Discovering Interpretable Feature Directions in the Embedding Space of Face Recognition Models

Richard Plesh*[1] Janez Križaj*[2] Keivan Bahmani[1] Mahesh Banavar[1]
Vitomir Štruc[2] Stephanie Schuckers[1]
[1]Clarkson University, United States [2]University of Ljubljana, Slovenia

Abstract

Modern face recognition (FR) models, particularly their convolutional neural network based implementations, often raise concerns regarding privacy and ethics due to their "black-box" nature. To enhance the explainability of FR models and the interpretability of their embedding space, we introduce in this paper three novel techniques for discovering semantically meaningful feature directions (or axes). The first technique uses a dedicated facial-region blending procedure together with principal component analysis to discover embedding space direction that correspond to spatially isolated semantic face areas, providing a new perspective on facial feature interpretation. The other two proposed techniques exploit attribute labels to discern feature directions that correspond to intra-identity variations, such as pose, illumination angle, and expression, but do so either through a cluster analysis or a dedicated regression procedure. To validate the capabilities of the developed techniques, we utilize a powerful template decoder that inverts the image embedding back into the pixel space. Using the decoder, we visualize linear movements along the discovered directions, enabling a clearer understanding of the internal representations within face recognition models. The source code will be made publicly available.

1. Introduction

Face recognition (FR) technology has proven itself beneficial across various domains. The benefits of face recognition stem from its ability to efficiently and unobtrusively determine identity. Leveraging this capability, face recognition technology has become ubiquitous in personal devices, border controls, and law enforcement [5, 26]. While it is not difficult to find positive applications for face recognition, it has also raised concerns about the opacity of recognition decisions in contemporary FR models [16], underscoring the need for improved interpretability and explainability to ensure their trustworthiness. Despite considerable efforts towards better understanding the mechanisms behind today's deep learning based FR techniques, it remains difficult to precisely explain the inner workings of FR models in a human-interpretable way. This is because most modern models are based on heavily parameterized neural networks with elaborate architectures that implement the input-output mapping in a complex non-interpretable way and are therefore often treated as *black-box* models, where only the input images and output results bear semantic meaning. In response to these challenges, researchers are continually studying the inner workings of FR models to better explain their behaviour [19]. Such explanations are critical for the transparency of automated decision-making, the trustworthiness of face recognition technology and, not least, are also expected to be available by default by various privacy laws and regulations, such as GDPR[1].

Central to the operation of FR models is the concept of template similarity. When two faces are subjected to a comparison within a face recognition system, a comparison score is typically computed that captures the similarity of the faces in the embedding (or template) space of the FR model. This comparison score, in a sense, encodes how similar two faces are in terms of their visual features and overall appearance. Typically, the comparison score is the extent of the explanation that face recognition systems provide during the recognition process. Unfortunately, this singular number leaves practitioners wondering: What visual features were used to determine the similarity and come to an identity conclusion? To answer such questions, various explainability techniques have emerged in the literature over the years that aim to provide insight into the internal mechanisms governing face recognition models [19].

Existing techniques towards the explainability of face recognition models generally fall into one of two categories: (i) *attribution techniques* that attempt to locate the most important pixels in an image given a recognition decision (or embedding comparison), and (ii) *embedding/template interpretation* techniques that assign human-interpretable meaning to the deep features used in modern face recognitions models. The first category of explainability techniques most often relies on the so-called *saliency maps* [2, 13, 18]. These maps give some indication of what region of the im-

*Equal contribution.

[1] https://gdpr-info.eu/

age was the most "important" for the recognition decision. Generally, their indications tend to highlight pixels that represent the eyes, nose, and mouth as important for a recognition outcome. While this result tells us that these regions are important to identity in a general sense, it does not give any further detail on what about the eyes, nose, and mouth distinguishes individuals. Furthermore, these predictions come with a high degree of variance, further increasing the uncertainty of results and the potential for selection bias [3, 6]. The second category of techniques is focused on deciphering what the deep features are specifically encoding about the face. These techniques have the potential to explain FR decisions in much greater detail as they are focused on the feature space rather than the image space. Prior work in this area has investigated the organization of the feature space by analyzing: (1) the similarity structure of the template codes [20], (2) the semantics of the greatest variance directions in the embedding space [21], or (3) feature hierarchies in the template space [11] to mention a few of the most impactful works. While these techniques provide insight into the organization and high-level characteristics of the FR embedding space, they are still limited in their ability to discover/interpret data attributes beyond a few basic classes (e.g., gender, illumination, viewpoint) and are challenging to apply with facial images captured in-the-wild.

In this work, we aim to expand the explainability of face recognition decisions and the interpretability of the FR template space by developing multiple novel technique for discovering semantically meaningful deep features (and directions) in the embedding space of contemporary face recognition models. Specifically, we propose the following techniques that also present the main contributions of this work:

- **Semantic Spatially Isolated Deep Feature Discovery:** With this approach, we first introduce a targeted facial-region blending process (illustrated in Figure 1) that manipulates local semantic structures of the face and produces images with identical pixel values in all areas except the targeted semantic region. Using a large number of such manipulated faces, we then probe the template space of face recognition models and explore directions of greatest variations to identity deep features that correspond to spatially isolated semantic face structures.
- **Label-Guided Discovery with Centroid Modelling:** With the second proposed approach, we utilize attribute labels of the facial images to identify clusters of faces in the template space that share the same appearance characteristics. We then estimate the difference vector between the centroid of an observed attribute-cluster and the centroid of a selected reference cluster (e.g., canonical faces in neutral pose, expression and homogenous illumination) and use this vector (a deep feature) to explain selected direction in the embeddings space.
- **Label-Guided Discovery with Regression Modelling:** For the last approach, we model the relationship between selected attributes and the embeddings of facial images using linear regression. This process allows us to estimate regression coefficients that can be interpreted as embedding space directions and exploited for studying template variations along the estimated embedding space axis.

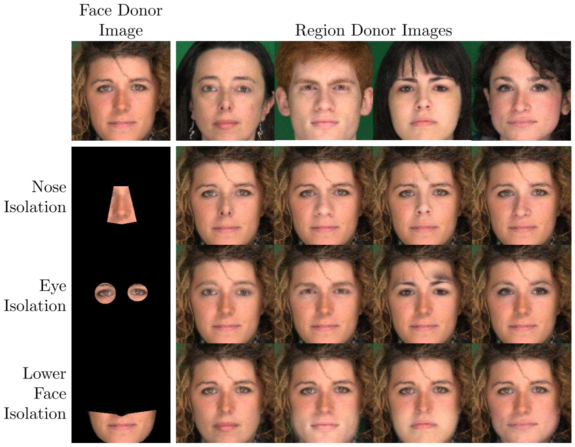


Figure 1. **Targeted facial-region blending.** An example set of the blended facial images used for discovering semantic-spatially isolated (SSI) deep features is presented. The face donor image is blended into the region donor images using facial region masks in the left-most column. After a set of blended images is created, they are embedded into the embedding space of a face recognition model and used for discovering SSI feature directions.

To visualize the results of the proposed feature discovery techniques, we utilize the recently introduced Deep Face Decoder [15] that allows us to invert FR templates back into the image space. We apply our techniques to three distinct face recognition models and demonstrate state-of-the-art semantic feature discovery with a number of interesting insights. For example, we show that: (1) certain directions in the template space encode local facial properties, (2) within-class variations can be encoded through feature directions and be consistently applied across various identities, (3) various facial features utilize distinctly-shaped feature manifold, and also point to differences and similarities among the considered 3 FR models.

2. Related Work

With advancements in face recognition capabilities and the deployment of FR models on a wider scale, a critical issue that received considerable attention recently has been enhancing the interpretability and transparency of face recognition models [19]. While a considerable number of conceptually different techniques have been proposed in the literature so far, the majority of existing work falls into two broad categories that are briefly presented below.

Attribution Techniques aim to identify informative image regions and commonly utilize saliency maps to elucidate the decision-making processes of the studied face recognition models. Castanon *et al.* [2], John *et al.* [13], and Xu *et* *al.* [30], for instance, employed saliency maps to visualize and quantify the critical features in facial images that influence the decisions made by deep learning-based face recognition systems. These methods not only provide insights into the features that are deemed important by the models but also contribute to a better understanding of how these models process and recognize facial features. Similarly, Domingo [17] presented an approach that used saliency maps for explaining facial analysis techniques in scenarios, where internal access to the model is limited. This methodology stands out for its ability to interpret recognition decisions in a true black-box scenario, emphasizing the changes in recognition probability when the images are perturbed.

Feature Interpretability Techniques, on the other hand, focus on different aspects of face recognition and often aim to understand the properties of the embedding space of face recognition models. Upchurch *et al.* [24], for example, studied the interpolation of features within deep neural networks to achieve controlled modifications in image attributes, such as age or expression. O'Toole *et al.* [20], Hill *et al.* [11], and Parde *et al.* [21] explored the organization of the embedding space, which is instrumental in understanding how these deep learning based face recognition models handle recognition across varied attributes. Wang *et al.* [28] focused on data augmentation techniques leveraging deep network feature linearization, and Williford *et al.* [29] and Knoche *et al.* [14] contributed to the field of explainable AI with innovative methods for explaining model predictions.

Our Contribution. The techniques, presented in this work, build on the research outlined above, but extend it in multiple aspects. Specifically, as we show in the experimental section, our techniques are capable of finding feature directions that correspond to much more complex facial structures/attributes with substantially less entanglement than what prior work was able to identify (i.e., global attributes, such as gender or ethnicity), and to determine feature axes that allow us to impact the encoded template properties, such as pose or illumination angle.

3. Semantic Spatially Isolated (SSI) Deep Feature Discovery

In this section, we present the first main contribution of this work, i.e., a novel technique for the discovery of feature directions in the embedding space of face recognition networks that correspond to semantically meaningful and spatially isolated visual facial features.

3.1. Problem formulation and method overview

Given an input face image $\mathbf{I} \in \mathbb{R}^{m\times n\times 3}$ and black-box facial recognition network R, our goal is to discover human-interpretable deep features (or directions), $\mathbf{t}$, in the abstract embedding space, $R(\mathbf{I})$, with the goal of gaining insight into the inner workings of deep face recognition models and the characteristics of its embedding space.

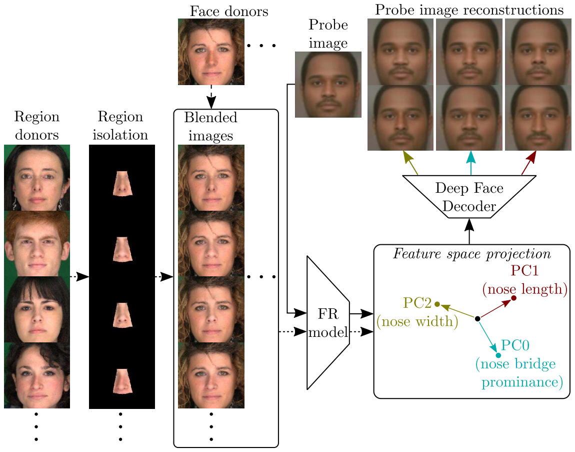


Figure 2. **The SSI deep feature discovery approach** leverages the variability in a specific facial region to discover principal directions within the embedding space that correspond to semantically meaningful changes of this region in the reconstructed images only. Semantic meaning is assigned to these directions through a visual analysis facilitated by a template inversion approach.

Central to the discovery of high-quality, disentangled, semantic and spatially isolated (SSI) features in the deep embedding space are two novel components, as illustrated in Figure 2. The first is a targeted facial-region blending (TFRB) module, which allows us to sample precise variations in different facial structures and then model them in the embedding space via principal component analysis (§3.2). The second is the feature decoding technique, which allows us to visualize the discovered deep feature directions with an arbitrary face sample and evaluate their level of entanglement and generalizability (§3.3).

3.2. Targeted blending and feature discovery

Facial region isolation. As different facial structures naturally correlate with one another in face data, we need to isolate the different face regions to learn disentangled facial recognition features. We begin by isolating the targeted face region, f, using a facial landmarking model S, for each sample in a dataset of N training images, $\{\mathbf{I}_i\}_{i=1}^N$. Here, a standard 68-point landmarking model is used, where the landmarks are used to isolate selected facial regions. For the experiments, we consider the eyes, nose and lower face region (see Figure 1), but the facial-region isolation process is general and can be applied to arbitrary facial structures.

Targeted blending. Next, we wish to blend the facial regions into donor faces to create a dataset of embeddings with selective variations. Using the previously computed facial region boundaries, we blend the facial region from a region donor image $\mathbf{I}_k$ with the surrounding face from a donor image $\mathbf{I}_m$ using a Gaussian kernel. This generates a dataset of blended images $\{\mathbf{I}^b_{m,k}|_{k\,=\,1\,:\,N\,-\,1}^{m\,=\,1\,:\,N}\} \in \mathbb{R}^{h\times w\times 3}$. Computing the corresponding face recognition embeddings for network, R, with embedding length d, results in the set of templates $\{\mathbf{t}^b_{m,k}|_{k\,=\,1\,:\,N\,-\,1}^{\,m\,=\,1\,:\,N}\} \in \mathbb{R}^d$.

Feature direction discovery. To learn appropriate (disentangled) feature directions, we need to first isolate the differences in the face embeddings attributed only to changes in the targeted face region f. Thus, we begin by computing the template center for each face donor image, i.e.:

$$\mu_m = \frac{1}{N-1} \sum_{k=1}^{N-1} \mathbf{t}^b_{k,m}. \tag{1}$$

Next, to remove the influence of each face donor, m, from the templates, we center the face embeddings with respect to their face donor images, as follows:

$$\mathbf{T} = [\mathbf{t}^b_{m,k} - \mu_m]\Big|_{k\,=\,1\,:\,N\,-\,1}^{\,m\,=\,1\,:\,N}. \tag{2}$$

This step ensures that the subsequent analysis focuses on region-specific features, rather than individual-specific traits. With the template scatter isolated to the target face region, f, we now center the face embeddings with respect to each feature, $\mathbf{T_c} = \mathbf{T} - \bar{\mathbf{T}}$, calculate the covariance matrix $\mathbf{C}$,

$$\mathbf{C} = \frac{1}{N(N-1)-1}\mathbf{T_c^T T_c}, \tag{3}$$

and then solve the following eigenproblem for the eigenvectors and eigenvalues:

$$\mathbf{CV} = \mathbf{V\Lambda}. \tag{4}$$

The eigenvectors $\mathbf{v}_i$ and eigenvalues λ_i are obtained from the columns of $\mathbf{V}$ and diagonal entries of $\mathbf{\Lambda}$, respectively. The leading eigenvectors corresponding to the largest d' eigenvalues, i.e., $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{d'}] \in \mathbb{R}^{d\times d'}$, $d' \leq d$ define the (orthonormal) principal axes of the face region subspace in the FR embedding space and represent the basis for meaningful features corresponding to the targeted face region f. The eigenvalues indicate the relative importance of each basis vector in describing the scatter. For our discovery procedure, we select the top d' eigenvectors, that jointly form a feature direction matrix $\mathbf{Z}$:

$$\mathbf{Z} = [\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_{d'}]. \tag{5}$$

3.3. Feature direction visualization

Based on the computed direction matrix $\mathbf{Z}$ it is possible to manipulate specific aspects of a provided face embedding by moving the embedding along the directions encoded in $\mathbf{Z}$. For instance, given a probe image $\mathbf{I}_p$, its template $\mathbf{t}_p$ can be transformed in a way that corresponds to spatially local appearance variations in the original input image $\mathbf{I}_p$, i.e.:

$$\mathbf{t}'_p = \mathbf{t}_p + \alpha\frac{\mathbf{v_i}}{|\mathbf{v_i}|}, \tag{6}$$

where $\mathbf{v_i}$ is the i-th eigenvector of the targeted facial region and α corresponds to the strength of the manipulation.

Finally, to evaluate the impact of the transformation induced by moving the embeding along the discovered feature directions, we use the state-of-the-art template inversion procedure from [15], called Deep Face Decoder (DFD). The DFD decoder is capable of inverting arbitrary face embeddings back into the visual domain, allowing for the interpretation of the encoded visual features. Using the decoder, the probe reconstruction $\mathbf{I}'_p$ can then be computed using:

$$\mathbf{I}'_p = \phi_D(\mathbf{t}'_p), \tag{7}$$

where, ϕ_D represents the DFD network mapping.

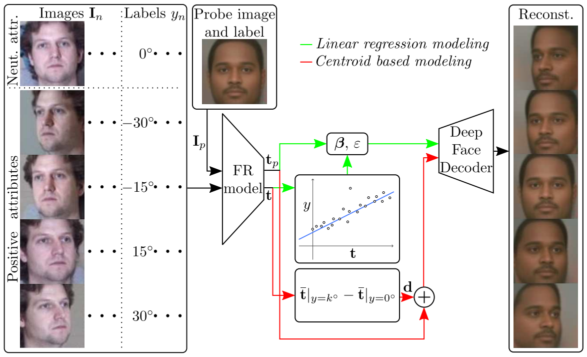


Figure 3. **Overview of the label-guided feature direction discovery techniques.** The flowchart illustrated the main ideas behind the centroid- and linear-regression-based modelling techniques to embedding-space direction discovery. Both techniques are capable of identifying linear directions that correspond to semantically meaningful within-identity face variations.

4. Label-Guided Feature Discovery

The discovery of semantic information in the template space can also be done by leveraging attribute labels associated with a dataset of facial images. In this section, we present two novel techniques, capable of discovering informative embedding space directions based on such labels. Formally, this task can be described as follows. Given a dataset of N facial images, $\{\mathbf{I}_n\}_{n=1}^N \in \mathbb{R}^{h\times w\times 3}$ with annotated attribute labels $\{y_n\}_{n=1}^N$ determine semantically meaningful embedding space directions corresponding to the labels. In accordance with this task, we design the first of our techniques based on a procedure build around centroid-based modelling and the second one based on linear regression modelling, as also illustrated in Figure 3. Details on the two techniques are provided below.

4.1. Centroid-based modeling

With the first proposed label-guided approach, based on centroid modelling, the initial requirement involves procuring a dataset comprising samples annotated with attribute labels. These samples must include both the target (positive) and baseline (neutral) manifestations of the attributes under investigation, such as *smile* and *neutral*, when studying facial expressions, for instance. This bifurcation is pivotal for isolating the attribute's effect on the corresponding embedding. Unfortunately, this also rules out many in-the-wild datasets, thus favoring the use of controlled datasets.

With the data requirement satisfied, we compute the mean template for a given attribute, denoted as κ. Similarly, the mean template for the corresponding neutral attribute, denoted as ν, is also computed. Next, we subtract the mean template pertaining to neutral attribute from the mean template of positive attribute, leaving behind the template difference d:

$$\mathbf{d}_\kappa = \bar{\mathbf{t}}|_{y_n=\kappa} - \bar{\mathbf{t}}|_{y_n=\nu} \tag{8}$$

This differential d embodies the deep features characteristic of the attribute κ, which can be incorporated into the given probe template $\mathbf{t}_p$ by

$$\mathbf{t}_p^\kappa = \mathbf{t}_p + \alpha \mathbf{d}_\kappa, \tag{9}$$

where factor α controls the level to which the attribute κ is considered. In our experiments we use $\alpha = 1$. The transformed probe template $\mathbf{t}_p^\kappa$ can then be reconstructed into the image domain, similarly to the procedure given in (7). Note that the above equation corresponds to a line in a vector space, where $\mathbf{d}_\kappa$ is the directional vector.

4.2. Linear Regression Modeling

For the second proposed label-guided discovery approach, we use linear regression to model the relationship between a particular label in the dataset (dependent variable y) and the deep features $\mathbf{t}$ that make up the facial recognition templates. The relationship can be expressed as:

$$y = \beta_0 + \beta_1 t_1 + \beta_2 t_2 + \beta_n t_n + \varepsilon. \tag{10}$$

Here, the coefficients $[\beta_1, \beta_2, \ldots \beta_n]$ represent the relative weights assigned to each deep feature to optimally predict the dataset label y. The intercept β_0 represents the predicted value when the independent variables are zero. The residuals ε represent the differences between the observed values of the labels y and the values predicted by the linear regression model. The coefficient vector $\boldsymbol{\beta}$ is fit by minimizing the sum of squared residuals:

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{N} (y_i - [\beta_0 \cdots \beta_n] \cdot \begin{bmatrix} 1 \\ \mathbf{t}_i \end{bmatrix})^2 \tag{11}$$

Given the coefficients $\boldsymbol{\beta} = [\beta_0, \beta_1, \ldots \beta_n]$ and the desired label y, the probe template $\mathbf{t}_p$ corresponding to neutral label can then be transformed as:

$$\mathbf{t}_p' = \mathbf{t}_p + \alpha[\beta_1 \cdots \beta_n], \tag{12}$$

where the weighting factor α is defined as:

$$\alpha = (y - \boldsymbol{\beta} \cdot \begin{bmatrix} 1 \\ \mathbf{t}_p \end{bmatrix}), \tag{13}$$

to ensure that the transformed template corresponds to the label y. Note that with the above setup, the regression coefficients are interpreted as a linear direction (or axis) in the embeddings space that can be traversed, similarly to a direction vector, to explore variations in the information content of the template corresponding to label y. The transformed template can again be decoded into the image space following the procedure given in (7).

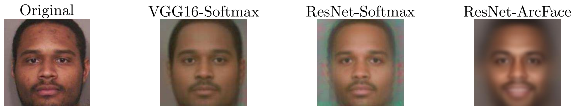


Figure 4. **Example decodings generated with DFD.** The probe image on the left (from NISTMedsII [7]) was first embedded using three different FR models and then decoded back into the image space using DFD [15]. Note what information is reconstructed.

5. Experiments

5.1. Experimental Settings

Face Recognition Models. We utilize three publicly available pre-trained face recognition models to evaluate the effectiveness of the proposed feature discovery techniques. The three models are chosen because they differ in the backbone architecture and learning objectives, and hence provide a solid cross-section of model variants for the evaluation. Additionally, these models are come with trained DFD inversion networks that are used to visualize results [15].

- **VGG16-Softmax**: The first model is based on the 16 layer convolutional neural network (ConvNet) VGG16 originally introduced in [23]. It consists of a set of convolutional layers with 3 × 3 filters, interspersed with max-pooling layers, and ending with fully connected layers with a softmax activation function to produce class probabilities. The model is trained on VGGFace2 [1] using a cross-entropy loss, followed by fine-tuning with a standard triplet loss. On LFW [12], VGG-16 attains a verification accuracy of 95.3%, as documented in [27].
- **ResNet-Softmax**: The second models uses a 50–layer residual ConvNet from [10], trained also on the VGGFace2 dataset, using a softmax loss. Unlike the VGG-16 model presented above, the ResNet-50 model contains skip connections, impacting the way the face images are encoded. Compared to VGG-16, the ResNet-50 model is also fairly lightweight with around 23M trainable parameters. The ResNet embeddings $\boldsymbol{e} \in \mathbb{R}^{2048}$ needed for the experiments are computed from the last global average pooling layer. The model has a verification accuracy of 97.3% on the LFW database, as reported in [27].
- **ResNet-ArcFace**: The last model is based on a 32–layer residual ConvNet trained on CASIA [31] using the implementation from [22]. While using the same type of backbone as the ResNet softmax implementation, ArcFace, [4], adds an angular margin to the softmax loss to enhance the discriminative power of the learned features. This modification is designed to encourage templates from the same identity to be near one another in an angular space, while increasing the gap between templates from different identities. Furthermore, feature embeddings are normalized before applying the ArcFace loss to encourage the templates to map to the surface of a hypersphere. This modification helps with robustness to changes in pose, il-

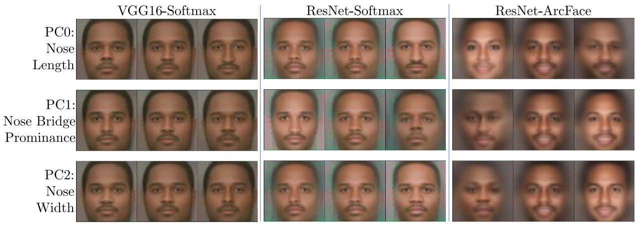


Figure 5. **Visualization of movements along the Nose directions.** Results are presented for 3 FR models (in columns) and the first 3 principal components of the PCA subspace (in rows).

lumination, and expression.

Decoder Models. For decoding the deep templates and inverting them back into the visual domain, we utilize the pretrained Deep Face Decoder models from [15]. The models are based on an inverted VGG architecture that maps face embedding to image reconstructions and were designed for deep feature-space exploration. The decoder models are pretrained for inverting the features of the VGG16-Softmax, ResNet-Softmax and ResNet-ArcFace FR models. Example decoding results using DFD for the three embedding models are shown in Figure 4.

Datasets We select three datasets for the experiments, each providing a unique environment to comprehend the nuances of the discovered deep feature directions, i.e.:

- **SiblingsDB-HQf.** [25] This dataset contains 184 frontal facial images of 92 sibling pairs captured at a resolution of 4256 × 2832. The dataset was acquired in front of a homogenous background and under diffuse illumination. This dataset is used exclusively for the discovery of semantic spatially isolated features. After removing duplicates and excluding problematic samples, 163 images are left for the quantitative part of the evaluation.
- **NISTMedsII.** [8] The NIST Multiple Encounter Dataset (MEDS) II dataset was compiled by the FBI Data Analysis Support Laboratory (DASL) and consists of persons with multiple frontal captures. The data is used exclusively for testing the discovered feature directions.
- **Multi-PIE.** [9] The Carnegie Mellon University Multiple Encounter Pose Illumination and Expression dataset is a face dataset consisting of over 750,000 images of 337 people recorded in up to four sessions over five months. Face images were collected in a laboratory environment and have controlled variation in viewpoint, illumination, and expression. This data is used exclusively for the learning of intra-subject deep FR features.

5.2. SSI Feature Direction Analysis

By applying the SSI feature discovery technique to particular regions of the face, we can discover feature directions that are corresponding to each of the targeted facial structures. As these directions represent a linear combination of the original deep features, we can consider them to be deep features themselves. Because the feature discovery process is unsupervised, we use a subjective visual analysis to associate semantic meaning to the discovered directions.

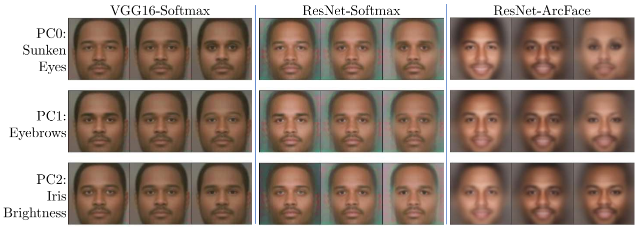


Figure 6. **Visualization of movements along the Eye Region directions.** Results are presented for 3 FR models (in columns) and the first 3 principal components of the PCA subspace (in rows).

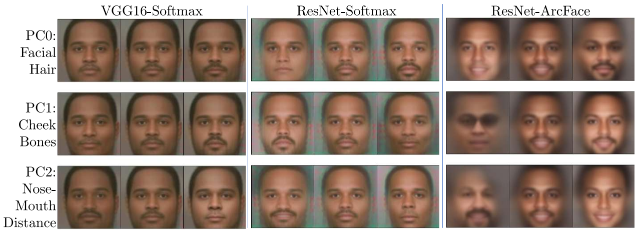


Figure 7. **Visualization of movements along the Lower-Face-Region directions.** Results are presented for 3 FR models (in columns) and the first 3 principal directions (in rows).

In Figures 5, 6, and 7, we demonstrate our technique in the nose, eye, and lower face regions, respectively, and visualize what impact movements along the first three principal axes have on the encoded template information. For brevity and ease of comparison between the 3 FR models, all results are visualized using the same probe image - from Figure 4:

- **Nose:** We observe some consistent themes in the discovered features, as illustrated in Figure 5. Consistently represented by principal component 0 (PC0), the nose length appears to be the most significant feature. Moving further down, PC1 appears to be tied to the prominence of the bridge of the nose. Last of all, PC2 appears to represent the nose width among all three FR model variations. The ResNet-ArcFace FR model has the weakest feature disentanglement from the feature vector, possibly due to the way with which the ArcFace loss encourages angular margins between different identities.
- **Eyes:** From Figure 6, we notice a similar trend in the common features discovered among the different FR models for the eye region. The most significant features consistently appear to relate to sunken or shadowed eyes, followed by eyebrow thickness and iris color. As each of these features can be modified using makeup, facial hair trimming, or colored contacts, these results suggest that modifying the eye region could be a successful method for obscuring identity. This also provides some evidence that the use of eye-shadow could be responsible for demographic variations in FR performance. Rather than being exploited as a vulnerability, this knowledge could also be used for the selective modification of these attributes dur-

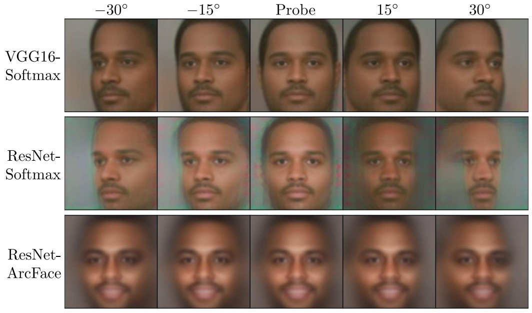
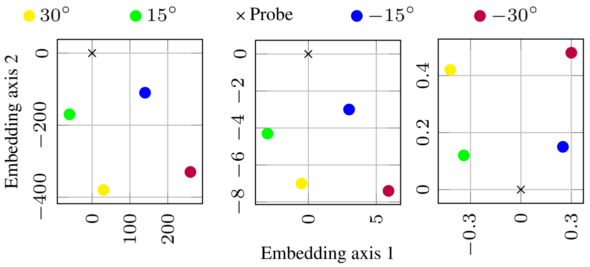


Figure 8. **Centroid-based Pose Features.** Using the proposed approach, we discovered the feature directions associated with the pose (in 15-degree increments) for three different FR models.



(a) VGGFace-Softmax (b) ResNet-Softmax (c) ResNet-ArcFace

Figure 9. **Pose-angle Centroid Visualization.** By using multidimensional scaling (MDS) we create a 2D representation of data labeled with different poses. We observe that despite the differing matcher designs (subfigures 9a, 9b, 9c), they each create a very similar horseshoe-type relationship for pose data.

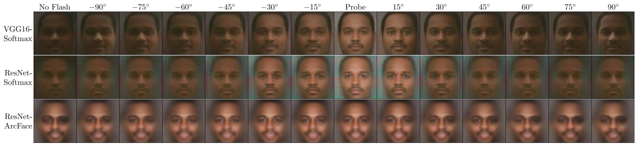
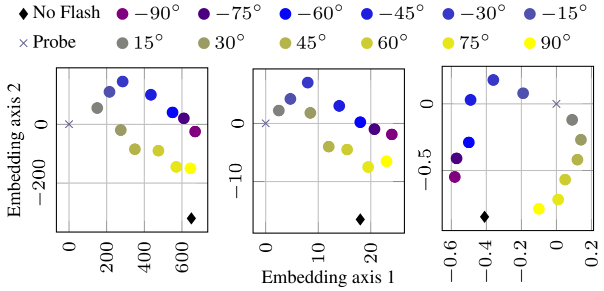


Figure 10. **Centroid-based Illumination Direction Features.** With the proposed approach, we are able to discover feature directions associated with the angle of illumination (in 15° increments) for three different face matchers. Best viewed zoomed-in.

ing training for increased model robustness.

- **Lower-Face:** Last of all, we test what deep features we discover when targeting the lower-face region. As the default cropping of the FR inputs typically cuts off the bottom of the chin, we opted to study the mouth and chin area together. As seen in Figure 7, we notice common features relating to facial hair, cheekbones, and nose-to-mouth distance. Unlike cheekbones and nose-to-mouth distance which have a strong inherent grounding to identity, we find it concerning that facial hair is found to be the most significant lower-face feature. This can likely lead to misidentification errors as facial hair color and style can change frequently, particularly among men. This result suggests that facial hair may be a source of differential performance among different demographic groups.



(a) VGGFace-Softmax (b) ResNet-Softmax (c) ResNet-ArcFace

Figure 11. **Illumination-angle Centroid Visualization.** By using multidimensional scaling (MDS), we create a 2D representation of data labeled with different illumination angles. We observe that each embedding network (subfigures 11a, 11b, 11c) has a unique path for positive and negative illumination angles, suggesting that the directional information is encoded in the templates. Additionally, we observe similar paths in the two softmax-based facial matchers with a unique circular path for the ArcFace-based matcher. This suggests that loss function design influences template feature organization more than matcher backbone design.

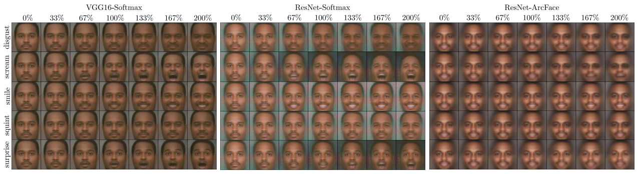
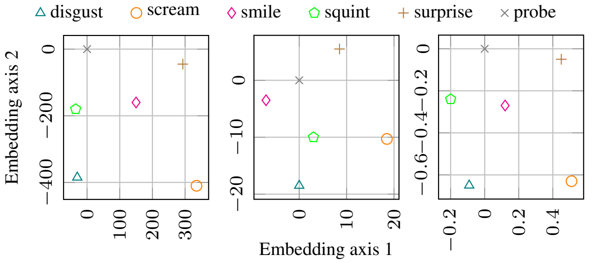


Figure 12. **Centroid-based Expression Features.** With the proposed methods, we discover the feature directions associated with 5 different facial expressions (disgust, scream, smile, squint, and surprise), for three different face matchers. The template is adjusted in the learned-expression direction by up to 200% of the average centroid difference. As can be seen in the columns right of 100%, the expression strength relationship carries beyond the average centroid distance. Best viewed zoomed-in.

5.3. Discovering Intra-identity Deep Features

In this section, we present experiments for discovering label-guided feature directions that correspond to intra-identity variation, specifically pose-angle, illumination-angle, and expression. We accomplish this using our two different label-guided feature discovery techniques on a controlled dataset. The first technique uses the relative positioning of the templates corresponding to different data labels to compute feature directions. This technique is flexible to different feature manifold shapes and compatible with multidimensional scaling (MDS) visualizations. Our second technique utilizes linear regression to compute the features directions that are predictive of the desired labels using linear combinations of the original deep features. Contrary to the prior technique, linear regression considers within-label variance during the fit procedure.

Pose Deep Features. When learning pose, we focused on five different pose angles, -30 degrees, -15 degrees, 0 degrees, 15 degrees, and 30 degrees. Empirically, we found



(a) VGGFace-Softmax (b) ResNet-Softmax (c) ResNet-ArcFace

Figure 13. **Expression Centroid Visualization.** With multidimensional scaling (MDS) we create a 2D representation of data labeled with different expressions in the dataset. We observe that the three embedding networks (subfigures 13a, 13b, 13c) have similar relative embedding distances between the different expression centroids, suggesting they each encode expressions similarly.

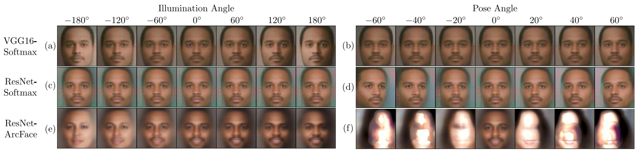


Figure 14. **Learned Features Directions using Linear Regression.** The fitted coefficients are used for visually evaluating the ability of the regression to model the relationship in rows a-f.

this to be the limit for left-to-right pose variation that is tolerated by the standard facial landmark detection pipelines. When using the centroid-learning approach, we computed the difference between the centroid for 0 degrees and the central embedding of either -30,-15,15, or 30. In Figure 8, we apply these learned vectors to the probe embedding to evaluate their semantic meaningfulness. For the two Softmax-based models, we see the learned feature able to control pose with minimal side effects. The ArcFace-based model has a more muted pose response, possibly illustrating better model robustness to pose variation. To visualize the shape of the feature manifold, we use MDS to create a low-dimensional representation of the relative distances between the different centroids (Figure 9). We find that, regardless of FR model, the relative distances between the label centroids appear to construct a similar horseshoe shape.

Next, we applied our linear regression pipeline to learn a single deep feature that can predict the pose label in the dataset. As shown in row (d) in Figure 14, this technique can learn a feature that has the desired effect on the pose angle for the ResNet-softmax network. In columns (b) and (f,) we see the features not having the desired effect for the other two FR models, likely due to the feature manifold shape.

Illumination Deep Features. Concerning illumination, we focused on the lighting angles -90 to 90 degrees (with 15-degree increments) in illumination along with a no-flash state. Using the same centroid learning approach used to detect pose features, we computed individual deep features for each illumination angle. Visualizing the deep features (Figure 10), we see varying illumination content in the embedding reconstructions. VGG16-Softmax is the most expressive in the visualization, able to encode both illumination intensity and direction. ResNet-Softmax and ResNet-ArcFace on the other hand, seem to primarily encode average light intensity. In Figure 11, we visualize the relative distances between the different light-angle centroids. We notice that VGG16-Softmax and ResNet-Softmax produce very similar mappings with two separate and somewhat parallel paths representing each illumination side. On the other hand, the ArcFace-based FR model encodes illumination angle and direction in a circular manifold. This suggests that the type of loss has a substantial effect on the structure of the feature space for these attributes.

The use of linear regression for learning the deep feature representing illumination has varying success among the different FR networks. In row (a) of Figure 14, we see reasonable success for the VGG16-based network. In row (c) however, we see that the learned feature for the ResNet-Softmax network does not have much effect. In row (e), we see that the ArcFace-based network can express the illumination direction, but is unable to do so without entanglement with other facial features.

Expression Deep Features. When learning deep features representing different facial expressions, we rely entirely on the centroid learning technique. This is because our dataset labels have only binary expression information, indicating the presence or absence of a particular facial expression. In Figure 12 we show the effect of the learned expression vectors on the probe image. We find VGG16-softmax to encode expression with the highest fidelity, followed by ResNet-Softmax, and then ResNet-ArcFace. When plotting the MDS visualization for the expression centroids (Figure 13), we observe that the models exhibit highly similar relative distances between different expressions. This suggests that the three embedding networks encode facial expressions in a comparable manner, indicating that despite differences in their architectures, these models share a similar approach to representing expressions in the embedding space.

6. Conclusion

We presented three novel techniques for learning semantically meaningful embedding space directions that can provide insights into the behavior of FR models. The techniques were tested in experiments with interesting findings.

Acknowledgement

The research presented in this work was supported in parts by the ARIS Research Programme P2-0250(B) “Metrology and Biometric Systems”, the ARIS Research Project J2-50069 (A) “MIXBAI”, the National Science Foundation under Grant No. 1650503, and the Center for Identification Technology and Research (CITeR).

References

[1] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2018)*, pages 67–74, Los Alamitos, CA, USA, may 2018. 5

[2] G. Castanon and J. Byrne. Visualizing and quantifying discriminative features for face recognition. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 16–23, 2018. 1, 2

[3] A. Das and P. Rad. Opportunities and challenges in explainable artificial intelligence (xai): A survey. *arXiv preprint arXiv:2006.11371*, 2020. 2

[4] J. Deng, J. Guo, N. Xue, and S. Zafeiriou. ArcFace: Additive angular margin loss for deep face recognition. In *Computer Vision and Pattern Recognition (CVPR)*, pages 4690–4699, 2019. 5

[5] H. Du, H. Shi, D. Zeng, X.-P. Zhang, and T. Mei. The elements of end-to-end deep face recognition: A survey of recent advances. *ACM Computing Surveys (CSUR)*, 54(10s):1–42, 2022. 1

[6] R. Dwivedi, D. Dave, H. Naik, S. Singhal, R. Omer, P. Patel, B. Qian, Z. Wen, T. Shah, G. Morgan, et al. Explainable ai (xai): Core ideas, techniques, and solutions. *ACM Computing Surveys*, 55(9):1–33, 2023. 2

[7] A. P. Founds, N. Orlans, W. Genevieve, and C. I. Watson. Nist special databse 32-multiple encounter dataset ii (meds-ii). Technical report, NIST, 2011. 5

[8] A. P. Founds, N. Orlans, G. Whiddon, and C. Watson. NIST special database 32 multiple encounter dataset II (MEDS-II) :: data description document. Technical Report NIST IR 7807, National Institute of Standards and Technology, 2011. Edition: 0. 6

[9] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker. Multi-pie. In *2008 8th IEEE International Conference on Automatic Face and Gesture Recognition*, pages 1–8, 2008. 6

[10] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 5

[11] M. Q. Hill, C. J. Parde, C. D. Castillo, Y. I. Colón, R. Ranjan, J.-C. Chen, V. Blanz, and A. J. O'Toole. Deep convolutional neural networks in the face of caricature. *Nature Machine Intelligence*, 1(11):522–529, 2019. Number: 11 Publisher: Nature Publishing Group. 2, 3

[12] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller. Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments. In *Workshop on Faces in 'Real-Life' Images: Detection, Alignment, and Recognition*, Marseille, France, Oct. 2008. 5

[13] T. A. John, V. N. Balasubramanian, and C. V. Jawahar. Canonical saliency maps: Decoding deep face models. In *IEEE Transactions on Biometrics, Behavior, and Identity Science*, volume 3, pages 561–572, 2021. 1, 2

[14] M. Knoche, T. Teepe, S. Hörmann, and G. Rigoll. Explainable model-agnostic similarity and confidence in face verification. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)*, pages 1–8, 2023. 3

[15] J. Križaj, R. O. Plesh, M. Banavar, S. Schuckers, and V. Štruc. Deep face decoder: Towards understanding the embedding space of convolutional networks through visual reconstruction of deep face templates. *Engineering Applications of Artificial Intelligence*, 132:107941, 2024. 2, 4, 5, 6

[16] B. Meden, P. Rot, P. Terhörst, N. Damer, A. Kuijper, W. J. Scheirer, A. Ross, P. Peer, and V. Štruc. Privacy–enhancing face biometrics: A comprehensive survey. *IEEE Transactions on Information Forensics and Security*, 16:4147–4183, 2021. 1

[17] D. Mery. True black-box explanation in facial analysis. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1595–1604, 2022. 3

[18] D. Mery and B. Morris. On black-box explanation for face verification. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1194–1203, 2022. ISSN: 2642-9381. 1

[19] P. C. Neto, T. Gonçalves, J. R. Pinto, W. Silva, A. F. Sequeira, A. Ross, and J. S. Cardoso. Explainable biometrics in the age of deep learning. *arXiv preprint arXiv:2208.09500*, 2022. 1, 2

[20] A. J. O'Toole, C. D. Castillo, C. J. Parde, M. Q. Hill, and R. Chellappa. Face space representations in deep convolutional neural networks. *Trends in Cognitive Sciences*, 22(9):794–809, 2018. 2, 3

[21] C. J. Parde, Y. I. Colón, M. Q. Hill, C. D. Castillo, P. Dhar, and A. J. O'Toole. Closing the gap between single-unit and neural population codes: Insights from deep learning in face recognition. *Journal of Vision*, 21(8):15, 2021. 2, 3

[22] S. I. Serengil and A. Ozpinar. Hyperextended lightface: A facial attribute analysis framework. In *Proceedings of the International Conference on Engineering and Emerging Technologies (ICEET)*, pages 1–4. IEEE, 2021. 5

[23] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *Proceddings of 3rd International Conference on Learning Representations, (ICLR)*, 2015. 5

[24] P. Upchurch, J. R. Gardner, G. Pleiss, R. Pless, N. Snavely, K. Bala, and K. Q. Weinberger. Deep feature interpolation for image content changes. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 6090–6099. IEEE Computer Society, 2017. 3

[25] T. F. Vieira, A. Bottino, A. Laurentini, and M. De Simone. Detecting siblings in image pairs. *The Visual Computer*, 30(12):1333–1345, 2014. 6

[26] M. Wang and W. Deng. Deep face recognition: A survey. *Neurocomputing*, 429:215–244, 2021. 1

[27] Q. Wang and G. Guo. Benchmarking deep learning techniques for face recognition. *J. Vis. Comun. Image Represent.*, 65(C), dec 2019. 5

[28] Y. Wang, X. Pan, S. Song, H. Zhang, G. Huang, and C. Wu. Implicit semantic data augmentation for deep networks. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc,

E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. 3

[29] J. R. Williford, B. B. May, and J. Byrne. Explainable face recognition. In A. Vedaldi, H. Bischof, T. Brox, and J. Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XI*, volume 12356 of *Lecture Notes in Computer Science*, pages 248–263. Springer, 2020. 3

[30] Z. Xu, Y. Lu, and T. Ebrahimi. Discriminative deep feature visualization for explainable face recognition. In *2023 IEEE 25th International Workshop on Multimedia Signal Processing (MMSP)*, pages 1–6, 2023. 3

[31] D. Yi, Z. Lei, S. Liao, and S. Z. Li. Learning face representation from scratch, 2014. 5