Controllable Gaze and Head-Pose Redirection via Latent Disentanglement in Convolutional Autoencoders

Master's thesis in Complex Adaptive Systems

Eric Blohm
Nadav Harari

DEPARTMENT OF ELECTRICAL ENGINEERING
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2026
www.chalmers.se

Master's thesis 2026

Controllable Gaze and Head-Pose Redirection via Latent Disentanglement in Convolutional Autoencoders

Eric Blohm
Nadav Harari

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Electrical Engineering
Chalmers University of Technology
Gothenburg, Sweden 2026

Controllable Gaze and Head-Pose Redirection via Latent Disentanglement in Convolutional Autoencoders

Supervisors: John Dahl, Andrecia Ramnath, Zenseact
Examiner: Jonas Fredriksson, Department of Electrical Engineering

Master's Thesis 2026
Department of Electrical Engineering
Chalmers University of Technology
SE-412 96 Gothenburg
Telephone +46 31 772 1000

Cover: A facial image from the multi-identity dataset, see Section 4.1, is used as input to the model. The image is encoded and reconstructed while controlling the target head pose. The resulting image grid shows the same face reconstructed at several different head orientations.

Typeset in LaTeX
Printed by Chalmers Reproservice
Gothenburg, Sweden 2026

Controllable Gaze and Head-Pose Redirection via Latent Disentanglement in Convolutional Autoencoders
Eric Blohm, Nadav Harari
Department of Electrical Engineering
Chalmers University of Technology

Abstract

Driver Monitoring Systems (DMS) increasingly rely on gaze and head pose estimation to assess driver attention and detect unsafe states. However, existing datasets are dominated by common driving patterns while rare yet safety-critical behaviors occur irregularly and are difficult to capture systematically. This motivates the use of synthetic and controllable image generation to improve robustness and validation.

This thesis investigates whether gaze direction and head pose can be controllably manipulated in image space through autoencoder-based latent disentanglement. A custom data collection procedure is developed to enable dense and geometrically consistent supervision of gaze and head pose, supporting controlled learning of latent factors. Based on this data, convolutional autoencoders are trained using a latent-swapping strategy and explicit label supervision to encode gaze and head pose into interpretable latent dimensions. In addition, a Laplacian-based edge loss is introduced to improve preservation of high-frequency image details.

The results demonstrate consistent and interpretable control of gaze and head pose within the training distribution. The model achieves high reconstruction quality and preserves fine-scale features such as corneal reflections, verified through a dedicated detection pipeline. For unseen identities, coherent eye-region structure and meaningful gaze and head pose variations are retained, though distortions in other image regions and lower evaluation scores reveal limited out-of-distribution generalization. The results highlight both the potential and the limitations of deterministic autoencoders, motivating future work on improved realism and generalization.

Keywords: Deep learning, Autoencoder, DMS, Latent Control, Gaze, Image Generation, Latent Disentanglement.

Acknowledgements

We would like to extend our thanks to Zenseact for the opportunity to do this project, and specifically to John Dahl, supervisor, and Andrecia Ramnath, co-supervisor, for their support and valuable insights. Furthermore, we want to thank our examiner at Chalmers, Jonas Fredriksson, for his work to support this project.

Eric Blohm, Nadav Harari, Gothenburg, June 2026

List of Acronyms

Below is the list of acronyms that have been used throughout this thesis listed in alphabetical order:

CR	Corneal Reflection
DiT	Diffusion Transformer
DMS	Driver Monitoring System
Euro NCAP	European New Car Assessment Program
FAZE	Few-Shot Adaptive Gaze Estimation
GAN	Generative Adversarial Network
GDPR	General Data Protection Regulation
GN	Group Normalization
GT	Ground Truth
ICMS	In-cabin Monitoring Systems
MAE	Mean Absolute Error
MSE	Mean Square Error
NeRF	Neural Radiance Fields
OMS	Occupant Monitoring Systems
PCCR	Pupil-Center and Corneal-Reflection
PoG	Point of Gaze
SOTA	State-Of-The-Art
ST-ED	Self-Learning Transformations for Improving Gaze and Head Redirection
WHO	World Health Organization

Contents

List of Figures

List of Tables

1
Introduction

Modern vehicle safety increasingly depends not only on understanding the external driving environment but also on accurately interpreting the state of the driver. In this context, in-cabin sensing, and in particular driver monitoring systems (DMS), has emerged as a critical component of advanced safety systems. According to the World Health Organization (WHO), approximately 1.19 million people die each year as a result of road traffic crashes [37]. Driver-related factors such as impairment, distraction, and fatigue substantially contribute to this crash risk. In Sweden, 48 people were killed in alcohol- or drug-related traffic crashes in 2024, corresponding to 23% of all road fatalities [4], while in the United States, alcohol-impaired driving accounted for 12,429 fatalities in 2023, representing 30% of all traffic deaths [3]. Moreover, fatigue is associated with elevated crash risk, with a meta-analysis summarized by the European Road Safety Observatory estimating an average increase in risk of about 29% [36]. Distraction is also a major contributor to crash risk. For example, using a mobile phone while driving is associated with a crash risk approximately four times higher than driving without phone use [37]. Given these statistics, recent focus has increasingly shifted toward in-cabin technologies such as driver monitoring systems (DMS), which observe driver behavior and state to detect unsafe conditions and provide timely warnings or assistance.

1.1 In-Cabin Safety and DMS

In-cabin safety refers to vehicle functions that sense and mitigate safety-critical conditions originating from within the passenger compartment rather than from the external traffic scene. In contrast to outward-facing perception, in-cabin systems address risks related to the human driver and the in-vehicle occupants. These risks include reduced driving ability due to fatigue, distraction, or impairment, as well as occupant-related factors such as seating position, restraint use, and presence of passengers. Euro NCAP's assessment framework reflects this division by explicitly covering both occupant status monitoring and driver state monitoring as relevant parts of Safety Assist evaluations, indicating that in-cabin sensing is treated as a distinct and increasingly important safety domain [5].

A DMS is an in-cabin sensing system designed to estimate driver state and driver behavior and, when necessary, trigger warnings or support actions. Conceptually, a DMS can be described as a pipeline consisting of in-cabin sensing, interpretation of driver cues, and a vehicle response strategy. Road-safety literature supports

the relevance of such warning and assistance functions. For example, reports from the European Road Safety Observatory show that fatigue increases crash risk and identify warning systems and vehicle technologies as promising countermeasures [36].

While DMS focuses on the driver, in-cabin safety also includes Occupant Monitoring Systems (OMS). This includes seat belt reminders, occupancy detection, and other occupant-status monitoring features that support passive safety decisions. Euro NCAP's protocols treat these areas separately from driver monitoring, yet they share technologies such as in-cabin cameras, illumination design, and robust perception under occluded or night-time conditions [5]. This shared technological foundation motivates a unified discussion of in-cabin sensing, even when the functional goals differ.

1.1.1 Increased Attention to DMS

The relevance of DMS has increased in recent years due to both safety incentives and regulatory developments. On the regulatory side, the European Union's General Safety Regulation introduces requirements for advanced safety systems, including driver-related warning functions [35]. The technical and test requirements for driver drowsiness and attention warning systems are specified in Commission Delegated Regulation (EU) 2021/1341 [9], and the corresponding requirements for advanced driver distraction warning systems are specified in Commission Delegated Regulation (EU) 2023/2590 [10]. Together, these documents show that driver state and attention are not only research topics but also part of formal safety requirements in modern vehicles [9, 10].

In parallel, consumer safety rating programs increase industry incentives to implement robust driver monitoring. With Euro NCAP including Driver State Monitoring as part of its safety assessment protocol, the availability and performance of such systems will influence the safety scoring [5]. Technical publications describing Euro NCAP's approach further indicate that direct driver-state monitoring is being promoted to detect distraction, fatigue, or unresponsiveness and to trigger appropriate vehicle responses [30].

1.1.2 DMS functions

Today, DMS-based safety functions are implemented using several sensing and inference approaches. In production vehicles, the system can be grouped into direct and indirect approaches from a methodological perspective. Indirect monitoring systems use vehicle-system signals and driver-vehicle behavior to infer reduced driving capability. Direct monitoring systems instead estimate the driver state using human cues, with the eye and head regions being particularly informative. The EU regulatory requirements for advanced driver distraction warning systems explicitly include monitoring gaze direction as part of determining distraction [10].

In practice, direct DMS uses camera-based sensing to estimate head pose, gaze,

eyelid closure, and blink behavior. Additionally, thermal imaging has been shown to effectively detect different levels of drowsiness by focusing on thermal patterns in the face, highlighting an emerging sensor that could enhance the effectiveness of DMS [2].

Finally, DMS is typically coupled to an interface connecting human and machine, and a response policy that determines how the vehicle communicates detected risk and how warnings escalate over time. Assessment protocols emphasize not only detection but also the appropriateness and timing of vehicle response strategies, as DMS is intended to support the driver before an unsafe state leads to a critical event [30].

1.2 Problem and Motivation

DMS have achieved impressive accuracy rates in detecting common distractions such as yawning, phone use, or prolonged eye closure, often exceeding 95% accuracy. However, fleet-collected data is dominated by common driving patterns, while rare but safety-critical behaviors occur irregularly and unpredictably, making them difficult to capture systematically. These long-tail scenarios, which may involve uncommon head poses or extended gaze deviations, remain underrepresented despite their importance for system robustness [41, 42]. As a result, models may achieve high performance while remaining vulnerable under realistic edge conditions, revealing a validation data gap that is challenging to address through additional real-world data collection alone.

Prior work in gaze and eye tracking has demonstrated the value of synthetically generated data. At the time, frameworks such as LEyes show that training on realistic synthetic eye images yields performance equal to or exceeding state-of-the-art methods for corneal reflection (CR) and pupil localization [7]. Similarly, 3DGazeNet relies on synthetic 3D eye geometry to supervise the prediction of dense eyeball coordinates and gaze direction from a single face image, demonstrating that enriching datasets with high-quality synthetic data improves robustness and generalization [46]. While these approaches are not DMS-specific, they provide evidence that synthetic data can effectively improve gaze- and eye-related modeling, thereby motivating its use for controlled manipulation of head pose and gaze in driver monitoring contexts.

1.3 Scope and Limitations

The scope of this thesis is to control the gaze direction and head pose of a given facial input image. There is a major focus on synthesizing photo-realistic images. That is, the synthesized images must remain physically plausible, including proper illumination, smooth edges, and the person's identity. The models explored prioritize encoder-decoder models in which the identity propagates through the network, and the only changing attributes of the output are gaze and head pose.

This thesis does not explicitly address the generation of unseen identities or new scene content, as such generation may introduce hallucinations and produce physically implausible results, for instance, in specular reflections that do not match the global illumination. The roll component of gaze and head pose is not explicitly controlled; instead, the work focuses on yaw and pitch, which capture the dominant variation in the recordings considered. In addition, gaze is represented as a single yaw and pitch angle, respectively, with origin between the eyes.

A key limitation in this project is the lack of commercial and open-source data with ground-truth gaze and head-pose labels across many identities and viewpoints; therefore, the model's performance will be limited to the data collected during the project. Additionally, the labels created for head pose are estimates from an open-source model; therefore, the performance of the trained model is, to some extent, limited by the precision of those estimates. Finally, the design and creation of the dataset is limited to the hardware available. Due to the aforementioned limitations in data availability and the complexity of a multi-view dataset collection setup, this project operates within the domain of an image-to-image-based solution.

The models trained in this thesis do not explicitly control identity or scene conditions; however, the proposed architectures could be extended to incorporate a broader set of control parameters. Such extensions could enable training on real DMS footage and improve compatibility with DMS-specific visual conditions. Evaluation was limited to the metrics described in Sections 3.3 and 4.3; future work could additionally include state-of-the-art (SOTA) gaze trackers and identity-preservation metrics.

Direct quantitative comparison with existing gaze and head-pose redirection methods was not performed, as these methods are evaluated on datasets licensed under non-commercial use and that differ substantially in acquisition setup and image domain from the data collected in this work, making fair cross-method comparison infeasible.

1.4 Project Outline

This thesis is structured as a two-stage investigation, focusing on disentangled representation learning for gaze and head-pose redirection in face images. The first stage focuses on isolating and controlling gaze direction under constrained conditions, while the second stage extends the problem to disentangle both gaze direction and head pose independently. Together, these stages progressively increase the complexity of the task and evaluate the limits of latent-space disentanglement under realistic variations.

In the first part of the project, a controlled single-identity dataset is designed and collected in which gaze direction is the only varying factor across frames. Using this dataset, a lean convolutional autoencoder is trained with latent swapping and label supervision to disentangle gaze from all remaining image semantics. This ex-

periment demonstrates that deterministic control of gaze direction can be achieved while preserving high-fidelity image reconstruction for a fixed identity.

In the second part of the project, the objective is extended to include independent control of both gaze direction and head pose. A new dataset is collected in which head motion in pitch and yaw is allowed, and experiments are conducted on both single-identity and multi-identity data. To address the increased ambiguity introduced by correlated factors, the training procedure is augmented with structured latent swapping constraints that explicitly enforce disentanglement between gaze and head pose. A more expressive encoder–decoder architecture is introduced to accommodate the increased variability in the data.

Finally, the models developed in both stages are evaluated using a combination of quantitative and qualitative measures. Reconstruction fidelity is assessed using standard reconstruction losses, while the preservation of physically meaningful features, such as CR, is evaluated using a dedicated CR detection pipeline. Qualitative evaluations further demonstrate the models' ability to generate controlled variations in gaze and head pose while maintaining visual realism.

2
Theory

This chapter presents the theoretical foundations used throughout the thesis, spanning anatomy, mathematical models, and deep learning–based methods. In addition, related work is reviewed to illustrate the evolution of the field, outline the current SOTA, and position this work's contribution. The discussion of prior research also motivates the design choices made.

2.1 Autoencoders

An autoencoder is a neural network trained to reconstruct its input. Although reconstruction itself is not particularly useful, the intermediate representation learned by the network is often highly informative. The model maps an input x to an intermediate representation z, known as the latent space, through an encoder function f, such that $z = f(x)$. The latent representation is then mapped to a reconstruction x' through a decoder function g, such that $x' = g(z)$. The complete mapping from input to reconstruction is illustrated in Figure 2.1. Rather than simply copying the input, the autoencoder must learn a compressed representation that retains the most important information in the data. When the latent space has a lower dimensionality than the input space, the autoencoder is referred to as undercomplete. This constraint forces the model to capture the most salient features of the training data [16].

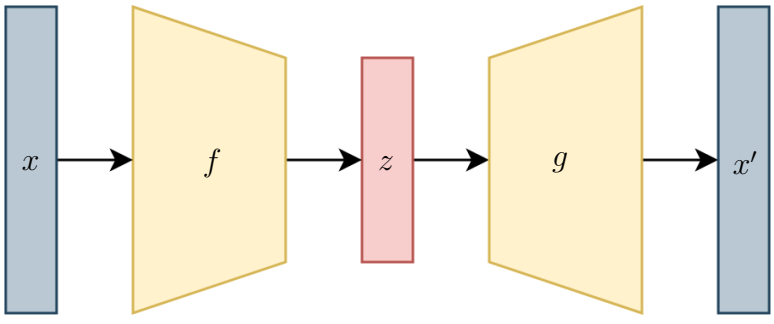


Figure 2.1: Illustration of the mapping $z = f(x)$ and reconstruction $x' = g(z)$ in an autoencoder architecture.

Autoencoders are typically trained with a reconstruction loss, where one common alternative is the mean squared error (MSE). MSE is suitable for continuous-valued image data and measures how faithful a reconstruction is compared to the input. It

directly measures the discrepancy at a pixel-level, as defined by:

$$L_{\text{MSE}}(x, x') = \frac{1}{n}\sum_{i=1}^{n}\|x_i - x_i'\|_2^2, \tag{2.1}$$

where $\|\cdot\|_2$ denotes the Euclidean norm (L_2), $\{x_i\}_{i=1}^n$ are the original images, and $\{x_i'\}_{i=1}^n$ are the corresponding reconstructions. As seen in Eq. (2.1), the reconstruction errors are squared, such that larger deviations are penalized more strongly, which makes MSE sensitive to outliers. Moreover, MSE may induce overly smooth or blurred reconstructions [48].

In the domain of gaze redirection, the 1-norm (L_1) loss is commonly used [52, 20, 32], with L_1 defined as:

$$L_1(x, x') = \frac{1}{n}\sum_{i=1}^{n}||x_i - x_i'||_1, \tag{2.2}$$

A likely reason is that the L_1 loss penalizes errors linearly and is less sensitive to outliers [48]. Compared to the MSE loss, it may reduce the tendency to produce overly smooth reconstructions, and He et al. found that it empirically performed better than MSE [20, 48].

Applications of autoencoders within computer vision include image classification and object detection [26]. In the classification case [26], it may be used to extract features within the areas of image recognition [33] and face recognition [1]. For the detection task, it enables the automatic extraction of target features and removes the limitations of manual extraction [26]. Within this domain, it can provide practical value and application to people's lives, such as pedestrian detection [49], video detection [14], and medical image detection [18].

2.2 Residual Neural Networks

Deep neural networks naturally integrate low-, mid-, and high-level features in a multilayer fashion, and the levels of features can be enriched by the number of stacked layers. The networks have led to many breakthroughs, mainly in image classification [22] and later in object detection. Methods like DETR and its variants use ResNet as the backbone to extract a feature representation before feeding it into a transformer-based network [8]. At the time, recent evidence showed that the network depth was of crucial importance [39, 44]. However, as network depth increases, the network's accuracy degrades, which was not due to overfitting [21, 22, 43]. Residual neural networks are designed to address the degradation in training accuracy that occurs as network depth increases, thereby enabling the effective training of deeper models. ResNet showed that deep neural networks with residual connections are easy to optimize and that the counterpart network without residuals exhibits higher training error. Moreover, ResNet can achieve accuracy gains with greatly increased depth, producing results substantially better than those of previous networks [22].

2.3 Human Eyes

The human eye is a biological optical system whose appearance and motion are determined by its anatomy and refractive structure. In computational modeling, the eye is often approximated using simple spherical geometry, for example as a single sphere or as a two-sphere system representing the eyeball and cornea. However, the true geometry and optics of the eye are more complex, which motivates more detailed approximations [6]. They rotate in three degrees of freedom, which can be represented in Euler angles as rotations about the three principal axes of a body relative to a fixed coordinate system: pitch, yaw, and roll [29]. The rotations around the three axes can be seen in Figure 2.2.

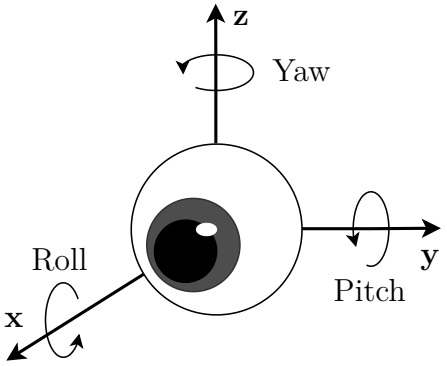


Figure 2.2: Coordinate system of the human eye showing rotational axes: pitch (rotation about the y-axis), yaw (rotation about the z-axis), and roll (rotation about the x-axis.

The rotations that make up the visual behavior take form in different groups of eye movements, which all affect how we track objects and react to visual stimuli. In addition, external conditions modulate observable eye characteristics, such as pupil size, which varies with luminance and accommodation demand, influencing both visual function and how the eye is measured by camera-based systems.

2.3.1 Human Eye Anatomy

The eye consists of an outer coat (sclera and cornea), a middle vascular layer (including the iris), and an inner neural layer (the retina) [24]. The sclera forms the visible white of the eye, while the transparent cornea provides most of the eye's refractive power. The iris controls pupil size and thereby regulates how much light enters the eye. The retina converts incoming light into neural signals; its central region, the fovea, supports the highest visual sharpness and is therefore the target region that gaze-stabilizing eye movements aim to keep aligned with objects of interest [24].

2.3.2 Human Gaze

Eye movements are commonly categorized into functional types that shift gaze and stabilize the retinal image, including saccades, smooth pursuit, vergence, the vestibulo-ocular reflex, and optokinetic nystagmus [24, 25]. In this thesis, we focus on two behaviors: saccades, which rapidly reorient the line of sight between targets, and smooth pursuit, which enables continuous tracking of a moving target to keep its image near the fovea [24]. The remaining movement types are not treated further, as they are not required for the data collection and modeling choices made in this work.

The human gaze can track a continuously visible moving target with smooth pursuit up to a characteristic angular velocity. In healthy observers, smooth pursuit can be performed for a stimulus moving at up to 100°/s, with the threshold being lower in many cases, and the ability decreases rapidly for higher velocities [24]. For higher velocities, saccades are performed to make up for the lack of pursuit. No human studies were found that establish absolute angular limits for yaw and pitch in saccade-free smooth pursuit. However, several human experiments have investigated smooth pursuit within explicitly defined central-field angle ranges. Typical paradigms use ±16° (32° peak-to-peak) both horizontally and vertically [27], and ±20° (40° peak-to-peak) at 20°/s in both yaw and pitch [45], while quantifying pursuit performance. Moreover, pursuit performance degrades at more eccentric eye positions in the orbit, supporting the practice of restricting experimental ranges to central gaze [51].

2.4 Point of Gaze

A foundational mathematical approach to computing the point of gaze (PoG) was introduced in 2006 by Guestrin and Eizenman. Their method estimates a gaze vector using pupil center and CR (PCCR) measurements. The simplest configuration enabling PoG estimation consists of a single camera and two light sources. However, this setup requires subject-specific calibration to estimate certain eye parameters, such as the radius of the cornea. In contrast, configurations using at least two cameras and two light sources allow reconstruction of the optic axis without such calibration. The authors identify the primary sources of PoG estimation error as inaccuracies in eye modeling, particularly deviations between the assumed and real corneal shape, as well as errors in pupil center and CR detection [17].

Building on the PCCR principle, modern commercial systems such as those of Smart-Eye and Tobii state that they employ near-infrared illumination to generate CRs, which are subsequently used for PoG estimation [23, 12]. Consequently, producing high-quality eye images with physically consistent CRs is crucial, as it enables the use of downstream gaze-tracking modules within DMS.

2.5 Related Works

The following section reviews prior work relevant to gaze and head-pose modeling, structured to both motivate the methodological choices of this thesis and position its contribution within the field. First, human eye modeling is discussed in the context of both analytical gaze tracking and synthetic data generation, highlighting how geometric formulations enable controllable eye appearance while introducing limitations in realism. Subsequently, work on gaze and head-pose redirection is presented chronologically, illustrating the evolution of the field from early warping-based approaches to learning-based 2D methods, and further to recent 3D-aware and generative frameworks.

2.5.1 Human Eye Modeling

Geometric eye models have historically been used both for analytical eye tracking and as a controllable basis for synthesizing eye images in computer graphics and learning-by-synthesis pipelines. A common abstraction is the double-sphere eye model, in which the eyeball and cornea are represented as two intersecting spheres with distinct centers and radii [17, 11]. This separation enables modeling of CRs under point illumination and supports classical PCCR reasoning. In modern eye-tracking pipelines, accurate localization of pupil and CR features remains a prerequisite for CR-based gaze estimation, and the PCCR principle has been described as the dominant approach in video-based eye tracking over the past decades [7].

For synthetic data generation, geometric eye models are attractive because they provide explicit control over variables such as gaze direction, head pose, and lighting geometry. However, the same simplifications that make these models tractable also limit realism. In computer graphics, the eye is often approximated by two spheres, sclera and cornea, and other components are simplified, while real eyes exhibit substantial individual variation and fine-scale appearance details [6]. This gap becomes critical when models are trained on synthetic images whose viewing geometry and illumination statistics do not match real conditions, since such mismatches can degrade transfer to real data [40].

2.5.2 Gaze and Head-Pose Redirection

Early methods of gaze redirection include warping-based approaches, in which a flow field is learned to spatially transform pixels in the input eye image toward a desired gaze direction. These methods operate by rearranging existing pixels in the input image, without synthesizing new image content. However, warping-based methods are limited when handling large redirection angles and cannot generate content outside the visible input. This limits their performance in cases of partial occlusion, such as eyelid closure or occluded scleral regions [15]. These limitations motivates the transition towards learning-based 2D gaze redirection methods that explicitly model image reconstruction, enabling the synthesis of unseen regions rather than merely rearranging visible pixels.

The authors of [20] propose a learning-based approach to gaze redirection that frames the task as conditional image generation with Generative Adversarial Networks (GANs). Given an input eye patch and a target gaze direction defined by yaw and pitch angles, the method synthesizes a new eye image with the gaze redirected. The model is trained using an adversarial loss, a gaze-direction loss implemented via an auxiliary gaze estimator embedded in the discriminator, a perceptual loss, and a cycle-consistency reconstruction loss to preserve subject-specific appearance. While this approach improves visual fidelity compared to warping-based methods, it is limited to eye patches.

While the above methods directly address gaze redirection, [32] introduce Few-Shot Adaptive Gaze Estimation (FAZE), which instead focuses on representation learning useful for disentanglement. They propose a disentangling encoder-decoder architecture that separates appearance, gaze direction, and head pose into distinct latent components. Rotation-equivariant structure is enforced by explicitly rotating gaze and head-pose components using known rotation matrices and reconstructing paired images. In addition to a reconstruction loss and a gaze-direction loss, FAZE introduces an embedding-consistency loss that encourages intra-subject consistency in gaze features. Gaze redirection is presented as a consequence of the learned equivariant representation rather than the primary objective of the method.

Building on the equivariant rotation-based latent representation introduced in FAZE, [52] extend this formulation in their method Self-Learning Transformations for Improving Gaze and Head Redirection (ST-ED), to full-face image generation. In contrast to earlier methods, ST-ED learns to represent additional factors such as illumination and color, alongside gaze and head orientation. Instead of relying solely on ground truth conditions, the model estimates internal control signals directly from the input images to guide how each factor is transformed during generation. The method combines adversarial training with an additional loss that explicitly penalizes errors in predicted gaze and head orientation between generated and target images, rather than relying only on pixel-level similarity. This loss is motivated by task-relevant inconsistencies, such as incorrect iris placement. In addition, ST-ED introduces constraints to separate gaze and head pose from other image variations, such as illumination, thereby reducing unintended interactions between factors.

Building on 2D-based methods, more recent works incorporate 3D modeling using Neural Radiance Fields (NeRFs) and Gaussian Splatting. GazeNeRF incorporates 3D awareness through two-stream MLPs that predict feature maps for the eyes and face region separately. The eye regions are transformed with a 3D rotation matrix before the two regions are composited via rendering. However, GazeNeRF shares the same limitations as other NeRF-based methods, namely long training times and high computational cost [38]. Building on the limitations of NeRFs and the idea of a two-stream method, DiT-Gaze incorporates 3D Gaussian Splatting and introduces Diffusion Transformers (DiT) to this field. DiT-Gaze introduces three novelties: a DiT renderer, an intermediate gaze sampler, and an orthogonality constraint loss

to enforce feature disentanglement. Their method represents the current SOTA for offline synthetic data generation in the field of head-pose and gaze redirection [31].

3
Gaze Redirection

In this first experiment, the disentanglement of gaze features from factors such as illumination and identity was investigated. This was done using a constrained dataset in which gaze is the only explicitly varying factor. The experiment tested whether it was possible to control gaze without any other visual attributes changing in the reconstructed image.

3.1 Dataset

Because no publicly or commercially available dataset met the requirements of this thesis, a dedicated gaze-tracking dataset was constructed. The principle was to fix every factor except the subject's eye movement, thereby enabling precise computation of gaze angles for each individual frame. This was accomplished by recording the subject while their gaze followed a moving stimulus on a screen and keeping a corresponding log of the stimulus's exact position. Consequently, the setup provided all the information needed to compute the eye's pitch and yaw angles using trigonometry. All data collection procedures were conducted in strict compliance with the General Data Protection Regulation (GDPR) and institutional ethics guidelines, with all participants providing informed consent prior to inclusion and additional consent for inclusion in any publication.

3.1.1 Coordinate System

All gaze and head pose angles in this project are defined using a right-handed coordinate system, according to Figure 2.2. The origin is located at the gaze reference point between the eyes, and rotations follow the standard right-hand rule.

Pitch corresponds to a rotation around the y-axis. In this coordinate frame, positive pitch is defined as a downward rotation. Consequently, gaze directed upward, or a head rotation upward, results in a negative pitch angle, while gaze or a head rotation downward produces a positive pitch angle.

Yaw corresponds to a rotation about the z-axis, with the z-axis defined as pointing upward. When viewed from above along the positive z-axis, a positive yaw corresponds to a counterclockwise rotation. As a result, gaze directed to the left or a head rotation to the left yields a positive yaw angle, whereas gaze or head rotation directed to the right yields a negative yaw angle.

3.1.2 Data Collection Setup

A camera was positioned to face the subject and aligned with the stimulus display, so that it was horizontally centered within the video frame. Vertically, the camera was offset upward to ensure that the resulting gaze angles remained within a predefined range during stimulus traversal. This was achieved by moving the stimulus along a predefined grid while keeping the participant at a fixed distance from the screen, resulting in yaw movements of $\pm 25°$ and pitch movements of $-20°$ to $+25°$. Moreover, the subject was oriented so that the eyes were at the same level as the camera and centered. To mitigate movement along all axes, the subject's head rested on a chin rest, thereby enabling more stable per-frame gaze estimation. A light source was placed directly beneath the camera to achieve uniform illumination and to ensure that a CR was present in the eyes in all frames where the eyelid did not occlude it.

3.1.3 Video of a Moving Stimulus

To generate interpretable gaze angles, the stimulus must have a known screen position at all times. A custom video was therefore created in which a small dot traverses a predefined pixel grid. The grid resolution was based on reported evidence of the functional range of smooth pursuit, described in 2.3.2, together with empirical estimates for the test subjects.

The duration of the stimulus video was set to 2 minutes, and the recording was captured at 30 frames per second, resulting in 3600 frames. This length balanced the desire for a dataset containing small-angle differences of adjacent frames, with the practical constraints of subject focus. The stimulus trajectory was chosen such that the resulting eye angular velocity reached a peak of 22.30°/s at the screen center and decreased to 18.32°/s at the grid boundary.

To ensure that the stimulus motion remained within the desired range of angular velocities, the mapping from linear screen motion to visual angle was derived. It was done using a pinhole-viewing approximation. Let $x(t)$ denote the horizontal position of the dot relative to the screen center (in pixels or centimeters), D the fixed viewing distance from the observer's eye to the screen (in centimeters), and $\theta(t)$ the subtended visual angle (in degrees) at time t. Under a pinhole-viewing approximation, this angle is:

$$\theta(t) = \arctan\left(\frac{x(t)}{D}\right). \tag{3.1}$$

Differentiating with respect to time yields the angular velocity,

$$\omega = \frac{d\theta}{dt} = \frac{1}{1 + \left(\frac{x}{D}\right)^2} \frac{1}{D} \frac{dx}{dt}. \tag{3.2}$$

Since $\frac{dx}{dt} = v$ is the linear speed of the dot on the screen, substitution and simplification give

$$\omega = \frac{v\,D}{D^2 + x^2}. \tag{3.3}$$

This expression shows that, even at constant linear speed v, the instantaneous angular velocity ω decreases as $|x|$ increases, i.e., as the stimulus moves farther from the screen center. The stimulus trajectory and speed were therefore selected such that the resulting ω remained below commonly cited limits for smooth pursuit across the entire grid.

3.1.4 Pre-processing

After data collection, the raw synchronized videos were processed into a final dataset. This pipeline consisted of frame extraction, image centering, gaze labeling, occlusion and noise filtering, and dataset splitting.

Frame extraction was performed by reading each frame from the raw video recording and saving it as an image. This enabled subsequent frame-wise processing. Moreover, to mitigate pixel-level drift between frames, each extracted frame was first centered on the subject's nose and subsequently cropped to a spatial resolution of 256×416 pixels. This resolution was selected to enable training an autoencoder with encoder and decoder depths of up to 5 layers each, where each layer reduces the spatial resolution by a factor of 2. Centering enforces that a fixed facial landmark remains at a consistent pixel location across the entire sequence, in this case the nose was used since it allows for a tight crop around the face where all facial attributes are visible. Although the subject's head was stabilized during recording, small movements may occur due to natural human motion, and this normalization step reduces unwanted positional variance that could otherwise degrade model performance.

Automatic annotation of the frames was performed using the known geometry of the data collection setup, namely the dot position in each frame, the grid size of the dot trajectory in pixel coordinates, the physical width of the dot trajectory, and the distance from the eyes to the screen. Consequently, the pitch and yaw for the eyes could be computed for every frame. A standardized ground-truth pair is thereby provided, consisting of the input image and its corresponding label.

Due to the camera and flashlight occluding part of the screen, a portion of the stimulus video was permanently unavailable to the subject. Frames in which the moving dot moved inside this occlusion region were therefore removed. The occluded area was defined using the known physical dimensions of the screen, the camera's geometric origin, and the measured size of the camera and flashlight region. Using the known physical dimensions of the screen and occluding elements, these measurements were converted into pixel coordinates that define the occlusion region, and the frames in which the dot was inside this region could then be removed.

In addition to the automatic removal of frames within the occlusion region, further frames were excluded due to eye blinks. Blink removal was performed manually by iterating through the dataset and identifying the start of each blink. For each detected blink, all frames from the first moment of eyelid closure until the eye had returned to the appearance before were removed. This ensured that no frames associated with incomplete or unstable eye appearance remained. Finally, the dataset was split into training, validation, and test sets, following the 80/10/10 convention.

3.2 Autoencoder with Latent Swapping

The first model, which was designed, implemented, and evaluated, leveraged the swapping of latent dimensions to disentangle gaze from the rest of the image features. To do so, a lean autoencoder with a convolutional encoder, a latent bottleneck, and a mirrored decoder was created with predefined dimensions for disentanglement.

3.2.1 Architecture

The model was built around a convolutional autoencoder that maps an input image to a compact latent representation and reconstructs the original image from this representation. Each input sample is represented as a tensor $\mathbf{x} \in \mathbb{R}^{B\times C\times H\times W}$, where B denotes the batch size, C the number of image channels, C=3 for RGB, and (H, W) the spatial resolution of the input, in this dataset $(H, W) = (416, 256)$. The decoder produces an output $\hat{\mathbf{x}}$ with the same shape as the input, $\hat{\mathbf{x}} \in \mathbb{R}^{B\times C\times H\times W}$, and a sigmoid activation in the last layer constrains reconstructed pixel intensities to the interval $[0, 1]$.

The encoder comprises five convolutional stages that, in each stage, reduce the spatial resolution while increasing the number of channels. Each stage uses a 3×3 convolution with stride 2 and padding 1, followed by ReLU. The channel progression through the encoder is

$$C \to 32 \to 64 \to 128 \to 256 \to 512$$

resulting in five downsampling operations.

To obtain the latent representation, the encoder output is flattened to a vector of length $512 \cdot 13 \cdot 8$ and projected to a latent vector $\mathbf{z} \in \mathbb{R}^{B\times d}$ using a fully connected layer. The latent dimensionality d was treated as a tunable hyperparameter and evaluated over

$$d \in \{4,\ 8,\ 16,\ 32,\ 64,\ 128,\ 256,\ 512\},$$

to identify a compact representation with sufficient reconstruction fidelity. A second fully connected layer maps $\mathbf{z}$ back to the flattened encoder feature size, after which the tensor is reshaped to $\mathbb{R}^{B\times 512\times 13\times 8}$ as input to the decoder.

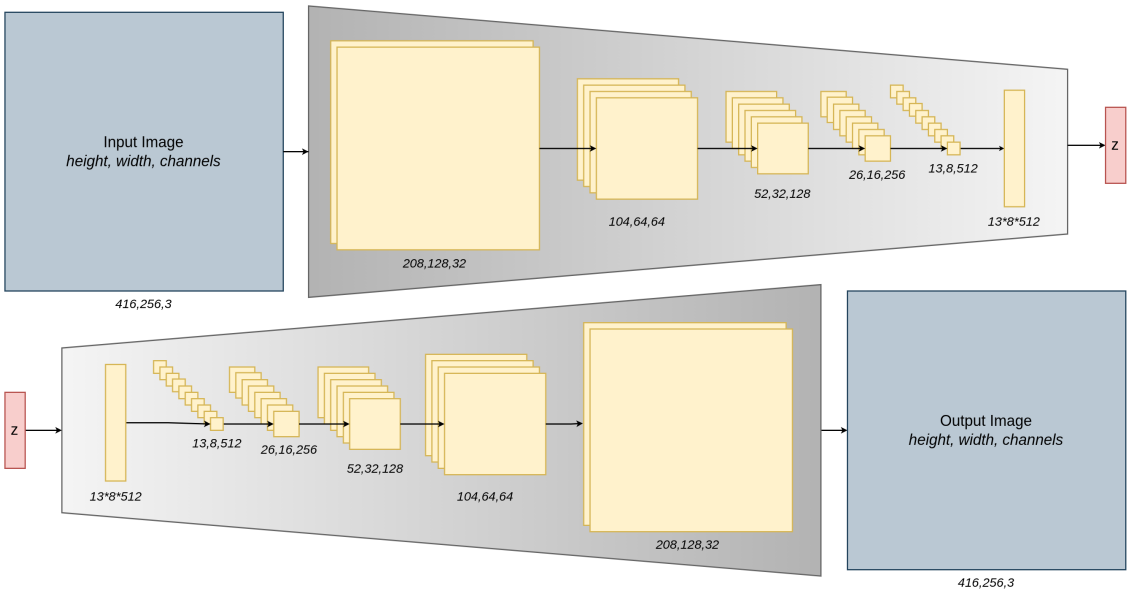


Figure 3.1: Overview of the lean autoencoder architecture, illustrating the data flow as well as the evolution of spatial resolution and channel depth throughout the network. The upper branch shows how the input image is encoded into a latent representation, while the lower branch illustrates how the latent representation is decoded back into a reconstructed image.

The decoder mirrors the encoder using five transposed convolution stages, without the ReLU at the final layer. The channel progression in the decoder is

$$512 \to 256 \to 128 \to 64 \to 32 \to C,$$

restoring the spatial resolution back to (H, W). The final transposed convolution is followed by a sigmoid activation, producing the reconstructed image $\hat{\mathbf{x}}$ in the same format and range as the input. The entire architecture is illustrated in Figure 3.1.

3.2.2 Training with Latent Swapping

The autoencoder was trained to disentangle gaze from the remaining image semantics by combining two complementary learning signals. A label loss, $\mathcal{L}_{\text{label}}$, encourages the gaze to be encoded into the predefined latent dimensions. In addition, a swap loss, $\mathcal{L}_{\text{swap}}$, inspired by VASA-1 [47], enforces that swapping only these dimensions between two samples is sufficient to reproduce the corresponding change in the reconstructed image. Additionally, each image was self-reconstructed without latent manipulation to preserve overall image fidelity by encouraging accurate reconstruction of each input without swapping. The self-reconstruction loss was defined as $\mathcal{L}_{\text{self}}$. These losses were then combined as a weighted sum to control the relative influence of anchoring the gaze dimensions and enforcing swap-consistency:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{label}}\mathcal{L}_{\text{label}} + \lambda_{\text{swap}}\mathcal{L}_{\text{swap}} + \lambda_{\text{self}}\mathcal{L}_{\text{self}},$$

where λ_{label}, λ_{swap} and λ_{self} are the weights for each loss.

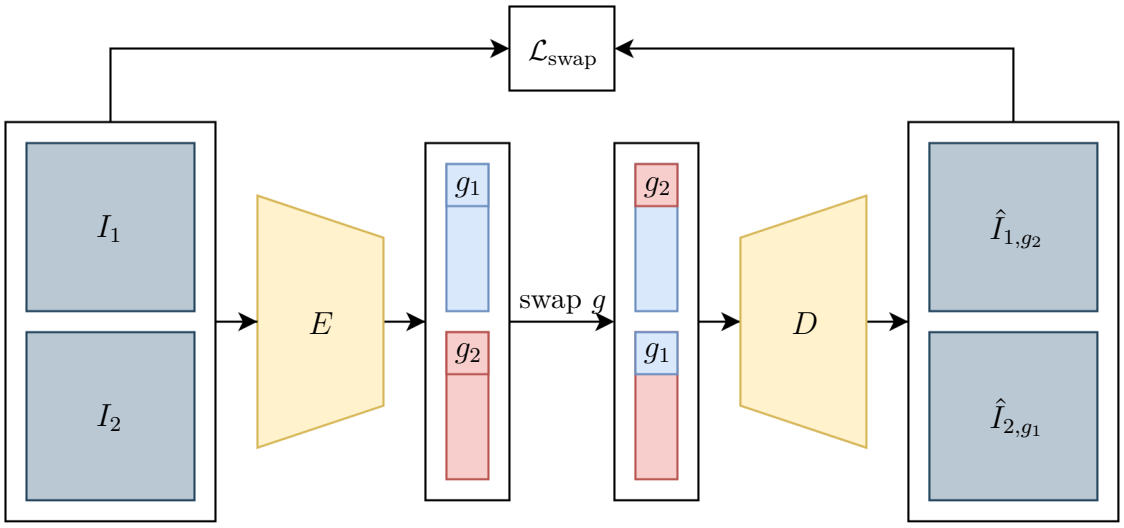


Figure 3.2: Illustration of latent swapping. Two images, I_1 and I_2, of the same identity are encoded into latent representations containing gaze (g_i). The gaze components from each latent representation, g_1 and g_2, are swapped while the rest of the latent representation remain fixed, and the modified latent vectors are decoded to reconstruct the corresponding images.

The label loss was computed as MSE between the encoded values in the predefined dimensions, one for yaw and pitch each, and the ground truth labels. Minimizing this label loss ensured that the encoder learned the gaze information from pixel space and encoded the correct labels into the predefined dimensions corresponding to the ground truth.

For each image pair, the swap loss was computed as the L_1 distance between a swapped reconstruction and the corresponding target image from the pair. When the gaze components are swapped between two latent representations, the resulting latent vector contains the gaze of the other image while retaining all remaining attributes. The decoder is therefore encouraged to reconstruct the target image, as the input images share identity and differ only in gaze. Finally, the self-reconstruction loss was also computed with L_1 as $\mathcal{L}_{\text{self}} = L_1(I_1, \hat{I}_1) + L_1(I_2, \hat{I}_2)$, where $\hat{I}_i = D(z_i)$.

In detail, latent swapping works as follows: given two images, I_1 and I_2, of the same identity, the images are encoded to obtain their respective latent representations, z_1 and z_2, where the gaze (g_1) of I_1, is defined as the first two dimensions of z_1. Similarly, g_2 corresponds to the first two dimensions of z_2. g_1 and g_2 are then swapped between z_1 and z_2, after which the modified latent vectors are decoded. This yields $\hat{I}_{1,g_2} = D(z_{1,g_2})$ and $\hat{I}_{2,g_1} = D(z_{2,g_1})$, where ideally $\hat{I}_{1,g_2} = I_2$ and $\hat{I}_{2,g_1} = I_1$, since the input images share the same identity and attributes and differ only in gaze. Therefore, $\mathcal{L}_{\text{swap}}$ is defined as $\mathcal{L}_{\text{swap}} = L_1(I_1, \hat{I}_{2,g_1}) + L_1(I_2, \hat{I}_{1,g_2})$. A schematic illustration of the swapping is included in Figure 3.2.

For this project, the learning rate was set to 0.0005, λ_{label} was set to 0.001, λ_{swap}

and λ_{self} were set to 1, the batch size was set to 8 and Group Normalization (GN) [50] was used. All hyperparameters were found empirically. The latent space sizes $d \in [4, 8, 16, 32, 64, 128, 256, 512]$ were used in an ablation study to evaluate how it affects the final performance and to provide a basis for training the models on the larger dataset in Section 4.

3.3 Evaluation

The model configurations were all evaluated qualitatively and quantitatively. Firstly, by observing the reconstructions, one could qualitatively assess the image quality and artifact presence. Secondly, the encoding accuracy of test images was calculated by computing the mean absolute error (MAE) between the ground truth (GT) and the encoded gaze. Lastly, CR presence and location were evaluated between the output image and the GT image. This was primarily used for evaluation between models to quantify performance differences, and further indicate reconstructions quality for each model.

The CR detections produced a confusion matrix for each eye in the swapped reconstructions. The CR detector was implemented by augmenting an image using cropping, gaussian blur, Laplace sharpening, and an additional gaussian blur for edge enhancement followed by a blob detector to locate the brightest pixel of the CR. The four different predictions in this project are classified according to Table 3.1.

Table 3.1: Interpretation of CR detection predictions.

Prediction	**CR in GT**	**No CR in GT**
CR in reconstruction	True Positive (TP)	False Positive (FP)
No CR in reconstruction	False Negative (FN)	True Negative (TN)

Additionally, the cases where TPs were found, the error in pixel coordinates between the brightest point of the CR in the output and GT were calculated. Finally, the detector performance was assessed on the ground truth test dataset to examine its reliability in model evaluation.

3.4 Results

The results shown in this section were generated with the autoencoder using a latent dimensionality of 32, with weights of 1 for the self-reconstructions and swap losses, and weights of 0.001 for the label loss. The model was trained for 200 epochs, with batch size 8 and learning rate 0.0005.

3.4.1 Dataset Coverage

The gaze-only dataset consisted of 3,249 images, uniformly distributed within $\pm 25°$ for yaw and from $-20°$ to $25°$ for pitch, as mentioned earlier. During pre-processing,

some samples were filtered out, thereby introducing sparsity into the distribution. Moreover, the empty region in the center corresponds to the area where the camera setup occluded the screen. Consequently, no GT image-gaze pairs exist in that region. The resulting gaze coverage is illustrated in Figure 3.3.

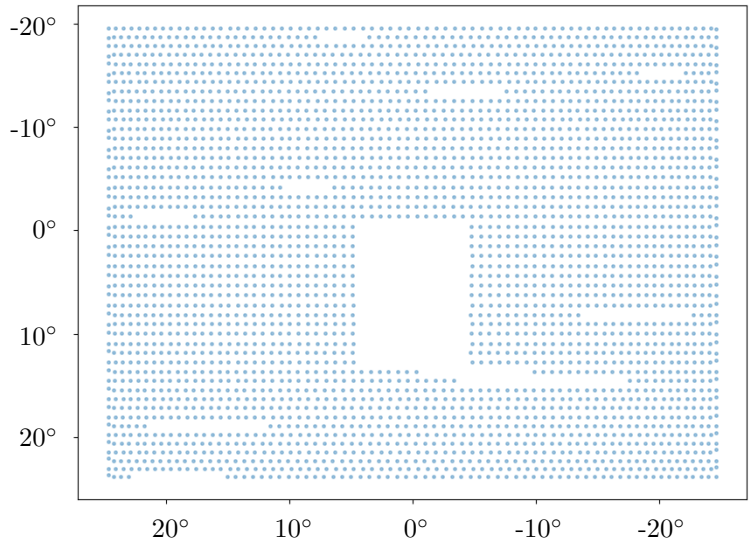


Figure 3.3: Dataset coverage over pitch (y-axis) and yaw (x-axis) for the gaze-only model after filtering for eyelid closures and camera setup occlusion.

3.4.2 Reconstruction Quality

The trained model produces high-quality images with little to no perceptual difference between the ground truth and the reconstructed images, as shown in Figure 3.4.

Figure 3.4: Reconstructions (Recon) of the GT images in the test set.

3.4.3 Disentanglement

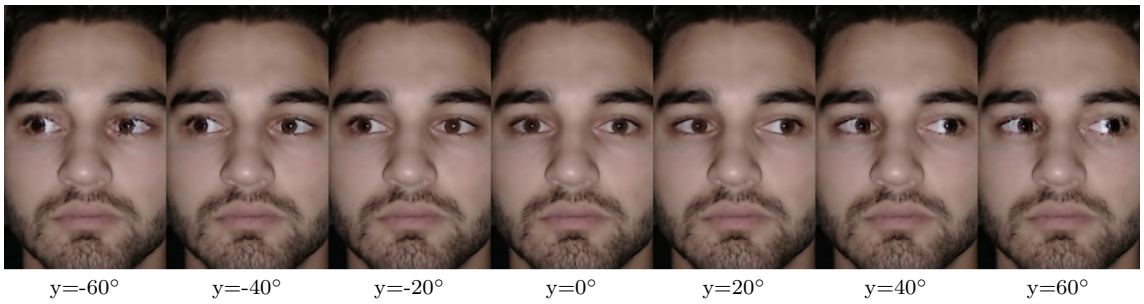
The results show that manipulation of the latent space via injection, after training a lean autoencoder under explicit supervision for predefined latent dimensions, yields controllability over the trained factors. In Figure 3.5, a random image from the test set is encoded, and for each combination of pitch and yaw in the image grid, the gaze is injected into the latent dimensions controlling the gaze, and the latent vector is then decoded to each of the output images in the grid. See Figure A.1 in the Appendix for a more extensive image grid.

Figure 3.5: Gaze direction sweep over pitch (rows) and yaw (columns).

The resulting model can encode any image from the dataset and redirect the gaze successfully to angles within the training distribution. Perturbations further outside the training distribution caused the model to produce images that were no longer aligned with the desired gaze direction and worse visual results, as can be seen by the generated outputs in Figures 3.6a and 3.6b.

Additionally, with this configuration, latent encoding accuracy was evaluated by comparing the first two gaze-related latent dimensions to their corresponding ground-truth labels. The resulting MAE was 0.75° and 0.23° for yaw and pitch, respectively.

(a) Redirection to different angles in pitch (p) with static angle in yaw of 0°.



(b) Redirection to different angles in yaw (y) with static angle in pitch of 0°.

Figure 3.6: Arbitrarily chosen test image encoded and redirected for different yaw and pitch angles showcasing model performance for angles outside of training distribution.

3.4.4 CR Detection

Prior to evaluating model performance, the reliability of the automated CR detector was validated against manual annotations of the test images. The detector achieved an F1 score of 0.99 and accuracy of 0.98, and its outputs are therefore treated as reliable proxies for glint presence in the subsequent model evaluation.

CR preservation of the model was assessed by applying the CR detection algorithm to redirected and reconstructed images versus ground-truth images and comparing the detected CR locations. Combined over the left and right eyes, the model achieved an accuracy of 0.97, an F1 score of 0.98, and a mean absolute CR localization error of 0.58 pixels. Finally, the corresponding confusion matrix for CR detection is shown in Figure 3.7.

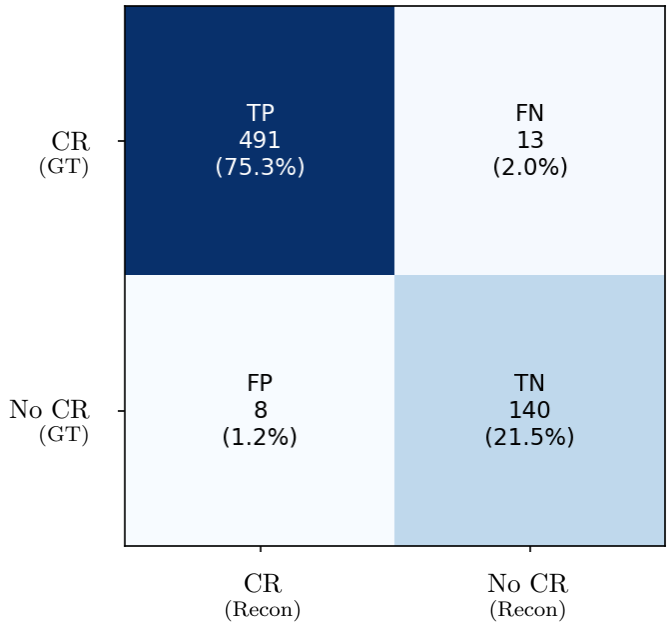


Figure 3.7: Combined confusion matrix from CR detections of the left and right eye respectively.

3.5 Discussion

The results of the gaze-only experiment demonstrate that a lean convolutional autoencoder, trained with explicit latent swapping and label supervision, together with a constrained, homogeneous dataset, is capable of learning a highly controllable representation of gaze direction. Under these conditions of dense, nearly uniform data sampling, reconstruction quality was visually indistinguishable from the ground truth across the test set, with only minor smoothing effects in certain regions. This high reconstruction fidelity is further reflected in the strong CR-detection performance, indicating near-perfect consistency between redirected and reconstructed versus ground-truth images. The preservation of fine-scale eye appearance, including the CR, suggests that the encoder–decoder architecture did not introduce perceptible artifacts under in-distribution conditions.

In addition to high reconstruction fidelity, the model achieved high accuracy in encoding gaze direction into the predefined latent dimensions, with low MAE in both yaw and pitch on the test set. This indicates that the swap loss successfully encouraged the encoder to extract gaze-related information from the image and place it into the intended subspace of the latent representation. Given the dense sampling of gaze angles in the dataset, such performance was expected as the test data lies close to the training data. The training data provides a smooth and dense manifold over gaze space, enabling effective interpolation between neighboring gaze directions.

Beyond reconstruction and encoding accuracy, the qualitative results show that the model gained strong controllability over gaze direction. Injecting target yaw and

pitch values into the designated latent dimensions resulted in consistent, predictable changes in eye orientation while leaving the remaining image content unchanged. This behavior confirms that disentanglement through latent swapping operated as intended in this controlled setting, and that the label loss effectively anchored the latent dimensions to a physically interpretable parameterization expressed in degrees. Together, these results validate the core hypothesis of this experiment: explicit swapping supervision can be used to enforce controllable disentanglement in an autoencoder when the underlying data distribution is sufficiently controlled.

However, the limitations of this setup are reflected when the model is evaluated outside the training distribution. Injecting gaze angles that were not observed during training caused the model to degrade rapidly, producing outputs that no longer aligned with the desired gaze direction and exhibiting reduced visual quality. This collapse is not unexpected and highlights that the learned representation does not extrapolate beyond the sampled gaze manifold. Rather than indicating a failure of the approach, this behavior reflects the strong inductive bias imposed by the homogeneous dataset. The model effectively learned to interpolate within the observed gaze range, but did not acquire a more general representation of eye appearance under arbitrary gaze directions.

Overall, the results show that controlled image reconstruction through latent disentanglement and parameter injection is feasible in a single-identity, gaze-only setting. The near-indistinguishable reconstruction quality and the stability of CR-related features demonstrate that such a model can preserve fine visual details while enabling deterministic control. These findings suggest that the approach is promising as a building block for synthetic data generation in gaze-sensitive domains, such as driver monitoring systems. However, the observed out-of-distribution collapse also underscores the need for more diverse datasets and more complex supervision when extending the approach to multiple factors. This motivates the subsequent experiment, in which head pose is introduced as an additional disentangled factor, and the limits of latent swapping are explored under increased data complexity.

4

Head and Gaze Redirection

Transitioning from the constrained gaze-only dataset, the project was extended to accommodate head pose variation and limited translational motion in the plane parallel to the stimulus screen. Gaze was computed using the same formulation as in the gaze-only dataset, with the addition of a correction for the gaze origin offset. Head pose was estimated using MediaPipe [13], which provides a transformation from a canonical face model to the detected facial landmarks. Finally, the model architecture was adapted to handle the increased complexity introduced by the additional control factor and image resolution.

4.1 Dataset

For the second experiment, which focused on joint disentanglement of head pose and gaze, a new dataset was collected using the same physical setup as described previously, but with relaxed constraints on head motion. In contrast to the first experiment, participants were allowed to rotate the head freely in yaw and pitch during recording.

Data collection was performed for two settings. First, a single-identity dataset was recorded, capturing a larger amount of data to obtain dense, smooth sampling of the joint head–gaze space across yaw and pitch. Second, a multi-identity dataset was recorded, in which the same procedure was repeated for multiple participants, but with fewer samples per identity. In both cases, the stimulus video described in Section 3.1.3 was used. Consequently, the experiment induced the same gaze-angle ranges as in the gaze-only experiment.

The recorded videos were post-processed using the same steps as in Section 3.1.4, with additional filtering to account for frames where the relative orientation between head pose and gaze was extreme. Specifically, extreme relative gaze was defined as the absolute difference between gaze angle and head pose angle exceeding predefined limits in yaw or pitch, where the thresholds were identical to those used in Section 3.1.2. In addition, head pose was labeled using MediaPipe, and the resulting transformation matrix was used to extract the head pose in degrees, which served as the pose labels for the data.

In practice, it is difficult to vary head pose in yaw and pitch without simultaneously shifting the gaze origin away from the camera. To address this, gaze angles were

computed using an offset-based formulation. Prior to recording, a reference configuration was established in which the gaze origin was aligned with the camera for each participant. This reference was used to estimate a constant offset between the camera coordinate frame and the gaze origin of the participant. During recording, any movement within the plane parallel to the screen was tracked, and the gaze angle was calculated by compensating for this offset. The offset was estimated using facial landmark detections provided by MediaPipe, ensuring consistent gaze-angle estimation despite translational head motion.

Due to the increased range of head rotations in this experiment, a larger portion of the image frame was occupied by the participant's face compared to the gaze-only dataset. Consequently, all images in this dataset were spatially cropped to a resolution of 416×416 pixels.

All data collection procedures were conducted in strict compliance with the GDPR and institutional ethics guidelines, with all participants providing informed consent prior to inclusion and additional consent for being included in any type of publication.

4.2 ResNet Autoencoder with Latent Swapping

The lean autoencoder used for gaze redirection with a static head pose was further developed into a more capable network as the data became more complex and the sampling density decreased. Inspired by the ResNet architecture and using a similar disentanglement methodology for latent swapping, training to disentangle head pose and gaze direction was performed with a more capable model and an additional loss term.

4.2.1 ResNet-based Autoencoder

Introducing head-pose control increases dataset complexity, which the network must accommodate. Therefore, the model implemented in Figure 3.1 was enhanced. Simply adding more layers has previously been shown to degrade performance. In contrast, architectures with residual connections achieve improved accuracy and training stability as discussed in 2.2. Consequently, the encoder consists of five ResNet blocks, each containing two convolutional layers. Within each block, the first convolution downsamples the spatial resolution while doubling the number of hidden channels, whereas the second convolution preserves both spatial resolution and channel dimensionality. Each block includes a residual skip connection that bypasses the two convolutional layers, enabling information flow and stabilizing training. In practice, this is performed using a convolutional layer with a 1×1 kernel with stride 2, and no norm or activation. All other convolutional layers use a 3×3 kernel size, group normalization, and ReLU activation. A schematic illustration of the ResNet block is shown in Figure 4.1.

The channel dimension is first increased from 3 (RGB) input channels to 32 hidden channels, after which the number of channels is doubled at each subsequent downsampling stage, and the hidden channel dimensionality evolves as follows:

$$C \to 32 \to 64 \to 128 \to 256 \to 512.$$

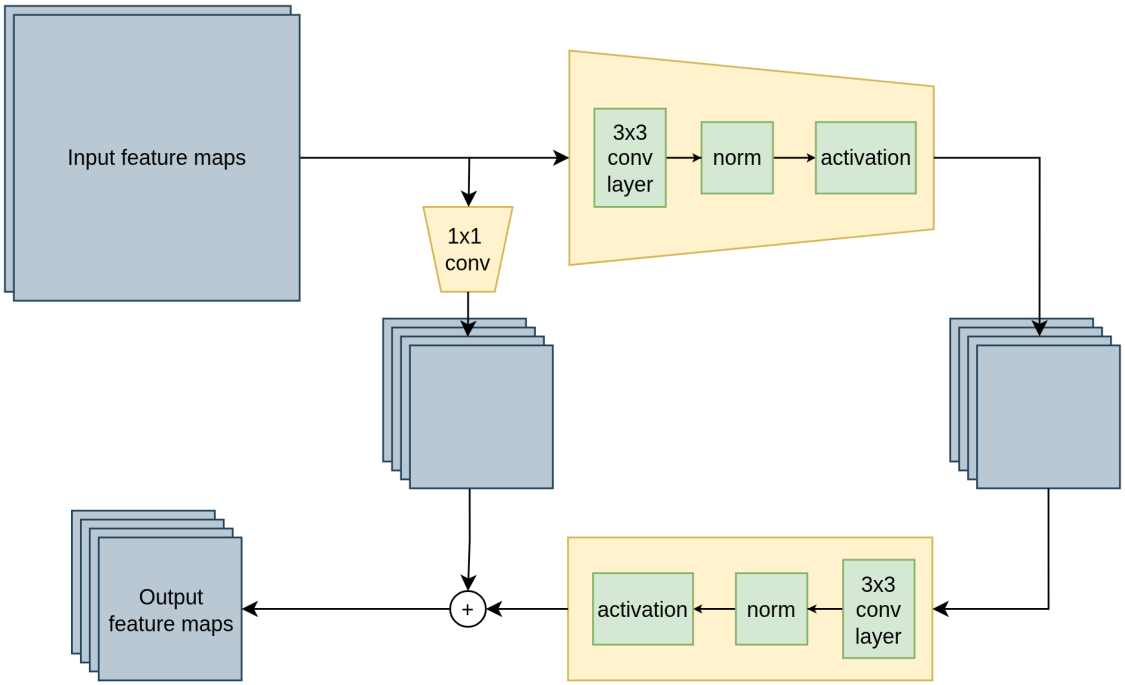


Figure 4.1: Architecture of a ResNet encoder block used in the autoencoder. The main branch consists of two 3×3 convolutional blocks, each include normalization and activation. Spatial downsampling and channel expansion are illustrated by the cone-shaped structure, while the rectangular structure preserves spatial dimensions. The residual connection passes features through a 1×1 convolution with a stride of 2 to match spatial resolution and channel dimensionality before addition. Grey blocks represent feature maps propagated through the network.

In addition to the stacked blocks, the final feature map is flattened into a one-dimensional vector and passed through a linear layer that projects it into a 128-dimensional latent space. Concretely, a feature map in $\mathbb{R}^{C \times H \times W}$ is reshaped into a vector in $\mathbb{R}^{C \cdot H \cdot W}$ and mapped as $\mathbb{R}^{C \cdot H \cdot W} \to \mathbb{R}^{128}$, with $(C, H, W) = (512, 13, 8)$. An overview of the full ResNet-based encoder architecture is shown in Figure 4.2.

The decoder was designed to mirror the encoder, using ResNet-based blocks with the same internal structure and residual skip connections. Spatial upsampling is performed using transposed convolutional layers. Within each block, the first convolution upsamples the spatial resolution while reducing the number of hidden channels, whereas the second convolution preserves both spatial resolution and channel dimensionality. Apart from the use of transposed convolutions for upsampling and asymmetric channel expansion, the decoder blocks are identical to those used in the

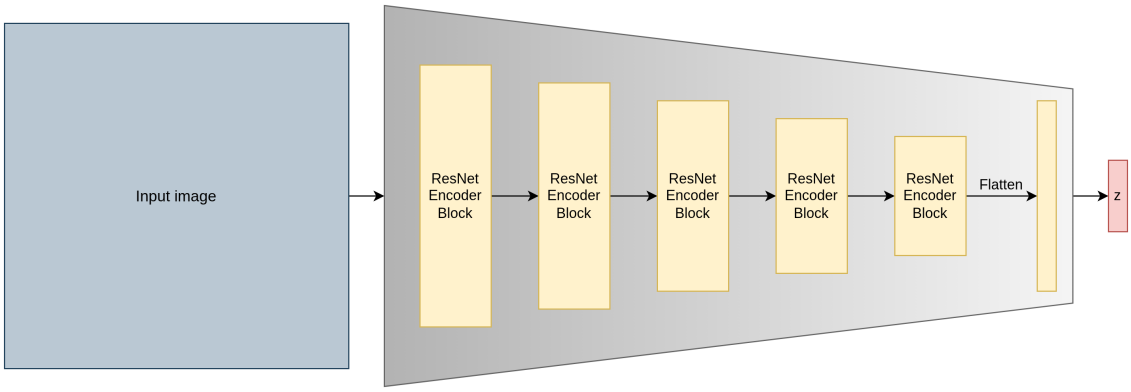


Figure 4.2: Encoder architecture composed of five stacked ResNet blocks, followed by flattening and a non-linear projection into the latent space z. The cone-shaped structure indicates progressive spatial downsampling of the feature maps across blocks.

encoder.

The 128-dimensional latent vector is first projected back to the same dimensionality as the flattened encoder feature map and reshaped into a tensor of size $C \times H \times W$. Specifically, the latent representation is mapped from $\mathbb{R}^{128} \rightarrow \mathbb{R}^{C \cdot H \cdot W}$, then reshaped and processed through the stacked decoder blocks. The hidden channel dimensionality is reduced asymmetrically to the encoder as follows:

$$512 \rightarrow 256 \rightarrow 128 \rightarrow 64 \rightarrow 32 \rightarrow C.$$

4.2.2 Training with Latent Swapping

The objective was to disentangle the head pose and gaze direction, two distinct but visually coupled factors. While the latent swapping strategy used in Section 3.2.2 was sufficient to separate gaze from other image semantics, directly extending the same approach to two correlated factors was not guaranteed to yield independently steerable representations. Joint swapping of gaze direction and head pose provides only implicit supervision, as the model may encode a combined factor that explains both variations simultaneously, thereby satisfying reconstruction objectives without enforcing true intra-factor independence.

As a baseline, the model was trained using the same joint swapping strategy as in Section 3.2.2, where both gaze-related and head-pose-related latent dimensions were swapped simultaneously between image pairs (joint swapping). This exposed the model to a wide range of gaze–head-pose combinations and encouraged disentanglement through combinatorial coverage of the data distribution. Such implicit intervention leverages the dataset's natural diversity but does not explicitly expose the model to independent control over the two factors. As a result, the model could still fail to support independent manipulation of one factor without affecting the other.

To explicitly enforce disentanglement, a structured supervision scheme based on 2×2 sub-batches was introduced. Each sub-batch consisted of four images, I_i with $i = 1, 2, 3, 4$, arranged to form all combinations of two gazes, g, and two head poses, h:

$$I_1 = (g_1, h_1), \quad I_2 = (g_1, h_2), \quad I_3 = (g_2, h_1), \quad I_4 = (g_2, h_2).$$

This construction ensured that, for each image, there existed one counterpart sharing the same gaze but a different head pose, another sharing the same head pose but a different gaze, and a fourth image sharing neither factor.

To form such sub-batches in practice, gaze and head-pose values were discretized into 1.5° bins. Candidate image pairs were selected such that samples intended to share a factor were grouped within the same bin, while samples intended to differ were separated by a minimum angular distance of 2.5° in the corresponding factor. This procedure ensured controlled similarity along one factor while maintaining sufficient separation along the other, of which the values of both factors were empirically chosen. However, 2x2 sub-batch training was performed only for the single-identity dataset due to limited coverage of head poses and gazes within each identity in the multi-identity dataset.

Within each 2×2 sub-batch, three types of latent swaps were performed: gaze-only swaps between samples sharing head pose, head-only swaps between samples sharing gaze, and joint swaps between diagonally opposite samples. The resulting swapped latent codes were decoded and compared against the corresponding target images using an L_1 loss. This imposed explicit counterfactual constraints, which cannot be satisfied unless gaze and head pose are encoded in separate latent subspaces.

To extend to the head-pose factor, an additional label loss was added for the predefined head-pose dimensions, analogous to the gaze anchoring label loss in Section 3.2.2. Moreover, a Laplacian-based edge loss, $\mathcal{L}_{\text{laplace}}$, was added to emphasize high-frequency structures, which was introduced in particular to discourage blurred CRs. The Laplacian-based loss was computed by convolving a discrete Laplacian operator to both the reconstructed image and the target image and minimizing an L_1 loss between the resulting edge responses. Consequently, the new total loss is computed as:

$$\mathcal{L}_{total} = \lambda_{\text{laplace}}\mathcal{L}_{\text{laplace}} + \lambda_{\text{label}}\mathcal{L}_{\text{label}} + \lambda_{\text{swap}}\mathcal{L}_{\text{swap}} + \lambda_{\text{self}}\mathcal{L}_{\text{self}},$$

where λ_{laplace}, λ_{label}, λ_{swap} and λ_{self} are the weights for each loss.

For the single-identity case, the model was primarily trained using self-reconstruction and joint-swap supervision on randomly sampled image pairs, ensuring coverage of the full data distribution while preserving high-frequency visual realism. In addition, a subset of training iterations (30% of batches) employed structured 2×2 sub-batches to encourage independent encoding of gaze and head pose. For the multi-identity dataset, image pairs were explicitly sampled within each identity so that the same supervision losses could be applied between ground-truth and generated images;

however, no explicit supervision through isolated gaze-only or head-only swapping was used.

4.3 Evaluation

The models were evaluated with the same procedure as in Section 3, with the addition of evaluating the controllability of head pose. Moreover, one identity was left out of the training set, allowing the model to be evaluated on how well it generalized to unseen identities. Finally, the baseline performance of the CR detector itself was only assessed on the single-identity but not on the multi-identity dataset, due to reliable results in the gaze-only and single-identity head and gaze datasets, as seen in Sections 3.4.4 and 4.4.4.

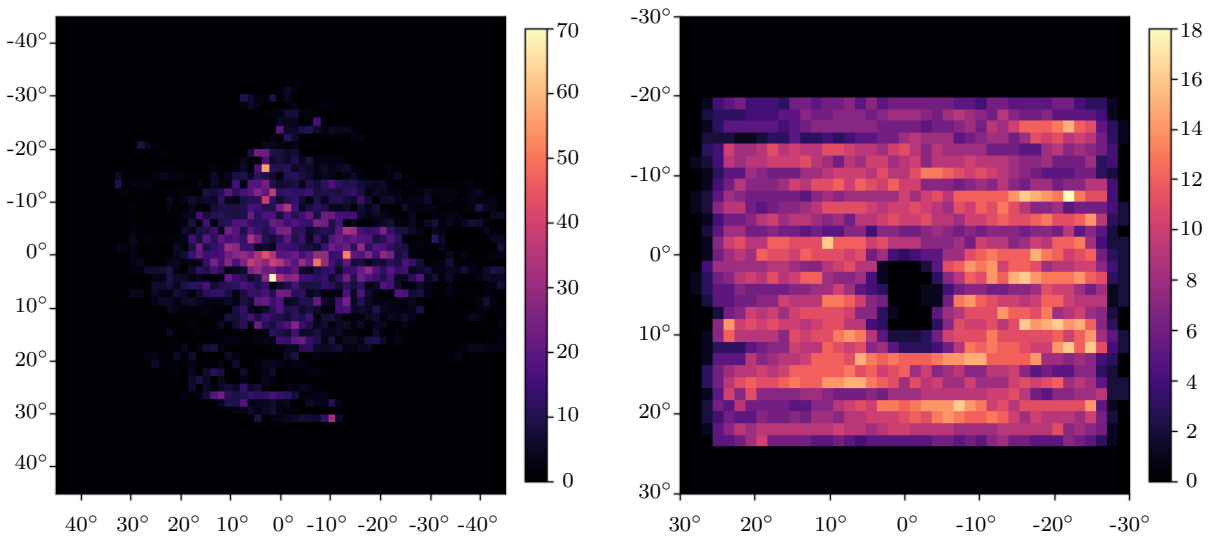
4.4 Results for Single-Identity Experiment

The results shown in this section were generated with the ResNet-based autoencoder using a latent dimensionality of 128, with λ_{self} and λ_{swap} set to 1, λ_{label} set to 0.001, and λ_{laplace} set to 0.25. The model was trained for 400 epochs with a batch size of 8 and a learning rate of 0.0005 on the single-identity dataset, using the 2×2 sub-batch intervention.

4.4.1 Dataset Coverage

The dataset consisted of 20 video recordings of a single participant performing the experiment. The raw dataset contains approximately 18 000 frames, of which 9 109 remain after filtering and post-processing.

The resulting head-pose distribution is concentrated around 0°, with decreasing coverage towards the tails. While head poses are observed near ±35° in pitch and approximately −35° to 45° in yaw, these regions remain sparse and asymmetric due to filtering of extreme relative angles between head pose and gaze. The overall coverage is shown in Figure 4.3a. Furthermore, the gaze distribution broadly follows that of the gaze-only dataset, with additional local density variations caused by overlapping gaze patterns across recordings. This distribution is illustrated in Figure 4.3b.



(a) Heatmap over head poses.

(b) Heatmap over gazes.

Figure 4.3: Coverage distributions shown as heat maps of the single-identity dataset for head pose (left) and gaze directions (right) across yaw (x-axes) and pitch (y-axes) angles with the colorbars describing the correspondence between number of frames for each combination and color in the heatmap.

4.4.2 Reconstruction Quality

The outputs produced were perceptually similar with slight smoothing of high-frequency details. Reconstructions of the participant in the single-identity dataset can be seen in Figure 4.4. Furthermore, the ablation in Section A.3.4 shows that introducing the Laplacian loss improved reconstruction quality quantitatively, given by the confusion matrices in Figure A.13 in the Appendix, together with the visual difference in Figure A.14. The CRs are visually present, with slight blurriness and inconsistencies.

Figure 4.4: Reconstructions (Recon) from the single-identity model of the GT images in the test set.

Furthermore, Figures A.7 and A.8 qualitatively show that the lean autoencoder pro-

duces low-quality images without CRs, indicating that the architecture is insufficient for the increased data complexity. Quantitative results in Table A.1 and Figure A.9 in Appendix A.3.2 further show that the ResNet-based autoencoder reduces the CR localization error from 1.53 to 0.88 pixels, while increasing accuracy from 0.40 to 0.93 and the F1 score from 0.50 to 0.96.

4.4.3 Disentanglement

The model yields smooth controllability within the training distribution as can be observed in Figures 4.5 and 4.6. However, artifacts such as eyelid inconsistencies and blurriness do appear, mainly in the tails of the training distribution for head pose, as seen in the images in the outer rows and columns of Figure 4.6.

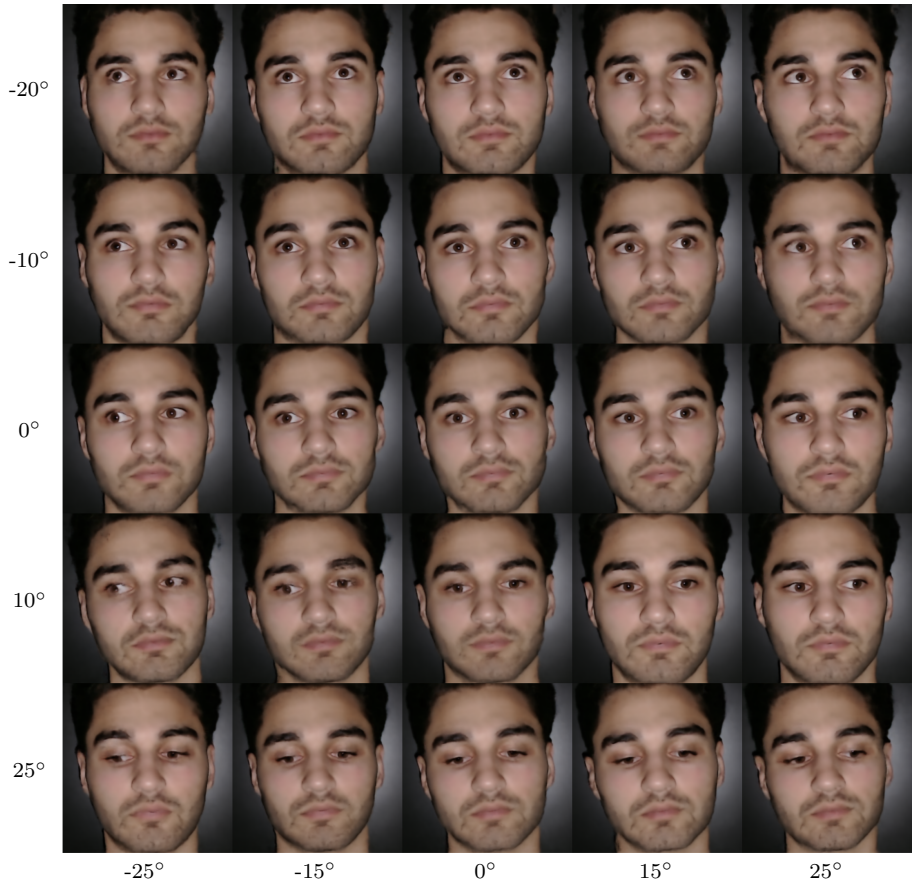


Figure 4.5: Gaze direction sweep for the single-identity model over pitch (rows) and yaw (columns).

For angles outside the defined training distribution, the single-identity model does not produce physically plausible redirections as can be observed in the example in Figure 4.7. The images quickly became noisy and grainy as redirection was performed closer to or beyond the tails of the dataset's coverage.

The latent encoding accuracy for control factors was evaluated by comparing the first four latent dimensions against the corresponding ground truth labels for yaw

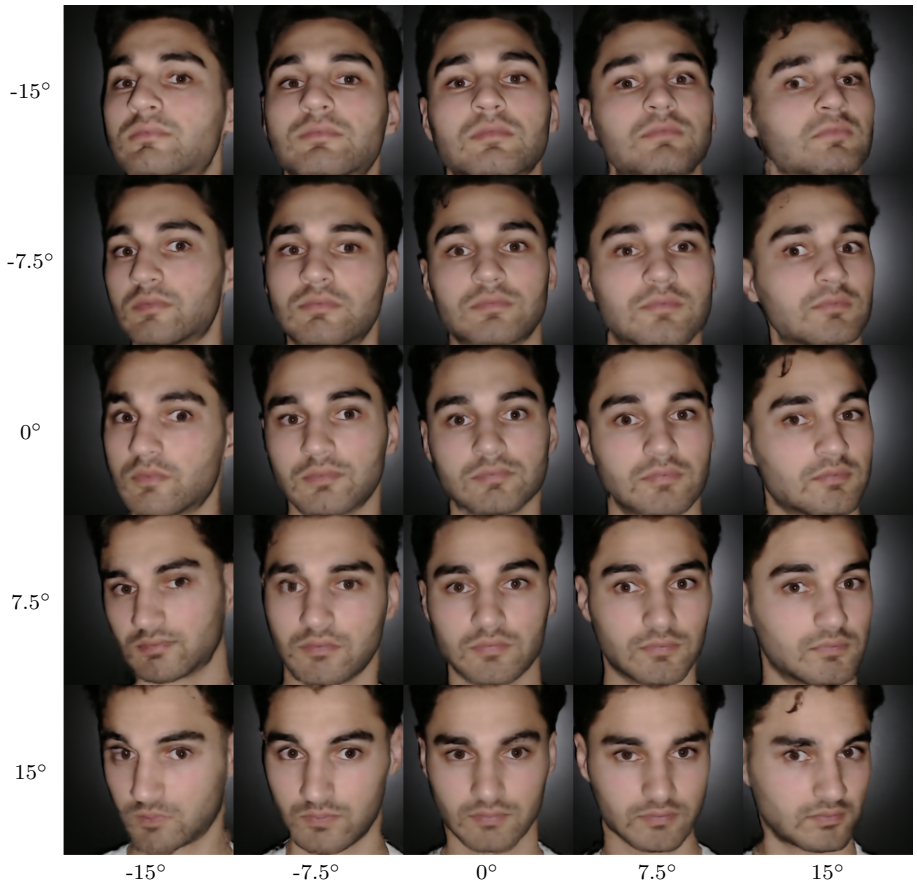


Figure 4.6: Head-pose sweep for the single-identity model over pitch (rows) and yaw (columns) angles ranging from $-15°$ to $15°$, respectively.

and pitch of gaze and head pose, respectively. The model achieves lower errors for head-pose parameters than for gaze parameters, with MAE for head-pose encoding of 0.21° for both yaw and pitch, and MAE for gaze encoding of 1.27° and 0.38° for yaw and pitch, respectively.

The impact of the structured 2×2 sub-batches, introduced in Section 4.2.2, was evaluated through an ablation study, see Appendix A.3.3. A model trained without the structured sampling resulted in comparable visual controllability and similar artifact patterns in both gaze and head-pose sweeps. No consistent improvement in reconstruction stability was observed across the evaluated range, and latent encoding errors remained similar between the two configurations.

4.4.4 CR Detection

Similarly to Section 3.4.4, the detector was validated against the ground truth test images in the head-gaze dataset, yielding identical performance with an accuracy of 0.98 and an F1 score of 0.99. This consistency indicates that the detector remains reliable for this new dataset.

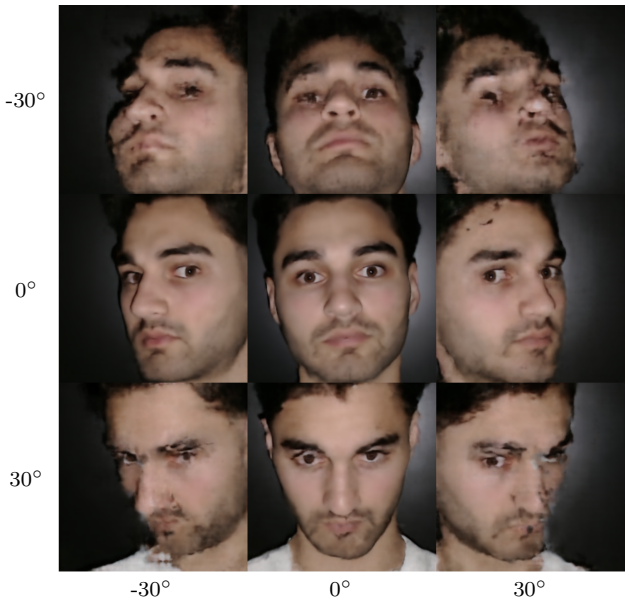


Figure 4.7: Head-pose sweep for the single-identity model over pitch (rows) and yaw (columns) angles outside of training distribution.

For the redirected and reconstructed images, the CR localization MAE was 0.88 pixels, with an accuracy of 0.93 and an F1 score of 0.96. The corresponding confusion matrix for CR detections is shown in Figure 4.8.

4.5 Results for Multi-Identity Experiment

The results shown in this section were generated with the same model architecture and training methodology as for Section 4.4, except for the sub-batch intervention. The model was trained on the multi-identity dataset, with one identity excluded.

4.5.1 Dataset Coverage

The multi-identity dataset consists of 25 participants, each recorded in five separate videos due to recording time constraints. After filtering and post-processing, the dataset contains 55 418 frames. Moreover, Figure 4.9a illustrates the overall head-pose coverage, and, as in the single-identity dataset, the distribution is centered around 0° with reduced coverage towards more extreme poses. However, a pronounced high-density region, of approximately 600 frames, is observed near 0° yaw and −10° pitch. This concentration is primarily caused by subjects exhibiting limited head movement. For instance, identity 7 shows minimal variation in head pose, as illustrated in Figure A.3 in the Appendix, where a large proportion of samples are clustered around 0° yaw and −10° pitch. Comparing these counts to the overall distribution indicates that identity 7 contributes significantly to this high-density region.

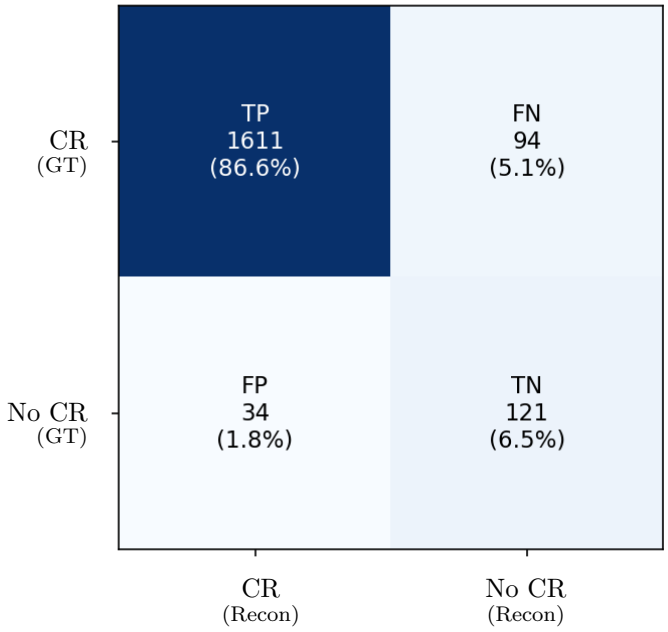
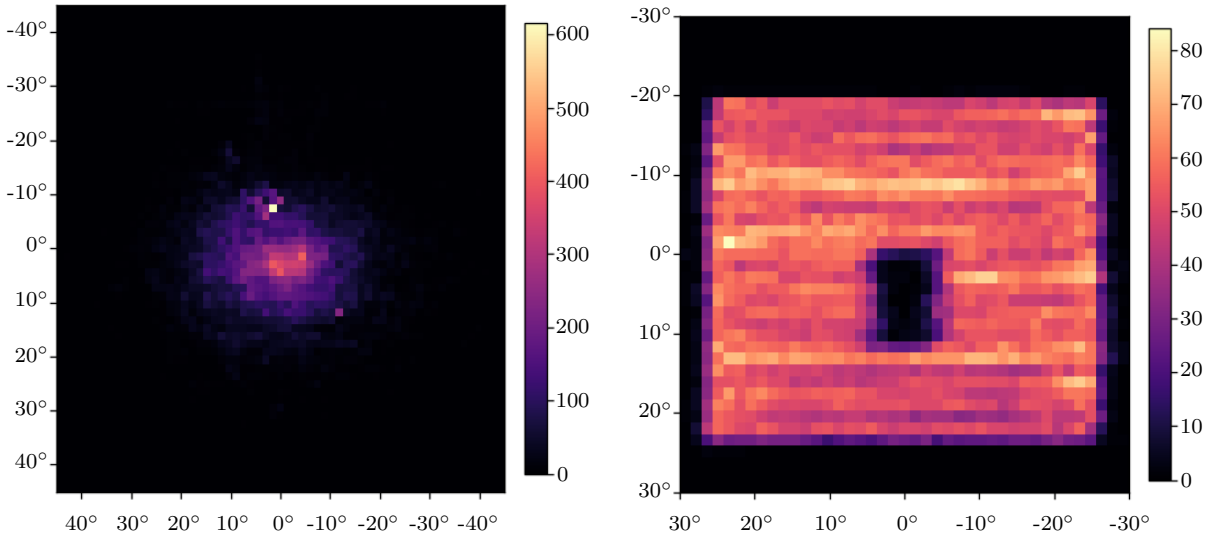


Figure 4.8: Confusion matrix of CR detections combined over left and right eyes on the ResNet autoencoder trained on the single-identity dataset.



(a) Heatmap over head poses.

(b) Heatmap over gazes.

Figure 4.9: Coverage distributions shown as heat maps of the multi-identity dataset for head pose (left) and gaze directions (right) across yaw (x-axes) and pitch (y-axes) angles with the colorbars describing the correspondence between number of frames for each combination and color in the heatmap. The heatmaps show the data collected from all participants combined.

In addition to head pose, the gaze coverage of the multi-identity dataset is approximately uniform, with a central region lacking samples due to camera-setup occlusion, as described in Section 3.4.1. The overall gaze distribution is shown in Figure 4.9b.

Local variations in density are observed due to overlapping gaze patterns across subjects. Furthermore, filtering introduces irregular sparsity, making these variations less predictable.

4.5.2 Reconstruction Quality

Image reconstruction is perceptually similar to that of the single-identity model, aside from the smoothing becoming slightly more prominent, as can be seen in Figures 4.10. Nonetheless, CRs are well-defined, and the model learns to reconstruct all identities in the dataset to a similar level.

Figure 4.10: Reconstructions using the multi-identity model on arbitrarily chosen test images of seen identities.

Additionally, the model was evaluated over the reconstruction and redirection for an identity not exposed to the model during training. The reconstructions of this identity can be observed in Figure 4.11, which showcases identity interpolation. The eye region and other facial features are visually clear with blurry artifacts, particularly around the head and background.

Figure 4.11: Multi-identity model producing reconstructions (bottom row) of the unseen identity (top row).

4.5.3 Disentanglement

The multi-identity model achieved similar controllability to the single-identity model for each identity, with slightly narrower controllability for head pose while maintaining reconstruction quality due to dataset coverage. This can be seen in Figure 4.12 and in Figure A.2 for sweeps over additional identities.

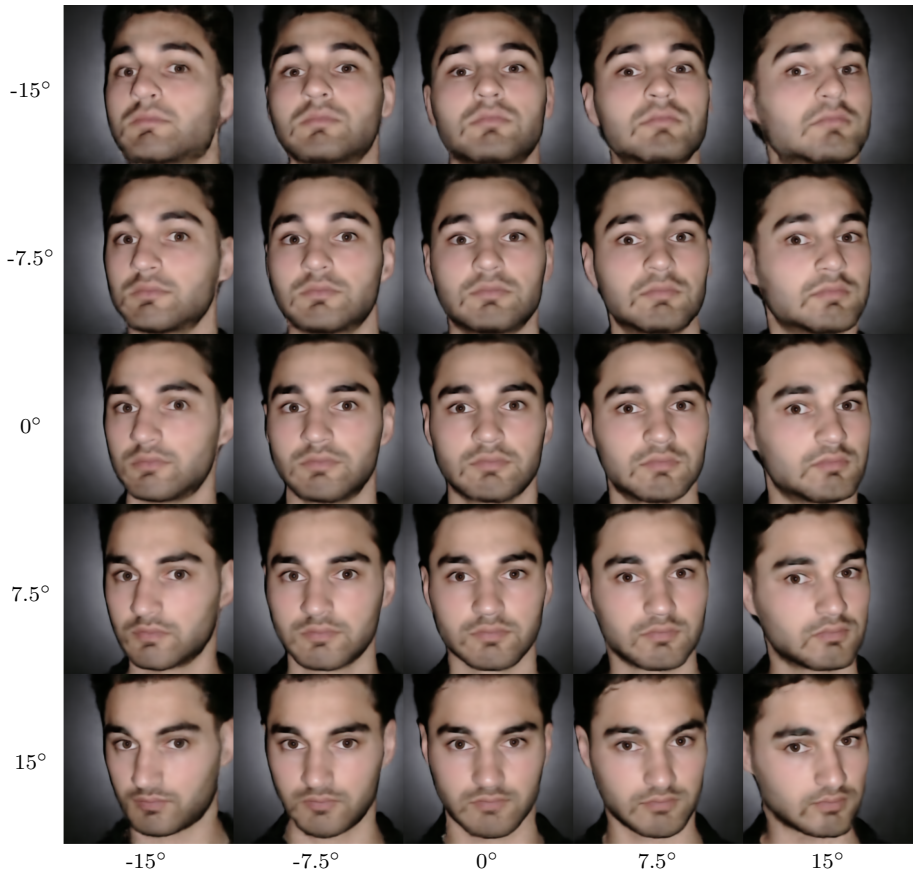


Figure 4.12: Head-pose sweep for the multi-identity model over pitch (rows) and yaw (columns) angles ranging from $-15°$ to $15°$, respectively.

The multi-identity model reaches MAE for gaze encoding of $0.76°$ and $0.41°$ for yaw and pitch, respectively, and MAE for head-pose encoding of $0.25°$ and $0.28°$ for yaw and pitch, respectively. Thereby, the multi-identity model lowers the latent encoding error of the control factors for gaze compared to the single-identity model.

For the unseen identity, while the model produces erroneous images, the identity remains visually consistent throughout the redirection, with predictable behavior over the gaze and head-pose sweeps seen in Figures 4.13 and 4.14. Moreover, the multi-identity model achieves MAEs of $2.98°$ and $3.56°$ for gaze encoding (yaw and pitch, respectively) and $0.93°$ and $1.98°$ for head-pose encoding (yaw and pitch, respectively).

4.5.4 CR Detection

The multi-identity dataset produced two types of results: from the in-distribution test sets and from the unseen identity. The in-distribution results were better than the results of the unseen person, achieving an accuracy of 0.87 and an F1 score of 0.92. Furthermore, CR localization yielded a MAE of 0.99 pixels. Conversely, the unseen test data achieved an accuracy of 0.67 and an F1 score of 0.79, with a CR localization MAE of 1.8 pixels.

Figure 4.13: Gaze direction sweep for the multi-identity model for the unseen identity over pitch (rows) and yaw (columns) angles ranging from $-20°$ to $20°$, respectively.

In addition, the confusion matrices for the in-distribution test sets and the unseen dataset are visualized in Figure 4.15.

4.6 Discussion

The results of the head and gaze redirection experiment demonstrate that an autoencoder-based disentanglement framework can be extended from a constrained gaze-only setting to a more complex scenario involving joint control of head pose and gaze. Compared with the first experiment, introducing head pose as an additional factor significantly increases both the visual complexity and the ambiguity in the data. Nevertheless, the model can learn a latent representation that enables controllable manipulation of both factors while preserving key aspects of facial appearance.

The inclusion of a Laplacian-based edge loss improved the preservation of high-frequency image structures, as reflected quantitatively in improved CR detection metrics in Section A.3.4. While the visual impact of this loss is subtle and not always immediately apparent in qualitative inspection, the improved F1 score and

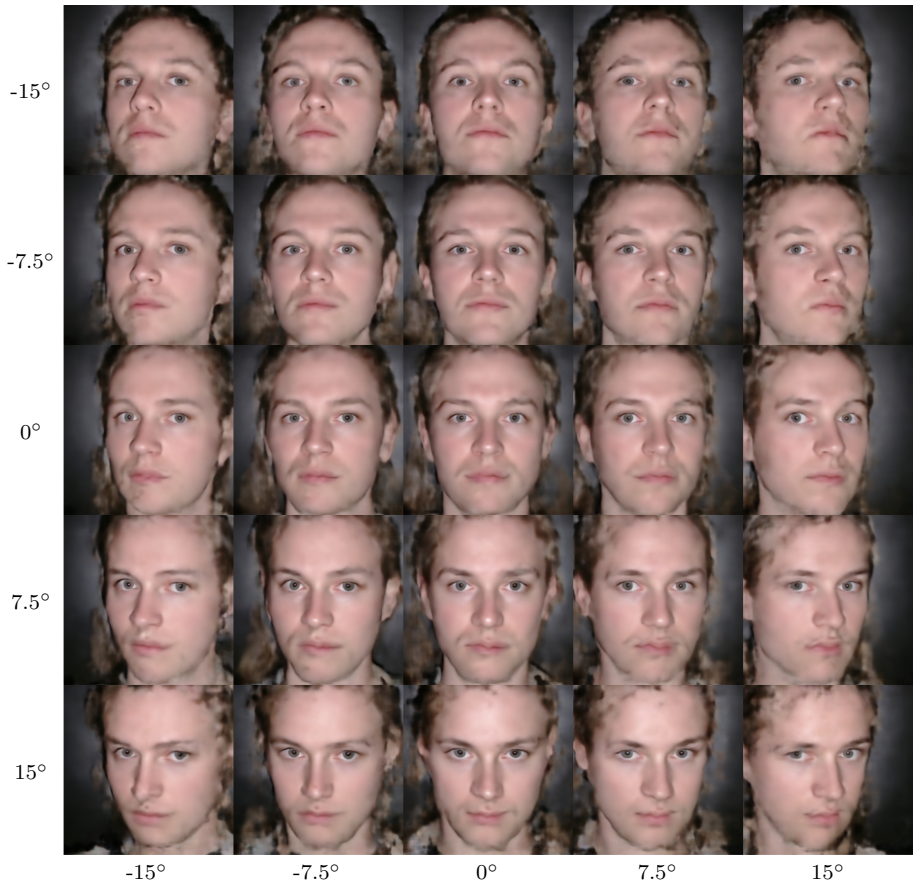
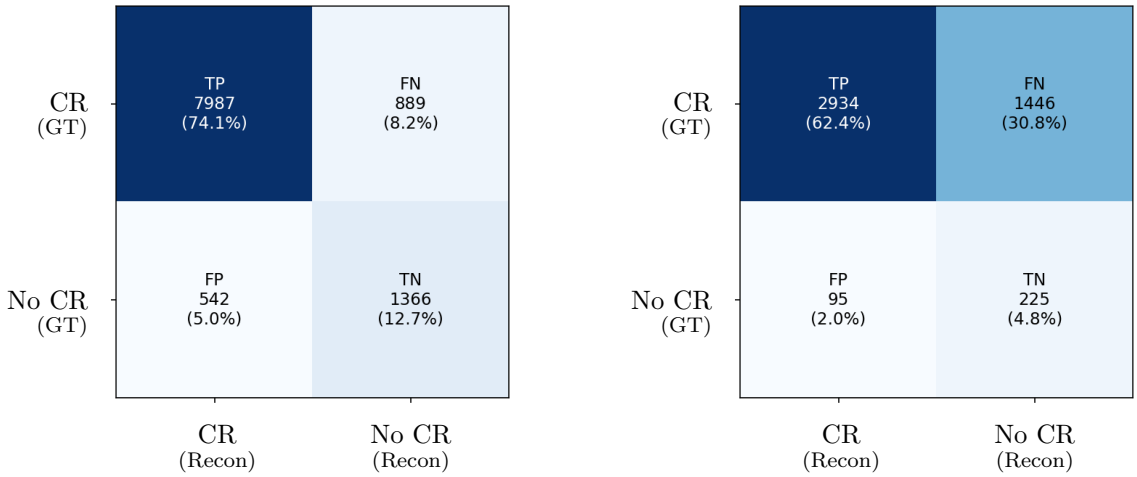


Figure 4.14: Head-pose sweep for the multi-identity model for the unseen identity over pitch (rows) and yaw (columns) angles ranging from $-15°$ to $15°$, respectively.

reduced localization errors indicate that edges and specular structures are more consistently preserved. From a practical perspective, this suggests that auxiliary edge-based losses can help stabilize fine-grained image details that are otherwise smoothed out by pixel-wise reconstruction objectives. Differences in F1 score primarily reflect changes in false negatives and false positives in CR detection, which, in turn, indicate whether the reconstructed images retain sufficiently sharp highlights for reliable detection. In this sense, CR detection serves as an indirect but informative proxy for the physical plausibility of eye appearance.

For the single-identity setting, the model exhibits a noticeable smoothing effect in the reconstructed and redirected images. This effect is more pronounced than in the gaze-only experiment and is particularly evident around hair, eyelids, and fine facial contours. A likely explanation is that multiple frames with nearly identical gaze and head pose labels originate from different video segments and therefore differ slightly in unlabeled factors such as eyelid state, facial expression, hair configuration, or small amounts of roll. When trained with pixel-based reconstruction losses, the autoencoder is encouraged to average over these variations, leading to a characteristic smoothing or "blending" effect. This behavior can also be interpreted as a form of lookup-table averaging, in which visually distinct inputs associated with



(a) CR detection over the test sets across identities seen during training.

(b) CR detection performed on the unseen identity.

Figure 4.15: Confusion matrix comparison between the multi-identity model performance over reconstructions redirected identities within the training distribution (4.15a) and for the unseen identity (4.15b). The rows indicate the GT and the columns are the reconstructions.

similar labels are mapped to a common latent representation and then decoded into an averaged output.

While the smoothing limits realism, it also implies that the model learns an interpolative representation over images sharing similar control parameters. As a result, the autoencoder can generalize across the label space, reconstructing plausible images even when the exact visual configuration was not observed during training. This trade-off highlights a fundamental property of autoencoder-based models trained with pixel losses, where they favor stable interpolation over sharp reproduction of characteristic details. In practice, this suggests that increasing data diversity or introducing perceptual or identity-preserving losses would be required to mitigate excessive smoothing without sacrificing controllability.

Although the structured 2×2 sub-batch scheme was motivated by a theoretically stronger disentanglement signal, its practical effect in this project was limited. In principle, such sub-batches provide explicit counterfactual supervision and strengthen the independent disentanglement of gaze and head pose. However, in the present setting, only two controllable factors were considered, and the dataset already provided substantial natural combinatorial coverage of these factors during standard training. This likely made the additional 2×2 supervision largely redundant in practice. At the same time, because the dataset did not contain exact combinatorial quadruplets, the 2×2 sub-batches had to be constructed through angular binning, such that samples intended to share gaze or head pose were only approximately matched. Combined with pixel-wise reconstruction losses, these residual mismatches introduced additional supervision noise, which likely contributed more

to smoothing and reduced image fidelity than to improved disentanglement. Nevertheless, explicit factor-isolating supervision is likely to become more important in more complex settings, for example, when disentangling a larger number of factors that co-vary in the training data or when the training data does not naturally provide sufficient combinatorial coverage. In such cases, especially when paired with perceptual or feature-based losses that are less sensitive to nuisance mismatches than strict pixel losses, structured supervision could offer a clearer advantage.

Despite these limitations, the single-identity model demonstrates controllability over both gaze and head pose, as evidenced by the sweeps across the training distribution. The smooth transitions observed in these sweeps indicate that disentanglement of multiple, visually coupled factors is feasible when sufficiently annotated data is available. Quantitatively, the latent encoding accuracy is higher for head pose than for gaze, which is consistent with intuition as head pose affects a larger portion of the image and introduces coarser, more global appearance changes. Meanwhile, gaze primarily alters fine-scale structures in the eye region.

As in the gaze-only experiment and as expected, the model fails to produce physically plausible outputs when queried outside the training distribution. For extreme head poses or gaze angles not represented in the data, the reconstructions degrade rapidly, exhibiting noise and artifacts. This collapse underscores the strong dependence of autoencoder-based models on the support of the training distribution and highlights their limited ability to extrapolate. Without explicit geometric priors or physically grounded constraints, the model cannot be expected to generalize reliably beyond the range of observed configurations.

The multi-identity setting exhibits qualitatively similar reconstruction behavior, with smoothing effects present across all training identities. Importantly, the model maintains consistent gaze and head-pose controllability for each identity in the training set, albeit with a narrower effective range for head pose due to uneven dataset coverage. Quantitative results show that the latent encoding accuracy for gaze improves compared to the single-identity case, suggesting that exposure to multiple identities helps regularize the latent representation and reduce overfitting to individual appearance cues.

For an identity not observed during training, the model produces reconstructions that retain coarse identity characteristics but exhibit substantial distortions, particularly outside the central facial region. The reconstructed face can be interpreted as an interpolation of identities in the training set that are perceptually similar in terms of skin tone, facial symmetry, and eye appearance. Hair and background regions, which are less constrained by the training data and more variable across identities, are especially prone to artifacts. Quantitatively, the latent encoding error for both gaze and head pose increases significantly for the unseen identity, yet the model still demonstrates a degree of transferable utility as gaze and head-pose sweeps remain predictable and structured, even if not fully accurate. This behavior suggests that identity generalization in autoencoder-based models improves gradually as the di-

versity of identities in the training set increases. With a sufficiently large and varied identity pool, the latent space could be expected to better capture identity-invariant structure while reducing reliance on averaging across dissimilar appearances. However, without explicit identity supervision or disentanglement, the model inherently interpolates within the training identity distribution rather than synthesizing truly novel identities.

From an application perspective, these results indicate that autoencoder-based disentanglement is a viable approach for generating controlled variations of gaze and head pose in domains such as DMS, provided that the operational conditions closely match the training distribution. The preservation of CRs and the ability to control gaze and head pose suggest potential utility for synthetic data generation and augmentation. At the same time, the observed smoothing effects, limited extrapolation, and challenges with unseen identities highlight the limitations of purely image-based autoencoder models. For deployment in safety-critical domains, additional constraints, richer supervision, or hybrid approaches incorporating geometric or temporal information would likely be required to achieve robust performance across the full range of real-world variability.

5
Future work

This project has primarily focused on the controllability of the latent space. However, several directions remain to improve both image quality and generalization. In particular, the distortions observed when introducing unseen identities are not present for identities within the training distribution. This suggests that the limitation is not primarily due to insufficient disentanglement, but rather to limited identity diversity in the training data, combined with the deterministic nature of the autoencoder. A natural direction, therefore, is to improve generalization to new identities. Inspired by FAZE, which, in addition to a disentangling autoencoder mentioned in Section 2.5.2, learns a person-specific gaze estimation model from only a few calibration samples. A similar lightweight adaptation step could be explored to adapt the model to new identities [32].

Regarding reconstruction quality, the current model exhibits smoothing artifacts commonly associated with pixel-wise losses. Theory and related works utilizing a PatchGAN refiner indicate improved high-frequency detail, although this comes at the cost of introducing minor artifacts [52]. In addition, adversarial refinement complicates pixel-level evaluation, such as CR localization, since the outputs are no longer strictly aligned with the input at a pixel level.

Another interesting direction is the use of more expressive decoders. Work combining autoencoders with diffusion-based decoders has demonstrated high-fidelity reconstructions in related domains [34]. These approaches typically rely on hybrid latent representations that combine deterministic encodings with stochastic components sampled from a prior distribution. Exploring similar hybrid formulations in this setting could improve reconstruction quality while preserving controllability, although their applicability to manipulations such as gaze and head-pose redirection remains unexplored to the best of our knowledge.

Due to data access restrictions, standard open-source datasets and commercial evaluation tools could not be used in this work. Future work should therefore incorporate task-specific losses, such as a gaze consistency loss, to ensure alignment between generated outputs and condition labels. In addition, incorporating identity-based losses using pre-trained identity encoders could reduce the reliance on pixel-wise supervision and improve perceptual quality while preserving identity.

From a data perspective, applying the proposed approach to real DMS footage is a necessary step toward evaluating its practical applicability. One possible direction is

to leverage state-of-the-art gaze and head-pose estimators to automatically annotate larger driver-cabin datasets. In addition, modern multi-camera DMS configurations enable richer forms of data acquisition. For example, synchronized observations from multiple viewpoints could capture the same subject simultaneously, facilitating the creation of geometrically consistent datasets spanning a broader range of head poses and gaze directions. Such datasets could support more realistic synthesis approaches that explicitly incorporate 3D spatial relationships and viewpoint consistency, as discussed in Section 2.5.2.

Overall, improving generalization to unseen identities, enhancing high-frequency reconstruction quality, and incorporating stronger supervision signals remain key challenges for future work.

6
Conclusion

This thesis showed that gaze direction and head pose can be controllably manipulated in image space using a disentangled encoder-decoder framework while producing photorealistic images. Within the training distribution, latent swapping combined with pixel-wise reconstruction losses enabled consistent control of gaze and head pose while preserving key visual features such as corneal reflections.

The main contribution of this work is a controlled experimental framework for studying image-based gaze and head-pose redirection across increasing levels of complexity, from a single-identity gaze-only setting to a multi-identity setting with joint head-pose and gaze variation. A central part of this contribution is the design of a dedicated data-collection pipeline in which participants perform regulated smooth pursuit under a known geometric setup, enabling dense and continuous gaze labels. In addition, the thesis contributed a task-relevant evaluation strategy based on CR detection, which complements standard reconstruction and encoding metrics, and showed that a Laplacian-based edge loss improves preservation of high-frequency details even when the visual effect is subtle.

At the same time, the results highlight clear limitations of deterministic autoencoder-based models. The models relied strongly on the support of the training distribution, exhibited smoothing due to pixel-wise losses, and degraded for unseen identities and out-of-distribution configurations. Overall, this work establishes image-based disentanglement as a promising framework for controllable gaze and head-pose manipulation and a meaningful step toward synthetic data generation for driver monitoring systems.

Bibliography

[1] M. Abdolahnejad and P. X. Liu, "A deep autoencoder with novel adaptive resolution reconstruction loss for disentanglement of concepts in face images," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–13, 2022. doi:10.1109/TIM.2022.3165261.

[2] A. Aghamalizadeh, A. Mazloumi, A. Nikabadi, A. Nahvi, F. Khanehshenas, and S. Ebrahimian. *Driver drowsiness is associated with altered facial thermal patterns: Machine learning insights from a thermal imaging approach*. Physiology & Behavior, vol. 283, p. 114619, 2024. doi: https://doi.org/10.1016/j.physbeh.2024.114619. Available at: https://www.sciencedirect.com/science/article/pii/S0031938424001641 (Accessed: 24 Feb. 2026).

[3] National Highway Traffic Safety Administration, *Alcohol-Impaired Driving: 2023 Data*, U.S. Department of Transportation, Report DOT HS 813 713, 2025. Available: https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/813713 (Accessed: 26 May 2026).

[4] Trafikverket (Swedish Transport Administration). *Alkohol och narkotika i vägtrafiken*. Web page, last updated 22 Sep. 2025. Available at: https://www.trafikverket.se/resa-och-trafik/trafiksakerhet/sakerhet-pa-vag/alkohol-och-narkotika-i-vagtrafiken/ (Accessed: 18 Feb. 2026).

[5] Euro NCAP. *Assessment Protocol – Safety Assist: Safe Driving (v10.4), Implementation 2023*. Version 10.4, February 2024. Available at: https://www.euroncap.com/media/80158/euro-ncap-assessment-protocol-sa-safe-driving-v104.pdf (Accessed: 18 Feb. 2026).

[6] P. Bérard, D. Bradley, M. Nitti, T. Beeler, and M. Gross. *High-Quality Capture of Eyes*. ACM Transactions on Graphics (Proc. SIGGRAPH Asia), vol. 33, no. 6, Article 223, 2014. Available at: https://la.disneyresearch.com/wp-content/uploads/High-Quality-Capture-of-Eyes-Pub-Paper.pdf (Accessed: 18 Feb. 2026).

[7] S. A. Byrne, M. Nyström, V. Maquiling, E. Kasneci, and D. C. Niehorster. *Precise localization of corneal reflections in eye images using deep learning trained on synthetic data*. Behavior Research Methods, vol. 56, no. 4, pp. 3226–3241, 2023. doi: https://doi.org/10.3758/s13428-023-02297-w. Available at: http://dx.doi.org/10.3758/s13428-023-02297-w (Accessed: 23 Mar. 2026).

[8] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko. *End-to-end object detection with transformers*. CoRR, vol. abs/2005.12872, 2020. Available at: https://arxiv.org/abs/2005.12872 (Accessed: 17 Apr. 2026).

[9] European Commission. *Commission Delegated Regulation (EU) 2021/1341 of 23 April 2021 supplementing Regulation (EU) 2019/2144 by laying down detailed rules concerning the specific test procedures and technical requirements for the type-approval of motor vehicles with regard to their driver drowsiness and attention warning systems and amending Annex II to that Regulation*. Official Journal of the European Union, L 292, 16 Aug. 2021. Available at: `https://eur-lex.europa.eu/eli/reg_del/2021/1341/oj/eng` (Accessed: 18 Feb. 2026).

[10] European Commission. *Commission Delegated Regulation (EU) 2023/2590 of 13 July 2023 supplementing Regulation (EU) 2019/2144 by laying down detailed rules concerning the specific test procedures and technical requirements for the type-approval of certain motor vehicles with regard to their advanced driver distraction warning systems and amending that Regulation*. Official Journal of the European Union, 2023/2590, 22 Nov. 2023. Available at: `https://eur-lex.europa.eu/eli/reg_del/2023/2590/oj/eng` (Accessed: 18 Feb. 2026).

[11] M. Eizenman, E. D. Guestrin, and J. H. Chin. *Eye tracking using a hybrid eye model*. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2009.

[12] Smart Eye. *Eye tracking*. Smart Eye Technology Overview. Available at: `https://smarteye.se/technology/eye-tracking/` (Accessed: 23 Apr. 2026).

[13] MediaPipe. *Face Mesh*. Read the Docs. Available at: `https://mediapipe.readthedocs.io/en/latest/solutions/face_mesh.html` (Accessed: 12 May 2026).

[14] Y. Feng, D. J. X. Ng, and A. Easwaran, "Improving variational autoencoder based out-of-distribution detection for embedded real-time applications," *ACM Trans. Embed. Comput. Syst.*, vol. 20, no. 5s, Art. no. 95 (October 2021), 2021. doi:10.1145/3477026.

[15] Y. Ganin, D. Kononenko, D. Sungatullina, and V. S. Lempitsky. *DeepWarp: Photorealistic image resynthesis for gaze manipulation*. CoRR, vol. abs/1607.07215, 2016. Available at: `http://arxiv.org/abs/1607.07215` (Accessed: 22 Apr. 2026).

[16] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016. Available at: `http://www.deeplearningbook.org` (Accessed: 23 Mar. 2026).

[17] E. D. Guestrin and M. Eizenman. *General theory of remote gaze estimation using the pupil center and corneal reflections*. IEEE Transactions on Biomedical Engineering, vol. 53, no. 6, pp. 1124–1133, 2006. doi: `https://doi.org/10.1109/TBME.2005.863952`.

[18] H. Hanafi, A. Pranolo, and Y. Mao, "CAE-COVIDX: automatic COVID-19 disease detection based on X-ray images using enhanced deep convolutional and autoencoder," *International Journal of Advances in Intelligent Informatics*, vol. 7, no. 1, pp. 49–62, 2021. doi:10.26555/ijain.v7i1.577.

[19] D. W. Hansen and Q. Ji. *In the eye of the beholder: A survey of models for eyes and gaze*. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 32, no. 3, pp. 478–500, 2010.

[20] Z. He, A. Spurr, X. Zhang, and O. Hilliges. *Photo-realistic monocular gaze redirection using generative adversarial networks*. CoRR, vol. abs/1903.12530,

2019. Available at: http://arxiv.org/abs/1903.12530 (Accessed: 22 Apr. 2026).

[21] K. He and J. Sun, "Convolutional neural networks at constrained time cost," arXiv preprint arXiv:1412.1710, 2014. Available at: https://arxiv.org/abs/1412.1710.

[22] K. He, X. Zhang, S. Ren, and J. Sun. *Deep residual learning for image recognition*. CoRR, vol. abs/1512.03385, 2015. Available at: http://arxiv.org/abs/1512.03385 (Accessed: 17 Apr. 2026).

[23] Tobii. *How eye tracking works and why it matters*. Tobii Blog, 2026. Available at: https://www.tobii.com/blog/how-eye-tracking-works (Accessed: 23 Apr. 2026).

[24] J.-J. Kang, S.-U. Lee, J.-M. Kim, and S.-Y. Oh, "Recording and interpretation of ocular movements: saccades, smooth pursuit, and optokinetic nystagmus," *Annals of Clinical Neurophysiology*, vol. 25, no. 2, pp. 55–65, 2023. doi:10.14253/acn.2023.25.2.55

[25] R. Lencer, A. Sprenger, and P. Trillenberg, "Smooth Eye Movements in Humans: Smooth Pursuit, Optokinetic Nystagmus and Vestibular Ocular Reflex," in *Eye Movement Research*, Springer, Cham, 2019, pp. 117–163. doi:10.1007/978-3-030-20085-5_4.

[26] P. Li, Y. Pei, and J. Li. *A comprehensive survey on design and application of autoencoder in deep learning*. Applied Soft Computing, vol. 138, p. 110176, 2023. doi: https://doi.org/10.1016/j.asoc.2023.110176. Available at: https://www.sciencedirect.com/science/article/pii/S1568494623001941 (Accessed: 23 Mar. 2026).

[27] L. McIlreavy, T. C. A. Freeman, and J. T. Erichsen, "Two-dimensional analysis of smooth pursuit eye movements reveals quantitative deficits in precision and accuracy," *Translational Vision Science & Technology*, vol. 8, no. 5, p. 7, 2019. doi:10.1167/tvst.8.5.7.

[28] T. Ohno, N. Mukawa, and S. Yoshikawa. *FreeGaze: A gaze tracking system for everyday gaze interaction*. Proceedings of the ACM Symposium on Eye Tracking Research & Applications, 2002.

[29] J. Otero-Millan, D. C. Roberts, A. Lasker, D. S. Zee, and A. Kheradmand. *Knowing what the brain is seeing in three dimensions: A novel, noninvasive, sensitive, accurate, and low-noise technique for measuring ocular torsion*. Journal of Vision, vol. 15, no. 14, p. 11, 2015. doi: https://doi.org/10.1167/15.14.11. Available at: https://jov.arvojournals.org/article.aspx?articleid=2466596 (Accessed: 27 Feb. 2026).

[30] A. Palao, R. Fredriksson, and M. Lenné. *Euro NCAP's current and future in-cabin monitoring systems assessment*. Proceedings of Enhanced Safety of Vehicles (ESV), Paper No. 23-0286, 2023. Available at: https://cdn.euroncap.com/media/77181/27esv-000286.pdf (Accessed: 18 Feb. 2026).

[31] A. Panchalingam, I. Bodala, and S. Middleton. *3D Gaussian and diffusion-based gaze redirection*. CoRR, vol. abs/2511.11231, 2025. Available at: https://arxiv.org/abs/2511.11231 (Accessed: 22 Apr. 2026).

[32] S. Park, S. De Mello, P. Molchanov, U. Iqbal, O. Hilliges, and J. Kautz. *Few-shot adaptive gaze estimation*. CoRR, vol. abs/1905.01941, 2019. Available at:

http://arxiv.org/abs/1905.01941 (Accessed: 22 Apr. 2026).

[33] E. Pintelas and P. Pintelas, "A 3D-CAE-CNN model for deep representation learning of 3D images," *Engineering Applications of Artificial Intelligence*, vol. 113, p. 104978, 2022. doi:10.1016/j.engappai.2022.104978.

[34] K. Preechakul, N. Chatthee, S. Wizadwongsa, and S. Suwajanakorn. *Diffusion Autoencoders: Toward a Meaningful and Decodable Representation*. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10619–10629, June 2022.

[35] European Parliament and Council of the European Union. *Regulation (EU) 2019/2144 of 27 November 2019 on type-approval requirements for motor vehicles and their trailers, and systems, components and separate technical units intended for such vehicles, as regards their general safety and the protection of vehicle occupants and vulnerable road users*. Official Journal of the European Union, L 325, 16 Dec. 2019. Available at: https://eur-lex.europa.eu/eli/reg/2019/2144/oj/eng (Accessed: 18 Feb. 2026).

[36] European Commission. *Road safety thematic report – Fatigue*. European Road Safety Observatory, Version 1.1, January 2021. Brussels: European Commission, Directorate-General for Transport. Available at: https://road-safety.transport.ec.europa.eu/system/files/2021-07/road_safety_thematic_report_fatigue_tc_final.pdf (Accessed: 18 Feb. 2026).

[37] World Health Organization (WHO). *Road traffic injuries*. Fact sheet, 13 December 2023. Available at: https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries/ (Accessed: 18 Feb. 2026).

[38] A. Ruzzi, X. Shi, X. Wang, G. Li, S. De Mello, H. J. Chang, X. Zhang, and O. Hilliges. *GazeNeRF: 3D-aware gaze redirection with neural radiance fields*. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 9676–9685, 2023.

[39] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014. Available at: https://arxiv.org/abs/1409.1556.

[40] A. D. Smith, B. Muthumanickam, Y. Feng, W. Ding, and K. Hauser. *UnityEyes 2: Open source synthetic eye generation for camera-based eye tracking with machine learning*. Proceedings of the ACM Symposium on Eye Tracking Research and Applications (ETRA), 2025. doi: https://doi.org/10.1145/3715669.3726838.

[41] Anyverse. *Solving false positives in DMS: Synthetic data for true attention validation*. Anyverse Blog, 2026. Available at: https://anyverse.ai/solving-false-positives-in-dms-using-synthetic-data-for-true-attention-validation/ (Accessed: 23 Apr. 2026).

[42] Anyverse. *Solving the DMS validation data gap: Synthetic data to master rare driver behaviours*. Anyverse Blog, 2025. Available at: https://anyverse.ai/solving-dms-validation-data-gap-synthetic-data-master-rare-driver-behaviors/ (Accessed: 23 Apr. 2026).

[43] R. K. Srivastava, K. Greff, and J. Schmidhuber, "Highway networks," arXiv preprint arXiv:1505.00387, 2015. Available at: https://arxiv.org/abs/1505.00387.

[44] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. E. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," arXiv preprint arXiv:1409.4842, 2014. Available at: https://arxiv.org/abs/1409.4842.

[45] M. J. Thurtell, A. C. Joshi, and M. F. Walker, "Three-dimensional kinematics of saccadic and pursuit eye movements in humans: Relationship between Donders' and Listing's laws," *Vision Research*, vol. 60, pp. 7–15, 2012. doi:10.1016/j.visres.2012.02.012.

[46] E. Ververas, P. Gkagkos, J. Deng, M. C. Doukas, J. Guo, and S. Zafeiriou. *3DGazeNet: Generalizing 3D Gaze Estimation with Weak-Supervision from Synthetic Views*. In: *Computer Vision – ECCV 2024*, pp. 387–404. Springer Nature Switzerland, Cham, 2025.

[47] S. Xu, G. Chen, Y.-X. Guo, J. Yang, C. Li, Z. Zang, Y. Zhang, X. Tong, and B. Guo. *VASA-1: Lifelike Audio-Driven Talking Faces Generated in Real Time*. arXiv preprint arXiv:2404.10667, 2024. Available at: https://arxiv.org/abs/2404.10667

[48] J. Terven, D.-M. Cordova-Esparza, J.-A. Romero-González, A. Ramírez-Pedraza, and E. A. Chávez-Urbiola, "A comprehensive survey of loss functions and metrics in deep learning," *Artificial Intelligence Review*, vol. 58, no. 7, p. 195, 2025. doi:10.1007/s10462-025-11198-7.

[49] Y. Wei and X. Hu, "Pedestrian anomaly detection method using autoencoder," In: *Proc. 2nd Int. Conf. on Intelligent Computing and Human-Computer Interaction (ICHCI)*, pp. 126–129, 2021. doi:10.1109/ICHCI54629.2021.00033.

[50] Y. Wu and K. He, "Group normalization," In: *Computer Vision – ECCV 2018*, pp. 3–19. Springer, 2018. doi:10.1007/978-3-030-01261-8_1.

[51] R. D. Yee, R. A. Goldberg, O. W. Jones, R. W. Baloh, and V. Honrubia, "Effect of eccentric gaze on pursuit," *Investigative Ophthalmology & Visual Science*, vol. 24, no. 8, pp. 1108–1114, 1983.

[52] Y. Zheng, S. Park, X. Zhang, S. De Mello, and O. Hilliges. *Self-learning transformations for improving gaze and head redirection*. CoRR, vol. abs/2010.12307, 2020. Available at: https://arxiv.org/abs/2010.12307 (Accessed: 22 Apr. 2026).

A
Appendix 1

A.1 Results

Figure A.1: Dense sweep over gaze angles for pitch (rows) and yaw (columns) over the full training distribution range for the gaze-only model trained in Chapter 3.

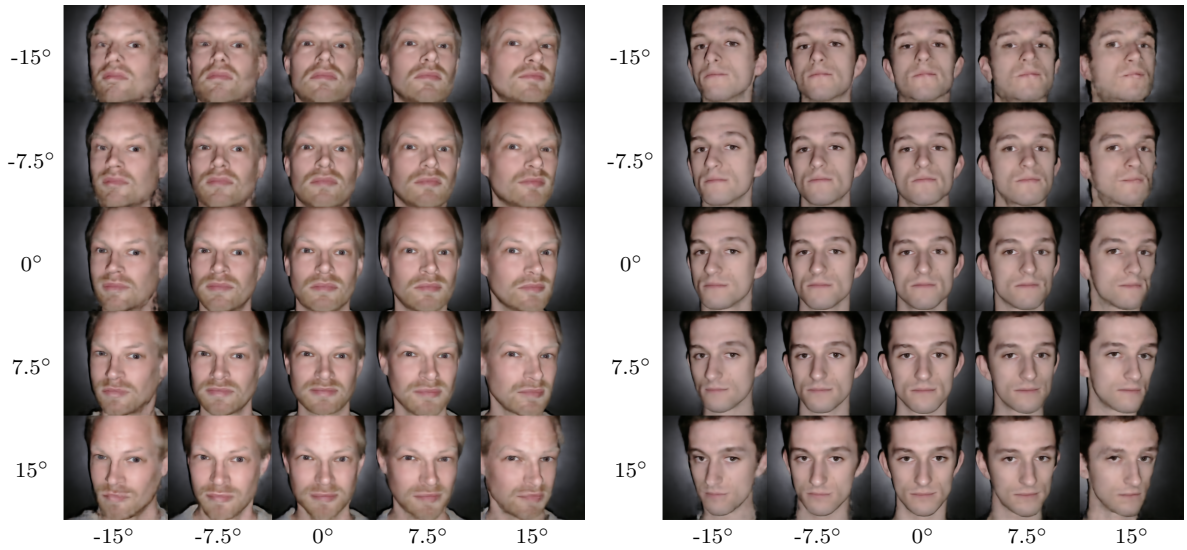


Figure A.2: Sweeps over head pose angles, for pitch (rows) and yaw (columns), from the multi-identity model trained in Chapter 4 for two of the participants in the dataset.

A.2 Multi-Identity Dataset

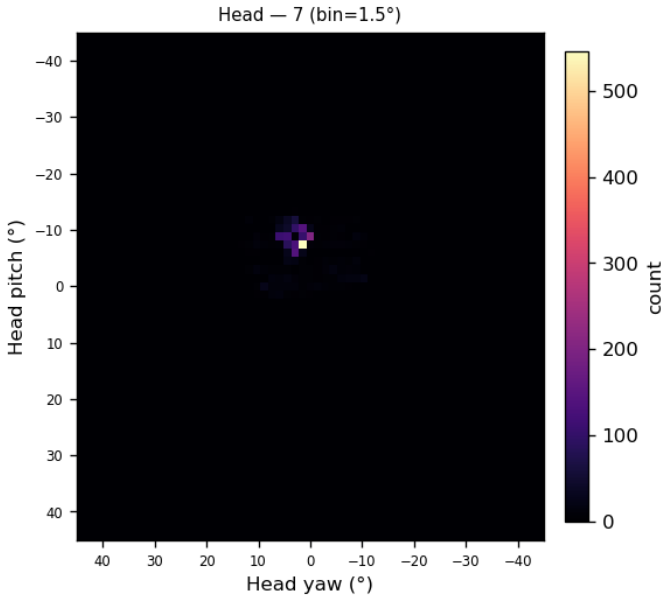


Figure A.3: Head pose coverage of identity 7 in the multi-identity dataset in degrees of yaw and pitch.

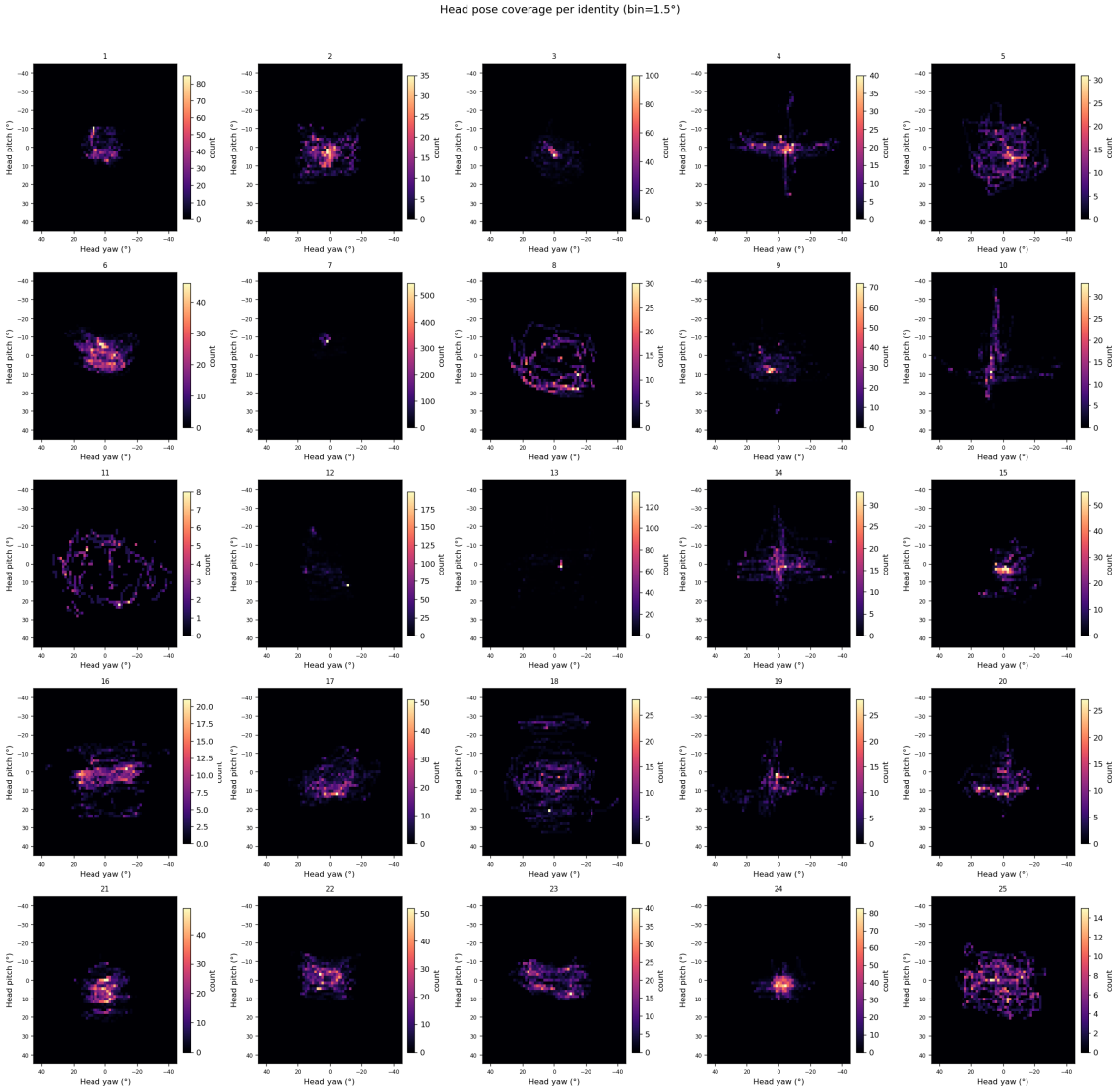


Figure A.4: Head pose coverage for each identity in the multi-identity dataset, visualized as heatmaps over yaw and pitch angles. The distributions illustrate the variability in head movement across identities and highlight differences in data density and coverage.

A.3 Ablation Studies

A.3.1 Latent Dimensionality

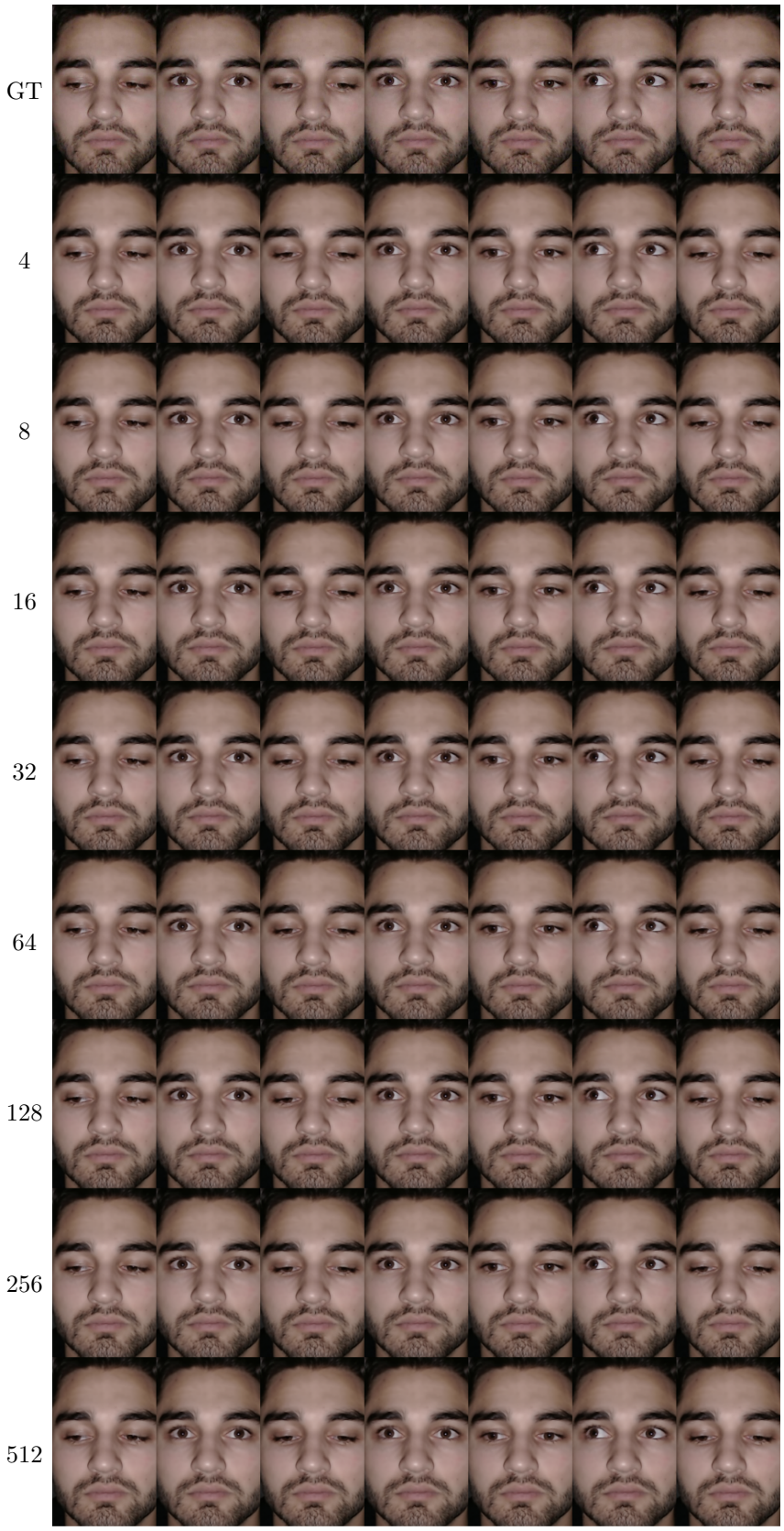


Figure A.5: Arbitrarily chosen test images (GT) reconstructed for several latent dimensionalities for the gaze-only model trained in Chapter 3.

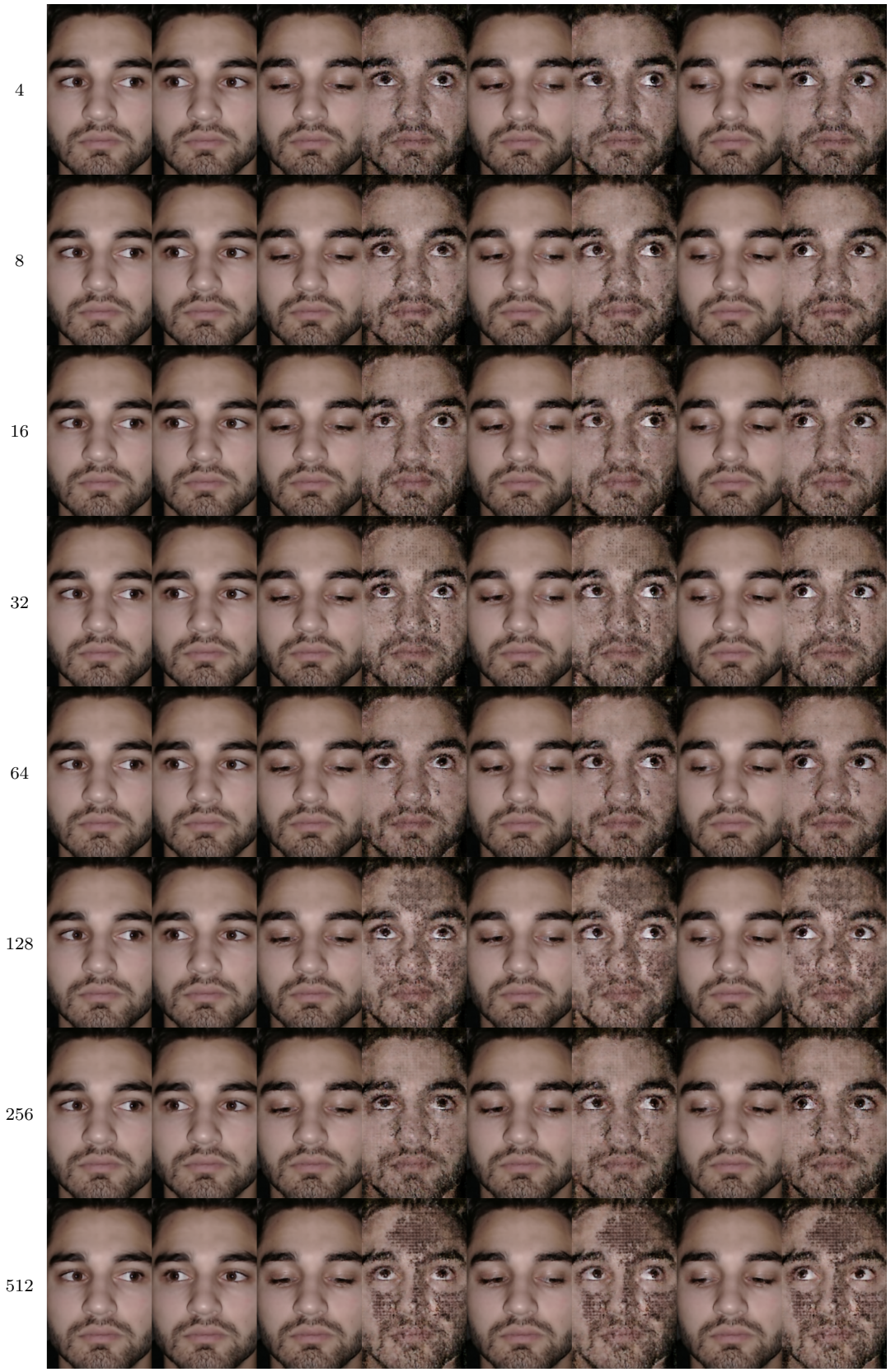


Figure A.6: Gaze redirections outside of training distribution over all ablated latent dimensionalities for the gaze-only model trained in Chapter 3.

A.3.2 Autoencoder Architecture

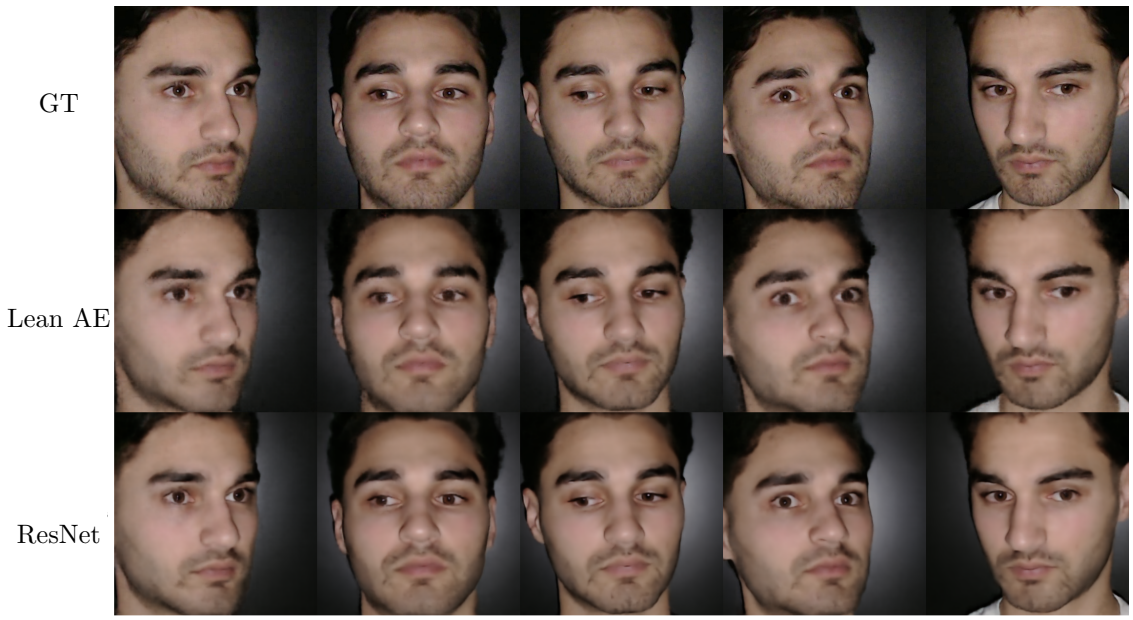
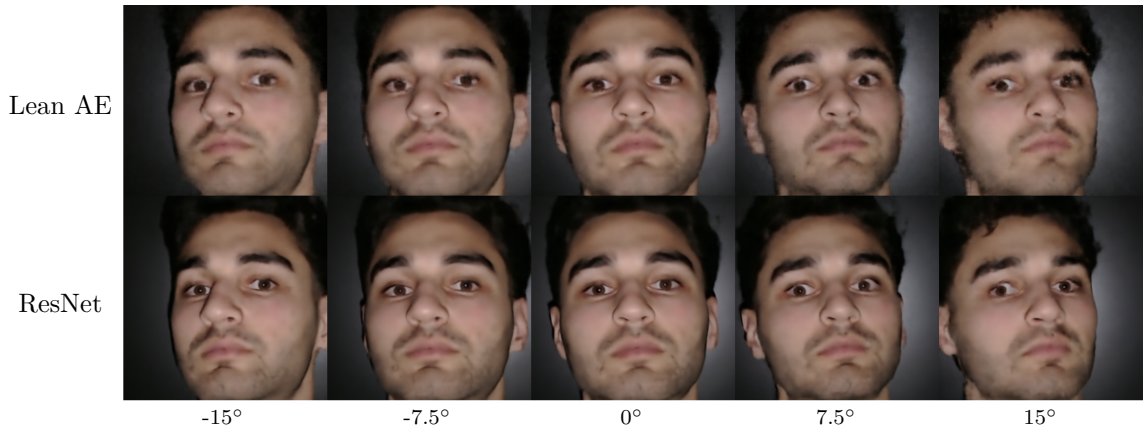
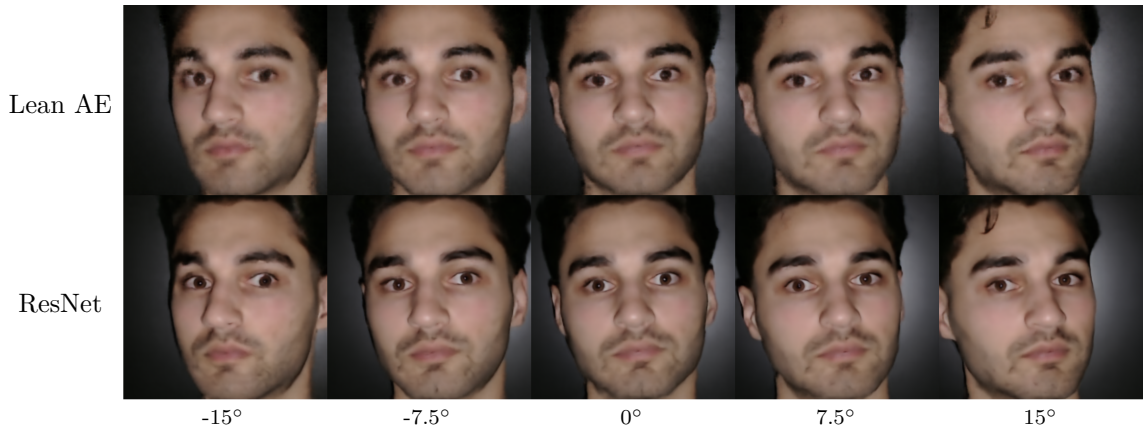


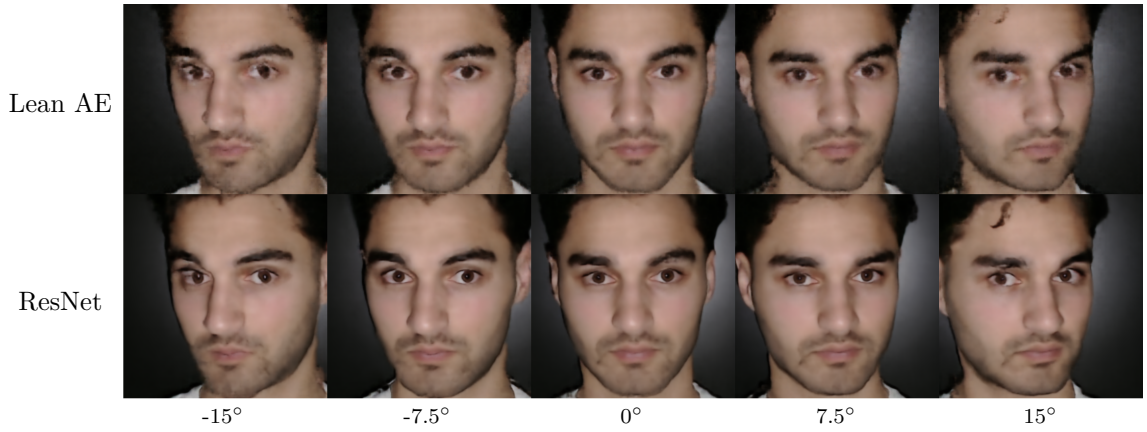
Figure A.7: Comparison of reconstructions of arbitrarily chosen test images (top row) using the lean autoencoder from Chapter 3 (middle row) and the ResNet-based autoencoder from Chapter 4 (bottom row), both trained on the single-identity dataset. The reconstruction quality difference clearly supports the introduction of a more complex and capable architecture when moving on from the gaze-only dataset.



(a) Head-pose sweep over yaw with pitch fixed at $-15°$.



(b) Head-pose sweep over yaw with pitch fixed at $0°$.

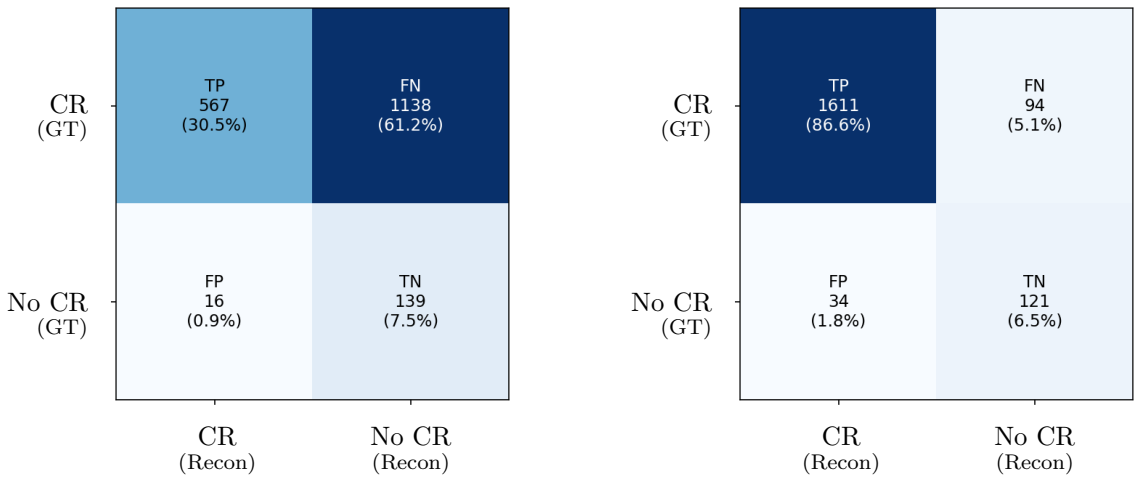


(c) Head-pose sweep over yaw with pitch fixed at $15°$.

Figure A.8: Sweeps over yaw angles for head pose for different fixed pitch angles using the lean autoencoder from Chapter 3 (first row) and the ResNet-based autoencoder from Chapter 4 (second row) in each of the subfigures, both trained on the single-identity dataset. While both equally disentangle the latent space and yield controllability over head pose as a factor, there is visible difference in image quality between the two, with the ResNet-based model outperforming the autoencoder from Chapter 3.

Table A.1: Summary of model performance on the single-identity dataset comparing the lean autoencoder from Chapter 3 and the ResNet-based autoencoder from Chapter 4. The rows signify which model was used, and the columns signify the evaluation metric. The encoding of head pose and gaze, and CR localization error is measured using MAE.

Model	Head Pose Yaw (°)	Head Pose Pitch (°)	Gaze Yaw (°)	Gaze Pitch (°)	CR (px)	Acc.	F1
Lean AE	0.23	0.24	1.35	0.47	1.53	0.40	0.50
ResNet AE	0.21	0.21	1.27	0.38	0.88	0.93	0.96



(a) Lean AE.

(b) ResNet AE.

Figure A.9: Confusion matrices of the lean autoencoder from Chapter 3 and the ResNet-based autoencoder from Chapter 4. The confusion matrix in Figure A.9b shows significantly better results compared to the matrix in Figure A.9a, with an increase in TPs and a reduction in FNs. Coupled with the results from Table A.1, the ResNet-based autoencoder yields better performance across all metrics, with only slight decrease in encoding error.

A.3.3 2x2 Sub-Batch Training

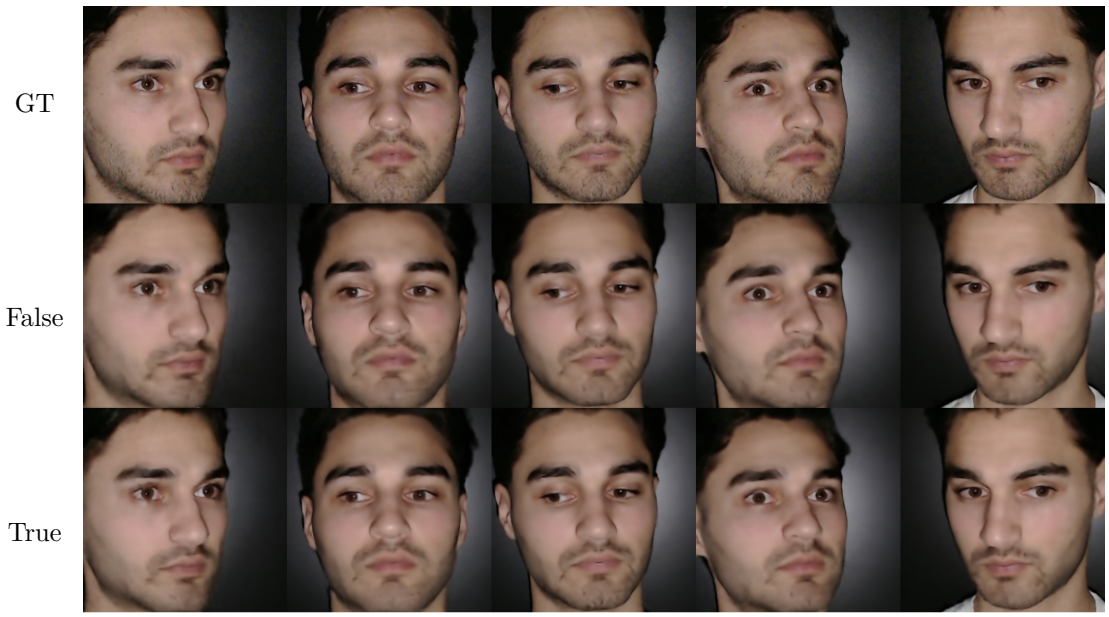
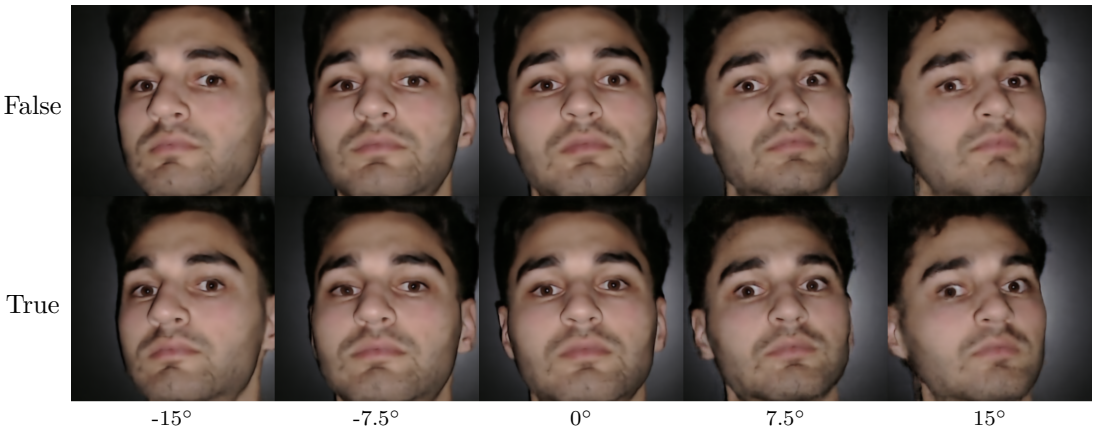
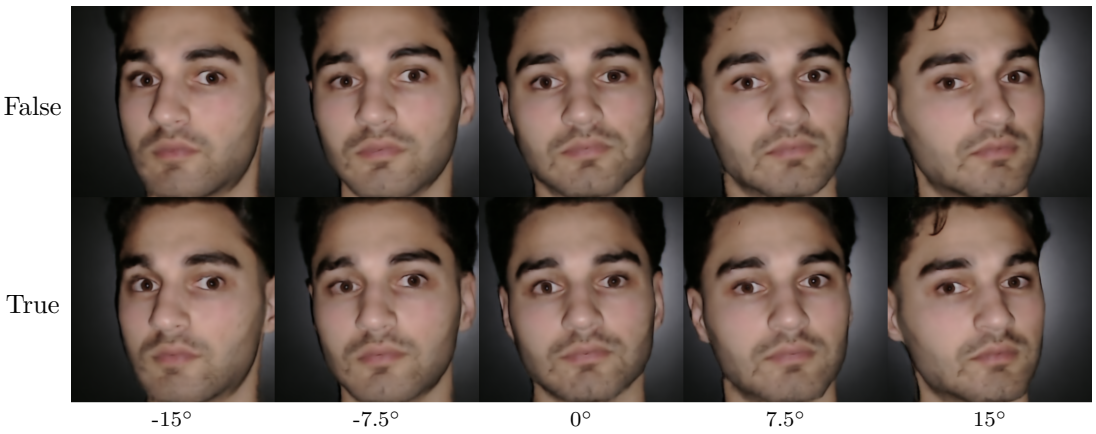


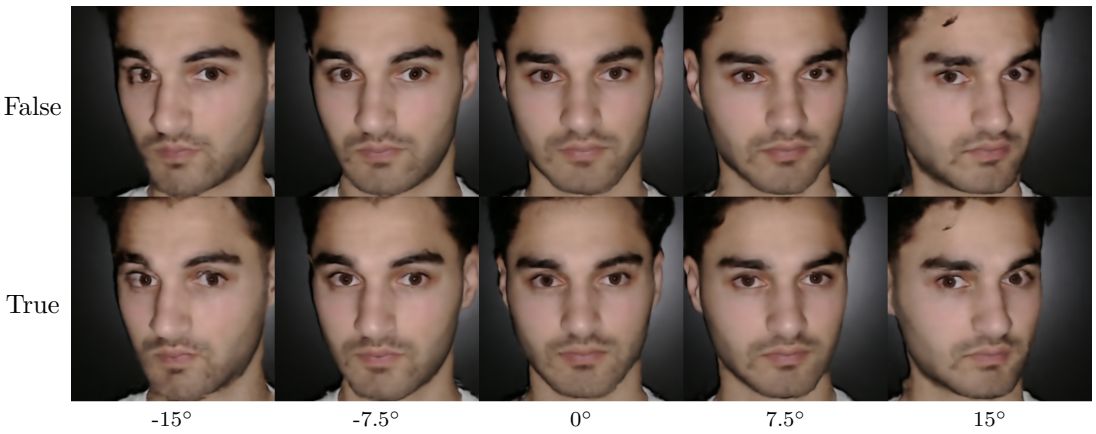
Figure A.10: Comparison of reconstructions of arbitrarily chosen test images (top row) using the ResNet-based autoencoder from Chapter 4 without the 2×2 sub-batch training (middle row) and with the 2×2 sub-batch training (bottom row), both trained on the single-identity dataset. The reconstruction quality difference is minimal, however, CRs are slightly more pronounced in this case for the model trained without the 2×2 sub-batch training.



(a) Head-pose sweep over yaw with pitch fixed at $-15°$.

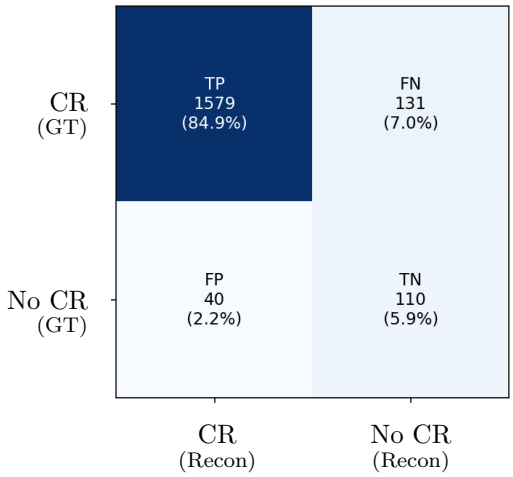


(b) Head-pose sweep over yaw with pitch fixed at $0°$.

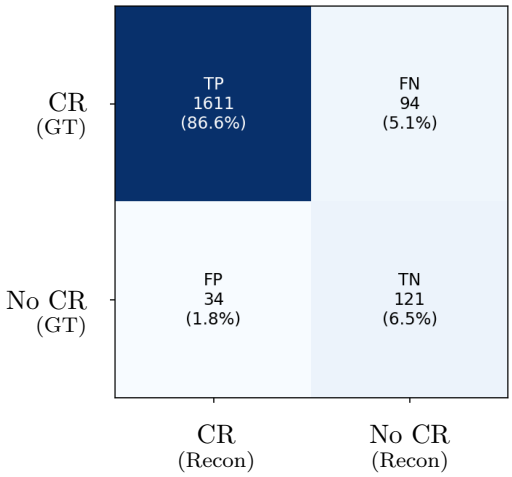


(c) Head-pose sweep over yaw with pitch fixed at $15°$.

Figure A.11: Sweeps over yaw angles for head pose for different fixed pitch angles using the ResNet-based autoencoder from Chapter 4 without the 2×2 sub-batch training (top row) and with the 2×2 sub-batch training (bottom row), both trained on the single-identity dataset. While the controllability can be observed to be qualitatively similar for the two models, there exists slightly better consistency and artifact-free reconstructions for the model trained without 2×2 sub-batch training. The models exhibit similar CR-reconstruction consistency as for Figure A.10.



(a) Trained **with** 2×2 sub-batch intervention.



(b) Trained **without** 2×2 sub-batch intervention.

Figure A.12: Confusion matrices of the autoencoder used in Chapter 4 trained with (left) and without (right) 2×2 sub-batch training. The confusion matrix in Figure A.12b shows slightly better results compared to the matrix in Figure A.12a, indicating that the explicit disentanglement intervention does not improve the performance of the CR detection, rather the method harms it.

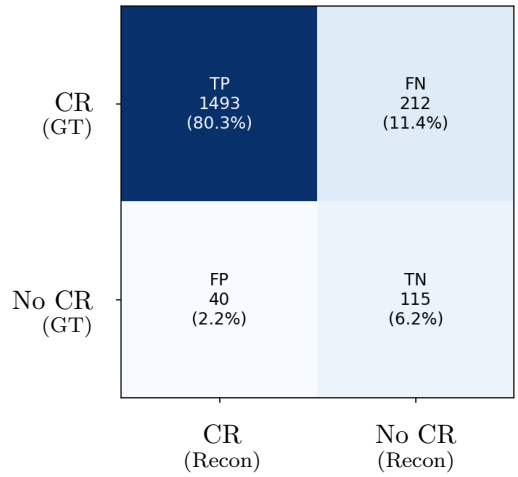
Table A.2: Comparison of the model used in Chapter 4 without the 2×2 sub-batch training (top row) and with the 2×2 sub-batch training (bottom row), both trained on the single-identity dataset. The columns signify the evaluation metric, and the encoding of head pose and gaze, and CR localization error is measured using MAE.

Model	Head Pose Yaw (°)	Head Pose Pitch (°)	Gaze Yaw (°)	Gaze Pitch (°)	CR (px)	Acc.	F1
2x2	0.21	0.20	1.28	0.40	0.96	0.91	0.95
No 2x2	0.21	0.21	1.27	0.38	0.88	0.93	0.96

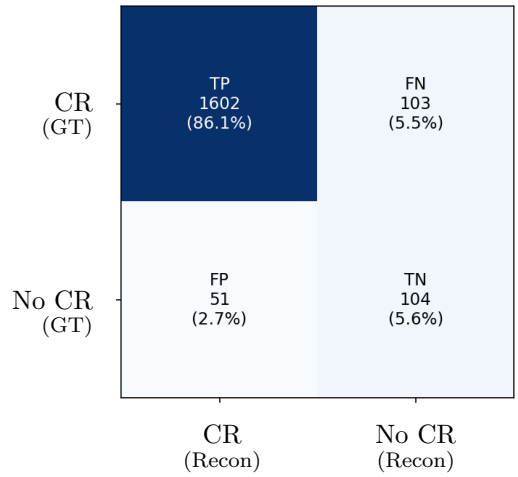
A.3.4 Laplacian Loss

Table A.3: Summary of reconstruction performance for the single-identity autoencoder model under different Laplace regularization weights. The rows signify which model was used, and the columns signify the evaluation metric. The CR localization error is measured using MAE. $\lambda_{\text{laplace}} = 0.25$ showcases best numerical performance for the metrics in the table.

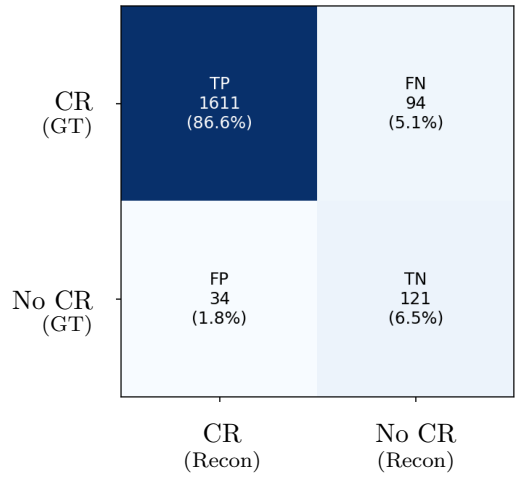
Model	CR (px)	Accuracy	F1
$\lambda_{\text{laplace}} = 0$	0.90	0.86	0.92
$\lambda_{\text{laplace}} = 0.1$	0.88	0.92	0.95
$\lambda_{\text{laplace}} = 0.25$	0.88	0.93	0.96
$\lambda_{\text{laplace}} = 0.5$	0.87	0.91	0.95



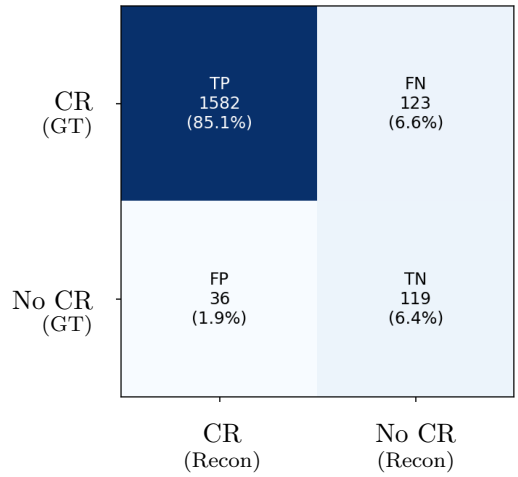
(a) $\lambda_{\text{laplacian}} = 0$



(b) $\lambda_{\text{laplacian}} = 0.1$



(c) $\lambda_{\text{laplacian}} = 0.25$



(d) $\lambda_{\text{laplacian}} = 0.5$

Figure A.13: Confusion matrices of the ResNet-based autoencoder trained with different weights, λ, for the Laplacian loss, $\mathcal{L}_{\text{laplacian}}$. The confusion matrix in Figure A.13c shows slightly better results compared to matrices in Figures A.13b and A.13d, where all three of them show improved performance, with an increase in TPs and TNs and a reduction in FNs, compared to the baseline in Figure A.13a.

Figure A.14: Reconstructions of test images (GT) over models trained with different weights for the Laplacian loss, $\lambda_{\text{laplacian}}$. Reconstructions shows definite qualitative proof of CR presence becoming more prominent and accurate with an introduced Laplacian loss, with $\lambda_{\text{laplacian}} = 0.25$ showing best visual results, noticeable in the eye regions in columns 1 and 2.

DEPARTMENT OF SOME SUBJECT OR TECHNOLOGY
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden
www.chalmers.se

CHALMERS
UNIVERSITY OF TECHNOLOGY