

Interpreting Face Recognition Templates using Natural Language Descriptions

Anastasija Manojlovska¹, Raghavendra Ramachandra², Georgios Spathoulas², Vitomir Štruc¹
Klemen Grm¹

¹University of Ljubljana, 1000 Ljubljana, Slovenia

²NTNU, Gjøvik, Norway

Abstract

Explainable artificial intelligence (XAI) aims to ensure an AI system’s decisions are transparent and understandable by humans, which is particularly important in potentially sensitive application scenarios in surveillance, security and law enforcement. In these and related areas, understanding the internal mechanisms governing the decision-making process of AI-based systems can increase trust and consequently user acceptance. While various methods have been developed to provide insights into the behavior of AI-based models, solutions capable of explaining different aspects of the models using Natural Language are still limited in the literature. In this paper, we therefore propose a novel approach for interpreting the information content encoded in face templates, produced by state-of-the-art (SOTA) face recognition models. Specifically, we utilize the Text Encoder from the Contrastive Language-Image Pretraining (CLIP) model and generate natural language descriptions of various face attributes present in the face templates. We implement two versions of our approach, with the off-the-shelf CLIP text-encoder and a fine-tuned version using the VGGFace2 and MAADFace datasets. Our experimental results indicate that the fine-tuned text encoder under the contrastive training paradigm increases the attribute-based explainability of face recognition templates, while both models provide valuable human-understandable insights into modern face recognition models.

1. Introduction

In the rapidly advancing field of Artificial Intelligence (AI), much of the focus has been on building deeper and more accurate models. However, this often leads to increased complexity, reducing their transparency. In sensitive real-world scenarios, AI models often process personal data in opaque and non-transparent ways, raising accountability concerns. Transparency and understanding of the internal mechanisms of such models are therefore crucial to address these issues. Regulations like the GDPR and the

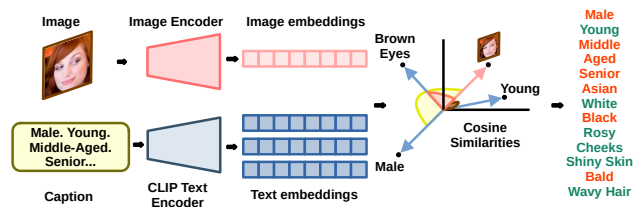


Figure 1. **Overview of the proposed CLIP-SMU (CLIP-based Symbolic Face Recognition Model Understanding) approach.** CLIP-SMU aims to interpret the information encoded in face templates using natural language descriptions. Given a face image I and a text description T we obtain the Image (I_{emb}) and Text (T_{emb}) Embeddings using Image and Text Encoders respectively. The I_{emb} are then compared to each of the T_{emb} and a cosine similarity score is generated for every attribute. Scores with high values indicate that certain attributes are encoded in the generated face templates (in green), while low scores represent attributes that are not encoded in the extracted face templates (in red).

AI Act emphasize the need for such clarity to ensure compliance and trust [8, 31]. Explainable Artificial Intelligence (XAI) aims to offer a solution to these issues by designing mechanisms capable of explaining AI-based decision-making processes and, consequently, ensuring that such decisions are transparent and understandable.

The importance of XAI is especially evident in the field of face recognition. Face recognition systems are widely used in a variety of applications, including surveillance, security, law enforcement, and personal device authentication. To be reliable and trustworthy, these systems must achieve high levels of accuracy. However, in order to further improve trust and comply with privacy laws and regulations their reasoning also needs to be interpretable and well-understood. Face recognition systems commonly represent facial images as high-dimensional feature vectors (face templates hereafter), in which compressed information about the appearance of the face is typically encoded. This includes identity information, but also cues about various facial attributes. Since different models produce face

templates with varying amounts of facial-attribute information, understanding what and how is encoded in the templates may provide insight into the internal workings of the models and consequently their decision-making process.

While many existing XAI techniques focus on local or visual explanations of face recognition systems [13, 22, 25], solutions capable of explaining the information encoded in face templates using Natural Language have received limited attention in the literature. To address this gap, we present in this paper **CLIP-SMU**, a CLIP-based Symbolic Face Recognition Model Understanding approach, as illustrated in Figure 1. The main idea behind CLIP-SMU is to describe the information content encoded in the extracted face templates using symbolic representations, such as language descriptions (or binary attribute labels). To facilitate this approach, we experiment with various face-image encoders (i.e., face recognition and face analysis models, such as AdaFace [11] and SwinFace [20]) and align the computed templates with a set of predefined text-description processed through CLIP’s text encoder. By reasoning over the joint image-text embedding space, we are able to identify attribute information present in the generated face templates and describe it using natural language. Through this approach, CLIP-SMU effectively translates a complex face templates into human-understandable form, making the internal mechanisms of modern face recognition models more transparent also to non-experts in this field.

In this paper, we make the following contributions:

- We introduce CLIP-SMU, a novel approach for interpreting the information content encoded in face templates using natural language. The proposed approach explores similarities in a joint image-text embedding space to produce human-understandable textual descriptions of the information encoded in the face templates.
- We explore two distinct implementations of CLIP-SMU, where the first uses CLIP’s off-the-shelf text encoder, whereas the second relies on a fine-tuned text encoder adapted to the particular task of interpreting face templates from SOTA face recognition models, i.e., AdaFace and SwinFace. The fine-tuned encoders will be made publicly available after review as another contribution.
- We investigate how different face recognition models influence the encoded information within face templates, offering valuable insights into their behavior. Furthermore, we demonstrate that the proposed CLIP-SMU framework generates human-understandable interpretations of face templates, making the insights accessible and meaningful even to non-experts.

2. Related work

Explainable AI. In standard “black-box” models, particularly deep learning systems, the internal processes that

lead to a decision or prediction tend to be difficult to interpret. Explainable AI (XAI) addresses this issue by making decision-making transparent, understandable, and justifiable allowing human users to understand and interpret the decisions made by AI models. A considerable number of techniques have been developed to advance the field of XAI, as detailed in [2] and [1], which provide extensive surveys on this field. Methods like Local Interpretable Model-agnostic Explanations (LIME) by Ribeiro *et al.* [24] and SHapley Additive exPlanations (SHAP) by Lundberg and Lee [14] offer model-agnostic approaches to interpret the predictions of complex models. LIME builds a local interpretable model around the classifiers’ predictions, while SHAP explains how each feature contributes to the final output by assigning an importance score to each prediction.

In addition to model-agnostic techniques, post-hoc explainability methods were designed specifically for image-based models. These methods aim to visually highlight the parts of the input, such as regions of an image, that were most influential in the model’s decision. Some of these methods include visualizing saliency maps as proposed by Simonyan *et al.* [26], Gradient-based methods [23, 27, 28] or CAM-based methods [6, 9, 10, 25] as explained by Zhang *et al.*, which proposed the Opti-CAM method [33].

Explainable Face Recognition (XFR) is a branch of XAI, which focuses on explaining and interpreting Face Recognition systems. Philips *et al.* [19] define four principles of XAI for biometrics and face recognition: *Explainability*, *Interpretability*, *Explanation Accuracy* and *Knowledge Limits*. They differentiate the terms *Explainable* and *Interpretable* in such a way that an Explainable system does not necessarily mean that it is interpretable. The system is explainable if it provides support and reasoning for each decision, which does not need to be correct. However, for a system to be interpretable it is required for the user to understand the explanations provided by the system.

Yin *et al.* [32] proposed one of the first approaches towards interpretable face recognition. They introduced two loss functions, i.e. Feature Activation Diversity (FAD) loss and Spatial Activation Diversity (SAD) loss. FAD enhances the robustness against occlusions, while SAD focuses on including semantic information during training. Williford *et al.* [30] proposed a face recognition system that uses triplets and an inpainting game to emphasize regions that distinguish between positive and negative matches.

Mechanistic interpretability is another subfield of XAI that tries to understand how the behavior of the neural network is affected by its individual components. In contrast to high-level interpretability approaches, which focus on abstract explanations or input-output relationships, mechanistic interpretability focuses on the model’s internal workings in order to explain how it processes information step by step. Most of the mechanistic interpretability research has

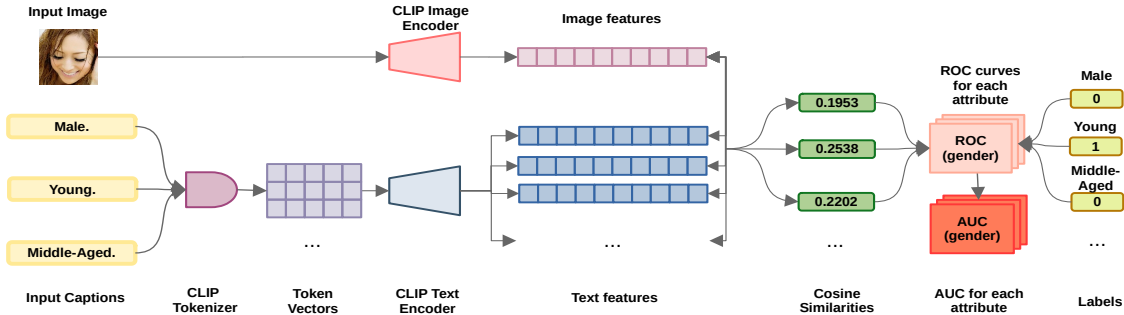


Figure 2. **Overview of the CLIP-SMU model for face template interpretation.** CLIP-SMU uses an image encoder to project the input face image into a face template space, where comparison with textual descriptions, processed through a suitably trained text-encoder are possible. Given a set of pre-defined text descriptions (possibly in the form binary attributes), CLIP-SMU then generates a ranked list of the most likely attributes encoded in the generated template. These attributes are hence naturally expressed as natural language descriptions.

focused on language models, leaving the vision domain relatively understudied [3, 7, 18]. For example, more recently Burns *et al.* [4] proposed an unsupervised method called Contrast-Consistent Search (CCS) for finding latent knowledge inside language models by answering “Yes” and “No” questions and mapping the representations of both answers to probabilities of being true.

Some of the research done in the vision domain includes the work of Zimmermann *et al.* [35], in which the authors investigated the mechanistic interpretability of nine SOTA deep learning models. They designed a psychophysics experiment to test how scaling of the models influences their mechanistic interpretability. Interestingly, they found no scaling effect for interpretability for any of the tested models. Moreover, a recent study by Palit *et al.* [18] implemented the causal mediation analysis (CMA) mechanistic interpretability method, originally developed for language models, for the language-image model BLIP [12].

Our Contribution. While many of the existing methods outlined above focus on local or visual explanations and can also be applied to face recognition models, our CLIP-SMU approach aims to interpret face templates using symbolic representations. Specifically, we attempt to leverage the information about facial attributes encoded in extracted face templates by using natural language descriptions and binary labels of facial attributes. While there have been previous attempts at generating natural language explanations for visual data [15, 17], they have, to the best of our knowledge, not been focused on the particularities of face recognition models, which is a unique aspect of CLIP-SMU [16].

3. Methodology

In this section, we present our CLIP-SMU approach to interpreting face templates using natural language description. The approach, illustrated in Figure 2, consists of the following two steps:

1. **Extracting face templates.** In this initial step, we utilize various face recognition models based on different model architectures and trained with different learning objectives to process facial images and extract face templates. Specifically, we implement two *state-of-the-art* face recognition and face analysis models, i.e., AdaFace [11] and SwinFace [20].
2. **Interpreting the extracted face templates.** After extracting the face templates, the next step is to decode and understand the information they contain. CLIP-SMU first analyzes these templates to identify the presence of encoded facial attributes and then generates natural language descriptions, which we associate with the extracted face templates by utilizing the language-image model CLIP [21]. To validate the results generated by CLIP-SMU, we train binary classifiers and identify the presence of absence of a specific attribute within the generated template.

This approach allows us to make the results of face recognition models more transparent and interpretable by presenting the information content of the face templates in a human-understandable form, which is easy to comprehend by non-experts. To achieve this goal, we develop two CLIP-SMU versions, which are presented in the following section.

3.1. CLIP-SMU for Face Template Interpretation

We consider two settings, when experimenting with the CLIP-SMU approach in Section 5. In the *first setting*, we employ CLIP’s image encoder together with its text encoder to evaluate how well CLIP’s off-the-shelf image encoder can extract meaningful face templates and represent facial data within a semantic space. This approach helps us to evaluate CLIP’s capability to map visual face features to natural language descriptions. In the *second setting*, we replace CLIP’s image encoder with the *state-of-the-art* face recognition and face analysis models AdaFace and Swin-

Face, which were specifically trained on face images. We adopt these models as backbones for feature (template) extraction, which we then try to interpret via natural language using CLIP’s text encoder. These two backbones are specifically selected, to allow us to study the differences in the convolutional and transformer-based architectures and their impact on the encoded information content.

In this second setting, we also fine-tune the different variants of the CLIP model, but instead of adapting the entire network’s weights, we only focus on fine-tuning the text encoder. The image encoders (AdaFace and SwinFace) are kept frozen, which allows us to leverage their pre-trained ability to extract face templates.

Zero-shot CLIP-SMU. For the Zero-shot CLIP-SMU approach, we use the pretrained CLIP model in a zero-shot fashion by connecting the face images’ embeddings with the text embeddings of the manually generated captions for the annotated attributes of each image. Specifically, we measure the similarity between a set of predefined text descriptions (captions) encoded with CLIP’s text encoder with image features (embeddings) extracted with CLIP’s image encoder. In the zero-shot setting, we use CLIP’s pretrained off-the-shelf encoders without modifying their weights. This approach is illustrated in Figure 2.

The images are first processed using the pre-trained ViT (Vision Transformer) based image encoder, which extracts relevant visual features from the facial images and generates the corresponding face templates. These templates encode the facial attributes in the image. Simultaneously, the captions corresponding to each facial attribute (such as age, gender, or facial characteristics) are tokenized using CLIP’s Tokenizer. This tokenization step converts the textual descriptions into a format that can be understood by the model. The tokenized captions are then passed through the pre-trained Transformer-based text encoder, which generates embeddings for each caption. These text embeddings represent the semantic content of the captions in a form that can be compared to the image embeddings, enabling the model to map the visual data to corresponding natural language descriptions effectively.

Fine-tuned CLIP-SMU. For the fine-tuned CLIP-SMU approach, we keep the image encoders (AdaFace and SwinFace) frozen and adapt the weights of the CLIP’s text encoder only. This allows the text encoder to adapt to the workings of the replaced image encoder.

For a batch of (image, text) pairs, the cosine similarities $sim(\cdot)$, between all the images’ embeddings I_{emb} and the corresponding texts’ embeddings T_{emb} , are maximized to fine-tune the CLIP model via a *Symmetric Cross Entropy (CE)* loss function. Image and text logits are obtained when the cosine similarity is multiplied by the temperature factor

t as shown in Eq (1), and multiplied by 100:

$$\begin{aligned} I_{logits} &= sim(I_{emb}, T_{emb}) \cdot e^t \\ T_{logits} &= I_{logits}^T, \end{aligned} \tag{1}$$

where I_{emb} and T_{emb} are the images’ and texts’ embeddings respectively, and t is the temperature factor, which is a learnable parameter adapted during training.

The logits are matrices of dimension ($batch\ size \times batch\ size$). Each row in the image matrix represents the similarity between a particular image in the batch and all of the captions in the batch, and vice versa for the text logits. Therefore, our goal is to maximize the similarity between the correct pairs of images and captions, i.e., the diagonal values of the logits matrix, and minimize all the other similarities. For that purpose, the *Cross Entropy Loss* is used as an objective function. The cross entropy loss for the pairs (image-text or text-image) is given by:

$$H(P, Q) = \sum_{i=1}^N -P_i \log Q_i, \tag{2}$$

where P_i is the probability of the i -th element in the true distribution P (it is either 1 for the correct pair or 0 for the others), Q_i is the probability of the i -th element in the predicted distribution P , which is calculated from the logits using the softmax function, and N represents the number of all possible pairs.

Therefore, the final *Symmetric Cross Entropy Loss (CE)* is calculated as a mean value of the image-to-text and text-to-image cross entropy losses, as shown in Eq (3), i.e.:

$$CE = \frac{H_I(P_I, Q_I) + H_T(P_T, Q_T)}{2}, \tag{3}$$

where $H_I(P_I, Q_I)$ is the image-to-text cross entropy loss, derived from the image logits, trying to maximize the diagonal values of the image logits matrix. Conversely, $H_T(P_T, Q_T)$ is the text-to-image cross entropy loss, trying to maximize the diagonal values of the text logits matrix, which is a transposed version of the image logits matrix.

3.2. Multi-label and multiple binary classifiers

To provide a reference for the interpretations generated by the CLIP-SMU approach, we use the AdaFace and SwinFace models as the backbones for a multi-label and a binary facial attribute classifier. Using this approach, our goal is to represent the encoded information content in the face templates in the form of binary attribute labels. Since such labels represent standard classifier outputs, this setting serves as a dual purpose. First and most importantly, it allows us to automatically generate a (approximate/weak) ground truth for the CLIP-SMU approaches, and second, it allows us to interpret the attributes encoded in the face templates using very basic symbolic representations.

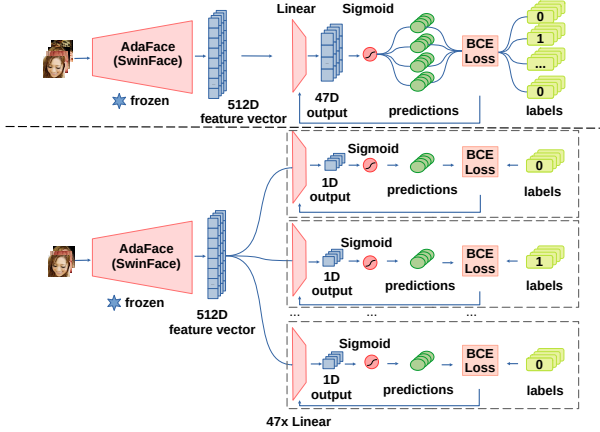


Figure 3. **Multi-label (top) and binary classifiers (bottom).** We implement binary classifiers for the prediction of facial attributes from the facial images to provide a (automatically generated) ground truth for the interpretations produced by CLIP-SMU.

Figure 3 illustrates the binary classifiers used for the prediction of 47 facial attributes. As can be seen, we rely on two configurations, one with a *multi-label classifier* that predicts all attributes within a single forward pass, and one with *multiple binary classifiers*, where a one classifier predicts one attribute label at the time.

The input to both of the classifiers is a 512-dimensional feature vector, which is extracted from the pre-trained face recognition models, AdaFace or SwinFace. The multi-label classifier processes the feature vector and produces a 47-dimensional output, where each dimension corresponds to one of the 47 facial attributes (e.g., "Male", "Young", "Black Hair", etc.), while the binary classifier produces only 1-dimensional output for one particular attribute. To ensure that each attribute is treated independently, a Sigmoid activation layer is applied just before the final output layer. By using a Sigmoid activation function, we ensure that the classifier treats each attribute prediction separately rather than as part of a mutually exclusive set of classes, which would be the case with a Softmax activation function. This is essential for facial attribute prediction, where multiple features (like "Smiling", "Eyeglasses" and "Wearing Hat") may occur together. Using both configurations in the experiments allows us to explore the trade-offs between shared learning of multiple attributes and the focused precision of individual attribute classifiers.

Training. In both the multi-label and binary classification setups, we use the *Binary cross-entropy (BCE)* loss function to train the classifiers, as shown in Eq. (4), i.e.:

$$BCE = -\frac{1}{N} \sum_{i=1}^N (y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)), \quad (4)$$

where y_i is the label of the particular attribute, \hat{y}_i is the prediction of the classifier and N is the number of samples.

BCE measures the difference between the predicted probability of an attribute (after applying the Sigmoid activation) and the actual label and allows for each facial attribute to be treated as a binary classification task—whether the attribute is present or absent.

4. Experimental Setup

4.1. Datasets

For fine-tuning and evaluation, we use facial images from the VGGFace2 dataset [5]. The text descriptions of the facial attributes are generated using the MAADFace dataset [29], which contains annotations of a large set of attributes for the VGGFace2 images.

VGGFace2 Dataset. The VGGFace2 [5] dataset is a large-scale face recognition dataset, containing 3.31 million images of 9131 identities, collected through Google Image Search. The images represent different pose, age, illumination, ethnicity and profession of the people. Example images are presented in Figure 4. The dataset is approximately gender-balanced, with 59.3% males, varying between 80 and 843 images for each identity, with 362.6 images on average.



Figure 4. **Example images from the VGGFace2 dataset.** The faces in VGGFace2 cover different attributes, from variations in hair color, presence of eyewear to facial hair and others.

VGGFace2 consists of a train and test split, where the train split features ~ 3.1 million images, representing 8631 identities, whereas the test split comprises $\sim 160,000$ images, corresponding to 500 identities.

MAADFace attribute annotations. The Massively Annotated Attribute Dataset (MAADFace) [29] contains face attributes annotations of the VGGFace2 dataset. It consists of 123.9 million attribute annotations of 47 different binary attributes. The MAAD-Face database was created by transferring the attribute annotations of the CelebA and LFW

datasets on the images of VGGFace2, by an annotation-transfer pipeline. An example of the face attribute annotation is depicted in Figure 5.



| Attribute | Value | Attribute | Value | Attribute | Value | Attribute | Value |
|-------------------|-------|------------------------|-------|------------------|-------|------------|-------|
| Male | 1 | Blond_Hair | -1 | Brown_Eyes | 0 | Eyeglasses | -1 |
| Young | 0 | Brown_Hair | 0 | Bags_Under_Eyes | 1 | Attractive | -1 |
| Middle_Aged | -1 | Cray_Hair | -1 | Bushy_Eyebrows | -1 | | |
| Senior | -1 | No_Beard | 1 | Arched_Eyebrows | -1 | | |
| Asian | -1 | Mustache | -1 | Mouth_Closed | 0 | | |
| White | 1 | No_Clock_Shadow | -1 | Smiling | -1 | | |
| Black | -1 | Goatee | -1 | Big_Lips | -1 | | |
| Rosy_Cheeks | -1 | Oval_Face | 0 | Big_Nose | 0 | | |
| Shiny_Skin | -1 | Square_Face | -1 | Pointy_Nose | -1 | | |
| Bald | -1 | Round_Face | -1 | Heavy_Makeup | -1 | | |
| Wavy_Hair | -1 | Double_Chin | -1 | Wearing_Hat | -1 | | |
| Receding_Hairline | 0 | High_Cheekbones | -1 | Wearing_Earrings | -1 | | |
| Bangs | -1 | Chubby | -1 | Wearing_Necktie | -1 | | |
| Sideburns | -1 | Obstructed_Forehead | -1 | Wearing_Lipstick | -1 | | |
| Black_Hair | -1 | Fully_Visible_Forehead | 0 | No_Eyewear | 1 | | |

Figure 5. **Face attribute annotation example.** Green indicates a positive attribute, red a negative, and yellow a missing annotation.

As the example shows, 1 indicates that the attribute is positively annotated, -1 means that the attribute is negative, and 0 implies that the attributes are not defined (there is no information whether the attribute is positive or negative).

4.2. Data preparation

Since we have different architectures that were implemented, each of them has different requirements regarding the input data preparation. Common to all visual models, images from the VGGFace2 dataset were first aligned using the *state-of-the-art* MTCNN algorithm [34] to produce images in which the faces are in the center. Since the VGGFace2 dataset consists of images of individuals with different poses in the wild which are not taken under controlled conditions, the MTCNN algorithm can not detect all the faces. For the experiments, we remove those images and consequently reduce the size of the dataset by ~ 10%. Moreover, since fine-tuning with the Cross Entropy Loss function requires unique (image, text) pairs in every batch, so it can associate a given text description to one particular image, there must not be repetitions of the same description for multiple images. Hence, we filter out the images, such that for every distinct caption we choose only one image. The filtering procedure further reduces the dataset, leaving us with ~260,000 training and ~15,000 test images. The images that remain are downsized to 112 × 112 pixels.

For the vision models (i.e., CLIP’s ViT, AdaFace and SwinFace), all images were first converted to RGB and then normalized before further processing. For CLIP’s Text Encoder the generated captions were tokenized using CLIP’s BPE Tokenizer and then converted to tensors. Finally, for

the binary classifier, a resampling technique was used. Since the positive and negative labels of the attributes in our dataset are highly imbalanced, we implemented an over-sampling technique on the training set, in which the minority class was randomly over sampled to match the number of samples of the majority class. For example, the attribute “Male” consists of ~ 1.8 million and ~ 1.2 million positive and negative training samples respectively, therefore ~ 600 000 random samples from the negative class were selected and added to training set in order to equal the number of positive samples.

4.3. Training details

Fine-tuned CLIP-SMU. In the fine-tuning process, the pre-trained weights for the image encoders (i.e., CLIP’s image encoder and the face recognition models) were kept frozen, and only the text encoder was fine-tuned. The batch size was 256 for the original CLIP model and 64 for the CLIP-SMU-AdaFace and CLIP-SMU-SwinFace variants. We used the *Adam Optimizer* for minimizing the loss.

Multi-label and binary classifiers. With the multi-label and binary classifiers, the backbone models were kept frozen and only the added layers were trained. The learning rate for both the multi-label and binary classifiers was set to 1×10^{-5} and the weight decay to 0. The parameters of the *Adam optimizer* were the same as described above. Both multi-label classifiers were trained for 15 epochs, while the binary classifiers for each attribute (AdaFace and SwinFace) were trained for 10 epochs each.

4.4. Performance metrics

To evaluate our models, we used Receiver Operating Characteristic (ROC) curves and the Area Under the Curve (AUC) as evaluation measures. The ROC curve is generated by plotting the True Positive Rate (TPR) against the False Positive Rate (FPR) at different threshold values. The TPR and FPR are defined by Eq. (5), i.e.:

$$\begin{aligned}
 TPR &= \frac{TP}{TP + FN}, \\
 FPR &= \frac{FP}{TN + FP},
 \end{aligned}
 \tag{5}$$

where TP, FN, FP and TN represent True Positives, False Negatives, False Positives and True Negatives, respectively.

5. Results

This section presents the results of our experiments by: (i) analyzing and comparing the ROC curves and AUC scores generated by various CLIP-SMU models and classifiers across 20 facial attributes, and (ii) providing a qualitative evaluation of the generated interpretations. It is worth noting that interpreting the encoded face attributes with

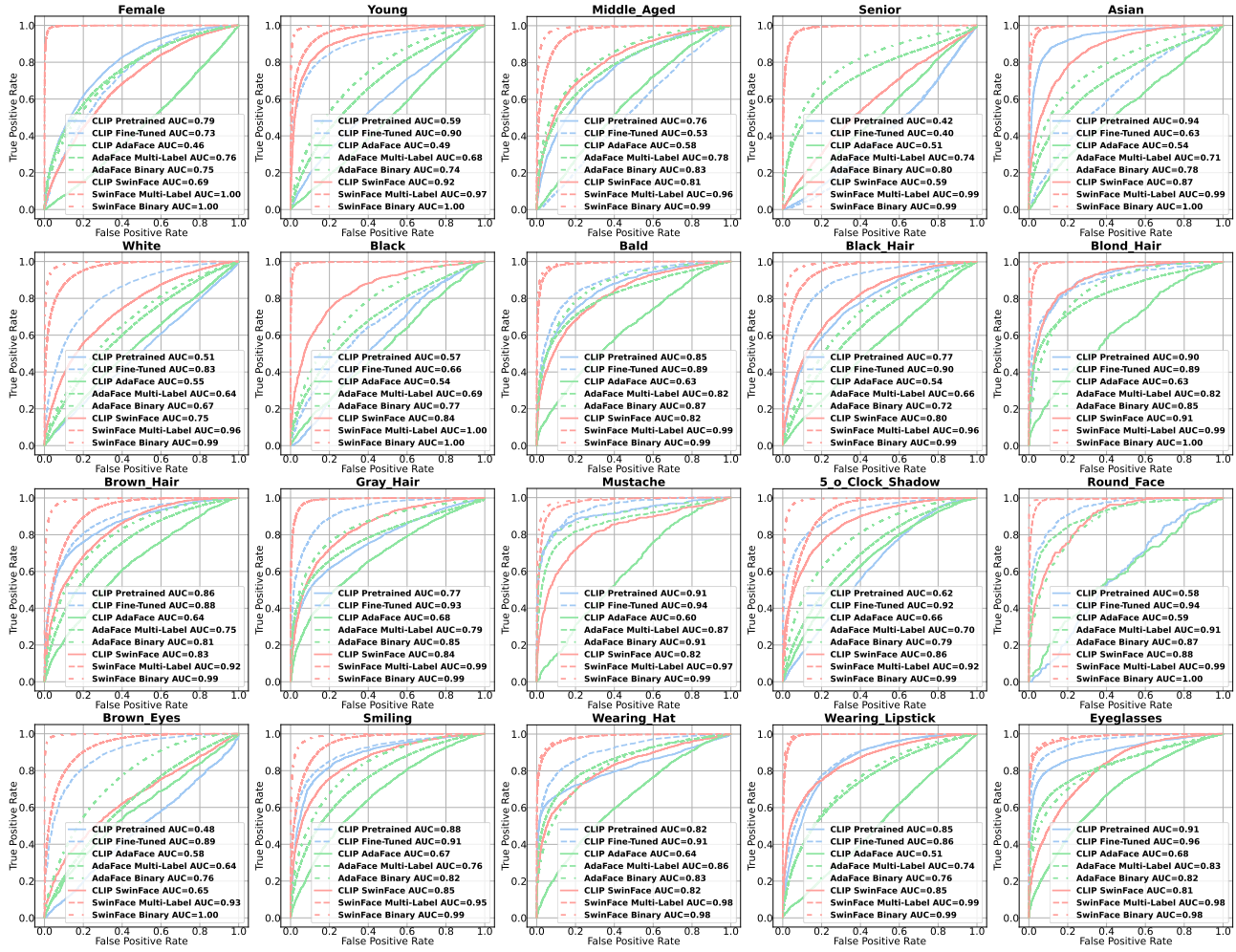


Figure 6. **ROC Curves for the attributes.** Dashed and dot-dashed lines represent the baselines the CLIP-based models are compared to (CLIP Fine-Tuned (blue dashed), AdaFace Multi-Label (green dashed), AdaFace Binary (green dot-dashed), SwinFace Multi-Label (pink dashed), SwinFace Binary (pink dot-dashed)), while filled lines refer to the different variants of the CLIP-based models (CLIP Pretrained (blue), CLIP-AdaFace (green), CLIP-SwinFace (pink)).

natural language descriptions produced by the CLIP-SMU models can be seen as a classification task, where the classification procedure is based on the similarities of the text descriptions and encoded visual features. ROC curves therefore provide insight into which specific attribute-based descriptions the models can most accurately align with the face templates. In other words, the ROC curves and associated AUC scores can be seen as a measure of confidence into the interpretations produced by the CLIP-SMU models.

Performance Evaluation. As shown in Figure 6, for most attributes, the fine-tuning of the original CLIP model helps to improve performance, except for the gender, age and race related attributes. The fine-tuned CLIP-SMU-SwinFace model outperforms the CLIP-SMU-AdaFace models in all of the cases, while both of them fall behind the performance of the classification-based approaches (SwinFace

Multi-Label/Binary and AdaFace Multi-Label/Binary) that provide reference information on the amount of attribute cues that can be inferred from the face templates.

Several findings can be made from the presented results:

- The AdaFace-based templates lead to the lowest classification performance across all considered attributes. Even with the most capable multi-label and multi-classifier approaches, the performance is commonly lower than that of the weakest SwinFace based approach, i.e., CLIP-SMU-SwinFace. This suggests that the AdaFace model abstracts away attribute information during the training process and leads to templates that should (to some extent) be robust to variations in these attributes.
- The template generation procedure of the SwinFace model differs significantly from that of AdaFace, as most of the attributes can be recognized very effectively us-



| | GROUND TRUTH | CLIP PRETRAINED | CLIP FINE-TUNED | CLIP - ADFACE | CLIP - SWINFACE |
|---|---|---|---|---|---|
|  | Man. Old. Shiny Skin. Gray Hair. Square Face. Double Chin. High Cheekbones. Chubby. Fully Visible Forehead. Brown Eyes. Bags Under Eyes. Bushy Eyebrows. Big Nose. Necktie. Eyeglasses. | Person. Middle age. Black. Receding Hairline. Oval Face. Square Face. Round Face. Chubby. Obstructed Forehead. Fully Visible Forehead. Mouth Closed. Smiling. Earrings. Eyeglasses. | Man. Old. Asian. Rosy Cheeks. Shiny Skin. Oval Face. Square Face. Double Chin. High Cheekbones. Obstructed Forehead. Brown Eyes. Bags Under Eyes. Smiling. Big Lips. Big Nose. Earrings. Necktie. Eyeglasses. | Man. Young. Black. Rosy Cheeks. Shiny Skin. | Man. Middle age. White. Shiny Skin. Black Hair. Square Face. Chubby. Bags Under Eyes. Big Nose. Heavy Makeup. Necktie. Lipstick. Eyeglasses. |
|  | Woman. White. Brown Hair. Oval Face. High Cheekbones. Arched Eyebrows. Smiling. Pointy Nose. Heavy Makeup. Earrings. Lipstick. Eyeglasses. | Woman. Middle age. White. Bald. Brown Hair. 5 o Clock Shadow. Oval Face. Square Face. Round Face. High Cheekbones. Obstructed Forehead. Fully Visible Forehead. Brown Eyes. Bushy Eyebrows. Arched Eyebrows. Mouth Closed. Smiling. Heavy Makeup. Earrings. Lipstick. Eyeglasses. | Woman. Old. White. Rosy Cheeks. Shiny Skin. Brown Hair. Mustache. Goatee. Oval Face. High Cheekbones. Chubby. Smiling. Pointy Nose. Heavy Makeup. Lipstick. Eyeglasses. | Man. Middle age. White. | Woman. Old. White. Rosy Cheeks. Brown Hair. Oval Face. High Cheekbones. Fully Visible Forehead. Bags Under Eyes. Smiling. Lipstick. Eyeglasses. |

Figure 7. **Qualitative representation of captions, generated with each of the models.** The green colored attributes are the true positives, i.e., attributes, which the model correctly predicted, while the attributes colored with red are false positives, i.e., attributes, for which it is incorrectly assumed that are present in the extracted face templates.

ing the reference classification-based approaches. Despite the presence of different attributes, the model still ensures competitive face recognition performance, as evidenced by the state-of-the-art results on the RAF-DB and CLAP2015 datasets, as reported in [20]. This observation points to the fact that the SwinFace model successfully exploits information from various tasks (face recognition, attribute recognition, expression recognition, etc.) when learning the model.

- The CLIP-SMU models in general perform weaker than the reference classification-based approaches, but still produce valid results in most cases, with CLIP-SMU-SwinFace outperforming CLIP-SMU-AdaFace results in most cases. This suggests that the natural language description of the SwinFace-based model are more reliable than those of the AdaFace model.

Qualitative Evaluation. Figure 7 shows a qualitative analysis of the generated captions using each of the CLIP-SMU models. Note that we show only binary attributes for easier interpretation, while the models in fact generate descriptions of the form “A photo of a [attributes A] face with [attributes B]”, where certain attributes would be placed before or after the word “face” depending on grammar. As illustrated, when using the pretrained, off-the-shelf CLIP model in a zero-shot fashion for generating attribute descriptions, almost all attributes increase the log likelihood of the predefined captions, which in turn leads to a high number of false positives, as shown in red in Figure 7. This suggests that the model struggles with interpreting the information content of the templates and “hallucinates” some of the attributes. Conversely, CLIP-SMU-AdaFace has the opposite problem and produces captions with fewer

attributes, leading to a larger number of false negatives, i.e., attributes that are present in the input image but are not included in the interpretations. It should be noted though that this behavior is expected and again points to the fact that AdaFace templates abstract away much of the attribute information from the input face images. The fine-tuned CLIP-SMU and CLIP-SMU-SwinFace fall somewhere in between and generate image interpretations that are true to the actual face attributes, but also miss some here and there.

6. Conclusion

In this paper, we presented a novel CLIP-SMU approach to explain the internal mechanisms of face recognition models by interpreting the information encoded in the extracted face templates through natural language descriptions. We implemented different versions of CLIP-SMU and explored its zero-shot capabilities, as well the feasibility of fine-tuned CLIP-SMU models. We evaluated the approaches with templates produced by two state-of-the-art face recognition models, i.e., AdaFace and SwinFace, and showed that the models are able to interpret the encoded information to varying extents – depending on the model considered.

Acknowledgments

The presented research was funded in parts by the ARIS project J2-50069 (MIXBAI), the ARIS research programme P2-0250, and the EC ENFIELD Project. The “Detecting and Explaining AI Using Language-Image Contrastive Insights” project has received funding from the European Union, via the oc1-2024-TES-01 issued and implemented by the ENFIELD project, under the grant agreement No 101120657.

References

- [1] Amina Adadi and Mohammed Berrada. Peeking inside the black-box: A survey on explainable artificial intelligence (xai). *IEEE Access*, 6:52138–52160, 2018. 2
- [2] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bénéttot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information fusion*, 58:82–115, 2020. 2
- [3] Leonard Bereska and Stratis Gavves. Mechanistic interpretability for AI safety - a review. *Transactions on Machine Learning Research*, 2024. Survey Certification, Expert Certification. 3
- [4] Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. In *The Eleventh International Conference on Learning Representations*, 2024. 3
- [5] Qiong Cao, Andrew Zisserman, et al. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 67–74. IEEE, 2018. 5
- [6] Aditya Chattopadhyay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE winter conference on applications of computer vision (WACV)*, pages 839–847. IEEE, 2018. 2
- [7] Arthur Conmy, Augustine Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. *Advances in Neural Information Processing Systems*, 36:16318–16352, 2023. 3
- [8] Benjamin Fresz, Elena Dubovitskaya, Danilo Brajovic, Marco F Huber, and Christian Horz. How should ai decisions be explained? requirements for explanations from the perspective of european law. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pages 438–450, 2024. 1
- [9] Ruigang Fu, Qingyong Hu, Xiaohu Dong, Yulan Guo, Yinghui Gao, and Biao Li. Axiom-based grad-cam: Towards accurate visualization and explanation of cnns. *British Machine Vision Conference (BMVC Oral)*, 2020. 2
- [10] Peng-Tao Jiang, Chang-Bin Zhang, Qibin Hou, Ming-Ming Cheng, and Yunchao Wei. Layercam: Exploring hierarchical class activation maps for localization. *IEEE Transactions on Image Processing*, 30:5875–5888, 2021. 2
- [11] Minchul Kim, Anil K Jain, and Xiaoming Liu. Adaface: Quality adaptive margin for face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18750–18759, 2022. 2, 3
- [12] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022. 3
- [13] Yuhang Lu, Zewei Xu, and Touradj Ebrahimi. Towards a comprehensive visual saliency explanation framework for ai-based face recognition systems. *arXiv preprint arXiv:2407.05983*, 2024. 2
- [14] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017. 2
- [15] Purushothaman Natarajan and Athira Nambiar. Vale: A multimodal visual and language explanation framework for image classifiers using explainable ai and language models. *arXiv preprint arXiv:2408.12808*, 2024. 3
- [16] Pedro C. Neto, Tiago Gonçalves, João Ribeiro Pinto, Wilson Silva, Ana F. Sequeira, Arun Ross, and Jaime S. Cardoso. Causality-Inspired Taxonomy for Explainable Artificial Intelligence. *arXiv e-prints*, page arXiv:2208.09500, Aug. 2022. 3
- [17] Truong Thanh Hung Nguyen, Tobias Clement, Phuc Truong Loc Nguyen, Nils Kemmerzell, Van Binh Truong, Vo Thanh Khang Nguyen, Mohamed Abdelaal, and Hung Cao. Langxai: Integrating large vision models for generating textual explanations to enhance explainability in visual perception tasks. *arXiv e-prints*, pages arXiv–2402, 2024. 3
- [18] Vedant Palit, Rohan Pandey, Aryaman Arora, and Paul Pu Liang. Towards vision-language mechanistic interpretability: A causal tracing tool for blip. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2856–2861, 2023. 3
- [19] P Jonathon Phillips and Mark Przybocki. Four principles of explainable ai as applied to biometrics and facial forensic algorithms. Technical report, National Institute of Standards and Technology Interagency or Internal Report 8312 Natl. Inst. Stand. Technol. Interag. Intern. Rep. 8312, 43 pages (September 2021), 2021. 2
- [20] Lixiong Qin, Mei Wang, Chao Deng, Ke Wang, Xi Chen, Jiani Hu, and Weihong Deng. Swinface: a multi-task transformer for face recognition, expression recognition, age estimation and attribute estimation. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023. 2, 3, 8
- [21] Alec Radford, Ilya Sutskever, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3
- [22] Ankit Rajpal, Khushwant Sehra, Rashika Bagri, and Pooja Sikka. Xai-fr: explainable ai-based face recognition using deep neural networks. *Wireless Personal Communications*, 129(1):663–680, 2023. 2
- [23] Sylvestre-Alvise Rebuffi, Ruth Fong, Xu Ji, and Andrea Vedaldi. There and back again: Revisiting backpropagation saliency methods. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8839–8848, 2020. 2
- [24] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "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*, pages 1135–1144, 2016. 2
- [25] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra.

- Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. [2](#)
- [26] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *CoRR*, abs/1312.6034, 2013. [2](#)
- [27] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. In *ICML, 2017 Workshop on Visualization for Deep Learning*, 2017. [2](#)
- [28] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3319–3328. PMLR, 06–11 Aug 2017. [2](#)
- [29] Philipp Terhörst, Arjan Kuijper, et al. Maad-face: A massively annotated attribute dataset for face images. *IEEE Transactions on Information Forensics and Security*, 16:3942–3957, 2021. [5](#)
- [30] Jonathan R Williford, Brandon B May, and Jeffrey Byrne. Explainable face recognition. In *European conference on computer vision*, pages 248–263. Springer, 2020. [2](#)
- [31] Alexander J Wulf and Ognyan Seizov. “please understand we cannot provide further information”: evaluating content and transparency of gdpr-mandated ai disclosures. *AI & SOCIETY*, 39(1):235–256, 2024. [1](#)
- [32] Bangjie Yin, Luan Tran, Haoxiang Li, Xiaohui Shen, and Xiaoming Liu. Towards interpretable face recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9348–9357, 2019. [2](#)
- [33] Hanwei Zhang, Felipe Torres, Ronan Sicre, Yannis Avrithis, and Stephane Ayache. Opti-cam: Optimizing saliency maps for interpretability. *Computer Vision and Image Understanding*, page 104101, 2024. [2](#)
- [34] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters*, 23(10):1499–1503, 2016. [6](#)
- [35] Roland S Zimmermann, Thomas Klein, and Wieland Brendel. Scale alone does not improve mechanistic interpretability in vision models. *Advances in Neural Information Processing Systems*, 36, 2024. [3](#)