FaceMINT: A Library for Gaining Insights into Biometric Face Recognition via Mechanistic Interpretability

Peter Rot[a,b], Robert Jutreša[a], Peter Peer[a], Vitomir Štruc[b], Walter Scheirer[c] and Klemen Grm[b]

[a]Faculty of Computer and Information Science, University of Ljubljana, Večna pot 113, 1000, Ljubljana, Slovenia
[b]Faculty of Electrical Engineering, University of Ljubljana, Tržaška c. 25, 1000, Ljubljana, Slovenia
[c]Department of Computer Science and Engineering, University of Notre Dame, IN 46556, Notre Dame, USA

ARTICLE INFO

Keywords:
face recognition
biometrics
mechanistic interpretability
sparse autoencoder
library

ABSTRACT

Deep-learning models, including those used in biometric recognition, have achieved remarkable performance on benchmark datasets as well as real-world recognition tasks. However, a major drawback of these models is their lack of transparency in decision-making. Mechanistic interpretability has emerged as a promising research field intended to help us gain insights into such models, but its application to biometric data remains limited. In this work, we bridge this gap by introducing the FaceMINT library, a publicly available Python library (build on top of Pytorch) that enables biometric researchers to inspect their models through mechanistic interpretability. It provides a plug-and-play solution that allows researchers to seamlessly switch between the analyzed biometric models, evaluate state-of-the-art sparse autoencoders, select from various image parametrizations, and fine-tune hyperparameters. Using a large scale Glint360K dataset, we demonstrate the usability of FaceMINT by applying its functionality to two state-of-the-art (deep-learning) face recognition models: AdaFace, based on Convolutional Neural Networks (CNN), and SwinFace, based on transformers. The proposed library implements various sparse auto-encoders (SAEs), including vanilla SAE, Gated SAE, JumpReLU SAE, and TopK SAE, which have achieved state-of-the-art results in the mechanistic interpretability of large language models. Our study highlights the promise of mechanistic interpretability in the biometric field, providing new avenues for researchers to explore model transparency and refine biometric recognition systems. The library is publicly available at `www.gitlab.com/peterrot/facemint`.

1. Introduction

Deep neural networks are often considered non-transparent as their decision-making processes are difficult for humans to fully understand [21]. This is particularly concerning for high-stakes applications such as face recognition, which is increasingly vital in areas like personal authentication and law enforcement [36]. While neural network-based biometric models achieve state-of-the-art accuracy, their practical utility is often limited to scenarios where detailed explainability is not strictly required. However, in applications where incorrect model decisions can have serious consequences (e.g., denial of border crossing or wrongful accusation), explainability becomes crucial. Face recognition templates, for instance, encode information in a highly entangled and compressed manner, where individual features encapsulate multiple factors of variation, making them difficult for humans to interpret [46]. Similarly, automated recognition decisions are usually the results of complex processing operations across hierarchies of model layers, leading to decision outcomes that are challenging to explain and understand. Privacy laws and regulations, such as GDPR, explicitly mandate that any automated decision affecting individuals must be interpretable, ensuring transparency, accountability, and fairness in AI-driven processes. Additionally, the EU AI Act emphasizes interpretability as a key prerequisite for ensuring human oversight and accountability in AI systems, which are recognized as essential components of trustworthy AI [12]. To address these challenges, the field of Explainable Artificial Intelligence (XAI) and Explainable Face Recognition (XFR) seeks to provide insights into the decision-making processes of these models [50].

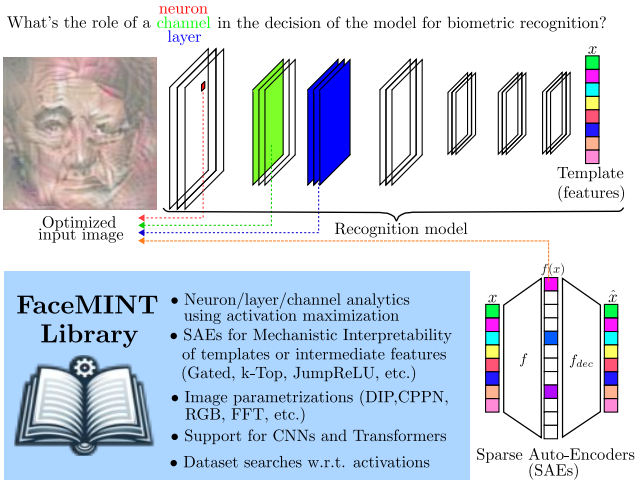


Figure 1: The FaceMINT library provides tools for the mechanistic interpretability of CNN and transformer-based biometric models, allowing researchers to analyze neurons, channels, and layers using activation maximization. It includes sparse autoencoders, advanced image parameterizations, and features like activation-based dataset searches to streamline analysis.

peter.rot@fe.uni-lj.si (P. Rot)
ORCID(s): 0000-0002-4491-2744 (P. Rot); 0000-0001-9744-4035 (P. Peer); 0000-0002-3385-5780 (V. Štruc); 0000-0001-9649-8074 (W. Scheirer); 0000-0002-3637-8182 (K. Grm)

While interpretability can be approached from different angles, *mechanistic interpretability* focuses on building a bottom-up understanding of how individual components of deep learning models function. This approach provides additional insights into how a model arrives at its decisions by analyzing the role of individual components (e.g., neurons, channels, or layers) in constituting the system as a whole [27]. Recent advances in mechanistic interpretability have shown promise in understanding complex models, particularly in the context of large language models [7, 4], yet this potential remains largely underexplored in biometric recognition. Despite the unique characteristics of biometric models, little effort has been made to tailor such interpretability methods to uncover insights specific to biometrics.

At the same time, researchers have proposed many approaches to explain the decision-making processes of biometric models. These proposed solutions are often designed for specific architectures, and adopting them for other biometric models and modalities requires extensive modifications. In [19], the authors propose an *explanation map* for biometric verification tasks, where each pixel of an image pair is assigned a similarity score, which allows a confidence score of the final biometric matching decision to be derived. This is a typical example of *post-hoc* explanation that does not account for the actual mechanism implemented by the AI model making the decision. In [20], the authors propose a method to train face recognition models to be more explainable to begin with. This is an exciting prospect for the future of explainable AI, however for the time being it is necessary to explain and interpret existing state-of-the-art biometric models that have been trained without inherent explainability of interpretability in mind. In [36], the authors show the LIME model for assigning feauture importance can highlight visually salient features in face recognition tasks. In [47], the authors use similar feature importance methodologie to derive a *face image quality* metric for biometric tasks. We note that most of these proposed methods are specific to either biometric models or training methodologies. In turn, this makes it challenging to interpret diverse biometric models effectively, limiting the broader applicability of current methods across different architectures. in contrast, the sparse autoencoder approach used here can be adapted to arbitrary models, does not require re-training or modification, and enables deep inspection of the underlying model mechanisms as opposed to explaining the model decisions alone.

To address these limitations, we introduce in this paper the FaceMINT library, a plug-and-play Python library (using PyTorch) that enables researchers to: (i) easily integrate, analyze and study different biometric models from an explainability point of view, (ii) experiment with methods for mechanistic interpretability, and (iii) visualize result of the explainability analysis. The library supports state-of-the-art techniques for mechanistic interpretability and facilitates analysis at different model levels, i.e., neuron, channel, and layer levels, as illustrated in Figure 1. Using the library's functionality, we demonstrate in the experimental section the effectiveness of mechanistic interpretability in analyzing two distinct state-of-the-art deep-learning models for face recognition: AdaFace [18], based on CNNs, and SwinFace [32], based on transformers. Specifically, FaceMINT implements state-of-the-art sparse autoencoders (SAEs) that have demonstrated strong performance in mechanistic interpretability for large language models. These include Gated SAE (G-SAE) [34], TopK SAE [14], and JumpReLU SAE [35], along with standard baseline models such as vanilla SAE and PCA for reference. In summary, the following major contributions in this paper are:

- We introduce a novel publicly available Python library, called FaceMINT, that implements mechanistic interpretability methods for biometric recognition. It provides researchers with a convenient way to analyze biometric models using state-of-the-art approaches from this field.
- We present a comprehensive set of experiments using state-of-the-art sparse autoencoders combined with activation maximization on biometric models. This is the first study to explore these approaches in the context of biometric recognition, offering new insights into their behavior and effectiveness when applied to face recognition systems. Our results contribute to the broader understanding of how interpretability tools can be adapted and applied within the biometric domain.

In the Section 2, we provide an overview of related work on explainable face recognition and mechanistic interpretability. In the Section 3 we introduce the FaceMINT library, outlining its features for analyzing neuron, layer, and channel activations, supported image parametrizations, and regularization techniques. We also introduce state-of-the-art sparse autoencoders integrated into FaceMINT, which underpin our experiments. In the Section 4 and the Section 5, we describe our experimental setup and report the results considering two state-of-the-art face recognition architectures: the CNN-based AdaFace and the transformer-based SwinFace. In the Section 6, we discuss our results and summarize our conclusions.

2. Related Work

2.1. Explainable Biometrics

The primary goal of explainable biometrics is to enable understanding of the fundamental components of a biometric system and to interpret the decisions it makes [25, 50]. One branch of methods for interpreting the decisions of biometric models aims to explain which facial features from two different face images lead to the decision that they do (or do not) belong to the same identity. These methods are usually based on intensity map optimization, such as Grad-CAM [26, 55] and Integrated Gradients (IG) [31].

The second branch, which is the focus of this work, aims to understand and interpret the decision-making process of the recognition model itself. Several approaches have been

proposed in the literature, with LIME (Local Interpretable Model-agnostic Explanations) [6] and SHAP (SHapley Additive exPlanations) [5] being popular techniques for understanding why the model made specific decisions, based on the idea of attributing *feature importance* to different segments of the input image. Based on these ideas, several developments have been proposed. ALIME [41] adds a denoising autoencoder step to better approximate the perturbed data manifold. It then uses distance in the latent space as opposed to directly comparing data samples. This has been shown to increase local fidelity of the generated explanations. Similarly, SLICE [6] seeks to stabilize the generated explanations using sign-entropy-based feature elimination. Meanwhile, BayLIME [54] treats the local linear coefficients of input features as random variables and performs Bayesian updating based on expert-defined priors to produce explanations.

While these SHAP and LIME - based methods provide post hoc explanations of model decisions, they do not reveal how internal layers process information or how specific neurons interact to form higher-level concepts. Additionally, these methods typically operate only at the input level, whereas real-world face recognition models rely on latent (embedding-space) representations that SHAP-based explanations fail to capture. Different from the methods discussed above, mechanistic interpretability follows a bottom-up approach, aiming to understand the system by analyzing individual components and their interactions.

2.2. Mechanistic Interpretability

The idea behind mechanistic interpretability approaches is that by analyzing and understanding individual components, we can infer how the entire system operates [4]. Mechanistic interpretability has provided valuable insights into explaining large language models, such as identifying neurons whose activations directly correspond to specific concepts [44]. Similarly, efforts to interpret vision models have also gained attention [2, 57]. However, despite its significant potential, mechanistic interpretability remains largely underexplored in the field of biometrics. Mechanistic interpretability could potentially enhance various fields of biometrics, such as cancelable biometrics [28], by enabling more precise transformation or masking of facial features through interpretable representations.

Among the different solutions available to interpret models in a mechanistic manner, various sparse autoencoders have emerged as a powerful solution for this task. By enforcing sparsity, these models can disentangle complex representations, making it easier to identify and analyze the role of individual features in decision-making processes. Instead of directly analyzing final or intermediate features, sparse autoencoders are commonly used to generate sparse representations, aiming to encode factors of variation in the data in a less entangled manner. The base sparse autoencoder was first proposed for use in the mechanistic interpretability of large language models in [14]. Different formulations of activation and loss functions were subsequently proposed to improve their reconstruction fidelity and loss explained [34, 35]. The Sparse Autoencoder methodology, originally developed for large language models, has also been successfully applied to interpretability tasks in vision models [15, 39]. The advantage of sparse autoencoders lies in their ability to learn meaningful representations in an unsupervised manner, which have also been shown to be interpretable.

In standard sparse autoencoders (SAEs) [9], the L1 penalty used to enforce sparsity often introduces unwanted behavior, such as shrinkage, where smaller feature activations are systematically underestimated. The L1 weight parameter also requires extensive fine tuning while training the autoencoder. To address these issues, alternate formulations of the SAE have been proposed.

Gated SAEs (G-SAEs) [34] were proposed to reduce undesirable biases caused by the L1 penalty, such as shrinkage, by separating the tasks of selecting features to activate and estimating their magnitudes. By applying the L1 penalty only to the selection process, Gated SAEs reduce shrinkage, enabling more accurate and interpretable representations.

TopK SAE [14] sparse autoencoders were proposed to address the shrinkage and bias introduced by L1-based sparsity methods, by enforcing sparsity through the explicit selection of the top-k activations while setting all others to zero. TopK SAE explicitly controls the number of active neurons, addressing shrinkage bias by selecting the top-k largest activations and avoiding the need for L_1 regularization, thereby preserving activation magnitudes. Dead latents are minimized through an initialization procedure that aligns encoder and decoder weights, and an auxiliary loss (AuxK) that encourages inactive latents to contribute to reconstruction.

JumpReLU SAE [35] offers an alternative approach to training sparse representations by addressing the problem of suppressed negative activations, which typically occurs in vanilla SAEs. Standard activation functions like ReLU permanently suppress negative activations, which contributes to polysemanticity and prevents neurons from specializing, ultimately degrading interpretability and modularity in deep networks. JumpReLU introduces a deterministic, learnable jump branch that reroutes negative inputs through a low-rank transformation.

All the aforementioned sparse autoencoders have primarily been evaluated on language-related tasks, with limited exploration of their performance in computer vision or biometrics. To address this gap, our library includes a comprehensive collection of state-of-the-art SAE implementations for comparison on these tasks.

2.3. Interpretation via Input Image Optimization

A straightforward way to obtain a possible interpretation of the role of a specific component (e.g., a neuron) in a network's decision-making process is to examine which images from the dataset maximally activate its response and look at their common properties. Another common approach is *Activation Maximization* [11, 17, 56], where the interpretation

is obtained by optimizing the input (starting, for example, from noise) to maximize the component's activation. Since gradients are typically available in such networks, the input image can be optimized according to a specific criterion (e.g., maximizing the activation of a selected neuron or layer).

For biometric recognition models it is often assumed that the input is an aligned image of a biometric sample (e.g., a face, eye, etc.) [18]. Therefore, this optimization process can be further refined by incorporating priors, such as prior knowledge about the expected spatial location of certain facial features [48]. A positive aspect of this approach is that the results are in the form of images (i.e., in a format interpretable by humans), from which we can infer their role in the neural network after optimization. Google developed the library Lucid[1] for such model exploration in the TensorFlow environment. A partial reimplementation also exists for models in the PyTorch environment, called Lucent[2]. In our work, we used these libraries as a foundation for developing a library in the PyTorch environment, where we extended and adapted its functionality for convenient use in the field of biometrics.

2.4. Toolkits for Explainable Biometrics

Several toolboxes, such as Xplique [13], PiML-Toolbox [43], and OmniXAI [52], have been proposed for explaining AI models. However, biometric systems, particularly in face recognition, require tailored interpretability approaches due to the need to distinguish and highlight subtle changes in facial features and regions that differentiate similar individuals. XAIface [23] provides implementations for post-hoc explanations in deep face recognition systems. It uses layer-wise relevance propagation (LRP) for feature attribution, along with saliency-based techniques like Grad-CAM++ [8] and Score-CAM [49] to highlight critical facial regions, and model-agnostic methods like LIME, SHAP, and RISE [30] for instance-level interpretability.

While the aforementioned toolkits and libraries provide valuable insights into a model's predictions and some analysis of salient input features, they do not incorporate mechanistic interpretability approaches. Mechanistic interpretability tools have the potential to offer a deeper understanding of the model's internal workings. With FaceMINT we aim to address this gap and make a powerful software library for mechanistic interpretability available to the biometric community. Our library extends the input optimization approach to sparse autoencoders, and provides pretrained SAEs for use with state-of-the-art face recognition models.

3. The FaceMINT Library

FaceMINT is an open-source Python library built on the popular PyTorch deep learning framework. It provides a suite of tools for mechanistic interpretability, enabling researchers and practitioners to analyze and interpret the inner workings of various biometric models. The library is designed to enhance the transparency and understanding of biometric systems and is publicly available at www.gitlab.com/peterrot/facemint.

Mechanistic interpretability of biometric models can be approached in multiple ways, many of which are supported within FaceMINT, as illustrated in Figure 2. Specifically, the library allows to explore various models characteristics at the neuron, channel and layer levels, and their arbitrary combinations. FaceMINT supports three key components that leverage conceptually different functionalities, which are described below and elaborated upon in the reminder of this section. These include:

- **Activation Maximization with Input Parametrizations:** The first key component of FaceMINT are techniques for optimization of input images to maximize the activation of a specific component (e.g., neuron, layer, channel, or arbitrary combination of them). FaceMINT techniques for this task can be applied to either intermediate components or final biometric templates (i.e., extracted features used for recognition). In the remainder of this section, we also outline the supported image parametrizations in FaceMINT that help to enhance interpretability.
- **Sparse Autoencoders:** The second component of FaceMINT are tools that extend activation maximization to sparse representations by incorporating sparse autoencoders (SAEs), a technique commonly used in mechanistic interpretability. The idea behind these interpretability mechanisms is to append SAEs to intermediate features or biometric templates to obtain more disentangled, sparse encodings. In the following sections, we describe the implemented SAEs and their role in supporting interpretability.
- **Dataset Image Search:** The library also provides functionality to identify images from the (training) dataset that maximize the activation of specific components, enabling quick comparison with images generated through input optimization techniques (e.g., Deep Image Prior). This aids in gaining insights into the model's decision-making process.

3.1. Activation Maximization with Input Parametrizations

In this section, we describe the process of input image optimization to achieve maximal activation and present the individual functionalities available in the FaceMINT library. These functionalities are summarized in Table 2 and are further detailed in the following paragraphs.

Input Image Optimization. For a selected biometric model with pre-trained weights, the goal is to create an optimized input image $\mathbf{I}^*$ that induces maximum activation in a specific component a_i of the model, such as a neuron, layer, channel, or their combination, while maintaining interpretability or

[1] Lucid: https://github.com/tensorflow/lucid
[2] Lucent: https://github.com/greentfrapp/lucent

Table 1
Taxonomy of selected explanation approaches discussed in this work. The first block groups local feature-importance surrogates; the second lists sparse-concept methods built on auto-encoders; the third covers prototype generation via activation maximization with differentiable image parameterizations.

Family	Representative methods	Typical modality	Core mechanism / distinguishing idea
Local feature importance (surrogate-based)	LIME [37], KernelSHAP [22]	tabular, vision	Perturb the neighbourhood of the target instance and fit a kernel-weighted linear surrogate; KernelSHAP chooses the Shapley kernel for axiomatic guarantees.
	SLICE [6], ALIME [41], BayLIME [54]	tabular, vision	*SLICE*: sign-entropy filtering & adaptive perturbations for run-to-run stability. *ALIME*: latent-space sampling via a denoising auto-encoder for manifold-aware locality. *BayLIME*: Bayesian linear surrogate with priors and posterior uncertainty estimates.
Sparse Auto-Encoder (concept-library)	SAE [7], G-SAE [34], topk-SAE [14], JumpReLU-SAE [35]	LLMs, vision models	Train sparsity-regularised auto-encoders (group, top-k, gated, JumpReLU variants); the resulting sparse codes act as disentangled, semantically coherent concepts that explain network activations.
Activation maximization (differentiable image parametrizations)	Feature Visualization [27], Lucid [45], DIP-AM [24]	vision	Optimise a differentiable image parameterization (e.g. Fourier texture, CPPN, DIP-net) to maximise neuron or class activation, producing visual prototypes/explanatory images.

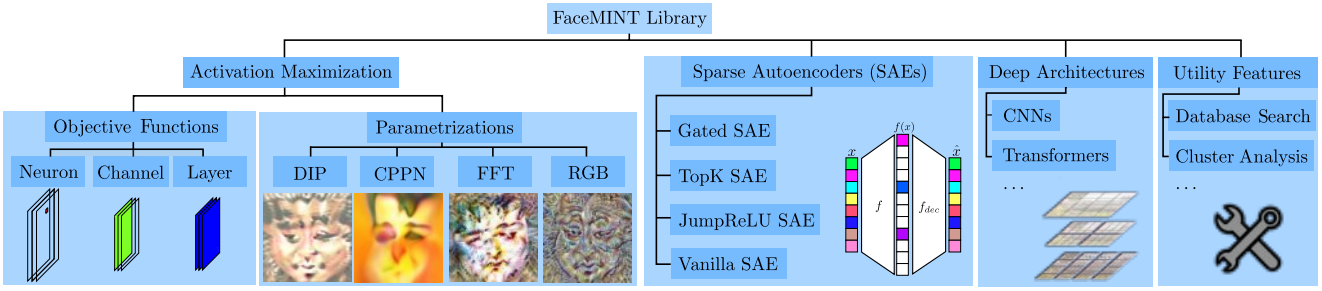


Figure 2: Block diagram of FaceMINT features, enabling mechanistic interpretability in biometrics. The library supports activation maximization on arbitrary model components, such as neurons, channels, layers, or their combinations. It incorporates state-of-the-art image parametrizations, like DIP and CPPN, to guide input image optimization. FaceMINT also includes recent sparse autoencoders (e.g., Gated SAE, TopK SAE, JumpReLU) that can be optimized at any point in the network. It is compatible with versatile deep architectures, such as CNNs and Transformers, and offers advanced utilities to facilitate research on mechanistic interpretability in biometric models.

realism through regularization. This objective is formally defined in Equation 1:

$$\mathbf{I}^* = \arg\max(a_i(\mathbf{I}) - R_\theta(\mathbf{I})), \tag{1}$$

where $a_i(\mathbf{I})$ is the activation of the selected component, and $R_\theta(\mathbf{I})$ is a regularization term controlled by parameters θ that guides the optimization process.

Biometric models are differentiable with respect to their input, and the parameterization of the image must also be differentiable. Through gradient descent, we can modify its parameters (pixel values, weights, etc.) to ensure optimal visualization. Therefore, we implemented several image parameterizations as defined below. The added value of the FaceMINT library is that it allows the recognition model and the choice of components to be analyzed to be modified in a simple, modular way.

RGB Space Parameterization. The simplest way to parameterize the input image is to take a three-channel image as the basis, where the values for each RGB channel are randomly sampled from a normal distribution. From an interpretability perspective, RGB images are convenient, as humans intuitively understand them.

Parameterization in Frequency Space (FFT). It is also possible to sample the RGB channel values from the frequency space. This is done by sampling the x and y coordinates from the frequencies of the discrete Fourier transform, which are then scaled to achieve uniform transformation. Using the inverse N-dimensional discrete Fourier transform for real inputs, we obtain values that are reshaped to the desired image size [27]. The advantage of this method is that it ensures spatial decorrelation of values within each channel. To ensure decorrelation between color channels, a Cholesky decomposition is performed on the covariance matrix of the RGB channels from the images on which the model was trained.

CPPN Parameterization [42]. A Compositional Pattern Producing Network (CPPN) [42] further constrains optimization by enforcing that adjacent pixels have similar colors. This parameterization modifies randomly initialized parameters of a multi-level neural network, which maps the (x, y) coordinates of the image to RGB values.

DIP Parameterization [48]. The Deep Image Prior (DIP) [48] assumes that a well-designed model architecture, without prior training, can capture the low-level features of images. DIP is widely used in the literature for image restoration. DIP is implemented using a UNet architecture [38] with

	Functionality	Equation
Objective Functions	Neuron Activation	$a_{i,c,u,v}(\mathbf{I}) = a_i[c,u,v]$
	Layer Activation	$a_i(\mathbf{I}) = \frac{1}{H \cdot W \cdot C} \sum_{l=0}^{H-1} \sum_{k=0}^{W-1} \sum_{h=0}^{C-1} a_i[h,l,k]$
	Channel Activation	$a_{i,c}(\mathbf{I}) = \frac{1}{H \cdot W} \sum_{l=0}^{H-1} \sum_{k=0}^{W-1} a_i[c,l,k]$
	Arbitrary Interactions	*n/a*
Parametrizations	RGB Space	$\mathbf{I}_{k,w,h} \sim \mathcal{N}(\mu, \sigma^2)$
	Frequency Space [27]	*n/a*
	CPPN [42]	*n/a*
	DIP [48]	*n/a*
Regularizations	α Channel	$\mathbf{I}'_{\text{RGB}} = \mathbf{I}_{\text{RGB}} \cdot \mathbf{I}_\alpha + \text{BG}_{\text{RGB}} \cdot (1 - \mathbf{I}_\alpha)$ $a_i(\mathbf{I}) = a_i(\mathbf{I}) \cdot (1 - \bar{\mathbf{I}}_\alpha)$
	L_1	$R_{L_1}(\mathbf{I}) = \sum_{l=0}^{H-1} \sum_{k=0}^{W-1} \sum_{h=0}^{C-1} \lvert\lvert\mathbf{I}[h,l,k]\rvert\rvert - \varepsilon$
	L_2	$R_{L_2}(\mathbf{I}) = \sum_{l=0}^{H-1} \sum_{k=0}^{W-1} \sum_{h=0}^{C-1} \sqrt{(\mathbf{I}[h,l,k] - \varepsilon)^2}$
	Diversity	$G_{i,j} = \sum_{u,v} a_i[k,u,v] \cdot a_j[l,u,v]$ $R_C(\mathbf{I}) = \sum_a \sum_{a \neq b} \frac{\text{vec}(G_a) \cdot \text{vec}(G_b)}{\lvert\lvert\text{vec}(G_a)\rvert\rvert \cdot \lvert\lvert\text{vec}(G_b)\rvert\rvert}$
	Total Variance	$R_{TV}(\mathbf{I}) = \sum_{i,j} \left((I_{i,j+1} - I_{i,j})^2 + (I_{i+1,j} - I_{i,j})^2 \right)^{\frac{\beta}{2}}$

$\mathbf{I}$ — parameterization of the input,
$\mathbf{I}'$ — new input image,
α — alpha channel,
a — activation,
i — layer index,
c — channel index,
u, v — spatial coordinates,
H — height of the conv. layer,
W — width of the conv. layer,
C — no. of channels,
$a_i[c,u,v]$ — activation index,
$I_{k,w,h}$ — image with k channels, height w, and width h,
$\mathbf{I}[h,l,k]$ — pixel index,
ε — small constant value,
BG — background,
R — regularization function,
G — Gram matrix,
$\bar{\mathbf{I}}_\alpha$ — average value of the α channel.

Table 2
Table of supported functionalities in the FaceMINT library.

randomly initialized weights. UNet combines information from different levels, enabling the generation of more interpretable images.

Optimization of the α Channel. The optimization of images with RGB or FFT parameterization has the drawback that it does not provide insight into the importance of individual features. By adding a fourth channel (α transparency) to the parameterization, we can also optimize depth in the input image (i.e., what is in the foreground and background of the image). This is achieved by merging the four-channel image back into a three-channel RGB image before passing it into the model, blending it with a random background based on the values of the α channel.

3.2. Sparse Autoencoders

While methods for maximizing activation can be applied directly to arbitrary representations in a recognition network, this approach is often suboptimal as such representations are typically highly entangled [29]. Mechanistic interpretability leverage SAEs to promote more sparse representations [4]. The FaceMINT library implements multiple SAEs that achieved state-of-the-art performance in interpreting large language models. The following subsections provide a detailed description of these methods.

Gated SAE [34]. G-SAEs improve sparse autoencoder performance by decoupling feature selection from magnitude estimation, effectively addressing the issue of shrinkage bias. This is a phenomenon where the L_1 penalty drives feature activations toward smaller values, resulting in underestimation to promote sparsity. By applying the L_1 penalty selectively to the feature selection process, G-SAEs achieve a better balance between sparsity and reconstruction accuracy, yielding sparser but more faithful decompositions. Key hyperparameters in G-SAE training include the L_1 sparsity coefficient and the weight scaling factor in the magnitude estimation path, both of which are critical for optimizing performance.

TopK SAE [14]. The TopK SAE [14] explicitly controls the number of active neurons by selecting the top-k largest activations, to preserve activation magnitudes. Key hyperparameters include the sparsity level k, which directly controls the number of active neurons, the auxiliary loss coefficient α, and the learning rate, all of which critically impact sparsity, reconstruction, and training stability.

JumpReLU SAE [35]. JumpReLU SAE [35] improves reconstruction of the input using a discontinuous activation function that zeroes out pre-activations below a threshold, enhancing sparsity while preserving activation magnitudes. The key hyperparameters are the threshold θ and kernel bandwidth ϵ, which balance between sparsity and accuracy of the reconstruction during training.

Vanilla SAE [1]. We also evaluate the vanilla SAE [1] as a baseline to assess the improvements made by the more advanced SAE architectures. This comparison helps highlight the performance gains of the enhanced SAEs.

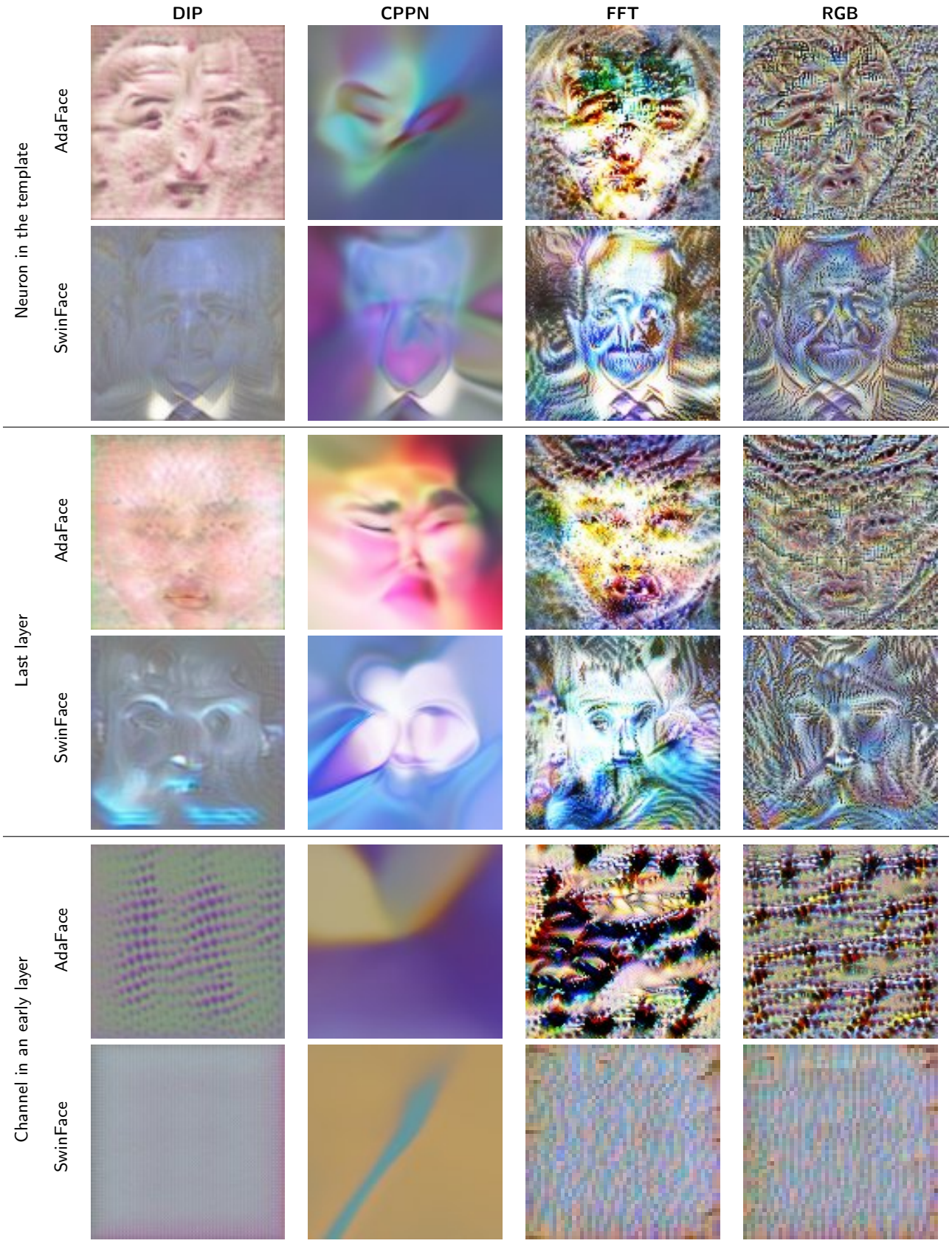
3.3. Dataset Image Search

The library also provides functionality to identify images from the (training) dataset that maximally activate specific components and compare them to optimized inputs generated through activation maximization using image priors (e.g., DIP, CPPN). Inspecting both sets allows users to reason about semantic alignment, i.e., whether real-world images that maximally activate a given component depict the same high-level features (e.g., baldness, eyeglasses) as those artificially produced using optimization techniques such as DIP or CPPN. While this feature is not novel in a methodological sense, it can significantly streamline the process for researchers seeking insights into what the component may represent in the model's decision-making process.

4. Experiments

In the previous section, we introduced the FaceMINT Library and the functionality it offers to biometric researchers. In this section, we demonstrate the usability of the library by evaluating it on two state-of-the-art models: the CNN-based AdaFace and the transformer-based SwinFace. We provide a comprehensive evaluation of how different image parameterizations affect the interpretability of activation maximization techniques. Next, we assess the efficiency of sparse autoencoders in obtaining sparse representations, and how interpretable they are. Finally, we present insightful interpretations of neurons in those recognition models.

Table 3
Demonstration of the usability of FaceMINT in visualizing various components of biometric models, specifically AdaFace and SwinFace. The visualization focuses on the last layer, an early channel, and an arbitrary neuron, employing different image parameterizations (DIP, CPPN, FFT, and RGB) to guide the visualization.



4.1. Datasets and Image Preprocessing

In our experiments, we use the largest publicly available face recognition dataset, Glint360K, which is composed of several large-scale datasets [1]. State-of-the-art face recognition models are typically trained on pre-aligned face images and are known to be sensitive to misalignment. To extract templates x, we first align all face images using the MTCNN keypoint detector [53] and downscale them to 112×112 pixels. We then extract the face templates for all images in advance using two state-of-the-art face recognition models: the CNN-based AdaFace and the transformer-based SwinFace. We selected these models to demonstrate the ability of mechanistic interpretability to generate meaningful insights for both types of model architectures: CNN-based and transformer-based. This approach significantly speeds up the training of the autoencoders, as it eliminates the need to run the face recognition model during training.

4.2. Input Parametrizations

When optimizing activation maximization, researchers can choose from various image parameterizations, which guide the optimization (e.g., by imposing specific constraints based on prior knowledge, such as facial structure) [42, 48]. In this set of experiments, we qualitatively evaluate these methods in the context of face biometrics by manually inspecting the produced images. We visualize the input images by considering activation maximization with respect to an individual neuron, a channel in the early layer, and the entire layer in a face recognition model. We inspect how closely the shapes in the generated images resemble those of a face, and whether expected facial features—such as the eyes, nose, and mouth—are present and whether the structural relationships typically found in faces (e.g., positioning of the eyes above the nose, etc.), are maintained in the optimized images.

4.3. Interpretability of Optimized Input Images

In this set of experiments, we qualitatively evaluate these parameterizations in the context of face biometrics by manually inspecting the produced images. In a qualitative manner, we assess: (i) how closely the shapes in the generated images resemble those of a face, and whether expected facial features—such as the eyes, nose, and mouth—are present; (ii) whether the structural relationships typically found in faces (e.g. positioning of the eyes above the nose, etc.), are maintained in the optimized images. This approach provides insight into which parameterizations produce images that are more or less interpretable within the domain of face biometrics.

4.4. Comparison of SAEs

Each SAE has its own hyperparameters, so we ensure a fair comparison by comparing best performing models, which we obtain using a logarithmic grid search. An individual autoencoder training run taxes approximately 2 hours on an RTX 4090 GPU, and the total training time for the entire grid search took approximately 3 days. For each optimized SAE, we report $\|L\|_0$, $\|L\|_1$, the number of alive features, and the reconstruction error (mean square error, MSE). The $\|L\|_0$ norm counts the non-zero activations, showing how many features are active at a given time. The $\|L\|_1$ norm sums the absolute activations, reflecting their overall strength. Alive features track how many features have been activated at least once during training. Finally, MSE measures how accurately the model reconstructs the original input from its learned representation. After obtaining tuned SAEs, we compare them based on visualized features and demonstrate the usability of our approaches.

4.5. Interpretability of Original Templates Vs. their Sparse Representations

When considering mechanistic interpretability, it is crucial to assess whether training sparse representations actually helps produce more interpretable features compared to those already present in the biometric model, which are generally known to be highly compressed, making them difficult to interpret. This has also been shown to be true for language models [7]. To confirm this, we have conducted an experiment to determine if the same applies to vision-based face recognition models (or biometric models). We begin by analyzing how many images from the original dataset activate a specific neuron in the following templates: (i) the original template, (ii) the PCA-transformed template (baseline), and (iii) the template transformed using the SAE. For each, we generate corresponding density histograms. We then look for clusters of activations and try to interpret individual neurons by examining dataset examples that maximize the activation of a particular feature or by analyzing their activation maximizations.

4.6. Feature Interpretation Protocol

Once the SAEs are trained, we search for promising interpretable features in individual sparse neurons by examining which images from Glint360k most strongly activate each neuron. This follows established mechanistic interpretability methodology from LLMs [7], where concrete dataset examples are used to identify candidate features. We first select neurons that are activated by a sufficiently large number of images, as indicated by the histograms in Figure 3. Since the total number of neurons is too large for manual inspection (each SAE produces a sparse representation $f(x) \in \mathbb{R}^{65,536}$), we sample 200 neurons from each. The neurons to inspect are picked as described in Section 5.4: we sample them from the medium density cluster of the feature histograms shown in Figure 3, namely, neurons activated $[10^{-5}, 10^{-3}]$ of the dataset (i.e., hundreds to thousands of images). Neurons like this have enough activating images for human annotators to discern meaningful semantic patterns, whereas those from lower or higher density clusters tend to have either too many or too few. Furthermore, we reject those neurons which are solely activated by images of a single subject. While training set memorization is a known problem in machine learning, it is not the subject of this study, as we are looking for generalizable learned features. Once selected, the experimental procedure for a given neuron i is then as follows:

1. Pass the Glint360k images through the face recognition model to extract their biometric templates x.
2. Feed the templates x into the trained sparse autoencoders to obtain their sparse representations $f(x)$.
3. Identify the dataset images that activate the neuron of interest, i.e., those for which $f(x)_i \neq 0$.
4. Rank the activating images by the magnitude of their activations $f(x)_i$ in descending order.
5. Manually inspect up to 256 of the most strongly activating images (all, if fewer are available) and check whether the majority share a common pattern. Additionally, visualize the Deep Image Prior (DIP)–parametrized image obtained through activation maximization, which should ideally semantically align with the recognized interpretation.

The manual inspection was independently performed by two of the examiners, who each examined the image grids and attempted to identify biometric features or other patterns exhibited by the activating images of a given neuron. A neuron was deemed interpretable when both examiners agreed on its interpretation. Examples of interpretable neurons, along with our assigned interpretations, are presented in Table 5.

4.7. Quantitative Feature Validation

Once human-interpretable features are discovered in the latent space of sparse autoencoders, it is necessary to evaluate whether these interpretations truly reflect the decision-making process of the face recognition model. To this end, we conduct two quantitative experiments: a **feature ablation study** and a **contrastive examples study**.

In the feature ablation study, we semantically edit images to remove a discovered feature while keeping the rest of the image intact, and then measure the response of the corresponding sparse feature. For example, if the identified feature is "Thick glasses", removing the glasses from a face should cause the activation of this feature to be zero, while all other facial attributes remain unchanged. To quantify the effect of such semantic modifications, we report: *(i)* the change in activation magnitude of the corresponding neuron, *(ii)* the change in the ℓ_2 norm of the face recognition template x, to assess whether the recognition embedding is significantly altered, *(iii)* the change in the ℓ_0 norm of the sparse representation $f(x)$, to evaluate the impact on other sparse features, which should ideally remain unaltered, and *(iv)* the cosine similarity between biometric templates, to ensure that the overall face recognition score remains high.

In the contrastive examples study, we verify whether the discovered interpretable features align with semantic attributes predicted by pretrained classifiers. Specifically, we compare images that maximally activate a given feature with those that minimally activate it, using pretrained classifiers for facial attributes such as gender, race, or age. If the sparse feature corresponds to a meaningful concept, the classifier should confirm it: for example, for the feature "Infants", the age classifier should assign ages below two years to the maximally activating images, whereas the minimally activating images should receive higher predicted ages.

5. Results

5.1. Interpretability of Activation Maximizations

In this set of experiments, we tested various image parameterizations to assess their interpretability, namely DIP, CPPN, FFT and RGB. In Table 3, we visualize activation maximizations for both models to observe whether we obtain meaningful results for a CNN and transformer. Specifically, we visualize: (*i*) an arbitrary selected neuron in the biometric template (first two rows), *ii*) last layer in the face recognition model (third and fourth row), and *iii*) channel in one of the first layers in the model (last two rows).

We observe that RGB and FFT image parametrizations are especially helpful for early layers, which primarily encode low-level features like edges and simple shapes (see last two rows). For the last layers, parametrizations that consider face shapes (DIP and CPPN) often yield more interpretable results in terms of face shapes. This is also expected, as deep architectures encode high-level concepts in the later layers, meaning that more interpretable features (e.g., concrete shapes of the face) are encoded in the later layers.

Note that despite the architectural differences between the models, the considered visualizations demonstrate the ability to gain insights into both CNN and transformer-based models. We observe that the semantic information in the optimized images is somewhat consistent when comparing different image parametrizations (see individual rows); however, the image quality differs significantly. The DIP parametrization typically returns the best quality in terms of the structure of the face, while the CPPN parametrization usually gives more smeared-out results. FFT and RGB typically yield similar results in terms of the shape of the face and facial features; however, the RGB values differ.

From an interpretability perspective, these methods allow us to understand what specific components of the network activate in response to, and what they are sensitive to. For example, channels in the early layers might be sensitive to repeating patterns, while later layers capture more semantic features, such as a big nose or a small mouth.

5.2. Optimization of SAEs

For all considered sparse autoencoders, we conducted a grid search to optimize their respective sparsity-related hyperparameters. For SAE and G-SAE, we varied the regularization term λ over a logarithmic scale from 10^{-2} to 10^{-12}, running each experiment for 10 epochs on the Glint360k dataset with both AdaFace and SwinFace features. For JumpReLU SAE, we performed a similar search by adjusting the sparsity constraint imposed by the activation function. For TopK, the sparsity was controlled by varying the hyperparameter k, which dictates the number of retained

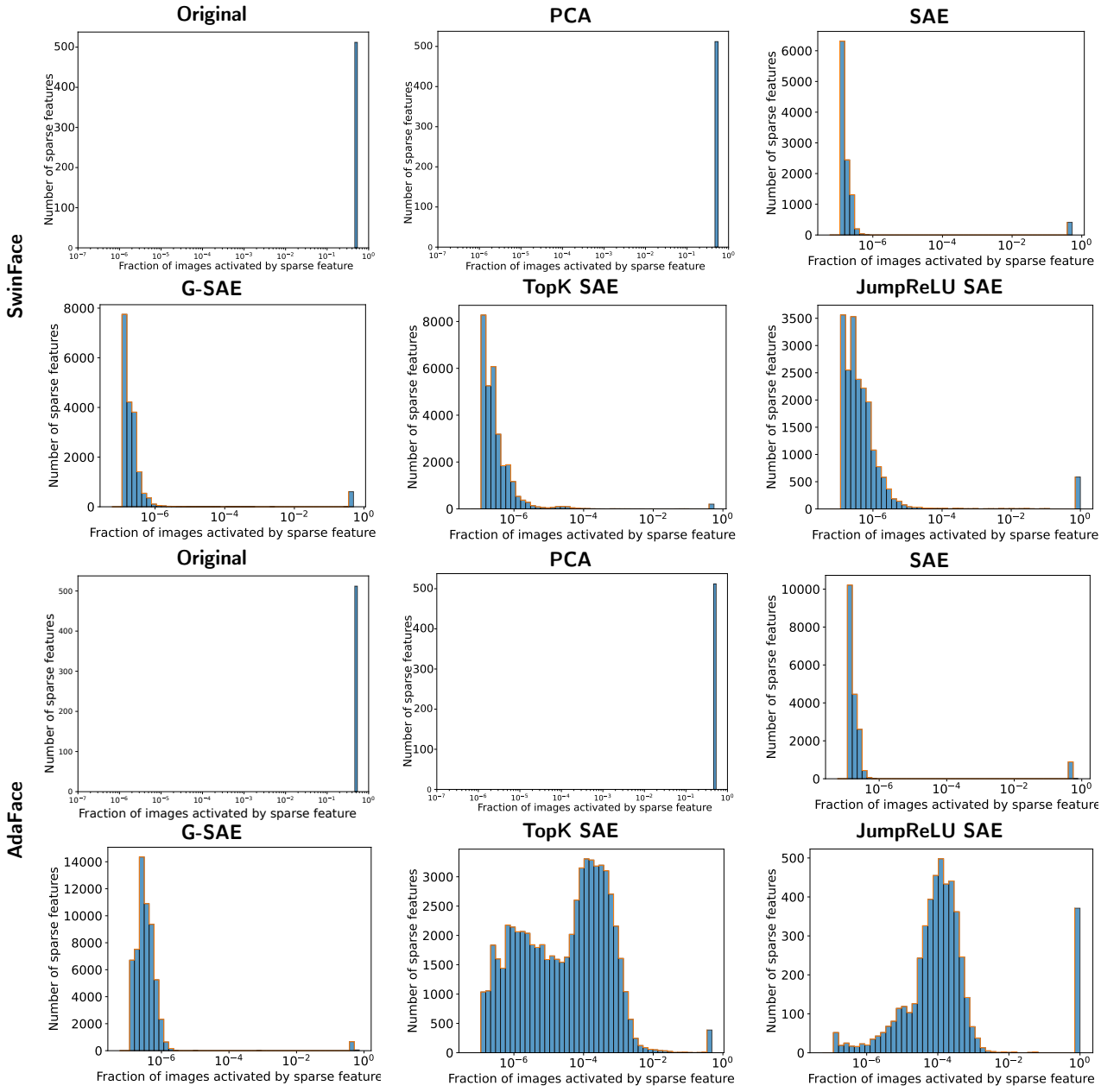


Figure 3: Feature density histograms. Note the logarithmic x-axis. The original template neurons, as well as their PCA, are fully dense. Sparse autoencoders translate this encoding into a much higher-dimensional space, which is more sparse as a result. Most sparse autoencoders considered here are able to encode arbitrary face templates as a linear combination of ≈ 100 feature directions.

elements in the sparse representation. We tested a range of k values from 50 to 10,000 to analyze its impact on model performance. Results from the entire grid search are available in the Appendix, while in Table 4, we report only the best-performing models, focusing on near-perfect reconstruction. The metrics included in this table are the $\|f(x)\|_0$, $\|f(x)\|_1$, the percentage of active neurons in the sparse representation (% Alive), and the reconstruction MSE. The table shows that all sparse autoencoders enforce sparsity while maintaining low MSE, demonstrating their ability to learn compact yet informative representations. This suggests that sparse autoencoders effectively reduce dimensionality while preserving essential features.

5.3. Feature Activation Comparison

In this experiment, we compare feature activation densities across three scenarios: original templates, sparse representations obtained after training (e.g., with G-SAE), and PCA-transformed templates as a baseline. We use the original templates as a baseline because we do not expect them to exhibit any sparsity — in fact, according to the superposition hypothesis [10], the learned templates represent a much larger set of features compressed into the limited space of (in

Table 4
Comparison of Different Sparse Autoencoders with Different Feature Extractors. All sparse autoencoders achieve near-perfect representation of the face templates, with varying degrees of sparsity and utilization of the feature dictionary space.

Feature Extractor	Autoencoder	$\|f(x)\|_0$	$\|f(x)\|_1$	% Alive	MSE
SwinFace	**SAE**	206	1.2117e+01	64.26	0.0000
	G-SAE	301	2.4032e+02	78.12	0.0000
	TopK SAE	200	1.7782e+02	94.23	0.0000
	JumpReLU SAE	245	9.9051e+02	14.62	0.0000
AdaFace	**SAE**	445	1.9744e+01	94.86	0.0000
	G-SAE	342	3.1897e+02	96.28	0.0026
	TopK SAE	400	2.8341e+02	98.98	0.0003
	JumpReLU SAE	941	1.8338e+03	66.52	0.0001

our case) the 512-dimensional space of the face recognition models. Therefore, we expect on average each dimension of the original and PCA spaces to be activated, on average, by half of the dataset images.

The results are presented in Figure 3. For each scenario, we measure the number of activated features by applying a threshold and plot density histograms to visualize the activation distributions. Sparse representations are trained to produce compact encodings, while PCA offers a dimensionality-reduction baseline. The results are presented as histograms with the fraction of dataset images activating a given neuron (basis direction) on the x-axis, and the number of neurons (basis directions) activated by a given fraction of the dataset on the y-axis. To emphasize the sparsity of the learned G-SAE representation, the x-axis is scaled logarithmically.

We observe that both the original templates and PCA-transformed templates exhibit dense feature activations across images in the dataset. Specifically, as expected, each template neuron and each principal component is activated, on average, by half of the Glint360k dataset, with very low variance.

However, when using the gated sparse autoencoder, we successfully identify two distinct clusters: *(i)* a smaller dense cluster that behaves similarly to the original template neurons and their principal components, and *(ii)* a sparse cluster of neurons that are, on average, activated by only a few (10–100) images from the Glint360k dataset. This pattern is evident for both transformer-based and CNN-based models, making it a particularly noteworthy finding. We further analyze clusters of activated features by examining dataset examples that strongly activate specific features and inspecting their input activation maximizations to infer meaningful patterns. In the following experiments, we examine input images for features that belong to either the sparse or dense cluster.

5.4. Qualitative Feature Interpretation

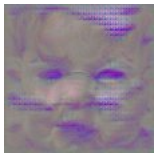
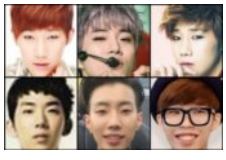
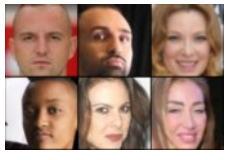
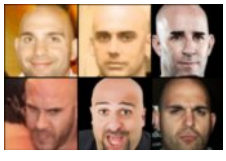
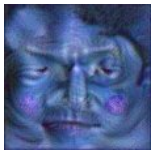
Inspecting images that activate specific sparse neurons, we offer the following observations. Sparse neurons in the low density clusters are commonly activated by too few images to be readily interpretable, except in cases where the neuron corresponds to a single person from the dataset. On the contrary, sparse neurons in the high density clusters (i.e., those activated by over 10% of the dataset) can also be difficult to interpret due to high amounts of distractors even among the images that activate the given neuron with the highest magnitude. That is to say, there are no discernible patterns in the activating images in the overwhelming majority of these neurons we investigated. Therefore, we focus our qualitative inspection on sparse neurons that have sufficient activating images (i.e., around 10^{-4} of the dataset or 1000 images), and which do not only correspond to a single person. We summarize our findings for the AdaFace and SwinFace sparse autoencoders in Table 5. We note that even within the limited set of sparse neurons we were able to interpret, there are many concepts shared between both face recognition networks (which do not share the same training set), which points to a degree of instrumental convergence [3] between the two given the same abstract training objective (face analysis) despite differences in network architecture, loss function, and training set.

For some of the highlighted neurons, we were able to produce images of the same concept using input image optimization with our differentiable image parametrization implementations. This serves as additional validation of the template feature direction actually encoding the concept in question. We note that qualitatively, the optimized input images for feature directions found by sparse autoencoders appear much more coherent that the input images obtained for individual basis neurons shown in Figure 3. However, for the neurons where no optimized input image is displayed, we were unable to generate a coherent image with the methods presented here. This was also apparent during the optimization procedure — during successful optimization runs, the neuron activation loss was far lower than unsuccessful ones. This also shows the input optimization approach is sound and meaningful, but not universally applicable. On average, this happened with around 20% of the (randomly sampled) AdaFace feature directions, and around 75% of the SwinFace feature directions. This implies that the CNN-based AdaFace model is more amenable to output maximization via input optimization than the transformer-based SwinFace model, which makes sense as the grid-based image prior is inherently represented in convolutional networks, whereas it is learned in transformers.

5.5. Survey

In order to quantitatively verify whether our interpretations of the latent neurons and DIP-optimized images presented in the previous section are valid, we have conducted a survey where we asked the participants to, in their own words, identify commonalities between the maximally activating images for each of the neurons presented in Table 5. Furthermore, the participants were also asked to evaluate whether the corresponding DIP-optimized image represents the same semantic concept as the one they identified as being common to the maximally identifying image. The number of questions in the survey was limited to 14 to encourage thorough participation.

Table 5
Qualitative interpretation of AdaFace and SwinFace feature directions found with sparse autoencoders.

AdaFace				SwinFace			
SAE	**Neuron**	**Activating images†**	**Optimized input**	**SAE**	**Neuron**	**Activating images†**	**Optimized input**
JumpReLU	3756	Infants 		JumpReLU	17016	Squinting 	
JumpReLU	16009	Caricatures, comics 		G-SAE	16694	Eyeglasses 	✗
JumpReLU	17654	Overweight (both genders) 	✗	G-SAE	24526	Dark hair, background 	✗
JumpReLU	25862	Older men, beards 		G-SAE	28406	Arab headdress 	✗
JumpReLU	38449	Baseball caps 	✗	G-SAE	19796	Nasolabial folds 	✗
JumpReLU	46303	Young male Asians 		G-SAE	22899	Large forehead 	✗
JumpReLU	46484	Bindi (Indian dot) 		G-SAE	34325	Mouth obstructed 	✗
TopK SAE	7912	Bald white men 		TopK SAE	5801	Red cheeks 	

†Labels were assigned manually based on visual inspection and serve as abbreviations in Section 5.5.

The survey had a total of $N = 89$ responses. Of those, 52 identified as women, and 37 identified as men. Furthermore, 9 of the participants identified as experts in the field of biometrics, whereas the rest did not. Participants in the survey were solicited within the authors' respective research groups and social media circles.

We present the results of the survey in Table 6. For each of the latent autoencoder features considered, we measured the following:

- **Concept agreement rate.** As the respondents had to describe the semantic concept represented by each set of images, we considered a given response to "agree" with our assessment if it mentioned a similar concept. As an example, we marked the AdaFace JumpReLU neuron 3756 as corresponding to the "Infants only" feature. A typical response to the same set of images which we marked as agreeing is, "They all look very young".
- **Conditional DIP agreement.** Of the respondents that identified the same underlying concept as our interpretation, we measure the share that agreed with the statement "the DIP optimized image represents the same semantic concept".

Based on the results of the survey, we note that for most of the features considered, the majority of the respondents agree and identify the images as representing the same semantic concepts. For all of the images considered, the rate of concept agreement is above random chance. We also note that there is a degree of correlation ($\rho = 0.455, p = 0.16$) between the rate of concept agreement for a given feature and whether we were able to obtain a DIP image for it. This suggests that human-identifiable features are also more amenable to optimization-based optimization by our library.

We noticed no statistically significant correlation between the responses and the age, gender, or biometrics expertise of the respondents.

5.6. Feature Ablation Study

In this set of experiments, we examine the extent to which the identified interpretations are meaningful and whether they correspond to actual factors in the model's decision-making process. We evaluate how the activation of an individual neuron in the sparse representation changes when the feature we identified it as encoding is altered through semantic image editing. We have edited the input images to this end, while ensuring that all other attributes remained unchanged. Concretely, for the feature *"Older men, beards"* we have removed the beards; for the *"Bald white men"* feature, we added hair; for the *"Baseball caps"* we removed the caps; and for the *"Eyeglasses"* feature we removed the glasses; for the *"Bindi"* feature we removed the bindi. This controlled manipulation enables us to directly test whether modifying the images leads to corresponding change in neuron activation, thereby providing evidence as to whether the discovered interpretations reflect the model's predictions. We observed how this changes the biometric template, and how the image manipulation affects the rest of the sparse representation. Concrete examples of semantic image editing are provided in Figure 4. Note that we edited only the attributes corresponding to the interpretation, while leaving all others intact.

Table 6
The results of our concept-agreement survey. The columns present the fraction of responses agreeing with our feature identification, and, conditional on concept agreement, the fraction of responses that consider the DIP-optimized image of a given concept to represent it well.

Feature name	$P(agree_C)$	$P(agree_{DIP} \mid agree_C)$
Infants	0.977	1.000
Caricatures, comics	0.989	1.000
Overweight	0.596	n/a
Older men, beards	0.977	0.865
Young male Asians	0.955	0.882
Bindi	0.888	0.468
Bald white men	0.977	0.701
Squinting	0.247	0.571
Large forehead	0.169	n/a
Mouth obstructed	0.955	n/a
Red cheeks	0.191	0.666

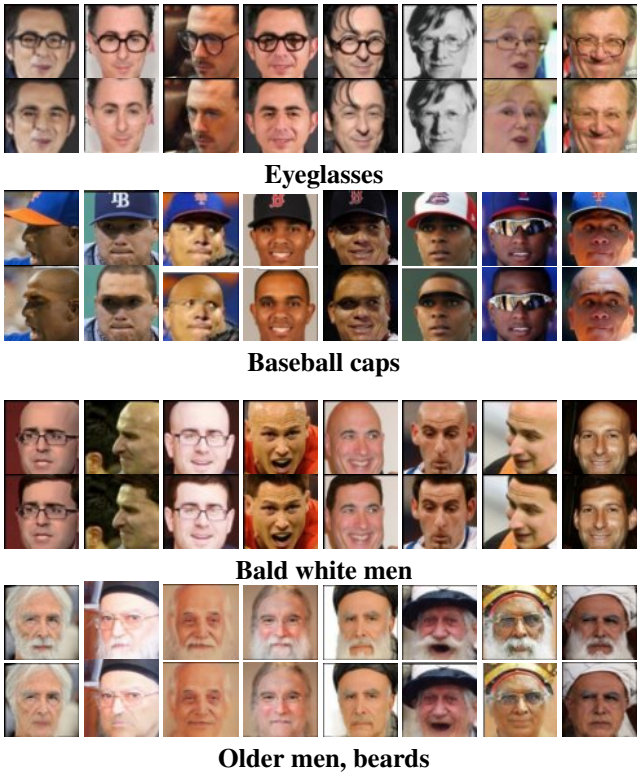


Figure 4: Semantic edits of input images for four discovered features. Each panel shows eight examples, with the original (top row) and edited images (bottom row).

The results with representative examples of these semantic edits, together with the presence or absence of each feature, are provided in Table 7.

Table 7
Results of our feature ablation study. We report the activation magnitude of the autoencoder associated with a given feature given its presence or artificial absence. We also report the norms of the face recognition templates derived from the original and edited images as the l_0 norms of the autoencoder latent vectors, and the cosine similarity between the face recognition templates of the original and edited images. The confidence intervals are reported as $\mu \pm \sigma$.

Feature	Presence	Activation	$\|x\|_2$	$\|f(x)\|_0$	Cos. sim.
Older men, beards		1.67 ± 0.41	21.8 ± 0.88	942 ± 1.0	0.73 ± 0.07
		0.00 ± 0.00	22.3 ± 1.3	940 ± 0.5	
Bald white men		0.47 ± 0.03	20.9 ± 2.4	199 ± 9	0.75 ± 0.08
		0.02 ± 0.04	20.3 ± 2.5	199 ± 10.2	
Baseball caps		1.52 ± 0.41	19.1 ± 3.9	942 ± 0.5	0.68 ± 0.22
		0.00 ± 0.00	19.3 ± 2.2	941 ± 0.3	
Eyeglasses		0.17 ± 0.04	1.36 ± 0.1	302 ± 2	0.88 ± 0.09
		0.002 ± 0.01	1.40 ± 0.11	300 ± 2	
Bindi		1.73 ± 0.34	24.8 ± 1.62	942 ± 0.6	0.58 ± 0.21
		0.39 ± 0.41	24.6 ± 1.67	941 ± 0.8	
Overall	✓ ✗	1.12 ± 0.66 0.08 ± 0.15	17.6 ± 8.32 17.6 ± 8.29	665 ± 340 664 ± 340	0.72 ± 0.10

For the *"Bindi"* feature, we used 62 original images and 62 corresponding edited images with the bindi removed. For the other features, which require more manual effort to edit, we used 16 original images and 16 corresponding edited images with the feature removed. We report the results in terms of the following quantitative metrics:

1. **Sparse neuron activation** of the original and edited images, which directly shows us the difference in activation magnitude if we remove the identified feature.
2. **Biometric template norm,** $||x||_2$, of both the original and edited images. In both AdaFace and SwinFace, the template norm correlates to image sample quality due to the nature of the training algorithm used.
3. L_0 **norm of the sparse representation** of the original and edited images. Significant change here indicates that the sparse space is more entangled than desired.
4. **Cosine similarities** of the biometric templates $f(x)$ between each pair of original and edited images. Here, a low cosine similarity would mean that the underlying face recognition models no longer recognize the edited image as representing the same person, which would indicate our image editing has ruined the image.

The results in Table 7 show that we can cause the interpreted neurons to deactivate using selective image editing, as the mean neuron activation is always lower over the edited images. Furthermore, the template norms show that the face recognition models consider the edited images to still represent high quality biometric samples, as the change is minimal. In addition, the cosine similarities between original and edited images are above the recognition thresholds for both models (0.25 for AdaFace and 0.2 for SwinFace), which means the edited images are still recognized as the same person, indicating that the edited images retained most of the identity-related information. Most importantly, the L_0 norms of the sparse representations show the autoencoders give us highly disentangled representations, as editing the images, in most cases, results in the L_0 norm dropping by 1. This means only the interpreted neuron has been deactivated, whereas the rest are still activated in the edited images.

5.7. Contrastive Examples

In this set of experiments, our goal is to analyze images that maximally and minimally activate a given discovered feature, and to validate whether human interpretations of these features align with labels predicted by pretrained classifiers. For each feature, we select 256 images that maximally activate it and 256 images that minimally activate it. We first present concrete examples those image sets for selected feature interpretations in Figures 5–10, arranged on 16×16 grids. Figure 5 presents images for the feature "Caricatures, comics." In all maximally activating images, there is not a single photorealistic example, whereas the minimally activating set contains only realistic images. Figure 6 shows images for the feature "Bald white men," where almost all images correspond to this description, with only rare distractors. The images for the feature "Overweights" are presented in Figure 7, where all maximally activating images exhibit a wider facial morphology, whereas the minimally activating examples show thinner faces. Figure 8 shows images for the feature "Baseball caps," where in all maximally activating images the person wears a baseball cap, whereas in the minimally activating images there is not a single baseball cap. Figure 9 shows images for the feature "Older men, beards," where in all maximally activating instances the description matches; in contrast, among the minimally activating images, no such combination occurs (there are cases of older men but without beards, and vice versa). For the feature "Infants" in Figure 10, we observe that all maximally activating images conform to this interpretation, while there is not a single infant among the minimally activating images.

Figure 5: “Caricatures, comics”: maximal (top) and minimal (bottom) activations.

Figure 6: “Bald white men”: maximal (top) and minimal (bottom) activations.

Figure 7: “Overweight”: maximal (left) and minimal (right) activations.

Figure 8: “Baseball caps”: maximal (top) and minimal (bottom) activations.

Figure 9: “Older men, beards”: maximal (top) and minimal (bottom) activations.

Figure 10: “Infants”: maximal (left) and minimal (right) activations.

Table 8
Inspecting variability of other discernible patterns within maximally and minimally activating images in the contrastive examples experiment. For each feature, we report the number of distinct subjects, *Max. ID repeats* (maximum images from a single identity), and yaw pose statistics.

Set	Feature	# subjects	Max. # imgs per ID	Yaw (deg.)
Maximally activating images	Caricatures, comics	132	16	$+2.14 \pm 13.88$
	Bald White men	246	3	$+1.10 \pm 17.39$
	Baseball caps	108	61	$+1.21 \pm 28.29$
	Older men, beards	65	77	-1.22 ± 13.37
	Bindi	41	92	-0.44 ± 18.53
	Overweight	87	31	$+2.32 \pm 15.66$
	Infants	122	8	$+0.46 \pm 14.01$
	Young male Asians	50	47	$+2.83 \pm 21.14$
	Red cheeks	230	6	$+1.43 \pm 19.43$
	Dark Hair	71	106	$+1.70 \pm 16.83$
	Arab headdress	165	15	$+1.07 \pm 17.30$
	Large Forehead	233	5	$+2.36 \pm 18.46$
	Mouth Obstructed	234	4	$+3.85 \pm 18.53$
	Nasolabial Folds	225	7	-0.02 ± 18.25
	Squinting	170	8	$+0.36 \pm 17.13$
	Eyeglasses	128	26	$+3.07 \pm 20.51$
	Aggregated ($\mu \pm \sigma$)	144.19 ± 69.80	32.00 ± 33.12	$+1.39 \pm 18.04$
Minimally activating images	Caricatures, comics	201	12	$+2.56 \pm 18.77$
	Bald White men	209	11	$+0.18 \pm 27.98$
	Baseball caps	211	6	$+0.27 \pm 20.11$
	Older men, beards	206	9	$+2.60 \pm 19.63$
	Young male Asians	213	5	$+1.65 \pm 18.48$
	Overweight	201	9	$+0.96 \pm 20.06$
	Infants	216	5	$+0.68 \pm 20.06$
	Young male Asians	199	10	$+3.92 \pm 16.46$
	Red cheeks	184	13	$+1.06 \pm 19.82$
	Dark Hair	199	13	$+1.35 \pm 18.56$
	Arab headdress	204	5	$+3.02 \pm 18.60$
	Large Forehead	163	25	$+0.35 \pm 17.76$
	Mouth Obstructed	205	6	$+0.68 \pm 22.01$
	Nasolabial Folds	171	15	$+4.70 \pm 16.81$
	Squinting	181	12	$+1.48 \pm 18.27$
	Eyeglasses	177	24	$+0.18 \pm 21.89$
	Aggregated ($\mu \pm \sigma$)	196.25 ± 15.49	11.25 ± 5.90	$+1.60 \pm 19.70$

,

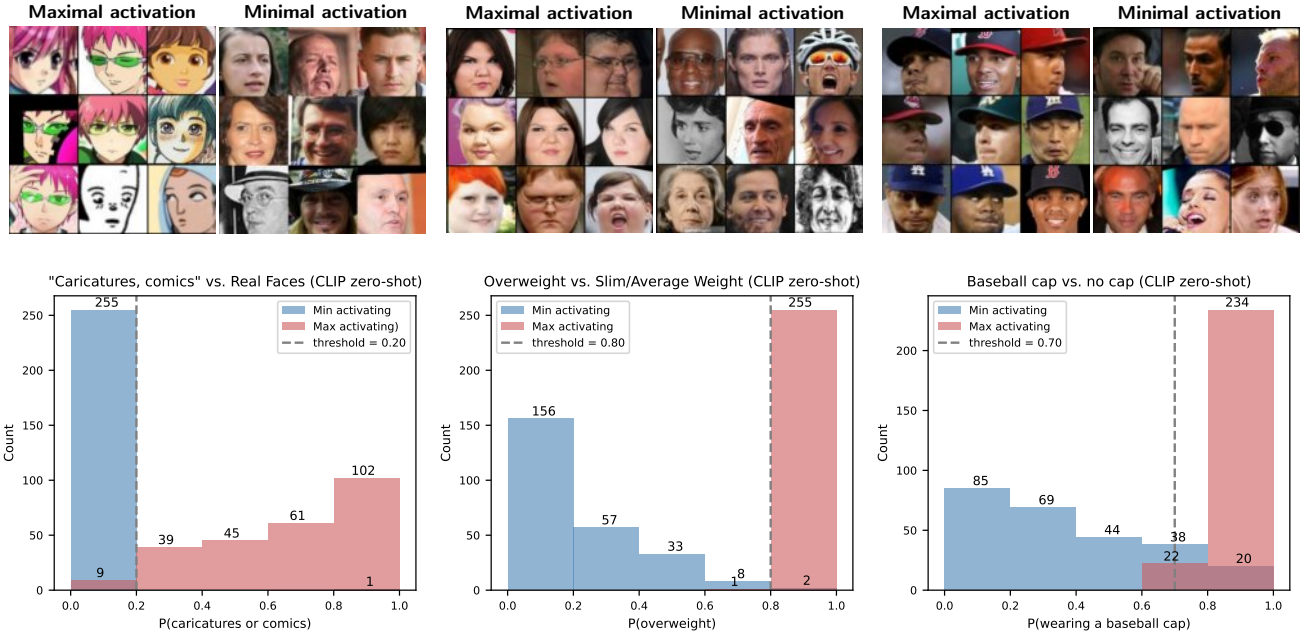
We then apply relevant open-source pretrained models depending on the semantic meaning of the feature. For features that are expected to encode age (e.g., "Infants", "Older men, beards", or "Young male Asians"), we use the FairFace [16] age classifier to label both sets of images and compare distributions via histograms. As the "Young male Asians" feature should also capture race and gender, we additionally apply the DeepFace [40] classifier to predict those attributes on both sets.

The results for demographic features, evaluated with age, gender, and ethnicity classifiers, are shown in Figure 12. For the feature "Young male Asian," panel (a) shows that 250 of 265 maximally activating images (red bars) were classified as Asian (using DeepFace), while the minimally activating images (blue bars) reflect the general Glint360k distribution (mostly White, fewer Asian, and others). Notably, the Asian predictions in the minimally activating images do not align with the other attributes (not young, not male), supporting the interpretation. Panel (b) confirms the gender aspect using DeepFace classifier: the maximally activating images are predominantly male, while the minimally activating images show a more balanced gender distribution. Panel (c) validates the age component using FairFace classifier: all maximally activating images are under 29, whereas only three of the minimally activating images are under 19. The remaining minimally activating images under 29 lack the additional attributes (gender and race), further confirming the feature's interpretation. Panel (d) shows results for the feature "Infants," evaluated with an age classifier. Among the maximally activating images, 239 are classified as 0–2 years old and the remaining 17 fall into the 3–9 years bucket; however, none of these images classified older than 3. No images with maximal activation are assigned to higher age groups, thereby confirming the interpretation. In contrast, none of the minimally activating images are classified as 0–2, with the majority assigned to ages above 20. Panel (e) presents age histograms for the feature "Older men, beards." All maximally activating images are classified as over 60 years, supporting the interpretation of "Old." Minimally activating images largely follow the background dataset distribution, and importantly, none of the few images classified as older than 60 contain beards. Panel (f) shows the race distribution for the feature "Bald white men," where the majority of maximally activating images are classified as White, again validating the assigned interpretation.

For attributes where no reliable pretrained classifier is available for tightly cropped face images, we use the CLIP model [33]. In this case, we formulate appropriate text prompts representing alternative classes and compute similarity scores between image embeddings and averaged text embeddings. In Figure 11, we present histograms for non-demographic attributes, where we employ the state-of-the-art CLIP model [33] in a zero-shot setting using descriptive prompts aligned with the interpretations. After applying a softmax, we obtain class probabilities. For the feature "Caricatures, comics," for instance, we use positive prompts such as *"a comic-style drawing of a face," "a cartoon drawing of a face," and "an illustration of a face"*, and contrast them with negative prompts such as *"a real photo of a human face," and "a photographic portrait of a person"*

Panel (a) shows the probability histogram for the feature "Caricatures, comics." With a threshold of 0.2, all maximally activating images fall below the cutoff, while only 9 minimally activating images do so, quantitatively confirming the interpretation. Panel (b) considers the feature "Overweight." A threshold of 0.8 separates all maximally activating images, with only 2 minimally activating images exceeding this value, again supporting the interpretation. Panel (c) presents the feature "Baseball caps." Using a threshold of 0.7, 234 maximally activating images reach probabilities between 0.8 and 1, though here the separation shows somewhat greater overlap between maximally and minimally activating images.

We have also investigated whether the discovered interpretable features also correlate with *other discernible patterns* such as pose, illumination, background, and identity within the positive-activating samples as opposed to the



(a) Probability histogram: real vs. drawn for "Caricatures, comics" feature.

(b) Probability histogram for "Overweight" feature.

(c) Probability histogram: presence of "Baseball caps" feature.

Figure 11: Examples of images that maximally and minimally activate discovered interpretable features, shown alongside histograms of the predicted probabilities for those features. Panel (a) illustrates the "Caricatures, comics" feature, panel (b) corresponds to "Overweight," and panel (c) to "Baseball caps." The probabilities were obtained using a pretrained CLIP model in a zero-shot setting, where descriptive prompts were aligned with interpretations from our human study, considering 256 maximal and 256 minimal images per feature.

dataset means and negative-activating samples. We recorded the number of different identities present in the 256 maximally and minimally activating images for each feature. Our results, which are presented in Table 8, show that a set of maximally activating images on average contains images of 144 ± 70 subjects, whereas the negative sets contain images of 196 ± 15 subjects, which is also close to the dataset mean for a random sample of 256 images. This shows that our interpretable features are marginally more ID-focused than a random image sample, but not significantly so. Furthermore, the negative activating samples are representative of a random dataset sample, as expected.

We have also used the SynergyNet 3D face model [51] for yaw pose regression. We found no statistically significant differences between the yaw pose distributions of maximally- and minimally- activating images for any of the features considered. We were also unable to identify any other apparent pattern regarding background or illumination.

6. Conclusion

In this paper, we presented FaceMINT Library, a tool designed to provide state-of-the-art methods for mechanistic interpretability in biometric research. The library facilitates rapid testing of (face) recognition models in a plug-and-play manner by offering objective functions for activation maximization, which can target neurons, layers, channels, or their combinations. To support this, it includes multiple image parametrizations, such as DIP, CPPN, FFT, and RGB, along with regularization functions to guide optimization. We observed that DIP and CPPN are particularly effective for interpreting neurons in the deeper layers of the network, as their priors favor input images resembling facial shapes. In contrast, RGB and FFT exhibit some structural similarities with DIP and CPPN but are more useful for understanding earlier layers, which typically encode low-level features such as basic shapes rather than high-level concepts.

We have also evaluated the library's sparse autoencoder implementations, namely G-SAE, TopK SAE, JumpReLU SAE, and vanilla SAE on face recognition models. We considered distinct architectures, the CNN-based AdaFace model and the transformer-based SwinFace model. Our experiments demonstrate that sparse autoencoders produces sparse representations with lower activation density compared to the denser activations observed in original templates and their PCA-transformed baselines. These results highlight the library's potential for advancing research in interpretable biometric systems using mechanistic interpretability.

Extensive experimentation attempting to interpret the SwinFace and AdaFace models using our library have shown AdaFace to be much more readily interpretable, measured by success in training sparse autoencoders, as well as our experiments with feature interpretation through dataset search and input image optimization. We hypothesize that this might be due to the inherent architectural differences between the

two models and the way they interact with the SAE and DIP optimization processes - i.e., the structural image prior present in the CNN-based Adaface as opposed to the learned attention structure of the transformer-based Swinface. However, whether this is a limitation of our library or inherent to the two face recognition models considered remains to be explored in future work.

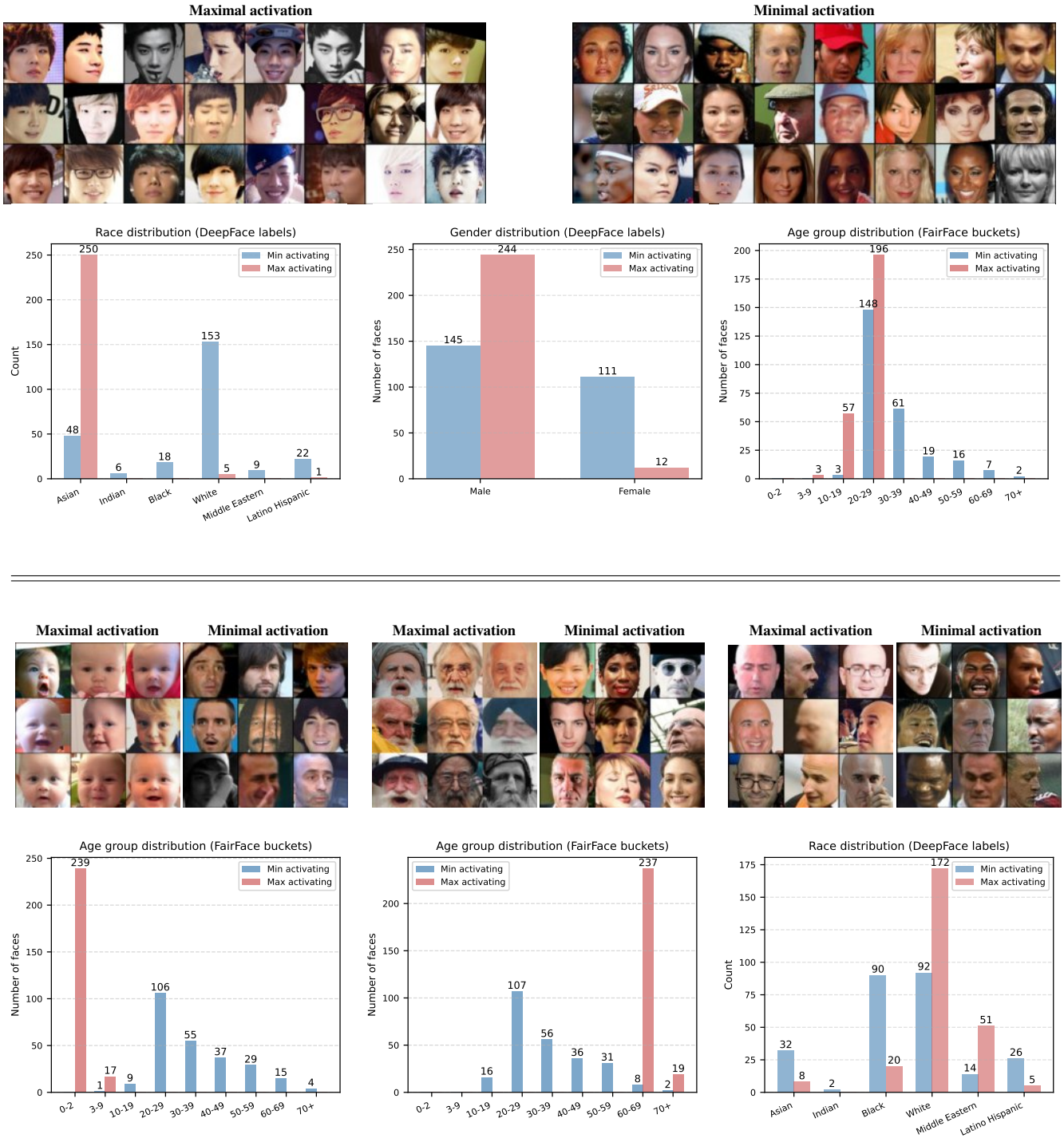


Figure 12: Examples of images that maximally and minimally activate selected sparse latent features, alongside demographic classifier outputs (shown as histograms). For each feature, we use 256 maximally and 256 minimally activating images, processed with DeepFace (race, gender) and FairFace (age). The top row illustrates the "Young Male Asians" feature with race, gender, and age predictions. The bottom row shows age distributions for "Infants" and "Older Men with Beards," real vs. drawn probabilities for "Caricatures/Comics," and race predictions for "Bald White Men."

The sparse autoencoders themselves capture effectively all of the information encoded in the templates of the face recognition models, as in the paper, we considered SAEs that achieve a zero reconstruction loss. However, as we have noted in the experimental protocol, relatively few of the neurons in the sparse autoencoder activation space have readily apparent interpretations. This is still an improvement on the original template space, where zero of the activation basis directions are interpretable by default, as the model architecture, the training objectives, and the optimization algorithms are known to encourage superposition [10] or polysemanticity, where each neuron is used to represent multiple concepts. Sparse autoencoders have been to a degree successfully introduced to reverse this process.

In the future, this library can be also extended to support additional biometric modalities, such as iris, sclera, periocular region, and palmprint recognition, as well as newer model architectures, including diffusion-based models and those designed for multi-modal fusion. We also plan to investigate cross-architecture robustness of discovered features. In particular, we plan to assess whether different face recognition models encode semantically aligned attributes (e.g., "beard") in a similar manner, and, if so, whether these features exhibit comparable neuron activations.

7. Acknowledgment

Research presented here was has been funded by the ARIS research project J2-50069 "Mechanistic Interpretability for Explainable Biometric AI" (MIXBAI), and the ARIS research programme P2-0250 "Metrology and Biometric Systems".

References

[1] An, X., Zhu, X., Gao, Y., Xiao, Y., Zhao, Y., Feng, Z., Wu, L., Qin, B., Zhang, M., Zhang, D., et al., 2021. Partial fc: Training 10 million identities on a single machine, in: IEEE/CVF International Conference on Computer Vision, pp. 1445–1449.

[2] Bau, D., Zhou, B., Khosla, A., Oliva, A., Torralba, A., 2017. Network dissection: Quantifying interpretability of deep visual representations, in: IEEE Conference on Computer Vision and Pattern Recognition, pp. 6541–6549.

[3] Benson-Tilsen, T., Soares, N., 2016. Formalizing convergent instrumental goals., in: AAAI Workshop: AI, Ethics, and Society, pp. 62–70.

[4] Bereska, L., Gavves, E., 2024. Mechanistic interpretability for ai safety–a review. arXiv preprint arXiv:2404.14082.

[5] Biagi, C., Rethfeld, L., Kuijper, A., Terhörst, P., 2023. Explaining face recognition through shap-based pixel-level face image quality assessment, in: IEEE International Joint Conference on Biometrics, IEEE. pp. 1–10.

[6] Bora, R.P., Terhörst, P., Veldhuis, R., Ramachandra, R., Raja, K., 2024. Slice: Stabilized lime for consistent explanations for image classification, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10988–10996.

[7] Bricken, T., Templeton, A., Batson, J., Chen, B., Jermyn, A., Conerly, T., Turner, N., Anil, C., Denison, C., Askell, A., et al., 2023. Towards monosemanticity: Decomposing language models with dictionary learning. Transformer Circuits Thread , 1–10.

[8] Chattopadhay, A., Sarkar, A., Howlader, P., Balasubramanian, V.N., 2018. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks, in: IEEE Winter Conference on Applications of Computer Vision, pp. 839–847.

[9] Cunningham, H., Ewart, A., Riggs, L., Huben, R., Sharkey, L., 2023. Sparse autoencoders find highly interpretable features in language models. arXiv preprint arXiv:2309.08600 .

[10] Elhage, N., Hume, T., Olsson, C., Schiefer, N., Henighan, T., Kravec, S., Hatfield-Dodds, Z., Lasenby, R., Drain, D., Chen, C., et al., 2022. Toy models of superposition. arXiv preprint arXiv:2209.10652 .

[11] Emam, A., Stomberg, T.T., Roscher, R., 2023. Leveraging activation maximization and generative adversarial training to recognize and explain patterns in natural areas in satellite imagery. IEEE Geoscience and Remote Sensing Letters , 1–5.

[12] European Union, 2024. Regulation (eu) 2024/1689 of the european parliament and of the council of 13 june 2024 on artificial intelligence and amending certain union legislative acts. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32021R0954. Official Journal of the European Union, L 168/1.

[13] Fel, T., Hervier, L., Vigouroux, D., Poche, A., Plakoo, J., Cadene, R., Chalvidal, M., Colin, J., Boissin, T., Bethune, L., Picard, A., Nicodeme, C., Gardes, L., Flandin, G., Serre, T., 2022. Xplique: A deep learning explainability toolbox. Workshop on Explainable Artificial Intelligence for Computer Vision , 1–4.

[14] Gao, L., la Tour, T.D., Tillman, H., Goh, G., Troll, R., Radford, A., Sutskever, I., Leike, J., Wu, J., 2024. Scaling and evaluating sparse autoencoders. arXiv preprint arXiv:2406.04093 .

[15] Gorton, L., 2024. The missing curve detectors of inceptionv1: Applying sparse autoencoders to inceptionv1 early vision. arXiv preprint arXiv:2406.03662 .

[16] Karkkainen, K., Joo, J., 2021. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 1548–1558.

[17] Katzmann, A., Taubmann, O., Ahmad, S., Mühlberg, A., Sühling, M., Groß, H.M., 2021. Explaining clinical decision support systems in medical imaging using cycle-consistent activation maximization. Neurocomputing 458, 141–156.

[18] Kim, M., Jain, A.K., Liu, X., 2023. Adaface: Quality adaptive margin for face recognition, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 18750–18759.

[19] Knoche, M., Teepe, T., Hörmann, S., Rigoll, G., 2023. Explainable model-agnostic similarity and confidence in face verification, in: IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 711–718.

[20] Lin, H., Liu, H., Li, Q., Shen, L., 2023. Activation template matching loss for explainable face recognition, in: IEEE International Conference on Automatic Face and Gesture Recognition (FG), pp. 1–8.

[21] Loyola-González, O., 2019. Black-box vs. white-box: Understanding their advantages and weaknesses from a practical point of view. IEEE Access 7, 154096–154113.

[22] Lundberg, S.M., Lee, S.I., 2017. A unified approach to interpreting model predictions. Advances in neural information processing systems 30.

[23] Mirabet-Herranz, N., Winter, M., Lu, Y., Bousnina, N., Pfıster, J., Galdi, C., Dugelay, J.L., Bailer, W., Ebrahimi, T., Correira, P.L., et al., 2024. Xaiface: a framework and toolkit for explainable face recognition, in: IEEE International Conference on Content-Based Multimedia Indexing (CBMI), pp. 1–7.

[24] Mordvintsev, A., Pezzotti, N., Schubert, L., Olah, C., 2018. Differentiable image parameterizations. Distill doi:10.23915/distill.00012.

[25] Neto, P.C., Gonçalves, T., Pinto, J.R., Silva, W., Sequeira, A.F., Ross, A., Cardoso, J.S., 2024. Causality-inspired taxonomy for explainable artificial intelligence. arXiv preprint arXiv:2208.09500.

[26] Oblak, T., Haraksim, R., Beslay, L., Peer, P., 2023. Probabilistic fingermark quality assessment with quality region localisation. Sensors 23, 4006.

[27] Olah, C., Mordvintsev, A., Schubert, L., 2017. Feature visualization. Distill .

[28] Patel, V.M., Ratha, N.K., Chellappa, R., 2015. Cancelable biometrics: A review. IEEE Signal Processing Magazine 32, 54–65.

[29] Peng, X., Yu, X., Sohn, K., Metaxas, D.N., Chandraker, M., 2017. Reconstruction-based disentanglement for pose-invariant face recognition, in: IEEE International Conference on Computer Vision, pp. 1623–1632.

[30] Petsiuk, V., Das, A., Saenko, K., 2018. Rise: Randomized input sampling for explanation of black-box models. arXiv preprint arXiv:1806.07421 .

[31] Qi, Z., Khorram, S., Li, F., 2019. Visualizing deep networks by optimizing with integrated gradients., in: CVPR Workshops, pp. 1–4.

[32] Qin, L., Wang, M., Deng, C., Wang, K., Chen, X., Hu, J., Deng, W., 2023. Swinface: a multi-task transformer for face recognition, expression recognition, age estimation and attribute estimation. IEEE Transactions on Circuits and Systems for Video Technology .

[33] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I., 2021. Learning transferable visual models from natural language supervision. CoRR abs/2103.00020. URL: https://arxiv.org/abs/2103.00020, arXiv:2103.00020.

[34] Rajamanoharan, S., Conmy, A., Smith, L., Lieberum, T., Varma, V., Kramár, J., Shah, R., Nanda, N., 2024a. Improving dictionary learning with gated sparse autoencoders. arXiv preprint arXiv:2404.16014 .

[35] Rajamanoharan, S., Lieberum, T., Sonnerat, N., Conmy, A., Varma, V., Kramár, J., Nanda, N., 2024b. Jumping ahead: Improving reconstruction fidelity with jumprelu sparse autoencoders. arXiv preprint arXiv:2407.14435 .

[36] Rajpal, A., Sehra, K., Bagri, R., Sikka, P., 2023. Xai-fr: explainable ai-based face recognition using deep neural networks. Wireless Personal Communications 129, 663–680.

[37] Ribeiro, M.T., Singh, S., Guestrin, C., 2016. " why should i trust you?" explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, pp. 1135–1144.

[38] Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation, in: Medical image computing and computer-assisted intervention (MICCAI 2015), Springer. pp. 234–241.

[39] Rot, P., Grm, K., 2024. Investigating the interpretability of biometric face templates using gated sparse autoencoders and differentiable image parametrizations, in: ICML 2024 Workshop on Mechanistic Interpretability.

[40] Serengil, S., Ozpinar, A., 2024. A benchmark of facial recognition pipelines and co-usability performances of modules. Journal of Information Technologies 17, 95–107. URL: https://dergipark.org.tr/en/pub/gazibtd/issue/84331/1399077, doi:10.17671/gazibtd.1399077.

[41] Shankaranarayana, S.M., Runje, D., 2019. Alime: Autoencoder based approach for local interpretability, in: Intelligent Data Engineering and Automated Learning–IDEAL 2019: 20th International Conference, Manchester, UK, November 14–16, 2019, Proceedings, Part I 20, Springer. pp. 454–463.

[42] Stanley, K.O., 2007. Compositional pattern producing networks: A novel abstraction of development. Genetic programming and evolvable machines 8, 131–162.

[43] Sudjianto, A., Zhang, A., Yang, Z., Su, Y., Zeng, N., 2023. Piml toolbox for interpretable machine learning model development and diagnostics , 1–8.

[44] Templeton, A., Conerly, T., Marcus, J., Lindsey, J., Bricken, T., Chen, B., Pearce, A., Citro, C., Ameisen, E., Jones, A., Cunningham, H., Turner, N.L., McDougall, C., MacDiarmid, M., Tamkin, A., Durmus, E., Hume, T., Mosconi, F., Freeman, C.D., Sumers, T.R., Rees, E., Batson, J., Jermyn, A., Carter, S., Olah, C., Henighan, T., 2024. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html. Affiliation: Anthropic.

[45] tensorflow, 2018. lucid. URL: https://github.com/tensorflow/lucid.

[46] Terhörst, P., Fährmann, D., Damer, N., Kirchbuchner, F., Kuijper, A., 2020. Beyond identity: What information is stored in biometric face templates?, in: IEEE International Joint Conference on Biometrics, pp. 1–10.

[47] Terhörst, P., Huber, M., Damer, N., Kirchbuchner, F., Raja, K., Kuijper, A., 2023. Pixel-level face image quality assessment for explainable face recognition. IEEE Transactions on Biometrics, Behavior, and Identity Science 5, 288–297.

[48] Ulyanov, D., Vedaldi, A., Lempitsky, V., 2020. Deep image prior. International Journal of Computer Vision 128, 1867–1888.

[49] Wang, H., Wang, Z., Du, M., Yang, F., Zhang, Z., Ding, S., Mardziel, P., Hu, X., 2020. Score-cam: Score-weighted visual explanations for convolutional neural networks, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 24–25.

[50] Williford, J.R., May, B.B., Byrne, J., 2020. Explainable face recognition, in: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M. (Eds.), Computer Vision – ECCV 2020, Springer International Publishing, Cham. pp. 248–263.

[51] Wu, C.Y., Xu, Q., Neumann, U., 2021. Synergy between 3dmm and 3d landmarks for accurate 3d facial geometry, in: 2021 International Conference on 3D Vision (3DV).

[52] Yang, W., Le, H., Savarese, S., Hoi, S., 2022. Omnixai: A library for explainable ai. arXiv preprint arXiv:2206.01612.

[53] Zhang, K., Zhang, Z., Li, Z., Qiao, Y., 2016. Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Processing Letters 23, 1499–1503.

[54] Zhao, X., Huang, W., Huang, X., Robu, V., Flynn, D., 2021. Baylime: Bayesian local interpretable model-agnostic explanations, in: Uncertainty in artificial intelligence, PMLR. pp. 887–896.

[55] Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A., 2016. Learning deep features for discriminative localization, in: IEEE Conference on Computer Vision and Pattern Recognition, pp. 2921–2929.

[56] Zhou, Z., Rong, S., Cai, H., Zhang, W., Yu, Y., Wang, J., 2017. Generative adversarial nets with labeled data by activation maximization. arXiv preprint arXiv:1703.02000.

[57] Zimmermann, R.S., Klindt, D.A., Brendel, W., 2024. Measuring mechanistic interpretability at scale without humans, in: ICLR 2024 Workshop on Representational Alignment.

Appendix

Grid Search to Obtain Optimal SAEs

This section presents the grid search results for optimizing the SAE, G-SAE, TopK SAE, and JumpReLU SAE models. Table 9 highlights key trends. For SAE, $\lambda = 10^{-4}$ offers the best balance of sparsity and reconstruction. G-SAE is more sensitive, with optimal λ at 10^{-8} (SwinFace) and 10^{-7} (AdaFace), beyond which activations vanishes. TopK SAE improves with increasing k, peaking at 200 (SwinFace) and 400 (AdaFace), indicating different optimal sparsity levels. JumpReLU SAE enforces the strongest sparsification, with best λ at 10^{-7} (SwinFace) and 10^{-8} (AdaFace). Overall, AdaFace embeddings require more active neurons than SwinFace, suggesting structural encoding differences.

Table 9: Grid search to obtain optimal SAE hyper parameters for SwinFace and AdaFace.

		SwinFace				**AdaFace**			
Autoencoder	**Hyperparameters**	$\|x\|_0$	$\|x\|_1$	**% Alive**	**MSE**	$\|x\|_0$	$\|x\|_1$	**% Alive**	**MSE**
SAE	$\lambda = 10^{-12}$	1101	5.6563e+02	97.55	0.0000	25280	4.0387e+03	46.03	0.0000
	$\lambda = 10^{-11}$	1034	5.5413e+02	98.58	0.0000	25364	4.0213e+03	47.06	0.0000
	$\lambda = 10^{-10}$	1093	5.6749e+02	99.37	0.0000	25002	3.7744e+03	46.81	0.0000
	$\lambda = 10^{-9}$	1048	5.5845e+02	99.00	0.0000	25240	3.7992e+03	46.90	0.0000
	$\lambda = 10^{-8}$	937	5.3978e+02	99.51	0.0000	25157	3.6978e+03	50.31	0.0000
	$\lambda = 10^{-7}$	628	4.4249e+02	99.67	0.0000	15386	1.2613e+03	100.00	0.0002
	$\lambda = 10^{-6}$	336	1.7482e+02	79.73	0.0000	1453	3.2418e+02	100.00	0.0000
	$\lambda = 10^{-5}$	242	3.3432e+01	41.82	0.0000	597	5.6230e+01	96.48	0.0000
	$\lambda = \mathbf{10^{-4}}$	**206**	**1.2117e+01**	**64.26**	**0.0000**	**445**	**1.9744e+01**	**94.86**	**0.0000**
	$\lambda = 10^{-3}$	205	5.8390e+00	65.40	0.0000	406	9.1209e+00	37.32	0.0002
	$\lambda = 10^{-2}$	52	4.0881e+00	7.71	0.5857	0	2.0000e-04	88.13	0.9109
GSAE	$\lambda = 10^{-12}$	8354	1.6660e+03	13.11	0.0000	20097	2.9403e+03	31.90	0.0000
	$\lambda = 10^{-11}$	8442	1.5754e+03	13.26	0.0000	20200	2.5368e+03	32.66	0.0000
	$\lambda = 10^{-10}$	7282	1.2878e+03	13.94	0.0000	19732	1.9847e+03	39.00	0.0000
	$\lambda = 10^{-9}$	4251	5.7744e+02	92.25	0.0002	7848	9.4433e+02	31.03	0.0003
	$\lambda = \mathbf{10^{-8}}$	**301**	**2.4032e+02**	**78.12**	**0.0000**	635	3.2999e+02	85.99	0.0000
	$\lambda = 10^{-7}$	235	2.3401e+02	71.25	0.0008	**342**	**3.1897e+02**	**96.28**	**0.0026**
	$\lambda = 10^{-6}$	178	1.3522e+02	35.41	0.0861	241	1.2739e+02	88.75	0.1051
	$\lambda = 10^{-5}$	0	3.7000e-03	93.83	0.9149	0	1.1300e-02	99.86	0.9111
	$\lambda = 10^{-4}$	0	2.8000e-03	92.86	0.9160	0	4.3000e-03	96.95	0.9142
	$\lambda = 10^{-3}$	0	7.0000e-04	64.09	0.9259	0	9.0000e-04	69.81	0.9352
	$\lambda = 10^{-2}$	0	1.0000e-04	14.04	0.9343	0	1.0000e-04	13.68	0.9460
TopK SAE	k = 50	50	6.5568e+01	52.49	0.1704	50	6.3381e+01	99.82	0.2189
	k = 100	100	1.7362e+02	0.44	0.0605	100	9.7253e+01	97.15	0.1485
	k = 200	**200**	**1.7782e+02**	**94.23**	**0.0000**	200	1.5276e+02	84.64	0.0961
	k = 300	300	2.1113e+02	95.97	0.0000	300	2.2491e+02	93.63	0.0304
	k = 400	400	2.2894e+02	96.17	0.0000	**400**	**2.8341e+02**	**98.98**	**0.0003**
	k = 500	500	2.4318e+02	87.58	0.0000	500	3.0519e+02	98.74	0.0000
	k = 600	600	2.5373e+02	88.48	0.0000	600	3.1671e+02	99.54	0.0000
	k = 800	800	2.7411e+02	96.30	0.0000	800	3.2948e+02	96.32	0.0000
	k = 1000	1000	2.9171e+02	97.88	0.0000	1000	3.3428e+02	95.94	0.0000
	k = 2000	2000	3.2802e+02	97.76	0.0000	2000	5.2316e+02	93.33	0.0000
	k = 5000	5000	4.2823e+02	97.16	0.0000	5000	5.7167e+02	94.70	0.0000
	k = 10000	10000	6.1051e+02	94.17	0.0000	10000	7.0667e+02	95.50	0.0000
JumpReLU SAE	$\lambda = 10^{-12}$	544	4.3137e+02	91.36	0.0000	16237	1.5694e+03	100.00	0.0002
	$\lambda = 10^{-11}$	480	4.2009e+02	76.76	0.0000	3733	8.2975e+02	99.92	0.0000
	$\lambda = 10^{-10}$	471	3.9628e+02	55.80	0.0000	970	5.8778e+02	93.76	0.0000
	$\lambda = 10^{-9}$	452	4.5127e+02	26.56	0.0000	660	5.6993e+02	59.04	0.0000
	$\lambda = 10^{-8}$	586	1.3118e+03	42.07	0.0000	**941**	**1.8338e+03**	**66.52**	**0.0001**
	$\lambda = \mathbf{10^{-7}}$	**245**	**9.9051e+02**	**14.62**	**0.0000**	371	1.2659e+03	7.79	0.0030
	$\lambda = 10^{-6}$	142	4.8077e+02	5.86	0.0624	130	3.5814e+02	4.76	0.2865
	$\lambda = 10^{-5}$	3	3.1228e+00	1.12	0.8411	1	4.7380e-01	0.91	0.8940
	$\lambda = 10^{-4}$	0	3.4100e-02	0.44	0.9105	0	1.7100e-02	0.36	0.9082
	$\lambda = 10^{-3}$	0	7.2000e-03	0.23	0.9131	0	3.4000e-03	0.14	0.9101
	$\lambda = 10^{-2}$	0	2.3000e-03	0.11	0.9139	0	1.2000e-03	0.08	0.9106

Survey Responses

In Table 10 and Table 13, we present examples of agreeing and disagreeing responses from the survey, corresponding to our annotations of concepts for the observed common features. For clarity, Table 10 contains primarily demographic-related attributes (e.g., age, gender, ethnicity), while Table 13 shows examples for non-demographic attributes.

Table 10: Examples of survey responses for the first set of interpreted feature directions.

	Infants	Comics	Overweight	Older men, beards	Young male Asians	Bindi
Agreeing	Babies; baby face; they are all kids; baby; very young kids (1.5 yrs)	Chinese cartoons; anime/drawn characters; drawings; digital drawings; comic characters	They all look fat; they are fat; too much body fat; people with more weight; overweight	Older man with beard; old man beard; old white skin + beard; white beards; older men with white beards	East Asian phenotype; asian boys; K-pop stars; young asian men; asian guys	Indian women; red dot on forehead; Indian older women; women with a bindi; symbol on forehead
Disagreeing	I don't know	Small nose	Women and feminine men; small eyes	N/A	Short hair	Unknown

Figure 13: Examples of survey responses for the remaining interpreted feature directions.

	Bald White men	Squinting	Large forehead	Mouth obstructed	Red cheeks
Agreeing	Baldness; bald; they are all bald; white bald men; bald white guys	Closed/almost closed eyes; eyes not seen; squinting into camera; eyes not fully visible; hidden eyes	Elliptical head shape	All have something in the mouth; instrument/mic near face; object near mouth; something in front/in mouth; something near mouths	They are all blushing; blushing; pronounced red cheeks
Disagreeing	N/A	Nonchalant; people in their 30s; light hair	Weird stare; I got nothing; direct eye contact; pointy nose; dark eyes	Same expression	Open mouth; thin eyebrows; white teenagers?; Anglo phenotype; uncanny valley