technische universität dortmund

fakultät statistik

Master Thesis

Enhancing facial attribute representation in autoencoder latent space using factor rotation techniques

Arindam Pal

June 20, 2025

Academic Advisors:
Prof. Dr. Philipp Doebler
M. Sc. Simon Klüttterman

Contents

List of Figures

List of Tables

1. Introduction

1.1. Motivation and background

In recent years, deep learning models like autoencoders have become an important tool in representation learning, particularly for complex visual data, such as facial images. Autoencoders are a type of neural networks capable of compressing high dimensional input data into low dimensional representations called the latent space or latent space representation that capture salient features of the input distribution [24, 32]. These representations are not only crucial for reconstruction tasks, but also serve as the basis for downstream applications such as attribute classification, facial editing, fairness analysis and identity disentanglement [23].

Despite their effectiveness, latent space representations learned by autoencoders often lack interpretability. Each latent space dimension can entangle several fundamental generative factors, making it difficult to isolate and manipulate specific attributes. Foundational work in deep learning has laid the groundwork for understanding the complexities of latent space and representation learning [20], while approaches such as β-VAE [23] and InfoGAN [10] have been used to disentangle latent space representations. This issue of disentanglement has been widely discussed and many solutions are based on architectural modifications or explicit disentanglement objectives [40, 7]. However, these approaches often involve intensive retraining and complex regularization terms.

A strategy to enhance the interpretability of latent space representations is through factor rotation, a technique traditionally used in psychometrics and multivariate statistics to improve the interpretability of the components of low dimensional representation extracted through a dimension reduction algorithm [55, 30]. Factor rotation involves applying a linear transformation to underlying factors that aligns them with more meaningful features. When adapted to autoencoders, rotation can realign the latent space to better reflect specific target attributes without modifying the original structure of the autoencoder [31, 15, 57].

1.2. Research question and contributions

The main research question to be addressed in this thesis is how factor rotation techniques can be applied to the latent space of autoencoders to enhance the interpretability and alignment of specific attributes. The basic approach is to create a more interpretable latent space by improving the alignment between latent space dimensions and identifiable attributes.

Extending factor rotation techniques to the latent space of autoencoders aims to highlight their effectiveness in structuring latent space representations beyond traditional applications. Building on this, a systematic method is proposed to align latent space dimensions with specific attributes, thereby enhancing interpretability without requiring modifications to the original autoencoder architecture. Additionally, a detailed evaluation strategy is developed to measure the impact of factor rotation on tasks such as attribute-based classification and latent space manipulation. Collectively, these contributions offer a versatile strategy for improving the interpretability and practical utility of autoencoder based feature representation tasks in machine learning and computer vision.

1.3. Structure of the thesis

As a starting point, section 2 introduces the concept of autoencoder and latent space representation. It discusses key architectural components such as encoder, decoder, latent space, different loss functions and autoencoder variants like convolutional autoencoder and variational autoencoder, including the β-variational autoencoder. This section also explores how the latent space encodes semantic features, forming the basis for further manipulation and interpretation.

Section 3 presents the concept of factor rotation, originally rooted in statistical analysis and explores its adaptation to the context of autoencoder latent spaces. Both orthogonal and oblique rotation are discussed, with a focus on aligning latent space dimensions with specific semantic attributes to enhance interpretability.

An overview of the datasets, used in this thesis, is given in section 4, highlighting the characteristics of CelebA and UTKFace datasets, including the diversity of annotated facial attributes provided by these datasets.

Section 5 details the methodology developed for this thesis. It describes the process of applying factor rotation to latent space representation using logistic regression coefficients and aligning rotated dimensions with selected attributes. The section further explains how these aligned latent space representations are used for downstream tasks, such as classification and fairness oriented evaluation.

A comprehensive experimental analysis is presented in section 6. The section provides a detailed description of the experimental setup using the CelebA dataset, including dataset preprocessing, model configuration and training procedure. This is followed by a summary of the baseline setup and the performance metrics used to assess reconstruction quality, disentanglement quality, attribute classification accuracy and fairness. Then details about latent space disentanglement are provided, including both quantitative results using disentanglement metrics and qualitative visualizations that highlight attribute alignment and separation. The effectiveness of latent space manipulation is examined

through classifier based predictions. This section also explores the potential of the proposed method for mitigating bias in attribute prediction. To confirm the robustness and generalizability of the method, additional experiments are conducted on the UTKFace dataset, reaffirming the consistency of findings across datasets and demographic variations.

Section 7 discusses the limitations of the proposed factor rotation method for enhancing semantic interpretability in autoencoder latent space. It highlights key challenges including the reliance on attribute classifiers as directions for rotation, the assumption of linear separability, the limited availability of attribute labels, the imperfect nature of attribute suppression and the dependency on the quality of learned latent space representation. These limitations highlight potential risks to fairness, interpretability and robustness of the method.

The thesis concludes in section 8, summarizing key findings and suggesting directions for future research in interpretable and fair representation learning through rotation of latent space.

1.4. Related work

Autoencoders are widely used for learning compact and structured latent space representations of high dimensional data such as facial images. The variantional autoencoder or VAE [32] imposes a probabilistic structure on the latent space, while β-VAE [23] encourages disentanglement by controlling the trade-off between reconstruction quality and prior regularization. Several extensions such as FactorVAE [31], β-TCVAE [9] and InfoGAN [10] aim to learn representations where individual latent space dimensions correspond to independent generative factors.

The FactorVAE, proposed by Kim and Mnih [31], extends the VAE framework by penalizing the total correlation among the latent space dimensions, a statistical dependence measure that captures dependence between them. To estimate total correlation in an unsupervised manner, an auxiliary discriminator trained to distinguish samples from the joint latent space distribution versus samples from the product of marginals was introduced, effectively encouraging well aligned and statistically independent latent space dimensions. This improves disentanglement compared to β-VAE while preserving reconstruction quality and also introduces a new disentanglement metric based on majority vote classification of latent space dimension.

Chen et al. [9] proposed the β-TCVAE or total correlation variational autoencoder, which improves disentanglement by explicitly penalizing the total correlation between the latent space dimensions, reducing statistical dependencies to produce disentangled representation. The literature also demonstrated that unsupervised disentanglement is a

challenging task and it serves as a motivation for supervised disentanglement of latent space dimensions using interpretable attributes.

To encourage disentanglement by matching the inferred latent space distribution to a factorized prior, Kumar et al. [33] proposed the DIP-VAE or disentangled inferred prior VAE. This model promotes independence among latent space dimensions by regularizing the covariance structure of the latent space distribution. In contrast, this thesis applies supervised post-training factor rotation to explicitly align selected facial attributes with individual latent space dimensions, providing a complementary approach to achieving interpretability in the latent space.

Wei et al. [57] proposed a method for controlling facial attribute synthesis by disentangling attribute feature axes directly in the latent space. Their approach allows targeted manipulation of facial attributes by aligning latent space dimensions with specific features, enabling better control over the generated images. This aligns closely with the goal of this thesis, which similarly seeks to enhance interpretability and control of latent space representations through supervised rotation techniques.

In the recent years post-training manipulation of latent space representation has has been explored for its potential to enhance interpretability, fairness or controllability. Classical rotation techniques [30, 29] have long been used in factor analysis to enhance the interpretability of underlying factors. In the deep learning context, such rotations can be applied to align latent space dimensions with known attributes or to separate correlated features, without retraining the model.

One such strategy for orthogonalizing latent space with respect to an attribute builds upon the general orthogonalization framework proposed by Rügamer et al. [51]. Given a matrix of input features, a projection matrix is computed, using the least squares formulation, to project an attribute vector onto the orthogonal complement of the input feature space instead of projecting directly onto the input feature space, thereby effectively removing the linear influence of the input features on this attribute. This strategy enables interpretable and fair post-training representation learning and is computationally simple to implement.

Furthermore, a method for learning orthogonally disentangled latent space within deterministic autoencoders was introduced by Cha and Thiyagalingam [8], using structured loss functions to enforce orthogonality constraints without relying on the probabilistic assumptions of a VAE or using supervised labels. Lample et al. [34] introduced a model for attribute editing of images that learns a disentangled representation by conditioning the decoder on target attributes while adversarially removing them from the latent space representation. Similarly, Zhang et al. [61] proposed an adversarial fairness model where a discriminator attempts to predict the protected attribute from the latent space

representation and the encoder is trained to suppress information about the protected attribute thereby promoting fairness. Sarhan et al. [52] proposed an orthogonal disentanglement method, where a neural network learns to separate latent space representation into sensitive and task relevant components, with the training objective enforcing mutual orthogonality to support fair prediction while retaining essential information.

A more domain specific approach is presented in the form of the Fairness-aware Disentangling VAE (FD-VAE) by Park et al. [46], which partitions the latent space into three interpretable subspaces, one for the target variable, one for protected attributes and one for mutual information. This allows controlled interventions on fairness sensitive attributes in facial attribute classification tasks. Another important direction in fairness aware representation learning is demonstrated by the framework proposed by Creager et al. [11], termed Flexibly Fair Representation Learning by Disentanglement or FairVAE, introduced a semi-supervised VAE that disentangles latent space representations into two complementary components, one capturing information relevant for the primary prediction task and another encoding sensitive attributes. The model is trained with a structured variational objective that enforces independence between these components, thereby supporting fair classification by explicitly suppressing information about protected variables in the task relevant latent space. This design enables flexible trade-off between fairness and utility and demonstrates competitive empirical performance across benchmark datasets. The FairVAE framework thus provides a principled and practical approach for fair representation learning in variational settings, contributing to the growing literature on disentangled and interpretable deep learning models.

A critical motivation for such fairness aware representation learning stems from findings like those in Buolamwini and Gebru [6], demonstrating significant intersectional accuracy disparities in commercial gender classification systems, particularly against darker skinned females. This study highlights the ethical and performance implications of biased representations in computer vision models. These concerns highlights the need for developing methods that explicitly address representation bias through latent space manipulation, thereby contributing to more equitable and transparent artificial intelligence systems.

The idea of aligning latent space dimensions with interpretable attributes is also explored in InterfaceGAN [54], which identifies linear directions in the latent space after training to enable attribute manipulation.

It has been shown that unsupervised disentanglement is inherently challenging without inductive biases or supervision [40]. To address this, several supervised or semi-supervised methods have been proposed, including works that explicitly seek to factor out protected or irrelevant variables from the latent space. The FactorVAE framework is especially relevant to this thesis, which similarly aims to promote interpretability by aligning latent space dimensions with human interpretable attributes. However, unlike FactorVAE, this thesis

explores post-training rotation techniques applied to the latent space of already trained autoencoders, offering a computationally simpler, modular alternative to adversarial disentanglement strategies. This highlights a key motivation of the current approach to retain the interpretability benefits of a well aligned latent space without introducing additional training complexity.

Higgins et al. [22] address a critical gap in representation learning by proposing a formal and quantitative definition of disentangled representations. Recognizing the ambiguity surrounding the term disentanglement, the literature argues that a truly disentangled representation should satisfy three criteria, modularity, compactness and explicitness. This literature also reveals that many methods that promote disentanglement in an unsupervised manner often fall short of producing interpretable latent space dimensions in a consistent way.

This thesis draws on these advances by combining factor rotation techniques along with targeted manipulation of latent space after rotation to enhance interpretability and control in generative models of facial images, contributing to the broader dialogue on fairness oriented and controllable generative modelling by offering a lightweight, post-training alternative to more complex adversarial or structured disentanglement methods. Specifically in this thesis, autoencoders are trained on facial images and then the latent space is rotated such that specific latent space dimensions are aligned with chosen facial attributes. Subsequently, the effect of suppressing certain attributes, like age, by adjusting the corresponding latent space dimension is also examined in this thesis. This approach builds on established mathematical foundations and recent fairness aware representation learning literature, while adapting them to the domain of image generation and manipulation using unsupervised deep learning methods.

2. Autoencoder and latent space representation

This section introduces the core concepts of autoencoders which are critical to the methodological improvements in this thesis. It begins with a general overview of autoencoders and their key architectural components, the encoder, decoder and latent space representation. The focus then shifts to convolutional autoencoders, which are particularly effective for learning spatially aware encodings from high dimensional visual data. Subsequently, the section explores variational autoencoders (VAE) and their extension, the β-VAE, which incorporate probabilistic modelling and disentanglement promoting regularization. Particular attention is given to the structure and semantics of the latent space, as it serves as the theoretical basis for the factor rotation techniques explained in section 3.

2.1. Introduction to autoencoders

Autoencoders are a class of unsupervised models that are commonly used for dimensionality reduction, feature learning, image compression, de-noising and generative modelling. They are neural networks trained to reproduce their input at the output, typically through a bottleneck architecture that forces the model to learn an intermediate representation. This intermediate representation, often referred to as the latent space or latent space representation, captures essential structures and patterns in the data. Fundamentally, an autoencoder comprises of two components, an encoder that compresses the input into a latent space, also called the bottleneck and a decoder that reconstructs the input from this latent space [20].

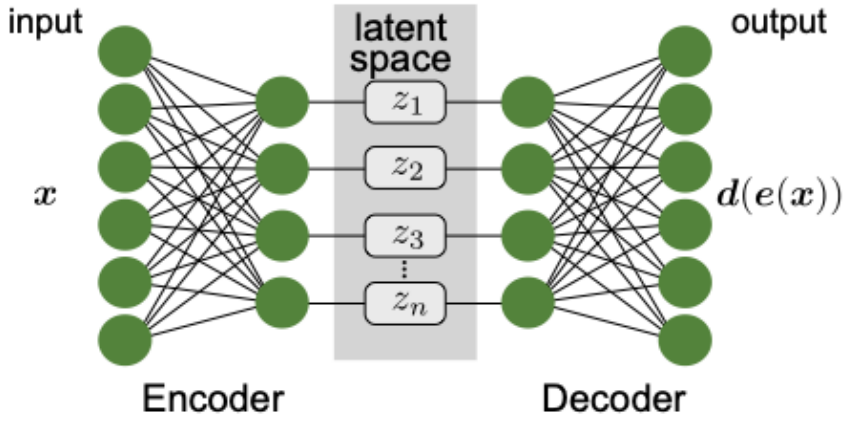


Figure 2.1: Illustration of a basic autoencoder architecture consisting of an encoder, a latent space and a decoder. The encoder transforms the input vector x into a latent space z, which captures essential features of the input data. The decoder then reconstructs the input as $d(e(x))$, aiming to minimize the difference between the original and reconstructed outputs. Figure reproduced from Neupert et al. [44].

The architecture of a basic autoencoder is illustrated in figure 2.1. The reconstruction objective encourages the network to learn salient features of the input data distribution instead of simply copying the input perfectly [20]. Autoencoders are particularly beneficial in fields where labelled data is rare but unlabelled data is plentiful, such as facial image analysis. Their ability to structure information within a latent space that captures the essence of visual data has enabled their application in areas like medical imaging and recommender systems [36, 4]. Autoencoders can be classified as undercomplete when the latent space has lower dimensionality than the input or overcomplete when the latent space has higher dimensionality than the input. Undercomplete autoencoders are mainly used for dimensionality reduction, whereas overcomplete autoencoders are capable of learning more complex representations [20].

2.1.1. Encoder

The encoder is a function $e_\omega : \mathbb{R}^s \to \mathbb{R}^n$, parametrized by a set of learnable weights ω, which maps the input data $\mathbf{x} \in \mathbb{R}^s$ into a latent space representation $\mathbf{z} \in \mathbb{R}^n$, where s is the dimension of the input data, n is the dimension of the latent space and typically $n \ll s$. Architecturally, the encoder is composed of layers of affine transformations like linear layers or convolutional layers, followed by non-linear activation functions such as ReLU or sigmoid. The output of the encoder, $\mathbf{z} = e_\omega(\mathbf{x})$, is referred as the latent space, which often serves as a compressed summary of the input [20].

2.1.2. Decoder

The decoder is a function $d_\theta : \mathbb{R}^n \to \mathbb{R}^s$, parametrized by another set of weights θ which attempts to reconstruct the original input $\mathbf{x}$ from the latent space representation $\mathbf{z}$. The decoder effectively reverses the operation of the encoder by transforming the latent space back into the original data space. Like the encoder, it is composed of learnable layers, but in the opposite direction, often employing deconvolution and upsampling layers in the convolutional autoencoder architecture [20, 32]. The output $\hat{\mathbf{x}} = d_\theta(\mathbf{z})$ should approximate the original input $\mathbf{x}$ and the network is trained to minimize the reconstruction loss between them [5].

2.1.3. Latent space

The latent space or latent space representation in an autoencoder refers to the intermediate representation $\mathbf{z} \in \mathbb{R}^n$ of the input data $\mathbf{x} \in \mathbb{R}^s$, produced by the encoder. It often serves as the bottleneck in the architecture, enforcing a dimensionality reduction that compels

the model to extract the most informative features of the data. The structure and interpretability of the latent space are critical, especially in applications such as representation learning, generative modelling, anomaly detection and feature disentanglement.

The latent space forms the backbone of the autoencoder and its learning capabilities. The encoder aims to preserve the essential characteristics of the input data distribution while discarding noise and irrelevant variance by often mapping complex high dimensional input data to a lower dimensional latent space. The decoder then attempts to reconstruct the input from the latent space. The effectiveness of the autoencoder depends largely on how well this latent space captures the underlying structure of the data [5].

Mathematically, the latent space is defined as:

$$\mathbf{z} = e_\omega(\mathbf{x}), \quad \hat{\mathbf{x}} = d_\theta(\mathbf{z}) \tag{1}$$

where e_ω is the encoder, d_θ is the decoder and $\mathbf{z}$ represents the latent space.

In an ideal setting, each dimension of the latent space $\mathbf{z}$ corresponds to a distinct factor of variation in the input data [23]. However, standard autoencoders do not enforce any particular structure or interpretability in the latent space. As a result, the latent space may encode entangled or correlated features that are difficult to interpret independently. To encourage meaningful organization in the latent space, techniques such as variational inference [32] and supervised disentanglement [40] are used. These methods aim to improve the alignment between latent space dimensions and semantically meaningful concepts.

Despite its utility, the latent space of an autoencoder can be highly entangled, especially when no explicit regularization is applied [23, 7]. This makes it difficult to interpret and manipulate individual latent space dimensions. Additionally, latent space may be sensitive to small perturbations or over fit to training data if the bottleneck is not well regularized or if the reconstruction loss dominates the training objective [5, 40].

2.1.4. Loss function

The loss function in an autoencoder plays a critical role in guiding the learning process by quantifying how well the reconstructed data approximates the original input. The encoder transforms the input into a latent space representation and the decoder tries to reconstruct the input. Training involves minimizing a composite loss comprising a reconstruction loss and in the case of variational models, a regularization term, explained later in subsection 2.3.

A commonly used reconstruction loss is the mean squared error (MSE), which penalizes the squared difference between each feature of the input and its reconstruction:

$$\mathcal{L}_{\mathbb{MSE}}(\mathbf{x}, \hat{\mathbf{x}}) = \frac{1}{s}\sum_{i=1}^{s}(x_i - \hat{x}_i)^2, \tag{2}$$

where x_i and $\hat{x}_i$ are the true value and the predicted or reconstructed value of the i-th feature of the input data $\mathbf{x}$, respectively and $i = 1, 2, \ldots, s$. This loss function assumes Gaussian reconstruction error and is particularly effective when the data consists of continuous values.

When the input features are normalized to the range $[0, 1]$, the binary cross-entropy (BCE) loss is often preferred and the decoder's outputs are treated as Bernoulli probabilities.

$$\mathcal{L}_{\mathbf{BCE}}(\mathbf{x}, \hat{\mathbf{x}}) = -\frac{1}{s}\sum_{i=1}^{s} x_i \log(\hat{x}_i) - (1 - x_i)log(1 - \hat{x}_i), \tag{3}$$

where x_i and $\hat{x}_i$ are the true value and the predicted or reconstructed value of the i-th feature of the input data $\mathbf{x}$, respectively and $i = 1, 2, \ldots, s$. This formulation arises under the assumption that each output pixel follows a Bernoulli distribution, effective for binary image data and is frequently used in variational autoencoders [32].

2.2. Convolutional autoencoder (CAE)

Convolutional autoencoder or CAE is a variant of autoencoder specifically designed to exploit the spatial structure of image data [43, 56]. CAE operates directly on two or three dimensional structure of the input, enabling the model to capture hierarchical patterns and spatial correlations more effectively. This architectural adaptation makes a CAE especially well suited for visual tasks such as image de-noising, super resolution and unsupervised feature extraction [58, 42].

In a CAE, the encoder consists of a series of convolutional and pooling layers that progressively reduce the spatial resolution of the input while increasing the depth of feature maps. This process extracts local patterns such as edges, textures and shapes in early layers and more abstract features in deeper layers. The result is a latent space representation that encodes essential information about the input image. The decoder mirrors the encoder structure using deconvolution and upsampling operations with the goal of reconstructing the original image from the latent space representation [60]. Figure 2.2 shows the architecture of a CAE.

One of the key advantages of a CAE is their ability to model translation invariance and spatial locality. By convolving a kernel over the input image, a CAE can share parameters across spatial locations, thereby reducing the number of learnable parameters and making

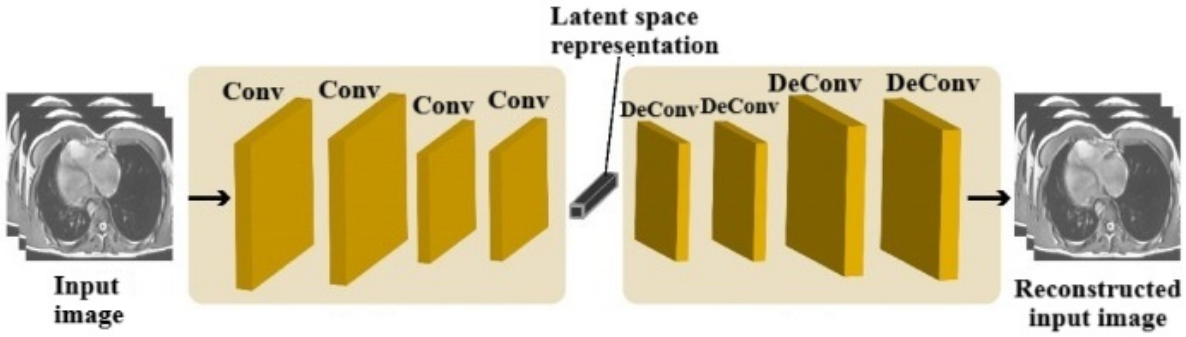


Figure 2.2: Architecture of a convolutional autoencoder (CAE). The encoder transforms the input image through successive convolutional (Conv) layers to a latent space representation. The decoder then reconstructs the image from this latent space representation using deconvolutional (DeConv) layers, aiming to preserve essential structural information in the reconstructed output. Figure adapted from Jafari et al. [26].

the model more robust to small shifts or distortions in the input. This parameter efficiency allows these autoencoders to scale well to high dimensional inputs, such as large greyscale or RGB images, compared to fully connected autoencoder variants that would require an infeasibly large number of weights [37].

The latent space representation learned by a CAE is typically a low resolution tensor rather than a flat vector, preserving some spatial structure even in the compressed form. This property has important implications for downstream tasks such as clustering, image retrieval and generative modelling, where spatial coherence of features plays a significant role. Moreover, because these autoencoders learn meaningful feature hierarchies in an unsupervised manner, they can be used for transfer learning or as a pre-training step before supervised fine tuning [24, 12].

CAE has been successfully applied in numerous domains. In face recognition, they have been used to extract invariant features from facial images [43]. In medical imaging, they help de-noise or enhance diagnostic scans while preserving anatomical detail. Their unsupervised learning capabilities also make them a valuable tool for anomaly detection in visual inspection and surveillance tasks, where reconstructive errors often highlight atypical or suspicious inputs.

However, CAE is not without limitations. Due to their local receptive fields, they may fail to capture long range dependencies unless the architecture is sufficiently deep or enhanced with additional mechanisms such as dilated convolutions or skip connections [59, 50]. Furthermore, their generative capacity is often more constrained compared to models like variational autoencoders or generative adversarial networks, as the reconstruction objective may lead to blurry outputs in the absence of more sophisticated loss functions [13, 35].

2.3. Variational autoencoder (VAE) and β-VAE

Variational autoencoder or VAE represents a significant advance in unsupervised deep generative modelling. Unlike classical autoencoder that learns a deterministic encoding of data, a VAE imposes a probabilistic framework on the latent space. This enables both principled regularization and generative capabilities, allowing the model to sample novel instances from the learned input data distribution [32, 48]. Figure 2.3 depicts the architecture of a VAE.

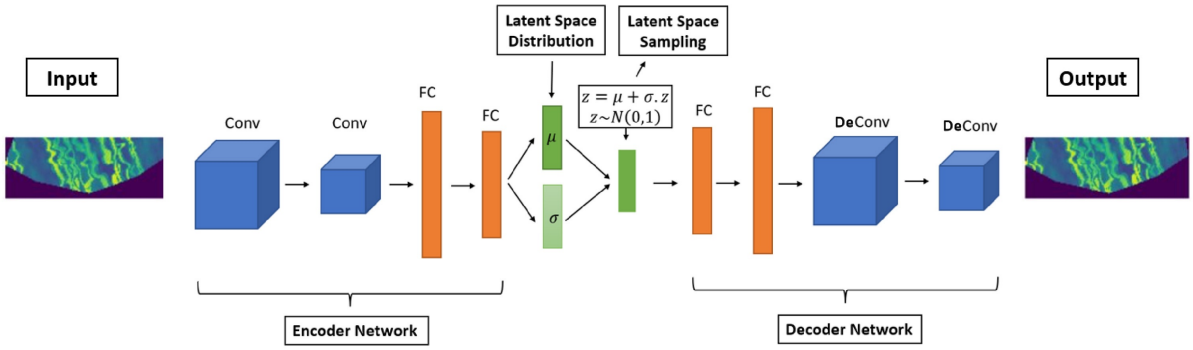


Figure 2.3: Architecture of a variational autoencoder (VAE) illustrating the encoding and decoding process. The encoder network, composed of convolutional (Conv) and fully connected (FC) layers, maps the input data to a latent space distribution defined by mean (μ) and standard deviation (σ). A latent space vector z is sampled using the reparametrization trick ($z = \mu + \sigma \odot \epsilon$, $\epsilon \sim \mathcal{N}(0, 1)$) and passed through the decoder network, which reconstructs the output using fully connected and deconvolutional (DeConv) layers (also called transposed convolutional layers). Figure adapted from Falola et al, [17].

At the core of a VAE is the idea of approximating the intractable posterior distribution $p(\mathbf{z} \mid \mathbf{x})$ over the latent space $\mathbf{z}$ given input data $\mathbf{x}$ using a learned variational distribution $q_\phi(\mathbf{z} \mid \mathbf{x})$, parametrized by a neural network, the encoder. The decoder is another neural network that models the likelihood $p_\theta(\mathbf{x} \mid \mathbf{z})$, which reconstructs the input from samples drawn from the latent space.

The training objective of a VAE is derived from maximizing the evidence lower bound (ELBO) of the marginal log-likelihood $\log p(\mathbf{x})$, given by:

$$\log p(\mathbf{x}) \geq \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x} \mid \mathbf{z})] \; - D_{KL}(q_\phi(\mathbf{z} \mid \mathbf{x}) \,||\, p(\mathbf{z})) \tag{4}$$

The first term is the expected log-likelihood or reconstruction term, which encourages the decoder to produce accurate reconstructions of the data. The second term introduces a constraint that conforms the latent space $q_\phi(\mathbf{z} \mid \mathbf{x})$ to remain close to a predefined prior distribution $p(\mathbf{z})$, typically chosen as a standard multivariate Gaussian $\mathcal{N}(0, \mathbf{I})$. This

serves as regularization and is implemented using the Kullback–Leibler (KL) divergence:

$$D_{\mathbf{KL}}(q_\phi(\mathbf{z} \mid \mathbf{x}) \mid\mid p(\mathbf{z})) = \int q_\phi(\mathbf{z} \mid \mathbf{x}) \log \frac{q_\phi(\mathbf{z} \mid \mathbf{x})}{p(\mathbf{z})} d\mathbf{z} \tag{5}$$

This regularization term measures the distance between the approximate posterior $q_\phi(\mathbf{z} \mid \mathbf{x})$ and predefined prior $p(\mathbf{z})$, enabling meaningful sampling and interpolation in the latent space. To allow back propagation through the stochastic sampling of $\mathbf{z} \sim q_\phi(\mathbf{z} \mid \mathbf{x})$, a VAE employs the reparametrization trick [32]. Instead of sampling $\mathbf{z}$ directly, the model samples from a standard normal distribution $\mathbf{z} = \mu + \sigma \odot \epsilon$, where μ and σ are outputs of the encoder network. This trick maintains randomness while preserving differentiability.

Despite the elegance of the VAE formulation, the trade-off between reconstruction quality and latent space regularization can lead to entangled latent space representations. The model may prioritize accurate reconstructions at the expense of learning disentangled latent space dimension, especially when the KL divergence term is weighted too lightly during training. To address this, the β-VAE was introduced [23], a modified formulation of the VAE with a tunable hyperparameter $\beta \geq 1$ that explicitly controls the strength of the regularization:

$$\mathcal{L}_{\beta-\mathbf{VAE}} = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x} \mid \mathbf{z})] \; - \beta D_{KL}(q_\phi(\mathbf{z} \mid \mathbf{x}) \mid\mid p(\mathbf{z})) \tag{6}$$

When $\beta > 1$, the model places more emphasis on the KL divergence term, encouraging the latent space to align more closely with the prior and pushing the encoder to produce more factorized and disentangled latent space representations. This enhanced regularization can come at the cost of lower reconstruction quality but it often yields latent space that correspond to semantically meaningful generative factors. In essence, β-VAE provides a controllable mechanism to navigate the trade-off between data reconstruction and latent space disentanglement, making them a powerful tool in representation learning. However, achieving optimal disentanglement often requires careful tuning of β and higher values can lead to posterior collapse, where the latent space is underutilized [3, 7].

Both VAE and β-VAE have been used extensively in applications ranging from semi-supervised learning and anomaly detection to controllable image synthesis and fairness aware representation learning [32, 40, 15]. In the context of this thesis, the β-VAE serve as the foundational model whose latent space is further refined through rotation techniques to align semantic attributes with specific latent space dimensions.

2.4. Structure and semantics of the latent space

The latent space of an autoencoder plays a pivotal role in determining the usefulness and interpretability of the learned representation. In the context of facial attribute modelling, a well structured latent space should ideally capture disentangled and semantically meaningful features that correspond to real world visual attributes. These representations are important not only for downstream tasks such as classification and generation, but also for promoting fairness and control in image synthesis and analysis.

In a standard CAE, the latent space is learned through purely reconstructive objectives and tends to be highly entangled, each latent space dimension may simultaneously encode multiple factors of variation, making interpretation and manipulation difficult. To address this, VAE introduces a probabilistic prior over the latent space, typically enforcing it to follow an uniformly spread multivariate Gaussian distribution. This regularization encourages a more continuous and structured latent space, enabling interpolation and sampling capabilities [32].

However, even with this structure, a VAE often falls short in achieving disentanglement, where each latent space dimension corresponds to an independent and interpretable factor of variation. To this end, β-VAE introduces a tunable hyperparameter β to weight the Kullback–Leibler divergence term in the VAE objective, thereby emphasizing the regularization of the latent space. By increasing β, the model encourages statistical independence among latent space dimensions, which promotes disentanglement at the cost of some reconstruction quality [23]. Empirical studies have shown that such disentangled representations not only improve interpretability but also enhance robustness and generalization across multiple downstream tasks [7, 14].

Nonetheless, it has been shown that unsupervised disentanglement is fundamentally underdetermined without inductive biases or labelled data [40]. As such, post-training techniques have become increasingly relevant. One such method involves rotating the latent space using factor rotation techniques borrowed from classical statistics [30, 29]. These methods can transform the latent space to better align individual dimensions with interpretable facial attributes, without altering the overall variance structure of the representations. Recent work has extended this idea to deep neural network latent spaces, showing that such rotations can enhance the alignment of latent space dimensions with semantic attributes while preserving reconstructive capacity [53, 57].

3. Factor rotation in latent space

This section introduces the concept of factor rotation as a classical technique from statistical analysis that has been increasingly adopted in representation learning to enhance the interpretability. Beginning with a conceptual overview of factor rotation, the section proceeds to describe the main categories of rotation, orthogonal and oblique and their mathematical properties. Finally, the section outlines how these methods are adapted to improve the structure of autoencoder latent space.

3.1. Overview of factor rotation

Factor rotation is a classical technique from exploratory factor analysis designed to improve the interpretability of learned representations by transforming the factor loading matrix [30]. Figure 3.1 shows the geometric illustration of factor rotation. The goal of rotation is to achieve a simpler and more interpretable structure in which each factor or dimension loads highly on a small number of variables and near zero on others. When applied to the latent space of an autoencoder, factor rotation can reveal latent space dimensions that align more distinctly with individual attributes [11]. Factor rotation serves as a linear

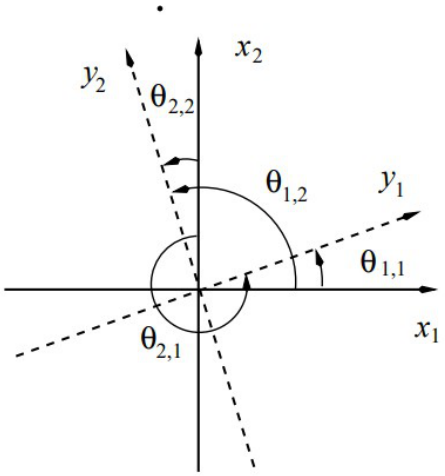


Figure 3.1: Geometric illustration of factor rotation. The original factor axes x_1 and x_2 are rotated to new factor axes y_1 and y_2 and the angle of rotation between an old axis x_i and new axis y_j is given by $\theta_{i,j}$, where $i, j \in \mathbb{Z}^+$. Figure reproduced from Abdi [1].

post-training operation that reorients the axes of underlying factors to better align them with more interpretable or disentangled features. Importantly, this transformation does not alter the fundamental geometry of the observed data, preserving pairwise relationships and reconstruction quality while allowing for more intuitive interpretation and manipulation.

In recent deep learning literature, rotation based transformations have been used to enhance disentanglement, where the aim is to ensure that each latent space dimension

captures a distinct and independent generative factor [7, 23]. This is especially useful in applications where control, fairness and interpretability are important, such as in generative modelling, domain adaptation or bias mitigation [11, 53]. Compared to other disentanglement methods, factor rotation has the advantage of being model independent and computationally efficient, requiring only the latent space representations and optionally some supervision to guide the alignment.

As such, factor rotation plays a central role in this thesis as a mechanism to restructure the latent space of autoencoders in a more interpretable and semantically meaningful manner. By aligning latent space dimensions with selected attributes while minimizing correlations with confounding variables, factor rotation supports downstream tasks such as targeted suppression, controlled generation and fair classification.

3.2. Mathematical foundations of factor rotation

Factor rotation originates from classical factor analysis, a statistical technique used to model the covariance structure among observed variables in terms of a smaller number of unobserved underlying factors.

In the classical setting, a s-dimensional input variable $\mathbf{x} \in \mathbb{R}^s$ is modelled as a linear combination of n underlying factors $\mathbf{f} \in \mathbb{R}^n$ and an additive noise term:

$$\mathbf{x} = \mathbf{\Lambda}\mathbf{f} + \boldsymbol{\epsilon}, \tag{7}$$

where, $\mathbf{\Lambda} \in \mathbb{R}^{s \times n}$ is the factor loading matrix and $\boldsymbol{\epsilon} \in \mathbb{R}^s$ is a vector of unique errors, typically assumed to be uncorrelated with the underlying factors and with each other, such that $\mathbb{E}[\boldsymbol{\epsilon}] = \mathbf{0}$, $\mathrm{Cov}(\boldsymbol{\epsilon}) = \mathbf{\Psi}$ and $\mathbf{\Psi}$ is a diagonal matrix.

Assuming $\mathbb{E}[\mathbf{f}] = \mathbf{0}$ and $\mathrm{Cov}(\mathbf{f}) = \mathbf{I}$, the total covariance matrix of the observed data is:

$$\mathbf{\Sigma}_x = \mathrm{Cov}(\mathbf{x}) = \mathbf{\Lambda}\mathbf{\Lambda}^\top + \mathbf{\Psi} \tag{8}$$

To enhance interpretability, a linear transformation matrix $\mathbf{R} \in \mathbb{R}^{n \times n}$ is applied to rotate the factor axes:

$$\mathbf{\Lambda}^* = \mathbf{\Lambda}\mathbf{R}, \quad \mathbf{f}^* = \mathbf{R}^{-1}\mathbf{f} \tag{9}$$

This shows that rotation does not alter the model's ability to explain the data but may improve the semantic alignment of individual factors with observable characteristics [27].

3.3. Common factor rotation techniques

Factor rotation techniques can be broadly classified into two categories, orthogonal and oblique. Figure 3.2 shows a comparison of orthogonal and oblique rotation. These approaches differ in whether they preserve the independence (orthogonality) of the rotated factors and each technique offers distinct advantages depending on the modelling goals.

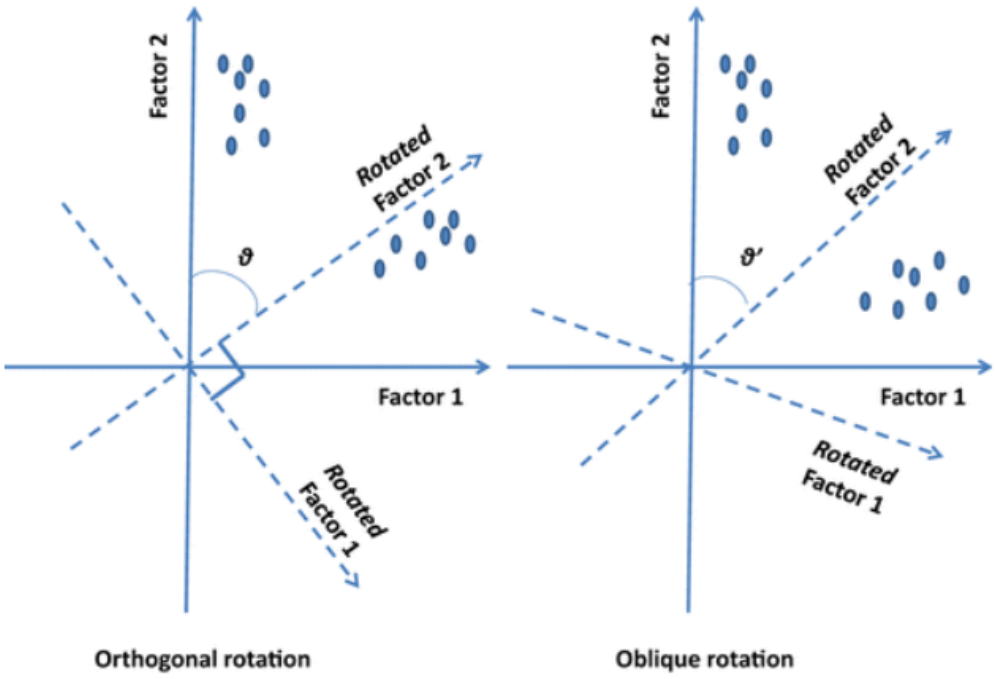


Figure 3.2: Comparison of orthogonal and oblique factor rotation. In orthogonal rotation (left), the rotated factors remain at 90° to each other, preserving independence between factors. In oblique rotation (right), the rotated factors are allowed to correlate, resulting in factor axes that are not orthogonal and potentially more interpretable factor solutions when underlying factors are expected to be correlated. Figure reproduced from Panaretos et al. [45].

3.3.1. Orthogonal rotation

Orthogonal rotation maintains the mutual independence between factors after the rotation. This transformation is defined by a square rotation matrix $\mathbf{R} \in \mathbb{R}^{n \times n}$ that satisfies the orthogonality constraint:

$$\mathbf{R}^\top \mathbf{R} = \mathbf{I}_n \tag{10}$$

where $\mathbf{I}_n$ is a identity matrix of dimension n. Given a representation matrix $\mathbf{X} \in \mathbb{R}^{m \times n}$, where m is the number of samples and n is the number of features, the rotated

representation of $\mathbf{X}$ is given as:

$$\tilde{\mathbf{X}} = \mathbf{X}\mathbf{R} \tag{11}$$

As $\mathbf{R}$ is orthogonal, the rotated model preserves the distances, angles and overall total variance structure from equation (8). It is particularly useful when independent factor axes are needed for interpretability, regularization or for maintaining theoretical properties such as decorrelation [2].

An orthogonal rotation matrix can be constructed through various techniques depending on the context and constraints of the application. A common method is to take an initial set of target directions like loading vectors or classifier coefficients and apply a **QR** decomposition, which decomposes a full rank matrix into an orthogonal matrix and an upper triangular matrix. The orthogonal matrix then serves as the desired orthogonal rotation matrix $\mathbf{R}$ [19]. In machine learning contexts, orthogonal rotation has been used to enhance clarity of internal representations without introducing redundancy.

3.3.2. Oblique rotation

Oblique rotation allows dependencies among the transformed factors, providing greater flexibility in aligning the representation with more interpretable or domain specific patterns. It is a generalization of orthogonal rotation [27]. The rotation matrix $\mathbf{R}' \in \mathbb{R}^{n\times n}$ in this case is not orthogonal:

$$\mathbf{R}'^{\top}\mathbf{R}' \neq \mathbb{I}_n \tag{12}$$

The rotated representation of $\mathbf{X}$ is again given as:

$$\tilde{\mathbf{X}} = \mathbf{X}\mathbf{R}' \tag{13}$$

As the rotation matrix $\mathbf{R}'$ is not orthogonal, the rotated factors become correlated. The covariance of the rotated factors is given by:

$$\Phi = (\mathbf{R}')^{-1}((\mathbf{R}')^{-1})^{\top} \tag{14}$$

The total covariance of the observed data after rotation is given by:

$$\mathbf{\Sigma^*}_x = \mathbf{\Lambda}^*\Phi\mathbf{\Lambda}^{*\top} + \mathbf{\Psi} = \mathbf{\Lambda}\mathbf{R}'(\mathbf{R}')^{-1}((\mathbf{R}')^{-1})^{\top}\mathbf{R}'^{\top}\mathbf{\Lambda}^{\top} = \mathbf{\Lambda}\mathbf{\Lambda}^{\top} + \mathbf{\Psi} = \mathbf{\Sigma}_x, \tag{15}$$

which remains unchanged from equation (8) [27].

This approach often yields factor axes that are more closely aligned with observable structure, even if some correlation between features is introduced [28].

Oblique rotation relaxes the orthogonality constraint and instead seeks an invertible matrix $\mathbf{R}'$ such that:

$$\det(\mathbf{R}') \neq 0 \tag{16}$$

This enables greater flexibility in aligning with target directions. In practice, oblique rotations can reveal subtler patterns or groupings in the data that are not apparent under strict orthogonality constraints [28].

3.4. Adapting factor rotation to autoencoder latent spaces

Instead of modifying the architecture of the autoencoder or the training procedure, this approach involves a post-training transformation of the latent space, guided by attribute specific direction. The goal is to reinterpret the learned latent space in a manner that enhances semantic alignment and disentanglement, while preserving the model's generative capacity [41].

In modern deep learning contexts such as autoencoders, there is no explicit factor loading matrix. Given a set of latent space representations $\mathbf{Z} \in \mathbb{R}^{m \times n}$ obtained from the encoder, where m denotes the number of samples and n is the latent space dimensionality, the first step involves identifying meaningful directions in the latent space. These directions correspond to interpretable attributes or features of interest and can be extracted through supervised classifiers trained on labelled data. A common approach is to train a logistic regression for each attribute and extract the weight vectors $\beta_1, \beta_2, \ldots, \beta_k$, each representing an attribute specific direction in the latent space and $k < n, k \in \mathbb{Z}^+$ [38].

These k vectors are then arranged as the first k columns of a matrix $\mathbf{B} \in \mathbb{R}^{n \times n}$, while the rest of columns of this matrix are filled with standard basis vector of $\mathbb{R}^n$ and $\mathbf{B}$ becomes the custom basis matrix for constructing the rotation matrix. For orthogonal rotation, $\mathbf{QR}$ decomposition is applied to $\mathbf{B}$ and the orthogonal matrix obtained from $\mathbf{QR}$ decomposition serves as the rotation matrix $\mathbf{R}$. For oblique rotation, the procedure remains the same except the first k columns of the orthogonal matrix is replaced by the k weight vectors. This yields a rotation matrix $\mathbf{R}'$ that is typically not orthogonal.

Once the rotation matrix is constructed, the transformed latent space representations are computed as $\tilde{\mathbf{Z}} = \mathbf{ZR}$ or $\tilde{\mathbf{Z}} = \mathbf{ZR}'$, where $\mathbf{Z}$ is the latent space. These rotated latent space representations are then passed through the decoder to reconstruct the input. Since the decoder was originally trained on latent space without any rotation, compatibility with the rotated latent space representations is maintained by transforming the decoder weights accordingly. Specifically, the decoder function d_θ is adapted to operate on $\tilde{\mathbf{Z}}$ by composing it with the inverse of the rotation matrix, $\mathbf{R^{-1}}$ or $(\mathbf{R}')^{-1}$ [41]. This preserves

the original reconstruction behaviour without the need for retraining, while enabling interpretability gains through latent space rotation.

The effectiveness of factor rotation is assessed by evaluating the semantic alignment of the rotated dimensions. Additionally, the degree to which specific dimensions can be manipulated such as by fixing or suppressing them to remove an attribute provides a qualitative indicator of disentanglement [38, 8]. At the same time, reconstruction quality must be monitored to ensure that the generative capacity of the autoencoder is not compromised by the transformation.

This method is modular and broadly applicable. It requires access to the latent space and labels and is independent of the encoder decoder architecture. Consequently, it serves as a lightweight interpretability technique that can be applied to a variety of autoencoder based models across datasets and tasks. Its simplicity, computational efficiency and generalizability make it a practical addition to representation learning pipeline that aims to promote structured and interpretable latent spaces.

4. Dataset overview

In this thesis, two publicly available facial image datasets, CelebA and UTKFace have been utilized. Both datasets contain rich attribute annotations and visual diversity, making them suitable for studying latent space transformations, interpretability and fairness in deep generative models.

4.1. CelebA

The CelebFaces Attributes Dataset (CelebA) is a large scale facial attributes dataset introduced by Liu et al. [39]. It consists of 202,599 facial images of 10,177 identities, each annotated with 40 binary attributes such as Male, Young, Smiling, Bald and more. These attributes offer semantic labelled data for training models aimed at disentangled representation learning. The dataset also includes five facial landmark locations per image, which allow for consistent alignment across images, an important step in ensuring spatial consistency in CAE. The images vary widely in terms of pose, background, lighting and expression, providing a challenging yet realistic benchmark for evaluating the generalizability of latent space representation techniques. In this thesis, CelebA is primarily used for experiments on attribute disentanglement and rotation alignment. The binary attribute labels facilitate supervised learning of rotation matrices, while the large number of samples supports training deep autoencoder architectures effectively.

4.2. UTKFace

The UTKFace dataset, introduced by Zhang et al. [62], consists of over 20,000 facial images with associated demographic labels including age, gender and ethnicity. Each image is annotated with a real valued age ranging from 0 to over 100, a binary gender label and a categorical ethnicity label covering five groups. The dataset includes a wide range of facial poses, expressions, lighting conditions and occlusions, contributing to its popularity in research on demographic fairness and bias in facial recognition systems. Unlike CelebA, which focuses on binary facial attributes, UTKFace provides both continuous and multi class labels, enabling more nuanced analysis of latent space transformations. In this thesis, UTKFace is used to evaluate whether the rotation based manipulations learned on one dataset generalize effectively across different label types and population distributions.

5. Methodology

This section outlines the methodological approaches proposed in this thesis for improving the interpretability and alignment of latent space representations in autoencoders using factor rotation techniques. A formal description of the latent space and its transformation is provided and then the theoretical foundations of orthogonal and oblique rotation in the context of autoencoder derived latent space representations are developed. The goal is to restructure the latent space such that meaningful attributes are aligned with specific directions in the latent space, thereby enhancing semantic disentanglement and facilitating downstream interpretability. The section includes mathematical derivations of the rotation matrices and their practical application to the representations learned by a CAE and a β-VAE.

5.1. Latent space representation and rotation

Let $\mathbf{Z} \in \mathbb{R}^{m \times n}$ denote the matrix of latent space representations extracted from the encoder, where m is the number of samples and n is the latent space dimensionality. Each row corresponds to the latent space for sample i and each column j denotes a latent space dimension over the dataset, where $i = 1, 2, \ldots, m$ and $j = 1, 2, \ldots, n$, while m is the number of samples and n denotes the dimensionality of the latent space. For most autoencoders trained with a standard normal prior or standard reconstruction loss, the empirical covariance matrix of $\mathbf{Z}$, denoted by $\Sigma = \frac{1}{m}\mathbf{Z}^\top\mathbf{Z}$, is approximately close to the identity matrix $\mathbf{I}_n$, assuming the latent space dimensions are roughly decorrelated.

Given this setup, consider a linear transformations of the latent space by multiplying $\mathbf{Z}$ with a full rank matrix $\mathbf{R} \in \mathbb{R}^{n \times n}$. The transformed latent space $\tilde{\mathbf{Z}} = \mathbf{ZR}$ represent a rotated version of the original latent space. This rotation is intended to align one or more of the transformed dimensions with meaningful attributes.

Crucially, the decoder remains valid under this transformation if the decoder weights are also adjusted accordingly. Suppose the decoder function is expressed as $d(\mathbf{v} + \mathbf{Z\Theta})$, where $d(\cdot)$ is a non-linear activation (e.g., ReLU or tanh), $\Theta \in \mathbb{R}^{n \times h}$ is the decoder weight matrix, $\mathbf{v} \in \mathbb{R}^h$ is a bias term and h is the dimensionality of the first hidden layer of the decoder. Then, the effect of the rotation is absorbed into the decoder as follows:

$$d(\mathbf{v} + \mathbf{ZRR^{-1}\Theta}) = d(\mathbf{v} + \tilde{\mathbf{Z}}\tilde{\mathbf{\Theta}}) \tag{17}$$

where $\tilde{\mathbf{Z}} = \mathbf{ZR}$ and $\tilde{\mathbf{\Theta}} = \mathbf{R^{-1}\Theta}$. This ensures that reconstruction performance remains intact and the decoder interprets the rotated features correctly.

5.2. Orthogonal rotation using logistic regression coefficients

The primary objective of this thesis is to align specific latent space dimensions with semantic attributes. This approach draws inspiration from methods used in interpretable representation learning, particularly in the context of post-training rotation techniques [51, 57].

Suppose a logistic regression is trained to predict a binary attribute $\mathbf{y} \in \{0,1\}^m$ using the latent space representation $\mathbf{Z}$, where 1 indicates the presence of the corresponding attribute in the images from which the latent space representation $\mathbf{Z}$ is derived. The model predicts the probability that the attribute $\mathbf{y} = 1$ given by:

$$P(\mathbf{y} = 1|\mathbf{Z}) = (1 + \exp(-\beta^\top \mathbf{Z} - b))^{-1}, \tag{18}$$

where $\beta \in \mathbb{R}^n$ and b are the learned coefficient vector and bias, respectively [16]. The coefficient vector β defines a linear decision boundary and its direction in the latent space defines how the attribute varies. The projection $\mathbf{Z}\beta \in \mathbb{R}^m$ defines a linear decision axis in latent space that best separates the two classes.

To align this direction with the first dimension in a rotated latent space, a method using **QR** decomposition is followed in this thesis, similar to techniques used in rotation based disentanglement. In the beginning, the vector β is standardized to have unit norm:

$$\bar{\beta} = \frac{\beta}{|| \beta ||_2} \tag{19}$$

where $|| \cdot ||_2$ is the Euclidean norm or ℓ_2 norm of a vector and then a matrix $\mathbf{B} \in \mathbb{R}^{n \times n}$ is constructed with $\bar{\beta}$ as its first column. The remaining columns are filled by standard basis vectors. Specifically, $\mathbf{B} = [\bar{\beta}, \mathbf{e}_2, \ldots, \mathbf{e}_n]$, where $\mathbf{e}_2, \ldots, \mathbf{e}_n$ are standard basis vectors of $\mathbb{R}^n$. A **QR** decomposition is performed on $\mathbf{B}$ to obtain the orthogonal matrix which is used for rotation as a rotation matrix $\mathbf{R}$. The rotation matrix $\mathbf{R} = [\bar{\beta}, \mathbf{q}_2, \ldots, \mathbf{q}_n]$ serves as an orthogonal rotation matrix, where $\mathbf{q}_2, \ldots, \mathbf{q}_n$ are the remaining orthonormal vectors from the orthogonal matrix obtained from the **QR** decomposition of $\mathbf{B}$. Since $\mathbf{R}^\top \mathbf{R} = \mathbf{I_n}$, this transformation preserves distances and variances in the latent space.

The new latent space dimensions $\tilde{\mathbf{Z}} = \mathbf{ZR}$ have their first component aligned with $\mathbf{Z}\beta$ (or $\mathbf{Z}\bar{\beta}$), while the others form an orthonormal basis that spans the remaining latent space orthogonal to $\mathbf{Z}\beta$.

The procedure described above naturally extends to multiple directions of interest. Given there are two models trained on the latent space for predicting two different binary attributes, say $\mathbf{y_1}$ and $\mathbf{y_2}$ such that $\mathbf{y_1}, \mathbf{y_2} \in \{0,1\}^m$, produces two coefficient vectors, $\beta_\mathbf{1}$ and $\beta_\mathbf{2}$. The vectors are standardized to $\bar{\beta_\mathbf{1}}$ and $\bar{\beta_\mathbf{2}}$ so that they have unit norm and

then the matrix $\mathbf{B}$ is constructed such that $\mathbf{B} = [\bar{\beta}_1, \bar{\beta}_2, \mathbf{e}_3, \ldots, \mathbf{e}_n]$, where $\mathbf{e}_3, \ldots, \mathbf{e}_n$ are standard basis vectors completing the span and then the **QR** decomposition of $\mathbf{B}$ is computed and the orthogonal matrix, obtained from this **QR** decomposition, forms the orthogonal rotation matrix $\mathbf{R}$. This ensures that the first few rotated dimensions are aligned with selected attributes, while preserving orthogonality across the full basis.

5.3. Oblique rotation with multiple aligned directions

Building upon the orthogonal rotation approach described before, oblique rotation relaxes the orthogonality constraint between latent space dimensions, allowing correlations between them. The goal is to transform the latent space dimensions such that the first few dimensions align with meaningful directions associated with selected attributes without enforcing orthogonality between them unlike the case of orthogonal rotation. The procedure introduced in this thesis is same as in the case of the orthogonal rotation, but with one major modification, after performing the **QR** decomposition of the custom basis matrix $\mathbf{B}$ to generate an orthogonal matrix, the original standardized coefficient vectors, that were obtained from logistic regression, are reinstated in place of their orthogonalized counterparts in the obtained orthogonal matrix to construct the final oblique rotation matrix $\mathbf{R}'$.

Similar to the orthogonal rotation, the directions for rotation are obtained from coefficient vectors $\beta_1, \beta_2, \ldots, \beta_k \in \mathbb{R}^n$, derived from logistic regression models trained to predict binary attributes from the latent space representation of an autoencoder and k is the number of selected attributes, $k < n$ and $k \in \mathbb{Z}^+$. The coefficient vectors are standardized to obtain $\bar{\beta}_{j_1}, \bar{\beta}_{j_2}, \ldots, \bar{\beta}_k$. To incorporate all directions simultaneously without enforcing orthogonality between them, a custom basis matrix $\mathbf{B} \in \mathbb{R}^{n \times n}$, with full rank, is constructed by placing $\bar{\beta}_1, \bar{\beta}_2, \ldots, \bar{\beta}_k$ in the first k columns and completing the basis with standard basis vectors such that,

$$\mathbf{B} = [\bar{\beta}_1, \bar{\beta}_2, \ldots, \bar{\beta}_k, \mathbf{e}_{k+1}, \ldots, \mathbf{e}_n] \tag{20}$$

A QR decomposition of $\mathbf{B}$ yields an orthogonal matrix $\mathbf{Q} = [\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_n]$, from the which the rotation matrix $\mathbf{R}' \in \mathbb{R}^{n \times n}$ is constructed as:

$$\mathbf{R}' = [\bar{\beta}_1, \bar{\beta}_2, \ldots, \bar{\beta}_k, \mathbf{q}_{k+1}, \ldots, \mathbf{q}_n] \tag{21}$$

The matrix $\mathbf{R}'$ retains the original directions in the first k dimensions while using orthogonal completion for the remaining ones, ensuring full rank and stability. However, since $\bar{\beta}_1, \bar{\beta}_2, \ldots, \bar{\beta}_k$ may be correlated, the overall transformation does not preserve orthogonality across all dimensions, making it an oblique rotation. This approach is

particularly useful when aligning multiple attribute related directions in the transformed representation, while still retaining compatibility with downstream models.

5.4. Analysis and downstream evaluation of rotated latent space representation

To assess the effectiveness of latent space rotation, using the procedure introduced in the subsection 5.2 and 5.3, for improving attribute alignment and reducing confounding influences, the transformed latent space representation is evaluated on downstream classification tasks. Additionally, the structure and disentanglement quality of the latent space itself are quantitatively analysed using a set of established representation metrics, the predictability matrix, modularity, separated attribute predictability (SAP) and the mutual information matrix. These metrics, drawn from disentanglement literature Ridgeway and Mozer [49], Kumar et al. [33], Chen et al. [9], provide complementary insights into how well individual latent space dimensions align with semantic attributes and whether the latent space exhibits meaningful factorization. The latent space representations are analysed using four metrics.

The predictability matrix $\mathbf{P} \in \mathbb{R}^{k \times n}$ is constructed by training a linear classifier to predict each ground truth attribute from each individual latent space dimension, where n is the dimensionality of the latent space and k is the number of semantic attributes and $k < n$, [33]. This matrix records classification performance for each pair of latent space dimension and selected attribute $(\mathbf{y}_i, \mathbf{z}_j)$, $\mathbf{y}_i$ denotes the i-th attribute selected for alignment, where $\mathbf{z}_j$ denotes the j-th latent space dimension, $i = 1, 2, \ldots, k$ and $j = 1, 2, \ldots, n$. A sparse and diagonally dominant predictability matrix suggests that specific latent space dimensions capture distinct attributes, indicative of properly aligned representations. In this thesis, only the latent space dimensions aligned with semantic attributes are taken into condiseration for computing the predictability matrix and thus $\mathbf{P} \in \mathbb{R}^{k \times k}$ is a square matrix.

The mutual information matrix, $\mathbf{M} \in \mathbb{R}^{k \times n}$, is used to measure the statistical dependency between each latent space dimension and each attribute. The entry $m_{ij} = I(\mathbf{y}_i, \mathbf{z}_j)$ represents the mutual information between attribute $\mathbf{y}_i$ and latent space dimension $\mathbf{z}_j$, estimated from empirical distribution of latent space samples and semantic attributes [9]. In this thesis, like the predictability matrix, only the latent space dimensions aligned with semantic attributes are taken into condiseration for computing the mutual info matrix and thus $\mathbf{M} \in \mathbb{R}^{k \times k}$ is a square matrix.

Modularity measures the degree to which each latent space dimension encodes information about a semantic attribute. Given a mutual information matrix $\mathbf{M} \in \mathbb{R}^{k \times n}$, where n is the number of latent space dimensions and k is the number of semantic attributes, each

element m_{ij} represents the mutual information between attribute $\mathbf{y}_i$ and latent space dimension $\mathbf{z}_j$. For each latent space dimension $\mathbf{z}_j$, a template vector $\mathbf{t}_j$ is constructed, which has the highest mutual information value $\theta_j = \max_i \mathbf{m}_{ij}$ at the position corresponding to the attribute with maximum mutual information and zero elsewhere. Mathematically, the i-th element of this template vector is given by,

$$t_{ij} = \begin{cases} \theta_j = \max_i m_{ij} & \text{if } i = \arg\max_i m_{ij} \\ 0 & \text{otherwise} \end{cases}$$

The deviation δ_j of the actual mutual information vector from this template vector is given by:

$$\delta_j = \frac{\sum_i (m_{ij} - \mathbf{t}_{ij})^2}{\theta_j^2 (k-1)}.$$

When $\delta_j = 0$, then it indicates perfect modularity, where latent space dimension $\mathbf{z}_j$ encodes only a single attribute, while $\delta_j = 1$ indicates equal mutual information among other dimensions. The modularity score for each dimension is given by

$$\mathbf{Mod}(\mathbf{z}_j) = 1 - \delta_j,$$

and the overall modularity of the representation is calculated as the average of these scores across all latent space dimensions [49].

The separated attribute predictability or SAP score measures the difference in predictive power between the top two latent space dimensions for each attribute. Let p_{ij} denote the predictability score of attribute $\mathbf{y}_i$ using dimension $\mathbf{z}_j$, then the SAP score for $\mathbf{y}_i$ is computed as

$$\mathbf{SAP}(\mathbf{y}_i) = p_{ij_1} - p_{ij_2}, \tag{22}$$

where j_1 and j_2 are the index correspond to the first and second most predictive latent space dimensions. A high SAP value indicates that one latent space dimension predominantly governs the variation in that attribute, which is desirable in disentangled representation [33]. Like modularity, the overall SAP score is reported as the average SAP score across all semantic attributes.

The downstream evaluation involves training a deep neural network (DNN) classifier to predict selected facial attributes based on reconstructed images obtained from the rotated latent space representations. The primary objective of this DNN classifier is to rate the reconstructed images generated from the rotated latent space based on the presence of the selected facial attribute. The DNN classifiers are trained on reconstructed images generated from the latent space without any rotation. After applying a rotation,

either orthogonal or oblique, to the latent space produced by the encoder, the rotated latent space is passed through the decoder to generate the reconstructed facial images. These images are then passed to the DNN classifier that predicts the presence of specific attribute of interest, in other words, the DNN classifiers rates these images based on the selected facial attribute.

To evaluate the impact of latent space rotation on fairness, one DNN classifier is trained on reconstructed images generated from the latent space without rotation, while two additional DNN classifiers are trained on reconstructed images generated from the latent space with rotation but with different attribute suppression strategies and finally, a comparison is made between these DNN classifiers. Model performance is evaluated using metrics, such as ROC AUC score. Additionally, subgroup specific metrics such as ROC AUC score grouped by selected attributes are reported to investigate fairness implications and to detect whether rotation and suppression strategies yield more balanced outcomes across subgroups.

The ROC AUC score, which stands for receiver operating characteristic area under the curve, is a comprehensive and widely adopted metric for evaluating the performance of binary classification models. It assesses the model's ability to distinguish between positive and negative classes across all possible classification thresholds. The ROC curve is a graphical plot that illustrates the relationship between the true positive rate and the false positive rate at various threshold settings [18]. The Area Under the Curve (AUC) provides a single scalar value summarizing the model's ability to rank positive instances higher than negative ones. A score of 1.0 indicates perfect discrimination, while a score of 0.5 suggests no discriminative power, equivalent to random guessing [21]. This metric is especially useful in scenarios involving imbalanced datasets, where accuracy alone can be misleading [47].

This extended evaluation pipeline, combining downstream and fairness evaluation with direct analysis of the latent space using representation metrics, provides a robust validation framework for the proposed latent space rotation and attribute alignment strategy. It not only assesses whether rotation preserve relevant attributes under suppression but also reveals how the underlying geometry of the latent space changes in terms of alignment, disentanglement and independence. These findings are crucial for determining whether the learned representations generalize across subgroups in a fair, interpretable and semantically meaningful way.

6. Evaluation and results

This section presents a comprehensive evaluation of the proposed framework for enhancing facial attribute representation through the application of factor rotation techniques on the latent space of autoencoders. The performance of the CAE and β-VAE models is assessed before and after applying orthogonal and oblique rotation techniques. By aligning specific latent space dimensions with selected facial attributes, the impact of rotation on attribute disentanglement, classification performance and bias mitigation is systematically analysed.

6.1. Experimental setup

To conduct the experiments, two types of autoencoders are trained on the CelebA dataset, a CAE and a β-VAE. The architecture of the CAE used in this thesis is depicted using figure 6.1 and figure 6.2 and that of the β-VAE is shown using figure 6.3 and figure 6.4. The training dataset comprised of 162,752 aligned and cropped RGB facial images, each resized to 64×64 pixels. The CAE and β-VAE are trained to minimize reconstruction loss, with the β-VAE additionally incorporating a β weighted Kullback–Leibler divergence term to promote disentangled representation. A value of $\beta = 1.5$ is chosen for the β-VAE to balance reconstruction quality and latent space disentanglement. The latent space dimension is set to 32 for both models. The CAE is trained for 10 epochs, while the β-VAE is trained for 50 epochs.

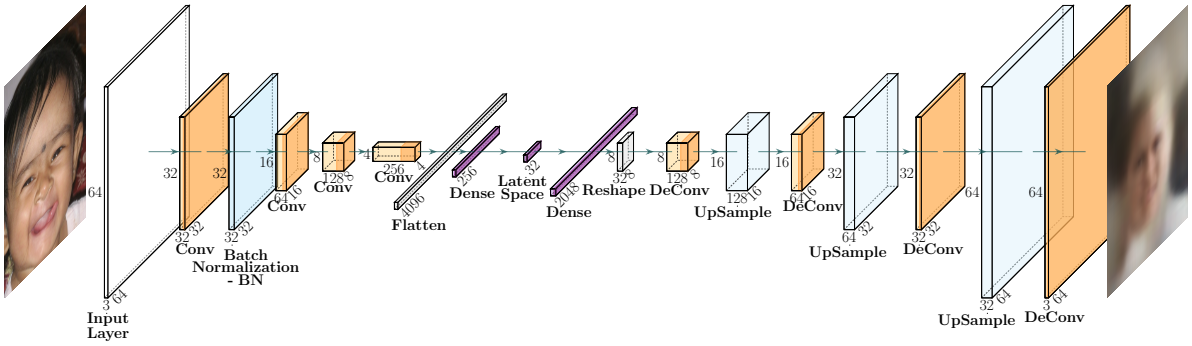


Figure 6.1: A schematic representation of the CAE architecture used in this thesis, detailing dimensional transformations from image to latent space and back. Image generated using the open source tool PlotNeuralNet by Iqbal [25].

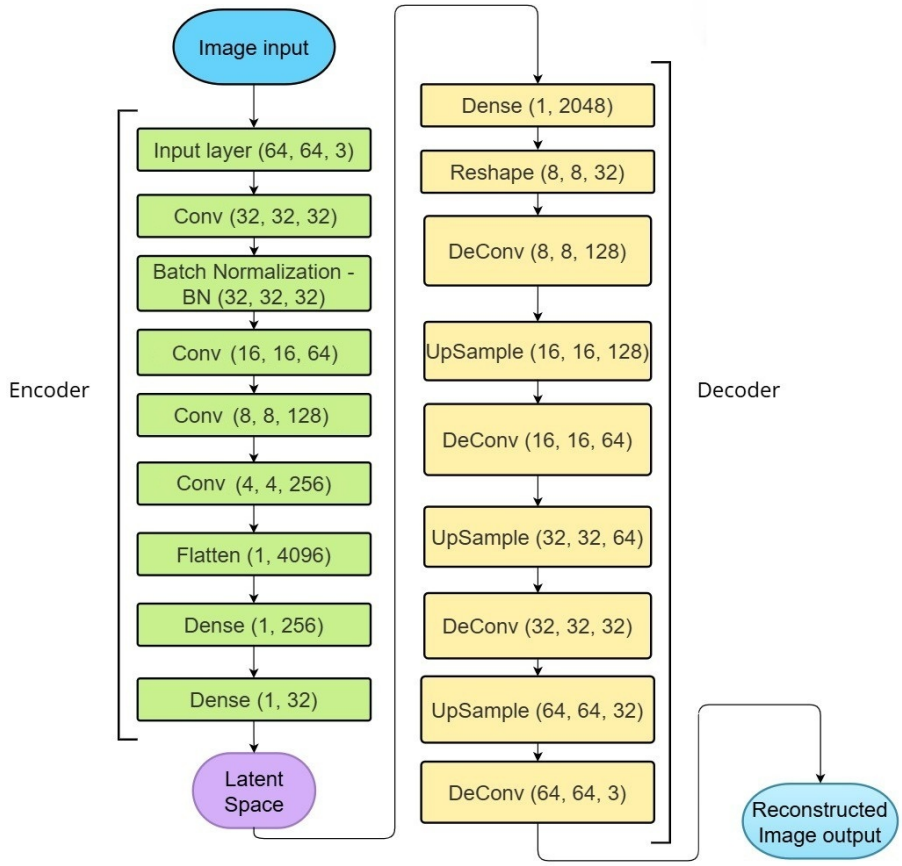


Figure 6.2: A flowchart representation of the CAE architecture used in this thesis.

After training, the latent space of both models are subjected to factor rotation techniques to improve the interpretability and alignment of latent space dimensions with semantic attributes. Specifically, logistic regression models are trained to predict selected facial attributes from the original latent space representation. The resulting coefficient vectors are used to construct orthogonal and oblique rotation matrices. These transformations are applied to the latent space to align selected dimensions with selected attributes while maintaining either orthogonality or allowing mild correlation between dimensions.

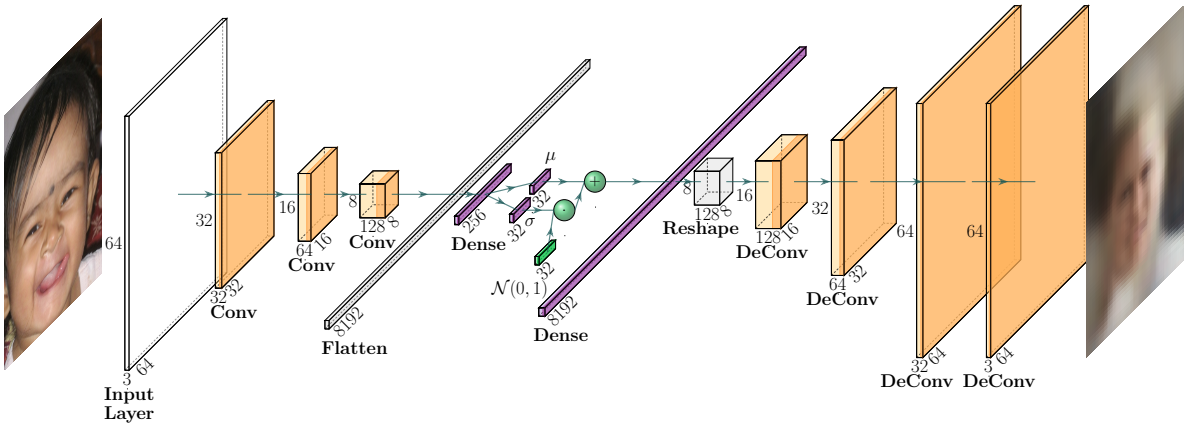


Figure 6.3: A schematic representation of the β-VAE architecture used in this thesis, detailing dimensional transformations from image to latent space and back. Image generated using the open source tool PlotNeuralNet by Iqbal [25].

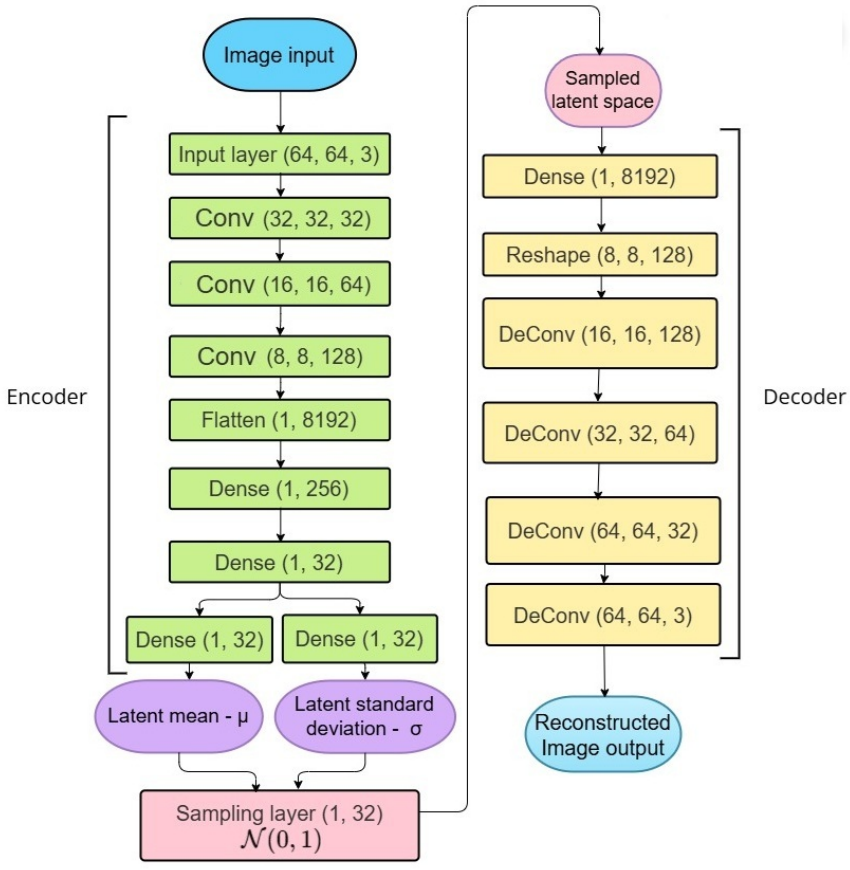


Figure 6.4: A flowchart representation of the β-VAE architecture used in this thesis.

Following rotation, attribute manipulation is performed by selectively modifying the aligned latent space dimensions, which were then passed through the decoder to reconstruct images. In addition to quantitative evaluation, latent space traversal is used to qualitatively

visualize disentanglement by varying selected latent space dimensions while keeping others fixed. The effect of these manipulations is also evaluated using a DNN classifier trained on reconstructed images generated using the latent space without any rotation. This classifier is used to predict presence of an attribute from images reconstructed from the latent space after rotation and manipulation, allowing for consistent evaluation of the visual and semantic interpretability of the rotated latent space dimensions. The architecture of this classifier is illustrated in figure 6.5 and figure 6.6. All neural network models used in this thesis are trained with early stopping.

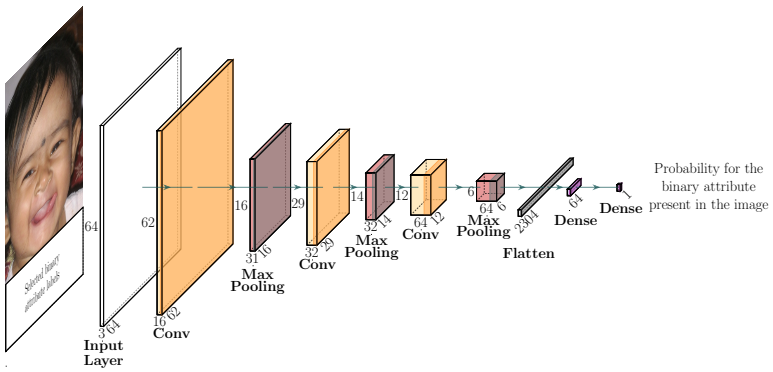


Figure 6.5: A schematic representation of the DNN classifier used in this thesis to evaluate the effect of latent space rotation on the autoencoder generated images. Image generated using the open source tool PlotNeuralNet by Iqbal [25].

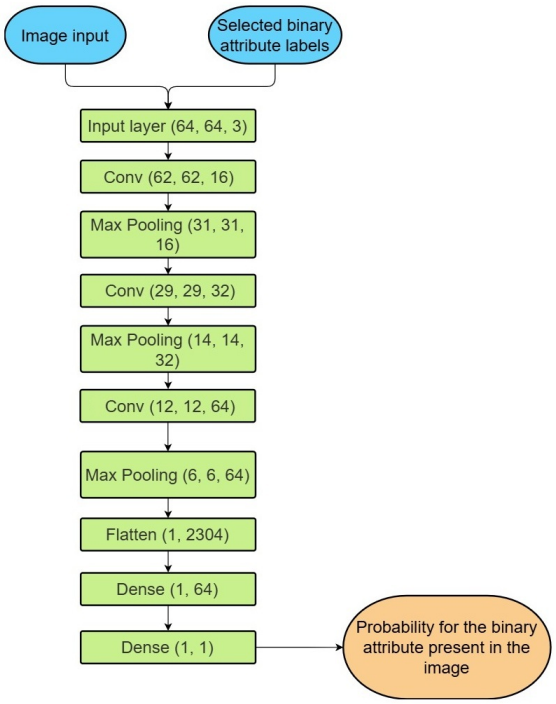


Figure 6.6: A flowchart representation of the architecture of the DNN classifier used in this thesis.

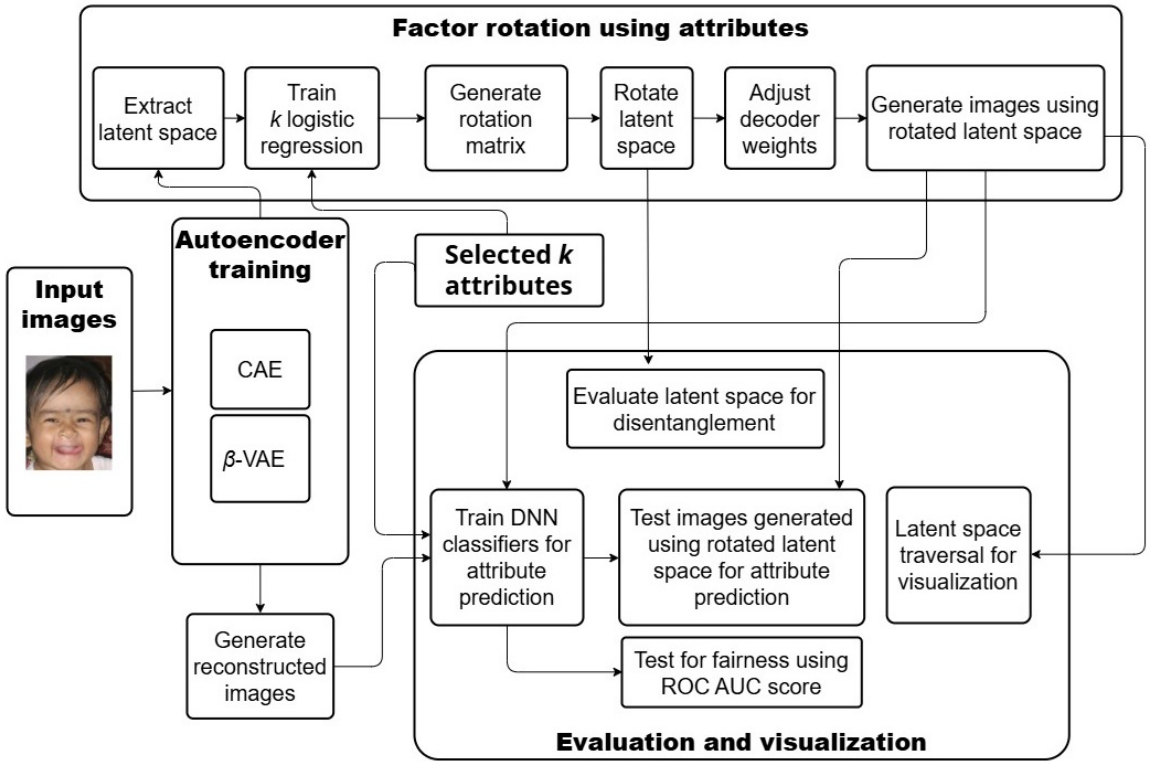


Figure 6.7: Overview of the proposed pipeline for enhancing facial attribute representation in autoencoder latent space using factor rotation. Input facial images are used to train a CAE and a β-VAE. Selected attributes are aligned with latent space dimensions through rotation matrices created using logistic regression coefficients. The rotated latent space is used to generate images, evaluate disentanglement, perform fairness testing and visualize latent space traversals. The pipeline supports both qualitative and quantitative evaluation of attribute separation and bias mitigation.

To analyse disentanglement, several metrics are computed, the predictability matrix, the mutual information matrix, the modularity and the separated attribute predictability (SAP) score are computed on the subset of latent space dimensions from the rotated latent space. Comparisons are made between the latent space representations with and without any rotation, to quantify improvements in attribute isolation and interpretability. An overview of the proposed pipeline is provided in figure 6.7.

6.2. Baseline and performance metrics

To evaluate the effects of latent space rotation on attribute classification and fairness, several performance metrics and baseline comparisons are carried out. These metrics cover reconstruction quality, attribute classification accuracy, disentanglement of latent space dimensions and fairness in downstream predictions. Evaluation is performed on a dataset which is not used for training both autoencoders and classifiers.

6.2.1. Baseline

For all tests, latent space without any rotation and its corresponding reconstructions are used as the baseline. This means that the original latent space representations produced by the CAE or β-VAE, without any factor rotation applied, served as the point of comparison when assessing classification performance, fairness and disentanglement.

6.2.2. Reconstruction quality

Reconstruction loss is used as a basic metric to monitor and compare the quality of reconstructed images across models. The CAE is trained using mean squared error (MSE) as its reconstruction loss, whereas the β-VAE used binary cross-entropy for the same purpose. These losses are computed on the validation set during training to guide optimization.

6.2.3. Disentanglement metrics

To evaluate the isolation of the individual attributes in the rotated latent space dimensions, disentanglement metrics are computed over the subset of latent space dimensions aligned with selected attributes. The predictability matrix, the mutual information matrix, the modularity and the separated attribute predictability (SAP) score are used for this purpose. They are used to compare the degree of attribute specific disentanglement introduced by orthogonal and oblique rotation relative to the latent space without rotation.

6.2.4. Attribute classification

To assess the extent to which the rotated latent space representation preserved attribute relevant information, a separate downstream classifier is trained to predict selected facial attributes from reconstructed images. Five binary attributes are selected from the CelebA dataset, Bald, Black_Hair, Pale_Skin, Smiling and Eyeglasses, for each attribute, a dedicated DNN classifier is trained on reconstructed images generated from the latent

space representation without rotation. The DNN classifiers are trained using a training-validation split consistent with the one used during autoencoder training.

After training, each DNN is evaluated on reconstructed images generated from rotated the latent space, including scenarios where a latent space dimension corresponding to the aligned attribute is modified or suppressed. Latent space suppression is applied by setting the rotated latent space dimension aligned with the selected attribute to a neutral value in the range of that dimension. This setup allowed the evaluation of how latent space control in the rotated space impacts the presence and separability of specific attributes in the regenerated images.

6.2.5. Fairness measures

To assess whether rotation and manipulation of latent space dimension can reduce group specific performance disparities, a targeted fairness evaluation is conducted using the β-VAE. Here, the latent space is rotated such that the first dimension is aligned with the Bald attribute, indicating baldness and the second with the Young attribute, indicating age, from the CelebA dataset. The effect of the Young attribute is then manipulated by shifting its corresponding latent space dimension to a value in the range of that dimension which enhances the accuracy of Bald attribute classification. DNN classifiers are used in this test for evaluation. The classifiers are trained on images reconstructed from the rotated latent space and while a baseline classifier is trained on images reconstructed from the latent space without any rotation. All classifiers are evaluated using ROC AUC score on the same test set. In addition to the overall ROC AUC score, group specific ROC AUC score is reported for both young and not young samples to assess potential fairness improvements.

6.3. Latent space disentanglement and visualization

To assess the quality of latent space disentanglement before and after applying factor rotation, predictability matrix, modularity, separated attribute predictability (SAP) score and mutual information gain matrix are evaluated and reported. These metrics serve as complementary indicators for measuring the extent to which individual latent space dimensions are aligned with specific semantic attributes.

6.3.1. Attribute predictability using predictability matrix

Attribute predictability quantifies how well a given attribute can be linearly predicted from individual latent space dimensions. For this purpose, a predictability matrix is

constructed using independent logistic regression models trained to predict each of five selected binary attributes, Bald, Black_Hair, Pale_Skin, Smiling and Eyeglasses, from each dimension of the latent space. Then the highest ROC AUC score obtained across all latent space dimensions for each attribute is reported, which reflects the maximum linear separability of that attribute in the latent space.

The results for both the CAE and the β-VAE models are presented in table 6.1. It is observed in the latent space with no rotation, all attributes show weak predictability, with values only slightly above 0.5. This suggests that in their raw form, the learned latent space representation is not well aligned with any individual attribute. However, after applying either orthogonal or oblique rotation, predictability improves significantly across all attributes and models. For instance, the predictability of the Bald attribute in the CAE increases from 0.520 to 0.910 after both orthogonal and oblique rotation, similar gains are observed for all other attributes, including in the case of the β-VAE. These results confirm that rotation effectively aligns specific attributes with individual latent space dimensions.

Model	**Attributes**	**No Rotation**	**Orthogonal**	**Oblique**
CAE	Bald	0.520	0.910	0.910
	Black_Hair	0.559	0.828	0.833
	Pale_Skin	0.629	0.902	0.917
	Smiling	0.538	0.796	0.833
	Eyeglasses	0.512	0.743	0.848
β-VAE	Bald	0.503	0.887	0.887
	Black_Hair	0.501	0.781	0.809
	Pale_Skin	0.501	0.874	0.895
	Smiling	0.602	0.777	0.789
	Eyeglasses	0.614	0.729	0.822

Table 6.1: Diagonal entries of the predictability matrix for each attribute under different rotation strategies for models trained on the CelebA dataset.

6.3.2. Mutual information score

To further verify disentanglement, a mutual information matrix is computed using the mutual information (MI) scores between each of the five attributes and the corresponding highest correlated latent space dimension. Higher MI values indicate stronger, more informative relationships between individual latent space dimensions and specific attributes.

Table 6.2 presents the diagonal values of the mutual information matrix, where each attribute is paired with its best aligned latent space dimension. Again, it is observed that rotation dramatically enhances the mutual information across both models and all attributes. For instance, the MI for Smiling increases from 0.002 (no rotation) to 0.190 (oblique rotation) in the CAE. These increases reinforce the conclusion that rotation clarifies and strengthens the relationship between latent space dimensions and semantic attributes, making these representations more interpretable and disentangled.

Model	Attributes	No Rotation	Orthogonal	Oblique
CAE	Bald	0.000	0.032	0.032
	Black_Hair	0.005	0.135	0.140
	Pale_Skin	0.005	0.057	0.065
	Smiling	0.002	0.144	0.190
	Eyeglasses	0.000	0.025	0.071
β-VAE	Bald	0.000	0.027	0.027
	Black_Hair	0.000	0.096	0.117
	Pale_Skin	0.001	0.047	0.055
	Smiling	0.017	0.126	0.138
	Eyeglasses	0.004	0.024	0.055

Table 6.2: Diagonal entries of the mutual information matrix for each attribute under different rotation strategies for models trained on the CelebA dataset. Higher values indicate stronger alignment of latent space dimensions with individual attributes.

6.3.3. Modularity and SAP score

Modularity and SAP score indicate how well each dimension in the latent space is aligned with one attribute. While modularity measures the extent to which each latent space dimension is associated with a single attribute, the SAP score reflects the difference in predictability between the top two most predictive latent space dimensions for each attribute, effectively quantifying the separation of attribute encoding.

The results in table 6.3 demonstrate that both metrics improve after rotation. For the CAE model, the average modularity increases from 0.896 (no rotation) to 0.952 (orthogonal rotation) and 0.933 (oblique rotation), while the average SAP score increases from 0.030 to 0.208 (orthogonal rotation) and 0.166 (oblique rotation). Similarly, for the β-VAE, the average modularity increases from 0.887 (no rotation) to 0.947 (orthogonal rotation) and 0.938 (oblique rotation), the average SAP score improves from 0.051 to 0.189 (orthogonal rotation) and 0.158 (oblique rotation). These improvements suggest

that factor rotation not only enhances alignment with attributes but also promotes greater disentanglement by reducing interference between latent space dimensions. Interestingly, the gains are consistent across both rotation types and models, indicating that even post-training transformations can meaningfully restructure the latent space.

Model	Rotation	Modularity	SAP score
CAE	No Rotation	0.896	0.030
	Orthogonal Rotation	**0.952**	**0.208**
	Oblique Rotation	0.933	0.166
β-VAE	No Rotation	0.887	0.051
	Orthogonal Rotation	**0.947**	**0.189**
	Oblique Rotation	0.938	0.158

Table 6.3: Average modularity and SAP score before and after factor rotation for the latent space of a CAE and a β-VAE trained on the CelebA dataset.

6.3.4. Visualization of latent space alignment

To qualitatively assess the degree of alignment between latent space dimensions and facial attributes, two complementary visualization techniques are utilized, latent space traversals, where individual latent space dimensions are systematically manipulated while keeping others constant and observing the effect on reconstructed facial images and correlation heat maps showing the correlation between each latent space dimension and the binary attribute labels,

The Pearson correlation coefficients between each latent space dimension and each of the five selected the binary attribute labels, in the case of both the CAE and β-VAE models are computed. The resulting heat maps, as shown in figure 6.8, reveals how different attributes are distributed across the latent space. In the CAE latent space without rotation, no single dimension is clearly aligned with any specific attribute. After applying orthogonal rotation, the heat map becomes significantly more structured, with attributes showing strong correlation with distinct latent space dimensions. This pattern becomes even more pronounced under oblique rotation, where attributes exhibit high correlation with specific rotated components while remaining minimally entangled with others. A pronounced improvement is observed figure 6.9, for the latent space of the β-VAE, before and after rotation due to the β-VAE's inductive bias toward disentanglement effect.

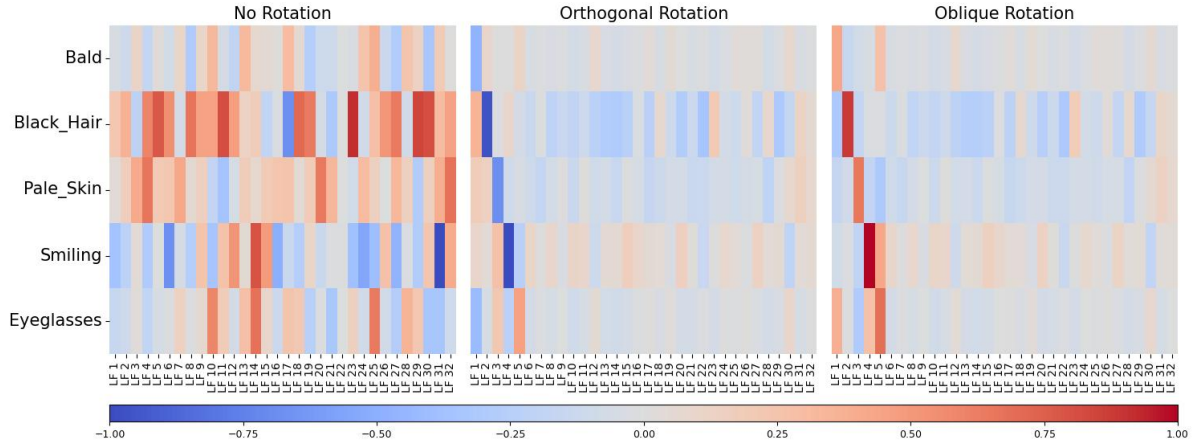


Figure 6.8: Correlation between the latent space dimensions of a CAE and five of the selected binary attributes before and after rotation

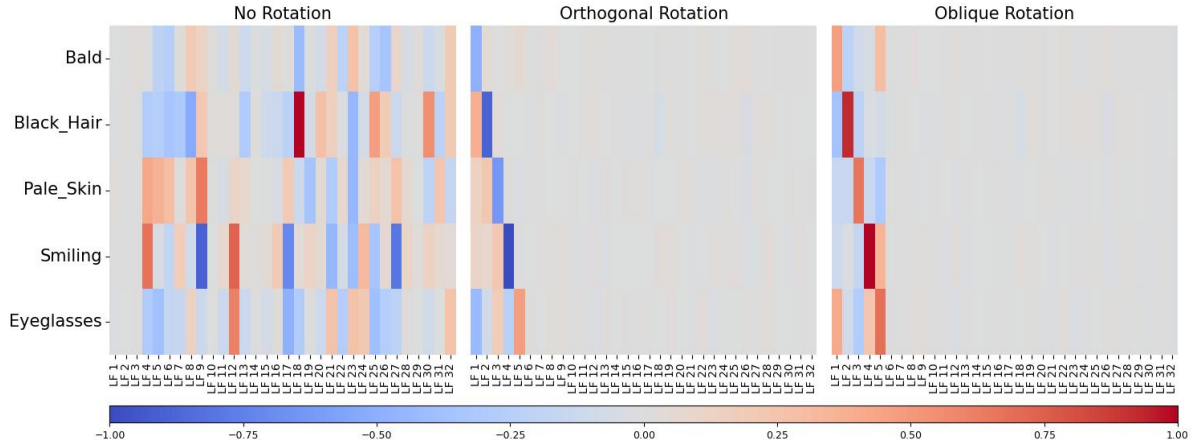
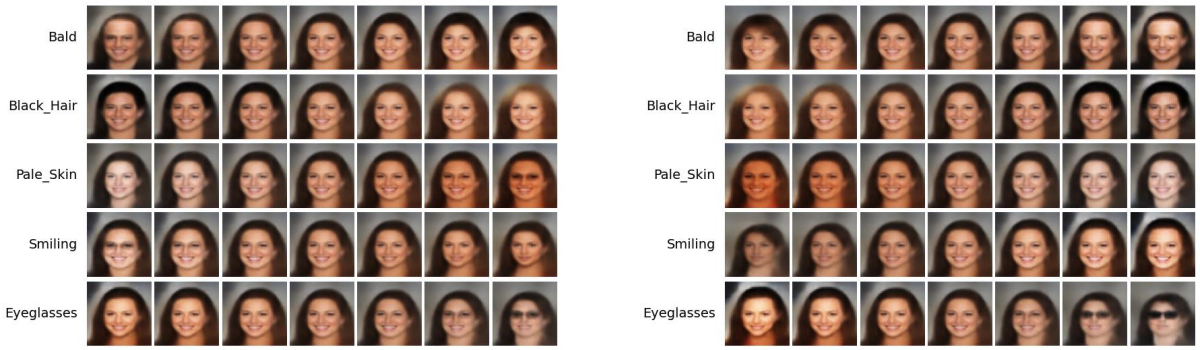


Figure 6.9: Correlation between the latent space dimensions of a β-VAE and five of the selected binary attributes before and after rotation

To further validate the disentanglement effect achieved through rotation, latent space traversals for the selected attributes using both the CAE and β-VAE models are generated. Specifically, the rotated dimensions aligned with Bald, Black_Hair, Pale_Skin, Smiling and Eyeglasses are selected and varied across a fixed range, keeping all other dimensions constant. The resulting images are displayed in figure 6.10 for CAE and figure 6.11 for β-VAE.

In the CAE model, latent space traversals after orthogonal and oblique rotation reveal the effectiveness of factor alignment. Following orthogonal rotation, traversals along selected dimensions, each aligned with a specific attribute such as bald or smiling, produce smoother and more isolated transformations in the reconstructed images. For example, increasing the value of the dimension aligned with bald progressively removes hair in a realistic manner, while the smiling dimension induces a gradual shift from neutral to

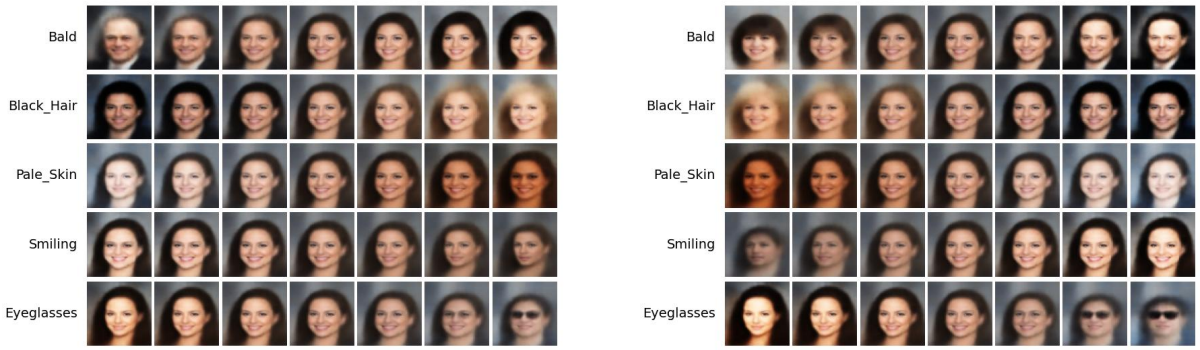
smiling expressions. With oblique rotation, the visual effect becomes even sharper and more attribute specific. The changes follow a clearer progression along the intended semantic axis, with minimal interference from unrelated facial features.



(a) Orthogonal rotation (b) Oblique rotation

Figure 6.10: Latent space traversal results after applying factor rotation techniques to the latent space of a CAE trained on the CelebA dataset. Each row corresponds to a latent space dimension aligned with a specific facial attribute and each column shows the reconstructed image as the value of that dimension is varied while others are held fixed. The traversal demonstrates how individual attributes are represented in the rotated latent space.

The β-VAE model exhibits a similar but even more distinct improvement in disentanglement after rotation. Although the β-VAE already imposes a degree of attribute separation due to its objective function, factor rotation significantly sharpens this structure. Traversals along rotated dimensions aligned with attributes produce highly consistent and semantically meaningful transformations, where changes are limited to the targeted attribute and other facial features remain largely unaffected. For instance, traversing the dimension aligned with Eyeglasses results in a clean interpolation between faces with and without eyeglasses. This demonstrates that factor rotation not only complements the disentanglement capabilities of the β-VAE but also enhances its practical utility for controlled generation and interpretation of facial features.



(a) Orthogonal rotation (b) Oblique rotation

Figure 6.11: Latent space traversal results after applying factor rotation techniques to the latent space of a β-VAE trained on the CelebA dataset.

These traversal plots serve as an intuitive visual confirmation of the increased interpretability and disentanglement achieved through factor rotation. Together with the correlation heat map analysis and the quantitative disentanglement metrics, they support the conclusion that both orthogonal and oblique rotation significantly enhance the semantic structure of the latent space. These results visually reinforce the quantitative findings from the four metrics used in subsection 6.3.1, 6.3.3 and 6.3.2.

6.4. Evaluating attribute manipulation with classifier predictions

To further assess the practical impact of factor rotation on the interpretability and controllability of the latent space, a focused evaluation is conducted by manipulating specific latent space dimensions and analysing downstream classifier predictions. To ensure that the effects of these manipulations are assessed independently of the latent space rotation pipeline. A separate DNN classifier is trained to predict each target attribute directly from the autoencoder reconstructed images. Specifically, five independent DNN classifiers are used, one for each of the selected facial attributes, Bald, Black_Hair, Pale_Skin, Smiling and Eyeglasses. These classifiers are trained on autoencoder reconstructed images, where each reconstructed image retained the original, unaltered latent space encoding of a real image from the CelebA dataset.

Each classifier is trained as a binary predictor using a standard convolutional deep neural network architecture as shown in figure 6.5 and figure 6.6, optimized to classify the presence of the target attribute from the autoencoder reconstructed images. The training process aims to replicate realistic attribute recognition in conditions unaffected by explicit latent space interventions.

After training, these DNN classifiers are evaluated on reconstructed images generated from an autoencoder, for three distinct cases, latent space with no rotation, with orthogonal

rotation and with oblique rotation. For the rotated cases, first the appropriate rotation is applied to the latent space representation of test images, to align the attributes with the respective latent space dimension. Then, for each sample, the latent space dimension that is highly correlated to the attribute being evaluated, is identified and the value of this dimension is set to a constant value like the average of the 10^{th} and 90^{th} percentile of this same dimension, effectively trying to suppress the expression of that attribute in the reconstructed image produced by the decoder. This manipulation is applied individually to each attribute aligned dimension for every attribute under evaluation, ensuring isolated intervention without modifying other components of the latent space.

The rotated and manipulated latent space is passed through the decoder to generate modified reconstructions, which are subsequently fed into the corresponding trained DNN classifiers. The goal of this evaluation is to observe how latent space manipulation under different rotation strategies affected the classifier's prediction scores for the presence of each attribute. If the rotation strategy has successfully disentangled and aligned specific latent space dimensions with interpretable attributes, then modifying those dimensions should lead to a shift in the classifier predictions.

To visualize the outcome of this procedure, histograms of the classifier output probabilities across the test dataset for each of the five attributes are plotted. These histograms are presented in Appendix A, figure C.1 and figure C.2 for the CAE and β-VAE models, respectively. For each attribute, the histogram plots enable a direct comparison of the classifier's predictions across, no rotation, orthogonal rotation and oblique rotation.

The results reveal that after rotation, a shift is observed between the manipulated distributions and distributions without any manipulation for all attributes, in both the CAE and β-VAE models. The distribution of predicted probabilities changes for all attributes, suggesting that the classifier correctly detects the manipulation of these attributes in the reconstructed images.

Overall, this classifier based evaluation demonstrates that factor rotation enhances the semantic alignment of individual latent space dimensions, revealing targeted manipulation results in coherent visual and predictive changes in the output. The histograms serve as both a qualitative and quantitative confirmation of the improved disentanglement.

6.5. Mitigating age bias in baldness prediction using latent space manipulation

In this subsection, a fairness based experiment aimed at mitigating demographic bias in the prediction of baldness from facial images is presented. Specifically, the potential confounding influence of the age on the prediction of baldness is investigated and evaluate

whether latent space disentanglement and targeted manipulation can improve fairness outcomes across age groups. This experiment is conducted using the CelebA dataset and serves as a concrete application of representation level fairness in autoencoder based models.

Baldness is generally more noticeable among older individuals, both in real world observations and within the CelebA dataset used in this thesis. This demographic imbalance can lead to biased representations in the latent space of an autoencoder, where features associated with baldness may be entangled with age related visual cues. Consequently, a classifier trained on reconstructed images might underperform on less represented groups, particularly young and bald individuals because their latent space representations deviate from the typical bald profile that includes age.

To evaluate and address this bias, two binary attributes from the CelebA dataset are selected, the Bald attribute, representing baldness, is selected as the primary attribute for prediction and the Young attribute, representing age, acting as a potential confounding factor. The impact of latent space rotation and manipulation on improving fairness in prediction of the Bald attribute is assessed.

The dataset is divided into training, validation and testing subsets. The attribute distribution is highly imbalanced across age and baldness categories, as shown in Table 6.4

Group	**Train**	**Validation**	**Test**
bald & young	895	52	106
bald & not young	2,816	371	305
not bald & not young	33,159	4,474	4,726
not bald & young	125,882	15,039	14,703

Table 6.4: Distribution of samples across Bald and Young attribute combinations in the CelebA dataset for training, validation and testing sets.

The training set comprises $162,752$ samples, out of which only 895 belong to the bald and young group, while $2,816$ correspond to bald and old individuals. In contrast, the majority of samples are not bald, with $33,159$ old and $125,882$ young individuals falling into this category. A similar distribution is maintained in the validation and test datasets, stressing on the challenge of predicting baldness in younger people due to their low representation. The extreme imbalance, particularly the low representation of bald and young samples in the dataset, highlights the need for fairness aware modelling strategies.

The experiment involves training three separate DNN classifiers using reconstructed images generated from a β-VAE and Bald attribute labels from the CelebA dataset, for

predicting the presence of the Bald attribute in these images. The first model, Classifier 1, is trained on images reconstructed from the latent space without any rotation and manipulation. This model serves as the baseline for comparison. The second model, Classifier 2, is trained on reconstructed images generated from an orthogonally rotated latent space. In this rotated space, the first latent space dimension is aligned with the Bald attribute, while the second dimension is aligned with the Young attribute. Before decoding, a targeted manipulation is applied to the second dimension of the orthogonally rotated latent space, which corresponds to age. Specifically, for samples that are both bald and young, this dimension is increased by the median value of the same dimension for samples that are bald but not young. Likewise, for samples that are not bald and not young, the second dimension of the orthogonally rotated latent space is incremented by the median value of the same dimension for samples that are not bald but young. The third model, Classifier 3, is identical to the Classifier 2 except follows a different targeted manipulation strategy for its second latent space dimension. In this strategy, the second dimension of the orthogonally rotated latent space, which corresponds to age, is globally shifted by the average of the median values of this same dimension for young and not young samples. The shift is applied in the direction that aligns with the sign of the correlation between the Young attribute and this latent space dimension after rotation. The results of this experiment are presented in Table 6.5 and figure 6.12

Group	**Classifier 1**	**Classifier 2**	**Classifier 3**
all	0.910	0.917	0.919
young	0.909	0.911	0.925
not young	0.859	0.864	0.860

Table 6.5: ROC AUC score of all classifiers predicting the Bald attribute before and after rotation of the latent space

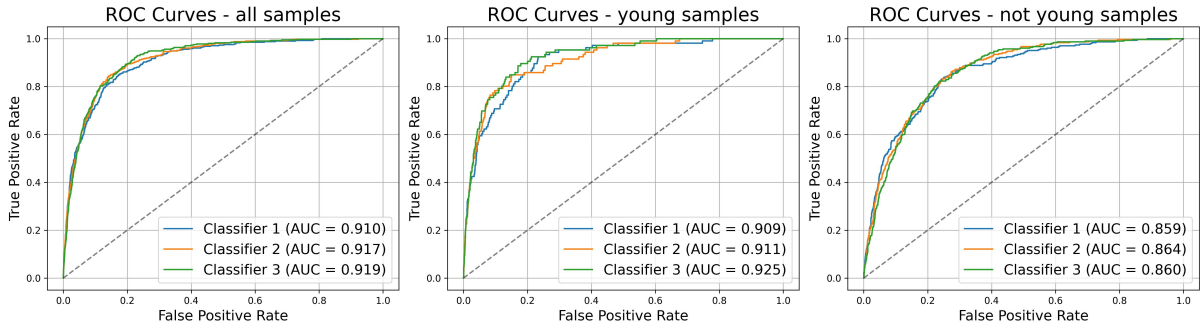


Figure 6.12: ROC curve for three classifiers predicting the Bald attribute across all samples (left), young samples (middle) and not young samples (right).

Across all samples, performance steadily improves from Classifier 1 to Classifier 2 and reaches its peak with Classifier 3, indicating that both rotation and age suppression enhance baldness prediction. For young samples, the ROC AUC score rises from 0.909 to 0.925, suggesting that the suppression of age related attribute improves performance on this subgroup. For not young samples, the ROC AUC score initially increases from 0.859 to 0.864, but slightly drops to 0.860. Despite this minor decrease, the final classifier still performs slightly better than the baseline. Also, the difference in ROC AUC score values between those of young and not young samples is reduced from 0.050 in Classifier 1 to 0.047 in Classifier 2, reflecting improved fairness.

These results demonstrate that orthogonal latent space rotation, followed by targeted or directionally consistent manipulation of a confounding attribute, can effectively enhance fairness and improve accuracy in facial attribute classification tasks. By reducing the latent space entanglement between baldness and age, the classifier generalizes better across age groups, especially for rare or under represented combinations such as young and bald. This experiment shows the potential of latent space disentanglement and manipulation techniques for achieving representation level fairness in generative models.

6.6. Validation on UTKFace dataset

To validate the generalizability of the proposed factor rotation methodology, both the CAE and the β-VAE models are evaluated on the UTKFace dataset. This dataset provides facial images annotated with age and gender, two salient and demographically meaningful attributes. The age attribute from the UTKFace dataset is converted to a binary attribute by using a threshold of 30 years, where samples with age below 30 is labelled as young (1) while those who are 30 and above are labelled as old (0). The focus of this evaluation is to assess how well the proposed methodology of applying factor rotation techniques to the latent space of an autoencoder generalizes across datasets and for this purpose, the impact of factor rotation is evaluated on three key criteria, disentanglement quality measured through modularity and SAP score, alignment of latent space dimensions with interpretable attributes using mutual information and the predictability of those attributes directly from latent space dimensions.

6.6.1. Attribute predictability using predictability matrix

Table 6.6 presents the predictability scores from the predictability matrix, computed as the classification accuracy for gender and age based on the rotated latent space dimensions. Without rotation, both models demonstrate moderate predictive power. For instance, the CAE achieves an attribute predictability of 0.566 and 0.500 for gender and age,

respectively. After orthogonal rotation, these scores notably improve to 0.842 and 0.747. The β-VAE follows a similar trend with improvements from 0.545 to 0.804 for gender and from 0.521 to 0.697 for age. Oblique rotation offers nearly identical gains for gender and age predictability. These results confirm that rotation enhances the semantic alignment of latent space dimensions, making attributes more separable.

Model	Attributes	No Rotation	Orthogonal	Oblique
CAE	gender	0.566	0.842	0.842
	age	0.500	0.747	0.803
β-VAE	gender	0.545	0.804	0.804
	age	0.521	0.697	0.749

Table 6.6: Diagonal entries of the predictability matrix for each attribute under different rotation strategies for models trained on the UTKFace dataset.

6.6.2. Mutual information score

To quantify the degree to which specific attributes are embedded into a particular latent space dimension, a mutual information matrix, comprising of the mutual information (MI) scores between the aligned latent space dimensions and each target attribute, are computed. As shown in Table 6.7, the diagonal values of mutual information matrix, indicating the correspondence between selected dimensions and their aligned attributes are substantially higher after rotation. For example, the MI between the latent space dimension aligned with the gender attribute and the gender attribute itself, increases from 0.012 to 0.203 and 0.149 for CAE and β-VAE, respectively, after orthogonal rotation. Similar improvements are observed for age attribute alignment, rising from 0.000 to 0.102 for CAE and from 0.003 to 0.067 for β-VAE, both after orthogonal rotation. Oblique rotation introduces a similar trend for both attributes, gender and age and for both CAE and β-VAE. This confirms the successful disentanglement and alignment introduced through rotation.

Model	Attributes	No Rotation	Orthogonal	Oblique
CAE	gender	0.012	0.203	0.203
	age	0.000	0.102	0.155
β-VAE	gender	0.012	0.149	0.149
	age	0.003	0.067	0.094

Table 6.7: Diagonal entries of the mutual information matrix for each attribute under different rotation strategies for models trained on the UTKFace dataset.

6.6.3. Modularity and SAP score

Table 6.8 reports the average modularity and SAP scores before and after rotation. For both CAE and β-VAE models, the application of orthogonal rotation substantially improved the disentanglement metrics. Specifically, for the CAE, average modularity increased from 0.456 to 0.974 and 0.878 using orthogonal and oblique rotation, respectively while average the SAP score rose from 0.115 to 0.204 and 0.137 with orthogonal and oblique rotation, respectively. Similarly, the β-VAE model exhibits an average modularity improvement from 0.468 to 0.975 and 0.851 using orthogonal and oblique rotation, respectively and the average SAP score shows a substantial jump from 0.018 to 0.173 and 0.118 for orthogonal and oblique rotation, respectively. These findings reflect that rotation of latent space dimensions, in particular, enhances its statistical independence and attribute separating capabilities. Although oblique rotation did not outperform orthogonal rotation, it maintained comparable modularity while offering moderate SAP improvements for β-VAE and stable results for CAE.

Model	Rotation	Modularity	SAP score
CAE	No Rotation	0.456	0.115
	Orthogonal Rotation	**0.974**	**0.204**
	Oblique Rotation	0.878	0.137
β-VAE	No Rotation	0.462	0.018
	Orthogonal Rotation	**0.975**	**0.176**
	Oblique Rotation	0.851	0.120

Table 6.8: Average modularity and SAP score before and after factor rotation for the latent space of a CAE and a β-VAE trained on the UTKFace dataset.

6.6.4. Visualization of latent space alignment

To qualitatively assess the effects of rotation on the semantic alignment of latent space dimensions, latent space traversal on the UTKFace dataset is visualised. In this procedure, the rotated latent space dimensions aligned with the selected target attributes are visualised and their values are varied while keeping the remaining dimensions fixed. This visual evidence further supports the claim that factor rotation facilitates disentangled and interpretable representations within the latent space.

(a) Orthogonal rotation (b) Oblique rotation

Figure 6.13: Latent space traversal results after applying factor rotation techniques to the latent space of a CAE on the UTKFace dataset. Each row corresponds to a latent space dimension aligned with a specific facial attribute and each column shows the reconstructed image as the value of that dimension is varied while others are held fixed. The traversal demonstrates how individual attributes are represented in the rotated latent space.

(a) Orthogonal rotation (b) Oblique rotation

Figure 6.14: Latent space traversal results after applying factor rotation techniques to the latent space of a β-VAE trained on the UTKFace dataset.

These experimental results on the UTKFace dataset reaffirm the effectiveness of latent space rotation strategies. Orthogonal rotation is particularly effective in enhancing disentanglement and attribute alignment, leading to improved interpretability and predictability of learned representations. Oblique rotation offers competitive performance, especially when considering multiple attributes simultaneously, due to its relaxed constraint on factor independence. Overall, the results suggest that the proposed methodology generalizes well across datasets and model architectures.

7. Limitations and discussion

In this thesis, the proposed method demonstrates that factor rotation techniques can enhance the semantic interpretability of latent space representation in autoencoders, however several important limitations need to be considered carefully.

The methodology proposed in this thesis relies on attribute classifiers trained using the original latent space to define rotation directions. This technique can introduce a dependency on the accuracy of these classifiers, which are used in determining semantic directions for rotating the latent space. If these classifiers fail to capture unbiased representation of the attributes, the resulting rotated latent space may misalign or distort the underlying semantics, undermining both interpretability and fairness goals.

A further limitation arises from the assumption of linear separability between attributes and latent space dimensions. Rotation directions are derived from the coefficients of linear models, while interpretable, it may oversimplify the true structure of the latent space. Certain attributes like age or gender may interact in non-linear ways with other visual cues like baldness and a linear alignment could miss these interactions or fail to fully disentangle correlated factors.

Another limitation is the availability of attribute labels. Datasets such as CelebA and UTKFace provide annotations for a limited set of facial attributes, which constrains the scope of disentanglement and fairness evaluation. Subtle variations that are not captured by these labels, may still be entangled in the latent space representation. As a result, the current evaluation metrics may not fully reflect the complexity of disentanglement or the presence of residual attribute leakage.

The strategy introduced in this thesis to analyse and enhance fairness involves suppressing specific attribute and depends on the assumption that the attribute is entirely encoded along a single latent space dimension after rotation. In practice, attributes tend to be distributed across multiple latent space dimensions, even after alignment. Suppressing one dimension may therefore not fully remove the influence of the attribute and in some cases, may lead to unintended artifacts or degradation of performance for certain subgroups.

The evaluation of robustness was limited to two relatively clean datasets with constrained variation in pose, lighting and occlusion. While promising results were obtained on CelebA and UTKFace, the method's applicability to more diverse, real world facial imagery remains uncertain. It is unclear how the rotation strategy would perform under more challenging conditions, such as low quality images, angled facial images or complex backgrounds.

8. Conclusion and outlook

In this thesis, the application of factor rotation techniques on the latent space of an autoencoder has been proposed to improve disentanglement and interpretability of facial attribute representation in the latent space. This thesis demonstrates that aligning latent space dimensions with specific attributes using factor rotation and followed by evaluation using disentanglement metrics, latent space traversals and classifier predictions, can significantly enhance the semantic clarity and controllability of the latent space of an autoencoder.

The latent space traversals showed that rotated latent space dimensions, yielded more realistic and isolated transformations of facial features with minimal interference from other attributes. Correlation heat maps before and after rotation revealed that rotation successfully decorrelated entangled features, especially in the β-VAE. Furthermore, manipulating attribute aligned latent space dimensions and evaluating classifier predictions highlighted that rotation enables more controlled and targeted editing of facial features without substantially degrading reconstruction quality. Both orthogonal and oblique rotation proved effective in aligning attributes with specific latent space dimension.

Rotation of latent space dimensions using factor rotation also demonstrated its usefulness in mitigating bias and enhancing fairness and accuracy across subgroups of samples. The results showed a reduction in prediction performance disparity between subgroups of samples, thereby indicating improvement in fairness. Overall, this thesis confirms that application of factor rotation on latent space of autoencoder provide a principled and interpretable framework for controlling bias in generative models, contributing to the development of fairer artificial intelligence systems in face analysis tasks.

There are several promising directions for extending this work. The rotation framework could be applied in combination with adversarial training or mutual information based regularization to further enhance disentanglement. Integrating dynamic or learnable rotation mechanisms directly into the autoencoder training pipeline could automate and optimize alignment for arbitrary downstream tasks. Additionally, exploring the use of these techniques on more diverse datasets or in domains beyond facial analysis such as medical imaging or human activity recognition could broaden their applicability. Finally, investigating fairness aware manipulation and suppression strategies using rotated latent spaces could contribute to building more transparent and equitable machine learning models.

References

[1] Abdi, H. Factor rotations in factor analyses. In *Encyclopedia of Research Methods for the Social Sciences* (Thousand Oaks, CA, 2003), M. Lewis-Beck, A. Bryman, and T. F. Liao, Eds., Sage, pp. 978–982. URL: `https://api.semanticscholar.org/CorpusID:14726876`.

[2] Abdi, H., and Williams, L. J. Principal component analysis. *WIREs Computational Statistics 2*, 4 (2010), 433–459. URL: `https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/wics.101`, arXiv:https://wires.onlinelibrary.wiley.com/doi/pdf/10.1002/wics.101, doi:10.1002/wics.101.

[3] Alemi, A., Poole, B., Fischer, I., Dillon, J., Saurous, R. A., and Murphy, K. Fixing a broken ELBO. In *Proceedings of the 35th International Conference on Machine Learning* (10–15 Jul 2018), J. Dy and A. Krause, Eds., vol. 80 of *Proceedings of Machine Learning Research*, PMLR, pp. 159–168. URL: `https://proceedings.mlr.press/v80/alemi18a.html`.

[4] Baur, C., Wiestler, B., Albarqouni, S., and Navab, N. *Deep Autoencoding Models for Unsupervised Anomaly Segmentation in Brain MR Images*. Springer International Publishing, 2019, pp. 161–169. doi:10.1007/978-3-030-11723-8_16.

[5] Bengio, Y., Courville, A., and Vincent, P. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence 35*, 8 (2013), 1798–1828. doi:10.1109/TPAMI.2013.50.

[6] Buolamwini, J., and Gebru, T. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency* (23–24 Feb 2018), S. A. Friedler and C. Wilson, Eds., vol. 81 of *Proceedings of Machine Learning Research*, PMLR, pp. 77–91. URL: `https://proceedings.mlr.press/v81/buolamwini18a.html`.

[7] Burgess, C. P., Higgins, I., Pal, A., Matthey, L., Watters, N., Desjardins, G., and Lerchner, A. Understanding disentangling in β-vae, 2018. URL: `https://arxiv.org/abs/1804.03599`, arXiv:1804.03599.

[8] Cha, J., and Thiyagalingam, J. Orthogonality-enforced latent space in autoencoders: An approach to learning disentangled representations. In *Proceedings of the 40th International Conference on Machine Learning* (23–29 Jul 2023), A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, Eds., vol. 202 of *Proceedings of Machine Learning Research*, PMLR, pp. 3913–3948. URL: `https://proceedings.mlr.press/v202/cha23b.html`.

[9] Chen, R. T. Q., Li, X., Grosse, R., and Duvenaud, D. Isolating sources of disentanglement in variational autoencoders, 2019. URL: https://arxiv.org/abs/1802.04942, arXiv:1802.04942.

[10] Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., and Abbeel, P. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in Neural Information Processing Systems* (2016), D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, Eds., vol. 29, Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper_files/paper/2016/file/7c9d0b1f96aebd7b5eca8c3edaa19ebb-Paper.pdf.

[11] Creager, E., Madras, D., Jacobsen, J.-H., Weis, M., Swersky, K., Pitassi, T., and Zemel, R. Flexibly fair representation learning by disentanglement. In *Proceedings of the 36th International Conference on Machine Learning* (09–15 Jun 2019), K. Chaudhuri and R. Salakhutdinov, Eds., vol. 97 of *Proceedings of Machine Learning Research*, PMLR, pp. 1436–1445. URL: https://proceedings.mlr.press/v97/creager19a.html.

[12] Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., and Darrell, T. Decaf: A deep convolutional activation feature for generic visual recognition. In *Proceedings of the 31st International Conference on Machine Learning* (Bejing, China, 22–24 Jun 2014), E. P. Xing and T. Jebara, Eds., vol. 32 of *Proceedings of Machine Learning Research*, PMLR, pp. 647–655. Number 1. URL: https://proceedings.mlr.press/v32/donahue14.html.

[13] Dosovitskiy, A., and Brox, T. Generating images with perceptual similarity metrics based on deep networks. In *Advances in Neural Information Processing Systems* (2016), D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, Eds., vol. 29, Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper_files/paper/2016/file/371bce7dc83817b7893bcdeed13799b5-Paper.pdf.

[14] Eastwood, C., and Williams, C. K. A framework for the quantitative evaluation of disentangled representations. In *6th International Conference on Learning Representations* (2018).

[15] Esmaeili, B., Wu, H., Jain, S., Bozkurt, A., Siddharth, N., Paige, B., Brooks, D. H., Dy, J., and van de Meent, J.-W. Structured disentangled representations. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics* (16–18 Apr 2019), K. Chaudhuri and M. Sugiyama, Eds., vol. 89 of *Proceedings of Machine Learning Research*, PMLR, pp. 2525–2534. URL: https://proceedings.mlr.press/v89/esmaeili19a.html.

[16] Fahrmeir, L., Kneib, T., Lang, S., and Marx, B. D. *Regression: Models, methods and applications*. Springer, Berlin, Heidelberg, 2013.

[17] Falola, Y., Churilova, P., Liu, R., Huang, C.-K., Delgado, J. F., and Misra, S. Generating extremely low-dimensional representation of subsurface earth models using vector quantization and deep autoencoder. *Petroleum Research 10*, 1 (2025), 28–44. URL: https://www.sciencedirect.com/science/article/pii/S2096249524000619, doi:10.1016/j.ptlrs.2024.07.001.

[18] Fawcett, T. An introduction to roc analysis. *Pattern Recognition Letters 27*, 8 (2006), 861–874. ROC Analysis in Pattern Recognition. URL: https://www.sciencedirect.com/science/article/pii/S016786550500303X, doi:10.1016/j.patrec.2005.10.010.

[19] Golub, G. H., and Van Loan, C. F. *Matrix Computations - 4th Edition*. Johns Hopkins University Press, Philadelphia, PA, 2013. URL: https://epubs.siam.org/doi/abs/10.1137/1.9781421407944, arXiv:https://epubs.siam.org/doi/pdf/10.1137/1.9781421407944, doi:10.1137/1.9781421407944.

[20] Goodfellow, I., Bengio, Y., and Courville, A. *Deep Learning*. MIT Press, 2016. Book in preparation for MIT Press. URL: http://www.deeplearningbook.org.

[21] Hanley, J. A., and McNeil, B. J. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology 143*, 1 (1982), 29–36. PMID: 7063747. arXiv:https://doi.org/10.1148/radiology.143.1.7063747, doi:10.1148/radiology.143.1.7063747.

[22] Higgins, I., Amos, D., Pfau, D., Racanière, S., Matthey, L., Rezende, D. J., and Lerchner, A. Towards a definition of disentangled representations. *CoRR abs/1812.02230* (2018). URL: http://arxiv.org/abs/1812.02230, arXiv:1812.02230.

[23] Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., and Lerchner, A. beta-VAE: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations* (2017). URL: https://openreview.net/forum?id=Sy2fzU9gl.

[24] Hinton, G. E., and Salakhutdinov, R. R. Reducing the dimensionality of data with neural networks. *Science 313*, 5786 (2006), 504–507. URL: https://www.science.org/doi/abs/10.1126/science.

1127647, arXiv:https://www.science.org/doi/pdf/10.1126/science.1127647, doi:10.1126/science.1127647.

[25] Iqbal, H. Harisiqbal88/plotneuralnet v1.0.0, Dec. 2018. doi:10.5281/zenodo.2526396.

[26] Jafari, M., Shoeibi, A., Khodatars, M., Ghassemi, N., Moridian, P., Delfan, N., Alizadehsani, R., Khosravi, A., Ling, S. H., Zhang, Y.-D., Wang, S.-H., Gorriz, J. M., Rokny, H. A., and Acharya, U. R. Automated diagnosis of cardiovascular diseases from cardiac magnetic resonance imaging using deep learning models: A review, 2022. URL: https://arxiv.org/abs/2210.14909, arXiv:2210.14909.

[27] Jennrich, R. I. *Rotation.* John Wiley & Sons, Ltd, 2018, ch. 10, pp. 279–304. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118489772.ch10, arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781118489772.ch10, doi:10.1002/9781118489772.ch10.

[28] Jennrich, R. I., and Bentler, P. M. Erratum to: Exploratory bi-factor analysis. *Psychometrika 78*, 3 (2013), 556–556. doi:10.1007/s11336-013-9346-0.

[29] Jennrich, R. I., and Sampson, P. F. Rotation for simple loadings. *Psychometrika 31*, 3 (1966), 313–323. doi:10.1007/BF02289465.

[30] Kaiser, H. F. The varimax criterion for analytic rotation in factor analysis. *Psychometrika 23*, 3 (1958), 187–200. doi:10.1007/BF02289233.

[31] Kim, H., and Mnih, A. Disentangling by factorising, 2019. URL: https://arxiv.org/abs/1802.05983, arXiv:1802.05983.

[32] Kingma, D. P., and Welling, M. Auto-encoding variational bayes, 2022. URL: https://arxiv.org/abs/1312.6114, arXiv:1312.6114.

[33] Kumar, A., Sattigeri, P., and Balakrishnan, A. Variational inference of disentangled latent concepts from unlabeled observations, 2018. URL: https://arxiv.org/abs/1711.00848, arXiv:1711.00848.

[34] Lample, G., Zeghidour, N., Usunier, N., Bordes, A., Denoyer, L., and Ranzato, M. Fader networks: Manipulating images by sliding attributes, 2018. URL: https://arxiv.org/abs/1706.00409, arXiv:1706.00409.

[35] Larsen, A. B. L., Sønderby, S. K., Larochelle, H., and Winther, O. Autoencoding beyond pixels using a learned similarity metric. In *Proceedings of The 33rd International Conference on Machine Learning* (New York, New York, USA, 20–22 Jun 2016), M. F. Balcan and K. Q. Weinberger, Eds., vol. 48 of *Proceedings of*

Machine Learning Research, PMLR, pp. 1558–1566. URL: https://proceedings.mlr.press/v48/larsen16.html.

[36] LeCun, Y., Bengio, Y., and Hinton, G. Deep learning. *Nature 521* (05 2015), 436–44. doi:10.1038/nature14539.

[37] Lecun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE 86*, 11 (1998), 2278–2324. doi:10.1109/5.726791.

[38] Li, X., Lin, C., Li, R., Wang, C., and Guerin, F. Latent space factorisation and manipulation via matrix subspace projection, 2020. URL: https://arxiv.org/abs/1907.12385, arXiv:1907.12385.

[39] Liu, Z., Luo, P., Wang, X., and Tang, X. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)* (December 2015).

[40] Locatello, F., Bauer, S., Lucic, M., Rätsch, G., Gelly, S., Schölkopf, B., and Bachem, O. Challenging common assumptions in the unsupervised learning of disentangled representations, 2019. URL: https://arxiv.org/abs/1811.12359, arXiv:1811.12359.

[41] Magri, L., and Doan, A. K. On interpretability and proper latent decomposition of autoencoders, 2022. URL: https://arxiv.org/abs/2211.08345, arXiv:2211.08345.

[42] Mao, X., Shen, C., and Yang, Y.-B. Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections. In *Advances in Neural Information Processing Systems* (2016), D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, Eds., vol. 29, Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper_files/paper/2016/file/0ed9422357395a0d4879191c66f4faa2-Paper.pdf.

[43] Masci, J., Meier, U., Cireşan, D., and Schmidhuber, J. Stacked convolutional auto-encoders for hierarchical feature extraction. In *Artificial Neural Networks and Machine Learning – ICANN 2011* (Berlin, Heidelberg, 2011), T. Honkela, W. Duch, M. Girolami, and S. Kaski, Eds., Springer Berlin Heidelberg, pp. 52–59.

[44] Neupert, T., Fischer, M. H., Greplova, E., Choo, K., and Denner, M. M. Introduction to machine learning for the sciences, 2022. URL: https://arxiv.org/abs/2102.04883, arXiv:2102.04883.

[45] Panaretos, D., Tzavelas, G., Vamvakari, M., and Panagiotakos, D. Investigating the role of orthogonal and non–orthogonal rotation in multivariate

factor analysis, in regard to the repeatability of the extracted factors: A simulation study. *Communications in Statistics - Simulation and Computation 48*, 7 (2018), 2165–2176. doi:10.1080/03610918.2018.1435803.

[46] Park, S., Hwang, S., Kim, D., and Byun, H. Learning disentangled representation for fair facial attribute classification via fairness-aware information alignment. *Proceedings of the AAAI Conference on Artificial Intelligence 35*, 3 (May 2021), 2403–2411. URL: https://ojs.aaai.org/index.php/AAAI/article/view/16341, doi:10.1609/aaai.v35i3.16341.

[47] Powers, D. M. W. Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation, 2020. URL: https://arxiv.org/abs/2010.16061, arXiv:2010.16061.

[48] Rezende, D. J., Mohamed, S., and Wierstra, D. Stochastic backpropagation and approximate inference in deep generative models. In *Proceedings of the 31st International Conference on Machine Learning* (Bejing, China, 22–24 Jun 2014), E. P. Xing and T. Jebara, Eds., vol. 32 of *Proceedings of Machine Learning Research*, PMLR, pp. 1278–1286. Number 2. URL: https://proceedings.mlr.press/v32/rezende14.html.

[49] Ridgeway, K., and Mozer, M. C. Learning deep disentangled embeddings with the f-statistic loss, 2018. URL: https://arxiv.org/abs/1802.05312, arXiv:1802.05312.

[50] Ronneberger, O., Fischer, P., and Brox, T. U-net: Convolutional networks for biomedical image segmentation, 2015. URL: https://arxiv.org/abs/1505.04597, arXiv:1505.04597.

[51] Rügamer, D., Kolb, C., Weber, T., Kook, L., and Nagler, T. Generalizing orthogonalization for models with non-linearities. In *Proceedings of the 41st International Conference on Machine Learning* (21–27 Jul 2024), vol. 235 of *Proceedings of Machine Learning Research*, PMLR, pp. 42796–42817. URL: https://proceedings.mlr.press/v235/rugamer24a.html.

[52] Sarhan, M. H., Navab, N., Eslami, A., and Albarqouni, S. Fairness by learning orthogonal disentangled representations. In *Computer Vision – ECCV 2020* (Cham, 2020), A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds., Springer International Publishing, pp. 746–761.

[53] Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., and Bengio, Y. Toward causal representation learning. *Proceedings of the IEEE 109*, 5 (2021), 612–634. doi:10.1109/JPROC.2021.3058954.

[54] Shen, Y., Yang, C., Tang, X., and Zhou, B. Interfacegan: Interpreting the disentangled face representation learned by gans, 2020. URL: `https://arxiv.org/abs/2005.09635`, arXiv:2005.09635.

[55] Thurstone, L. L. *Multiple-Factor Analysis: A Development and Expansion of the Vectors of Mind*. University of Chicago Press, Chicago, 1947.

[56] Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., and Manzagol, P.-A. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *J. Mach. Learn. Res. 11* (Dec. 2010), 3371–3408.

[57] Wei, Q., Zheng, W., Li, Y., Cheng, Z., Zeng, Z., and Yang, X. Controlling facial attribute synthesis by disentangling attribute feature axes in latent space. In *2023 IEEE International Conference on Image Processing (ICIP)* (2023), pp. 346–350. doi:10.1109/ICIP49359.2023.10223056.

[58] Xie, J., Xu, L., and Chen, E. Image denoising and inpainting with deep neural networks. In *Advances in Neural Information Processing Systems* (2012), F. Pereira, C. Burges, L. Bottou, and K. Weinberger, Eds., vol. 25, Curran Associates, Inc. URL: `https://proceedings.neurips.cc/paper_files/paper/2012/file/6cdd60ea0045eb7a6ec44c54d29ed402-Paper.pdf`.

[59] Yu, F., and Koltun, V. Multi-scale context aggregation by dilated convolutions, 2016. URL: `https://arxiv.org/abs/1511.07122`, arXiv:1511.07122.

[60] Zeiler, M. D., and Fergus, R. Visualizing and understanding convolutional networks, 2013. URL: `https://arxiv.org/abs/1311.2901`, arXiv:1311.2901.

[61] Zhang, B. H., Lemoine, B., and Mitchell, M. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (New York, NY, USA, 2018), AIES '18, Association for Computing Machinery, p. 335–340. doi:10.1145/3278721.3278779.

[62] Zhang, Z., Song, Y., and Qi, H. Age progression/regression by conditional adversarial autoencoder. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017), IEEE.

Appendix A Implementation details

All experiments were conducted using *Python* 3.11[1] within the open source platform *kaggle*[2], which also provides access to a large number of datasets including the CelebA and UTKFace datasets. All neural network models used were built and trained using *tensorflow* 2.18.0[3] and *keras* 3.5.0[4] on NVIDIA Tesla T4 GPU, while *numpy* 1.26.4[5] and *pandas* 2.2.3[6] were used for data manipulation and preprocessing, *matplotlb* 3.7.5[7] and *seaborn* 0.12.2[8] facilitated qualitative analysis through visualization, $scikit-learn$ 1.2.2[9] was employed for computing evaluation metrics as well as for training logistic regression models. All training and evaluation procedures, including latent space rotation and attribute suppression experiments, were executed within the constraints of this environment. To ensure reproducibility, random seeds were fixed across both *tensorflow* and *numpy*. Training times were optimized by utilizing data generators. To generate the neural network architecture diagrams in figure 6.1, figure 6.3 and figure 6.5, the open source tool PlotNeuralNet by Iqbal [25] was used. Flowchart diagrams in figure 6.2, figure 6.4, figure 6.6 and figure 6.7 were created using the online open source platform *visual paradigm – online productivity suite*[10]. The source code is available at: `https://github.com/Arindam240143/Autoencoder-factor-rotation`.

[1] `https://www.python.org/`
[2] `https://www.kaggle.com`
[3] `https://www.tensorflow.org/`
[4] `https://keras.io/`
[5] `https://numpy.org/`
[6] `https://pandas.pydata.org/`
[7] `https://matplotlib.org/`
[8] `hhttps://seaborn.pydata.org/`
[9] `https://scikit-learn.org/stable/`
[10] `https://online.visual-paradigm.com/`

Appendix B Model parametrization

Model	Parameter	Value
CAE	optimizer	"Adam"
	lr	1e-3
	loss	mean squared error (MSE)
	batch	32
	latent_size	32
	epochs	10 (CelebA) 30 (UTKFace)
	early_stopping_patience	5 (CelebA) 15 (UTKFace)
β-VAE	β	1.5 (CelebA) 2 (UTKFace)
	optimizer	"Adam"
	lr	1e-3
	loss	binary_crossentropy and KL Divergence
	batch	32
	latent_size	32
	epochs	50
	early_stopping_patience	20
	random_state	42
DNN classifier	optimizer	"Adam"
	lr	1e-3
	loss	binary_crossentropy
	batch	32
	epochs	30
	early_stopping_patience	20
Logistic regression	penalty	"l2"
	dual	False
	tol	1e-4
	C	1.0
	fit_intercept	True
	intercept_scaling	1
	class_weight	None
	random_state	None
	solver	"lbfgs"
	max_iter	100 (rotation matrix creation) 200 (disentanglement metrics)
	multi_class	"auto"
	verbose	0
	warm_start	False
	n_jobs	None
	l1_ratio	None

Table B.1: Training parameters for models used in this thesis

Appendix C Additional plots

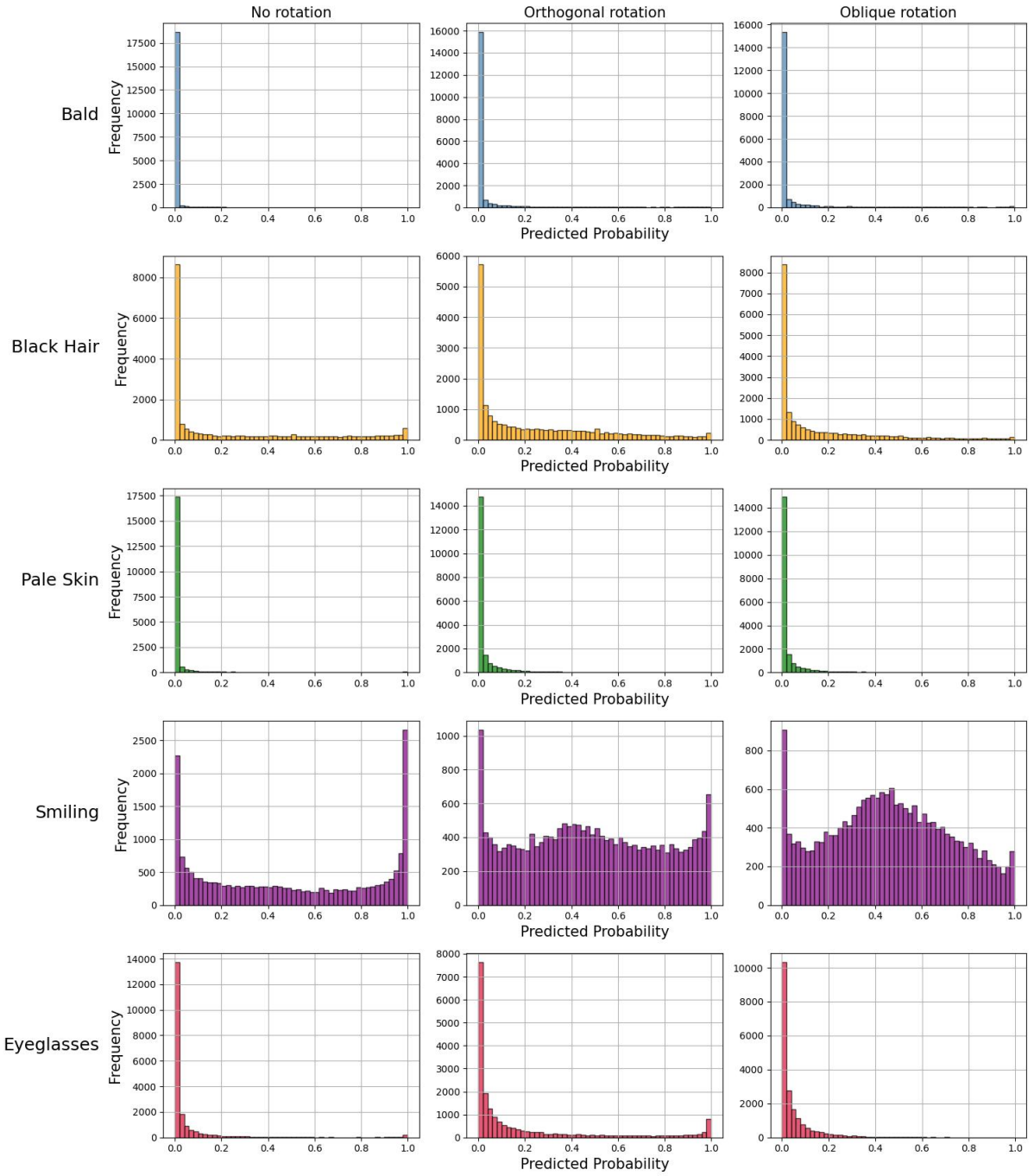


Figure C.1: Histogram of DNN classifier prediction probabilities for five selected facial attributes based on reconstructed images from CAE. Each subplot compares predictions under three conditions, no rotation, orthogonal rotation and oblique rotation. In rotation cases, the latent space dimensions aligned with the corresponding attribute is manipulated and the effects are visualised.

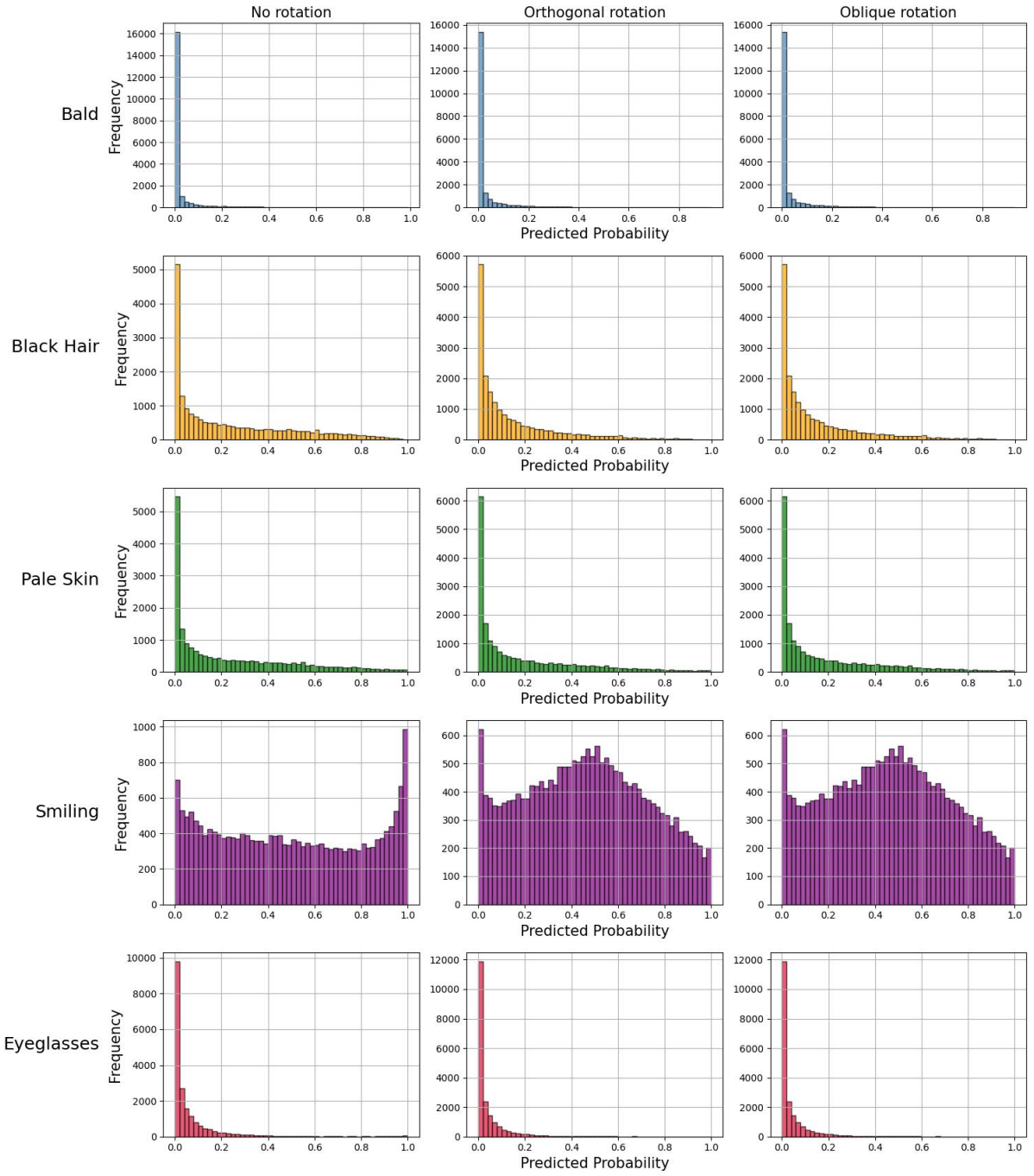


Figure C.2: Histogram of DNN classifier prediction probabilities for five selected facial attributes using reconstructed images from β-VAE. Similar to the case of the CAE, predictions are shown under no rotation, orthogonal rotation and oblique rotation, following manipulation of attribute aligned latent space dimensions. The results illustrate the effect of disentangled control on classifier responses.

Eidesstattliche Versicherung

(Affidavit)

PAL, ARINDAM

Name, Vorname
(surname, first name)

230143

Matrikelnummer
(student ID number)

☐ Bachelorarbeit
(Bachelor's thesis)

☑ Masterarbeit
(Master's thesis)

Titel
(Title)

ENHANCING FACIAL ATTRIBUTE REPRESENTATION IN AUTOENCODER LATENT SPACE USING FACTOR ROTATION TECHNIQUES

Ich versichere hiermit an Eides statt, dass ich die vorliegende Abschlussarbeit mit dem oben genannten Titel selbstständig und ohne unzulässige fremde Hilfe erbracht habe. Ich habe keine anderen als die angegebenen Quellen und Hilfsmittel benutzt sowie wörtliche und sinngemäße Zitate kenntlich gemacht. Die Arbeit hat in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde vorgelegen.

I declare in lieu of oath that I have completed the present thesis with the above-mentioned title independently and without any unauthorized assistance. I have not used any other sources or aids than the ones listed and have documented quotations and paraphrases as such. The thesis in its current or similar version has not been submitted to an auditing institution before.

DORTMUND, 20.06.2025

Ort, Datum
(place, date)

Arindam Pal

Unterschrift
(signature)

Belehrung:
Wer vorsätzlich gegen eine die Täuschung über Prüfungsleistungen betreffende Regelung einer Hochschulprüfungsordnung verstößt, handelt ordnungswidrig. Die Ordnungswidrigkeit kann mit einer Geldbuße von bis zu 50.000,00 € geahndet werden. Zuständige Verwaltungsbehörde für die Verfolgung und Ahndung von Ordnungswidrigkeiten ist der Kanzler/die Kanzlerin der Technischen Universität Dortmund. Im Falle eines mehrfachen oder sonstigen schwerwiegenden Täuschungsversuches kann der Prüfling zudem exmatrikuliert werden. (§ 63 Abs. 5 Hochschulgesetz - HG -).

Die Abgabe einer falschen Versicherung an Eides statt wird mit Freiheitsstrafe bis zu 3 Jahren oder mit Geldstrafe bestraft.

Die Technische Universität Dortmund wird ggf. elektronische Vergleichswerkzeuge (wie z.B. die Software „turnitin") zur Überprüfung von Ordnungswidrigkeiten in Prüfungsverfahren nutzen.

Die oben stehende Belehrung habe ich zur Kenntnis genommen:

Official notification:
Any person who intentionally breaches any regulation of university examination regulations relating to deception in examination performance is acting improperly. This offense can be punished with a fine of up to EUR 50,000.00. The competent administrative authority for the pursuit and prosecution of offenses of this type is the Chancellor of TU Dortmund University. In the case of multiple or other serious attempts at deception, the examinee can also be unenrolled, Section 63 (5) North Rhine-Westphalia Higher Education Act (*Hochschulgesetz, HG*).

The submission of a false affidavit will be punished with a prison sentence of up to three years or a fine.

As may be necessary, TU Dortmund University will make use of electronic plagiarism-prevention tools (e.g. the "turnitin" service) in order to monitor violations during the examination procedures.

I have taken note of the above official notification:*

DORTMUND, 20.06.2025

Ort, Datum
(place, date)

Arindam Pal

Unterschrift
(signature)

***Please be aware that solely the German version of the affidavit ("Eidesstattliche Versicherung") for the Bachelor's/ Master's thesis is the official and legally binding version.**