddings from the fully-connected layer in a similar manner to how images are stacked in the 3DiM framework. Moreover, to enhance the model's performance in task-specific learning, we introduced a novel gaze and head loss. These efforts successfully guided the model toward better convergence in noise loss to produce non-noisy images and improved performance in gaze redirection, as evidenced by our quantitative results. Nevertheless, the qualitative outcomes concerning the redirected gaze and head images do not meet the expected standards.

We hypothesize that the discrepancy between our method of deriving pose embeddings and the original technique used in 3DiM is contributing to the generation of qualitatively inferior images compared to the ground truth. Concerning that the 3DiM model is a novel architecture proven effective only when all components are correctly aligned with their initial settings, our modified method of embedding pose conditions has adversely impacted the model's understanding of spatial representation, resulting in images that are qualitatively less intuitive and lack 3D-consistency.

6.3 3D-Aware Conditional Diffusion Model

To integrate a 3D-aware model into our work, we adopted the viewpoint-conditioned diffusion model designed for novel view synthesis tasks. Specifically, we utilized the 3DiM [WCM+22] model, which is a novel architecture that understands 3D geometry without relying on any pre-trained 3D priors. Since the model learns from scratch, it is light-weight compared to other state-of-the-art approaches, making it more accomplishable to implement. Their proposed cross-attention and weight-sharing mechanism exploit symmetries between frames and poses, significantly aiding the alignment with the content of the conditioning view and resulting in highly consistent 3D outputs.

Although our methodological approach, based on this model, has achieved some success in conducting gaze redirection, there are inherent limitations in the image quality of the outputs. Upon closer examination, we hypothesize that the fundamental differences between the original intent of the 3DiM model and its adaptation for our specific task have led to a degradation in performance.

The 3DiM model was originally developed for the novel view synthesis and 3D reconstruction of general objects, having less complexity compared to our task that target gaze redirection for human faces. Human faces present a significantly higher level of complexity due to their exclusive person-specific features, which are far more intricate than static objects that exhibit predictable symmetries, colors, and shapes. In general, objects have limited degrees of freedom and can be broadly categorized into single classes making them somewhat predictable and static. For example, objects categorized as "chairs" have a common predictable structure, e.g., legs supporting a seat. While details can differ in terms of the number of legs, the shape of the seat, or whether it has back support, human faces exhibit greater unique characteristics and expressions. Even basic facial structures like eyes, noses, mouths, and ears are combined with subtle details such as eyelashes, eyebrows, and the details of skin features (e.g., wrinkles, freckles, moles). These are further complicated by dynamic facial expressions, ranging from subtle to highly distinct, making each human face uniquely complex. This fundamental difference in tasks means that the unique architecture of the 3DiM model, which excels

with general 3D objects, might not be as effective for the highly variable and detailed domain of human faces. Moreover, the original 3DiM model's performance is assessed using the ShapeNet dataset, both for training and evaluation, without cross-dataset testing. Despite it demonstrates examples using additional image sources from the internet, the image must correspond to specific ShapeNet classes, requiring white backgrounds and minimal shadows. Consequently, the model's performance on noisy, real-world images remains unexplored, highlighting a potential gap when adapting it to more dynamic and complex scenarios such as human faces.

6.4 Future Works

6.4.1 Modification of pose embeddings

In the original 3DiM methodology, pose embeddings are generated using the global rotation and translation matrices extracted from pairs of images. These matrices are then used to construct rays based on the relative rotations and translations between the image pairs. In contrast, our approach modifies this method by stacking the rotation matrices, as the way the images are processed.

To further refine this technique, we propose using the relative rotation between the two matrices as $\boldsymbol{R} = \boldsymbol{R}_{\text{target}} \cdot \boldsymbol{R}_{\text{input}}^{-1}$, subsequently transforming these into rotation embeddings. This adjustment aims to shift the learning focus of the fully-connected layer from absolute rotations to relative rotation embeddings. By doing so, the model can better understand and adapt to the relative differences in gaze direction between various images. This methodology not only aligns with our ideal objective of performing gaze redirection on any input image, regardless of whether it was included in the training set, but also facilitates redirection in scenarios where explicit pose data is unavailable.

6.4.2 Loss functions for latent facial parameters

In our experiments, directly associating latent codes with images resulted in discrepancies between the generated and target images. To address this, we propose utilizing latent parameters within a loss function framework instead.

Our empirical experiments demonstrated that incorporating task-specific gaze and head losses significantly improved outcomes within the GazeCapture dataset, compared to the ETH-XGaze dataset where only noise loss was employed. Despite lower image quality in GazeCapture, we observed that the gaze in the generated images closely resembled that

of the target images to a certain degree. This underscores the importance of targeted loss functions in steering model learning towards specific objectives.

Following the methodology of DECA [FFBB21], we can incorporate similar loss functions from their work. Given our existing algorithm 4.7 that reproduces the target image using the predicted noise, integrating additional losses by using this recovered image as input to the DECA model is feasible. We noticed these losses adaptable to our approach:

- Identity Loss: This measures the cosine similarity between the embeddings of the generated and target images. It encourages the model to capture essential attributes of a person's identity, ensuring that the rendered image resembles the input subject.
- Shape Consistency Loss: The latent parameters from the DECA encoder include data on shape, expression, texture, pose, and illumination settings. The shape parameter, in particular, should remain consistent if the images are from the same subject. DECA's approach involves swapping the shape parameter between the generated and ground truth images and further using this in their loss calculations. The intuition is that if the model has accurately estimated the face shape, then swapping these parameters between images of the same person should yield indistinguishable results. In our work, we can simply employ the MSE loss between the ground truth and the predicted shape parameters to enforce shape consistency.
- Landmark Re-projection Loss: DECA's framework also includes a decoder that uses the shape, expression, and pose parameters to predict 2D landmarks, 3D landmarks, and vertices for 3D face reconstruction. The landmark loss evaluates the discrepancy between the ground-truth 2D face landmarks and those projected onto the face image by an estimated camera model. Implementing this loss in our model would facilitate the learning of the face's 3D structure, enhancing the accuracy of our gaze redirection. Figure 6.2 illustrates the predicted landmarks from the DECA decoder to our GazeCapture dataset. It is noticeable that DECA, with its robust 3D priors of human facial geometry, predicts landmarks that generally align well with the facial features, even on unseen datasets. This confirms our decision of incorporating DECA to extract latent facial parameters.

By adopting these advanced loss functions, we aim to refine our model's ability to more accurately interpret and reconstruct complex facial features, ultimately enhancing the quality of the gaze redirection.

Figure 6.2: Predicted landmarks from the DECA [FFBB21] decoder. First column is the original input image, second column is the predicted 2D landmarks, third column is the predicted 3D landmarks projected to the input image.

6.4.3 Adopting pre-trained models with high 3D priors

Recognizing the limitations of the 3DiM [WCM+22] model, we explored alternative viewpoint-conditioned diffusion models to enhance our approach. Notably, Zero-1-to-3 [LWH+23] is another state-of-the-art method in novel view synthesis, leveraging the capabilities of large-scale pre-trained diffusion models. In particular, it utilizes Stable Diffusion [RBL+22], a groundbreaking model in the field of AI image generation. Stable Diffusion is a text-to-image model trained on billions of images, leading to a robust understanding of the 3D properties of the physical world. By fine-tuning on Stable Diffusion, we can benefit from their pre-trained prior knowledge of 3D geometrical information and better fulfill our gaze redirection as 3D-aware image generation.

7 Conclusion

This thesis began by identifying the limitations of existing gaze redirection approaches and aimed to address these limitations with a fully 3D-aware approach. We redefined the task of gaze redirection as a generative approach using conditional diffusion models, capitalizing on their well-documented strengths in producing detailed and diverse images. Specifically, we employed viewpoint-conditioned diffusion models that have previously been utilized in novel view synthesis tasks and can leverage 3D transformations as conditions, conducting image-to-image translations.

Taking advantage of this model, we introduced explicit 3D gaze rotations derived from gaze labels to enable the model to simulate the complex interactions between gaze directions across different images. Additionally, we extracted latent facial parameters using an existing framework designed for 3D face reconstruction from 2D images, which provides a deeper understanding of facial geometry. This was crucial as human faces present a higher level of complexity than standard 3D objects, requiring rich, feature-specific information to enhance model performance.

Through extensive experiments with our proposed method on both the ETH-XGaze and GazeCapture datasets, we demonstrated that using 3D transformations as conditions effectively addresses the task of gaze redirection. Furthermore, the incorporation of latent facial codes enhances the model's ability to understand and manipulate the 3D structure of the human face, resulting in images with improved clarity and accuracy of facial structures. However, acknowledging the inherent limitations of our proposed methods, we also suggest practical enhancements that could further refine this approach. In conclusion, our conceptual idea of interpreting the gaze redirection task as a 3D-aware conditional image generation task has proven to be valid and shows substantial potential for further refinement and development.

Bibliography

[ACC+21] M. Alcañiz Raya, I. Chicchi Giglioli, L. Carrasco Ribelles, J. Marín-Morales, M. E. Minissi, G. Teruel-García, M. Sirera, L. Abad. "Eye gaze as a biomarker in the recognition of autism spectrum disorder using virtual reality and machine learning: A proof of concept for diagnosis." In: *Autism Research* 15 (Nov. 2021). DOI: 10.1002/aur.2636 (cit. on p. 14).

[BSP+15] M. Borgestig, J. Sandqvist, R. Parsons, T. Falkmer, H. Hemmingsson. "Eye Gaze Performance for Children with Severe Physical Impairments Using Gaze-Based Assistive Technology-a Longitudinal Study." In: *Assistive technology : the official journal of RESNA* 28 (Oct. 2015). DOI: 10.1080/10400435.2015.1092182 (cit. on p. 14).

[CNC+23] E. R. Chan, K. Nagano, M. A. Chan, A. W. Bergman, J. J. Park, A. Levy, M. Aittala, S. D. Mello, T. Karras, G. Wetzstein. "GeNVS: Generative Novel View Synthesis with 3D-Aware Diffusion Models." In: *arXiv*. 2023 (cit. on p. 21).

[CST+22] L. Carelli, F. Solca, S. Tagini, S. Torre, F. Verde, N. Ticozzi, R. Ferrucci, G. Pravettoni, E. Aiello, V. Silani, B. Poletti. "Gaze-Contingent Eye-Tracking Training in Brain Disorders: A Systematic Review." In: *Brain Sciences* 12 (July 2022), p. 931. DOI: 10.3390/brainsci12070931 (cit. on p. 14).

[CZZ+20] N. Chen, Y. Zhang, H. Zen, R. J. Weiss, M. Norouzi, W. Chan. *WaveGrad: Estimating Gradients for Waveform Generation*. 2020. arXiv: 2009.00713 [eess.AS] (cit. on p. 25).

[DE16] Y. Deldjoo, R. Ebrahimi Atani. "A Low-Cost Infrared-Optical Head Tracking Solution for Virtual 3D Audio Environment Using the Nintendo Wii-Remote." In: *Entertainment Computing* 12 (Jan. 2016). DOI: 10.1016/j.entcom.2015.10.005 (cit. on p. 15).

[DN21] P. Dhariwal, A. Nichol. *Diffusion Models Beat GANs on Image Synthesis*. 2021. arXiv: 2105.05233 [cs.LG] (cit. on p. 27).

[FFBB21] Y. Feng, H. Feng, M. J. Black, T. Bolkart. *Learning an Animatable Detailed 3D Face Model from In-The-Wild Images*. 2021. arXiv: 2012.04012 [cs.CV] (cit. on pp. 32, 57, 58).

[GKSL16] Y. Ganin, D. Kononenko, D. Sungatullina, V. Lempitsky. *DeepWarp: Photorealistic Image Resynthesis for Gaze Manipulation*. 2016. arXiv: 1607.07215 [cs.CV] (cit. on p. 17).

[HJA20] J. Ho, A. Jain, P. Abbeel. *Denoising Diffusion Probabilistic Models*. 2020. arXiv: 2006.11239 [cs.LG] (cit. on pp. 23–26, 30, 33).

[HKW11] G. Hinton, A. Krizhevsky, S. Wang. "Transforming Auto-Encoders." In: vol. 6791. June 2011, pp. 44–51. ISBN: 978-3-642-21734-0. DOI: 10.1007/978-3-642-21735-7_6 (cit. on p. 18).

[HS22] J. Ho, T. Salimans. *Classifier-Free Diffusion Guidance*. 2022. arXiv: 2207.12598 [cs.LG] (cit. on pp. 27, 36, 43, 44).

[HSZH19] Z. He, A. Spurr, X. Zhang, O. Hilliges. *Photo-Realistic Monocular Gaze Redirection Using Generative Adversarial Networks*. 2019. arXiv: 1903.12530 [cs.CV] (cit. on p. 18).

[JTA21] A. Jain, M. Tancik, P. Abbeel. *Putting NeRF on a Diet: Semantically Consistent Few-Shot View Synthesis*. 2021. arXiv: 2104.00677 [cs.CV] (cit. on p. 20).

[KAY08] E. KAYYAM. "The Eyeball: a complete coordinate system for location and time." In: *Philica)* (2008), p. 7 (cit. on p. 15).

[KB17] D. P. Kingma, J. Ba. *Adam: A Method for Stochastic Optimization*. 2017. arXiv: 1412.6980 [cs.LG] (cit. on p. 44).

[KKK+16] K. Krafka, A. Khosla, P. Kellnhofer, H. Kannan, S. Bhandarkar, W. Matusik, A. Torralba. "Eye Tracking for Everyone." In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016 (cit. on pp. 43, 51, 54).

[Kle86] C. Kleinke. "Gaze and Eye Contact. A Research Review." In: *Psychological bulletin* 100 (July 1986), pp. 78–100. DOI: 10.1037/0033-2909.100.1.78 (cit. on p. 13).

[KW13] D. P. Kingma, M. Welling. "Auto-Encoding Variational Bayes." In: *CoRR* abs/1312.6114 (2013). URL: https://api.semanticscholar.org/CorpusID:216078090 (cit. on pp. 25, 33).

[LLW+23] L. Li, Q. Lian, L. Wang, N. Ma, Y.-C. Chen. *Lift3D: Synthesize 3D Training Data by Lifting 2D GAN to 3D Generative Radiance Field*. 2023. arXiv: 2304.03526 [cs.CV] (cit. on p. 21).

[LSP+23] D. Lee, Y. Shin, R. W. Park, S.-M. Cho, S. Han, C. Yoon, J. Choo, J. Shim, K. Kim, S.-W. Jeon, S.-J. Kim. "Use of eye tracking to improve the identification of attention-deficit/hyperactivity disorder in children." In: *Scientific Reports* 13 (Sept. 2023). DOI: 10.1038/s41598-023-41654-9 (cit. on p. 14).

[LWH+23] R. Liu, R. Wu, B. V. Hoorick, P. Tokmakov, S. Zakharov, C. Vondrick. *Zero-1-to-3: Zero-shot One Image to 3D Object*. 2023. arXiv: 2303.11328 [cs.CV] (cit. on pp. 21, 30, 58).

[MCL+21] Q. Meng, A. Chen, H. Luo, M. Wu, H. Su, L. Xu, X. He, J. Yu. *GNeRF: GAN-based Neural Radiance Field without Posed Camera*. 2021. arXiv: 2103.15606 [cs.CV] (cit. on p. 20).

[MRLV23] L. Melas-Kyriazi, C. Rupprecht, I. Laina, A. Vedaldi. *RealFusion: 360*deg $Reconstruction of Any Object from a Single Image$. 2023. arXiv: 2302.10663 [cs.CV] (cit. on p. 30).

[MST+20] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, R. Ng. "NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis." In: *ECCV*. 2020 (cit. on pp. 19, 20, 31).

[MTG+14] A. Moon, D. M. Troniak, B. T. Gleeson, M. K. X. J. Pan, M. Zheng, B. A. Blumer, K. E. Maclean, E. A. Croft. "Meet Me where I'm Gazing: How Shared Attention Gaze Affects Human-Robot Handover Timing." In: *2014 9th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (2014), pp. 334–341. URL: https://api.semanticscholar.org/CorpusID:13576005 (cit. on p. 14).

[Mut09] B. Mutlu. "Designing Gaze Behavior for Humanlike Robots." PhD thesis. Jan. 2009 (cit. on p. 13).

[NBM+21] M. Niemeyer, J. T. Barron, B. Mildenhall, M. S. M. Sajjadi, A. Geiger, N. Radwan. *RegNeRF: Regularizing Neural Radiance Fields for View Synthesis from Sparse Inputs*. 2021. arXiv: 2112.00724 [cs.CV] (cit. on p. 20).

[PMM+19] S. Park, S. D. Mello, P. Molchanov, U. Iqbal, O. Hilliges, J. Kautz. "Few-Shot Adaptive Gaze Estimation." In: *International Conference on Computer Vision (ICCV)*. Seoul, Korea, 2019 (cit. on p. 43).

[RBL+22] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, B. Ommer. *High-Resolution Image Synthesis with Latent Diffusion Models*. 2022. arXiv: 2112.10752 [cs.CV] (cit. on pp. 21, 28, 58).

[RFB15] O. Ronneberger, P. Fischer, T. Brox. *U-Net: Convolutional Networks for Biomedical Image Segmentation*. 2015. arXiv: 1505.04597 [cs.CV] (cit. on p. 26).

[RJ21] C. K. Ramachandra, A. Joseph. "IEyeGASE: An Intelligent Eye Gaze-Based Assessment System for Deeper Insights into Learner Performance." In: *Sensors* 21.20 (2021). ISSN: 1424-8220. DOI: 10.3390/s21206783. URL: https://www.mdpi.com/1424-8220/21/20/6783 (cit. on p. 13).

[RPG+21] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, I. Sutskever. *Zero-Shot Text-to-Image Generation*. 2021. arXiv: 2102.12092 [cs.CV] (cit. on pp. 21, 28).

[RSW+23] A. Ruzzi, X. Shi, X. Wang, G. Li, S. D. Mello, H. J. Chang, X. Zhang, O. Hilliges. *GazeNeRF: 3D-Aware Gaze Redirection with Neural Radiance Fields*. 2023. arXiv: 2212.04823 [cs.CV] (cit. on pp. 15, 19, 31, 39).

[SCS+22] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. Denton, S. K. S. Ghasemipour, B. K. Ayan, S. S. Mahdavi, R. G. Lopes, T. Salimans, J. Ho, D. J. Fleet, M. Norouzi. *Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding*. 2022. arXiv: 2205.11487 [cs.CV] (cit. on pp. 21, 28).

[SCV21] L. Shi, C. Copot, S. Vanlanduit. "GazeEMD: Detecting Visual Intention in Gaze-Based Human-Robot Interaction." In: *Robotics* 10.2 (2021). ISSN: 2218-6581. DOI: 10.3390/robotics10020068. URL: https://www.mdpi.com/2218-6581/10/2/68 (cit. on p. 13).

[SH22] T. Salimans, J. Ho. *Progressive Distillation for Fast Sampling of Diffusion Models*. 2022. arXiv: 2202.00512 [cs.LG] (cit. on pp. 25, 33).

[SHC+21] C. Saharia, J. Ho, W. Chan, T. Salimans, D. J. Fleet, M. Norouzi. *Image Super-Resolution via Iterative Refinement*. 2021. arXiv: 2104.07636 [eess.IV] (cit. on p. 32).

[SMP+22] M. S. M. Sajjadi, H. Meyer, E. Pot, U. Bergmann, K. Greff, N. Radwan, S. Vora, M. Lucic, D. Duckworth, A. Dosovitskiy, J. Uszkoreit, T. Funkhouser, A. Tagliasacchi. *Scene Representation Transformer: Geometry-Free Novel View Synthesis Through Set-Latent Scene Representations*. 2022. arXiv: 2111.13152 [cs.CV] (cit. on p. 54).

[SSK+21] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, B. Poole. *Score-Based Generative Modeling through Stochastic Differential Equations*. 2021. arXiv: 2011.13456 [cs.LG] (cit. on p. 25).

[SWMG15] J. Sohl-Dickstein, E. A. Weiss, N. Maheswaranathan, S. Ganguli. *Deep Unsupervised Learning using Nonequilibrium Thermodynamics*. 2015. arXiv: 1503.03585 [cs.LG] (cit. on pp. 23–25, 30).

[SZW19] V. Sitzmann, M. Zollhöfer, G. Wetzstein. "Scene Representation Networks: Continuous 3D-Structure-Aware Neural Scene Representations." In: *Advances in Neural Information Processing Systems*. 2019 (cit. on p. 53).

[WBM+17] E. Wood, T. Baltrusaitis, L.-P. Morency, P. Robinson, A. Bulling. *GazeDirector: Fully Articulated Eye Gaze Redirection in Video*. 2017. arXiv: 1704.08763 [cs.CV] (cit. on p. 17).

[WBSS04] Z. Wang, A. Bovik, H. Sheikh, E. Simoncelli. “Image quality assessment: from error visibility to structural similarity.” In: *IEEE Transactions on Image Processing* 13.4 (2004), pp. 600–612. DOI: 10.1109/TIP.2003.819861 (cit. on p. 46).

[WCM+22] D. Watson, W. Chan, R. Martin-Brualla, J. Ho, A. Tagliasacchi, M. Norouzi. *Novel View Synthesis with Diffusion Models*. 2022. arXiv: 2210.04628 [cs.CV] (cit. on pp. 21, 30, 31, 43, 55, 58).

[ZIE+18] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, O. Wang. *The Unreasonable Effectiveness of Deep Features as a Perceptual Metric*. 2018. arXiv: 1801.03924 [cs.CV] (cit. on p. 46).

[ZPB+20] X. Zhang, S. Park, T. Beeler, D. Bradley, S. Tang, O. Hilliges. “ETH-XGaze: A Large Scale Dataset for Gaze Estimation under Extreme Head Pose and Gaze Variation.” In: *European Conference on Computer Vision (ECCV)*. 2020 (cit. on pp. 39, 51, 54).

[ZPZ+20] Y. Zheng, S. Park, X. Zhang, S. D. Mello, O. Hilliges. *Self-Learning Transformations for Improving Gaze and Head Redirection*. 2020. arXiv: 2010.12307 [cs.CV] (cit. on pp. 18, 31, 34, 43, 45, 47).

[ZSB18] X. Zhang, Y. Sugano, A. Bulling. “Revisiting Data Normalization for Appearance-Based Gaze Estimation.” In: *Proc. International Symposium on Eye Tracking Research and Applications (ETRA)*. Mar. 28, 2018, 12:1–12:9 (cit. on p. 30).

[ZTB+18] M. Zollhöfer, J. Thies, D. Bradley, P. Garrido, T. Beeler, P. Péerez, M. Stamminger, M. Nießner, C. Theobalt. “State of the Art on Monocular 3D Face Reconstruction, Tracking, and Applications.” In: (2018) (cit. on p. 31).

All links were last followed on June 3, 2024.

Declaration

I hereby declare that the work presented in this thesis is entirely my own and that I did not use any other sources and references than the listed ones. I have marked all direct or indirect statements from other sources contained therein as quotations. Neither this work nor significant parts of it were part of another examination procedure. I have not published this work in whole or in part before. The electronic copy is consistent with all submitted copies.

---

place, date, signature