Explainable Face Recognition based on Accurate Facial Compositions

Haoran Jiang[1] Dan Zeng[1*]
[1]Shanghai University
dzeng@shu.edu.cn, jianghaoran@shu.edu.cn

Abstract

With impressive advances made in face recognition, the explainability has attracted more and more attentions in the community, which delves into traceable and well-founded clues behind the identifications in addition to the confidence scores. However, the current Explainable Face Recognition (XFR) methods are difficult to balance the explainability and the recognition performance. In this paper, we propose a framework based on Accurate Facial Compositions, namely AFC-XFR. The framework consists of three modules: the Backbone for feature extraction, the Local Feature Refine Module (LFRM) for semantic feature refining, and the Self-Attention based Reconstruction Module (SARM) for serialized feature interaction. Fifteen semantic features, which are accurately captured from local facial components via the proposed acquisition scheme, are conveyed in the latter two modules. Moreover, the LFRM allows us to verify three significant insights experimentally, obtaining the explainability from the perspective of model decisions. Inspired by the insight "Facial features are processed holistically", the SARM's internal feature interaction mechanism facilitates performance increase. Extensive experiments on varying loss functions and network architectures accomplish consistent advances on evaluation benchmarks.

1. Introduction

Deep convolutional neural networks (DCNNs) have dominated the mainstream methods of computer vision tasks for remarkable performance that can approach, even surpass the human level. However, lacking a thorough understanding of DCNNs not only limits the adaptability range especially in crucial fields, but also makes it possible to attack or deceive maliciously, *i.e.*, adversarial attacks [12, 5]. To mitigate such a dilemma, the Explainable AI (XAI) [28] systems and theories are springing up like mushrooms. DARPA [13] defines the XAI as endeavoring to create AI systems whose *learned models* and *decisions* can be understood and trusted by end-users. Here, we pay more attention to the explainability for the latter (decisions).

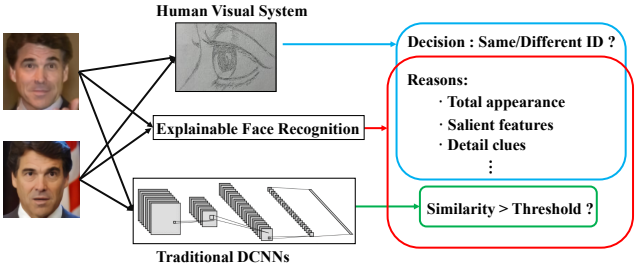


Figure 1. The differences and relevances among human visual systems, explainable face recognition, and traditional DCNNs.

There exists a similar situation in the field of face recognition. Currently, deep face recognition [24, 31] has gradually substituted the position of traditional face recognition in industry and academia due to its better generalization and robustness. However, explaining the deep face model itself and corresponding decisions remains unsolved, which pushes the progress of the explainable face recognition (XFR) [37]. To the best of our knowledge, depending on before or after the training stage, XFR methods can be mainly divided into two categories, searching for clues influencing recognizing and building inherently explainable models. The former can be regarded as a post-hoc procedure for the trained model. Most methods tend to explore learned products with the aid of network attention [45]. These products can be filter parameters, activation maps, classified weights, and output scores, *etc.*, which are followed by the analysis of the relations between them with input faces in static or dynamic patterns. The latter category keeps the network pipeline intact and guides the learning process to be explainable. After re-thinking the existing methods, we find some limitations that deserved to be solved. Although the former can help visualize the influence degree of different facial regions for identification, we are not satisfied with the rough positions but look forward to the subtler facial components. Moreover, splitting the face into components, such as mouth, eyes, brows, *etc.*, is

*Dan Zeng is the corresponding author.

consistent with human cognition and descriptive habits. As for the latter, most methods aim to design loss functions or network architectures [41] which may limit practical applicability. Therefore, our goal is to design a universal framework that can be readily plugged into existing networks and bring steady performance improvements.

In this paper, we propose an AFC-XFR framework which consists of three core modules, *i.e.* Backbone Network, Local Feature Refine Module (LFRM), and Self-Attention based Reconstruction Module (SARM). Considering that faces belonging to the same person share the same properties of local mappings and global associations because of inherent identity information, we randomly sample and generate positive pairs with the same subject as the input of the framework. Through the Backbone pretrained on conventional face datasets, we obtain feature maps from one middle layer and **Instance-based Representations** from the last FC layer. To divide and obtain local semantic features from feature maps, fine and robust component masks are demanded. Therefore, we propose an acquisition scheme that benefits from the strength and potential of 3D reconstruction and face parsing techniques. Here, we obtain fifteen facial semantic features which correspond to facial components, *i.e.*, *mouth, eyes, brows, nose, cheeks, forehead, nosewings, jaw, hair, regions between nose and mouth, and regions between brows*. In addition to these fifteen feature maps, we also retain the origin feature map to maintain global information. Then we feed these sixteen feature maps into the LFRM to refine and obtain **Component-based Representations**. Then we concatenate them, obtaining **Fused Representations**. **Final Representations** are obtained by concatenating **Fused Representations** with **Instance-based Representations**. Finally, to meet the input requirements of SARM, we serialize **Component-based Representations** to obtain two embedding sequences and achieve the reconstructing task among two input embedding sequences and one from the Transformer decoder. Moreover, for the 3D reconstruction, the core of the proposed acquisition scheme, we re-segment the vertexes of 3D Basel Face Model (BFM) [3], and perform the acquisition scheme on all used training and test sets to obtain their "-AFC" versions, which will be released to facilitate relevant studies.

Based on the proposed method, we can verify **three insights** experimentally about face recognition from [30] which are as follows: (1) Facial features are processed holistically; (2) Among the different facial components, eyebrows (eyes and brows) are the most important for recognition; (3) When suffering degradation, visual systems are more robust to familiar faces compared with unfamiliar faces. Although the aforementioned insights have been proved experimentally or empirically in human visual systems, we firstly explore them in deep face models. In summary, the paper includes the following contributions:

- To explore deep face recognition from the perspective of local semantics, we propose a novel composition-based framework called AFC-XFR with the self-attention style architecture for not only explainability but also better recognition capacity.
- To construct suitable samples for training and test, we design a novel facial component acquisition scheme, by taking full advantage of the strengths of 3D face reconstruction and face parsing. Besides, we will release the BFM vertexes re-segmentation and "-AFC" versions of some common face datasets to facilitate the latter researchers.
- Some insights of face recognition systems, well-known but lacking solid theory support, are verified experimentally, which can be treated as remarkable advances in the explainability from the perspective of model decisions. Moreover, AFC-XFR can be readily plugged into existing models, and comprehensive experiments show significant performance improvement on various architectures and loss functions.

2. Related Work

2.1. Explainable Face Recognition

The XFR researches can be grouped into two categories: one refers to the post-hoc procedure relying on elaborate perturbation on pipelines, then visualizing the impacts served as the basis for explainable insights. [27] first introduces visual psychophysics into the field of face recognition and provide metrics for output responses resulting from perturbations of input stimuli. [6] evaluates six saliency map methods by proposed metric "hiding game". In addition, similar works include direct visualization of the filters [43], deconvolutional networks to reconstruct inputs from different layers [44], gradient-based methods to generate novel inputs that maximize certain neurons [22], etc. Though plausible cause and effect results can be obtained, these heuristic methods may fail in providing a solid theoretical basis. [37] defines XFR as "why face matching system matches faces" and provide a quantitative metric "inpainting game" for objectively comparing XFR systems. The nature of these approaches is producing coarse localization maps highlighting the important regions in the image for making decisions. [42] and [49] all focus on the special scene, similar-looking face recognition, aiming at improving human visual accuracy with the aid of networks. [39] visualizes the features of shape and texture that underlie subject identity decisions. On the whole, these methods are either achieving visualizations on rough and ambiguous locations or drawing conclusions by means of heuris-

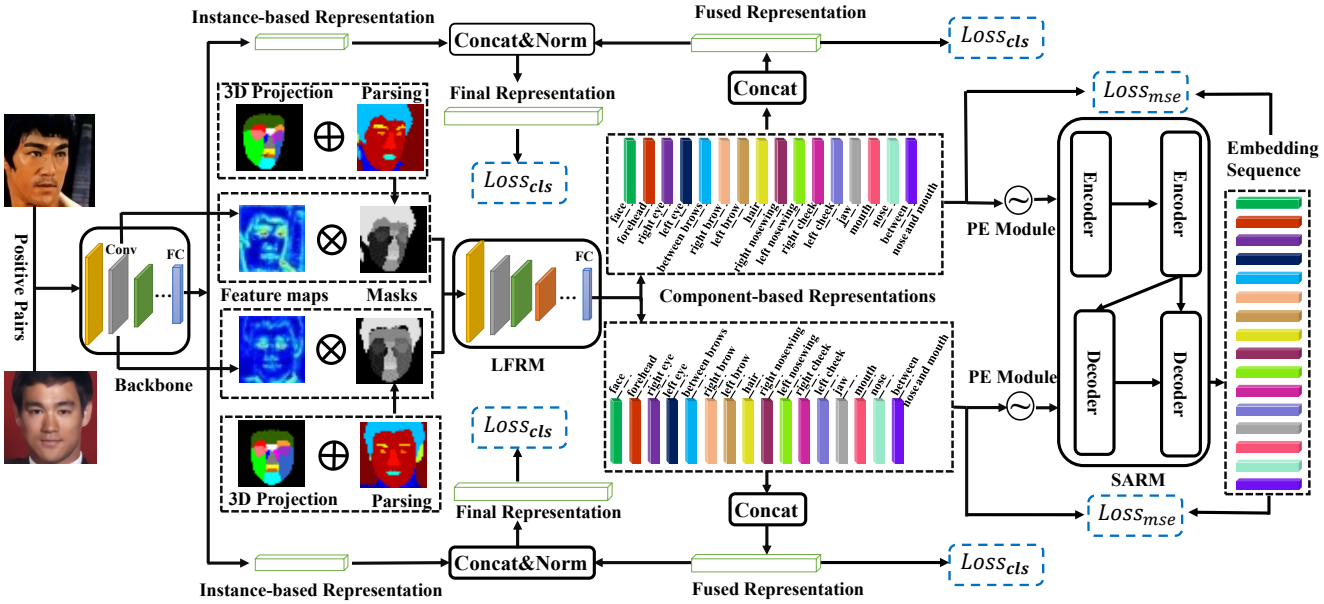


Figure 2. Overview of our method. $\oplus$ denotes covering parsing masks on reconstruction masks to get final masks, and $\otimes$ denotes multiplying final mask and feature maps to get local features. The overview of our AFC-XFR are as follows: (1) Using the classic ResNet-like Backbone to extract **Instance-based Representations**, and feature maps from one Conv layer of the Backbone; (2) Loading masks to get fifteen bounding-boxes for each component, then refining semantic features via the LFRM (including six Convs, one ROI pooling, and one FC), finally fusing these **Component-based Representations** with **Instance-based Representations** as the **Final Representations**; (3) Performing representation sequence reconstruction of positive face pairs based on the SARM.

tics. Considering such issues, we make the finer partition and directly verify some insights from [30] experimentally.

The other category is leading the representation learning to be explainable during the training stage. In deep face recognition, inspired by AnchorNet [23], [41] define and achieve explainability as each dimension of the representation activates on consistent semantic face structure without sacrificing accuracy. However, this approach needs the design of siamese architectures and complicated part-based occlusions, limiting practical applicability. By contrast, our approach directly brings steady performance improvements based on existing architectures.

2.2. Visual Tasks with Transformer Architecture

Transformer [32] architecture based attention mechanism has become the de-facto standard for natural language processing (NLP) tasks. Due to its versatile and powerful relation modeling capability for sequences, recent researches attempt to combine its core self-attention mechanism with dominant convolution networks [34, 2], or directly replacing convolution with self-attention [26]. Meanwhile, the applications of Transformer-like structures have been extended to vision tasks [38, 25, 7, 4, 51, 9]. e.g. DETR [4] and its modified version Deformable DETR [51] introduce Transformer architectures into object detection and achieve comparable performance with the bipartite matching and parallel decoding.

Our approach is particularly similar to ViT [9] used in image recognition, and its nearly direct application in face recognition [50]. They all split an image into patches and feed linear embeddings of these patches into a Transformer encoder. What we have in common is that explicitly modeling all pairwise interactions between elements in a sequence. However, the differences between us include two aspects: one is that our SARM consists complete encoder and decoder architectures, not just encoders. The more important one is those feature embeddings extracted from precise facial components are treated the same way as tokens in NLP applications, not simply image patches. As the first attempt to apply Transformer into deep face recognition, we hope this method inspires future researchers.

3. Proposed Method

3.1. Facial Components Acquisition Scheme

The proposed facial components acquisition scheme realizes the construction of the datasets for training and test. At first, considering the demand for performing fine segmentation on the face, a natural idea, utilizing the face parsing technique, the counterpart of image segmentation, hits us. Yet, the face parsing mainly focuses on major components like mouth, eyes, nose, and hair, or global components like skin, not fine components, which may not entirely satisfy our demands. Therefore, we reflect that with the aid of

mature face detection technique, the fine partition may rely on detected facial landmarks manually. Yet, when making attempts, we encounter two difficulties: (1) After locating the facial landmarks, we need to connect landmarks to form enclosed areas, which do not match the fifteen facial components we have defined in general. (2) Faces captured in the wild inevitably encounter the cases of pose variations and occlusions, leading to the loss or shift of landmarks and invalidating the attempt.

Thus, we re-think the scheme demand, in short, generating the position mask corresponding completely to the original face or feature map. The mask consists of values from 0 to 15, where 1 to 15 represent the fifteen facial components and 0 corresponds to the non-face background. The pixel-level degree precision attracts us to consider the 3D mesh. The 3D mesh can be utilized to depict the face structure spatially, whose basic units vertexes correspond to facial position points finely. So the one-to-one correlations between pixels of the 2D plane and vertexes of the 3D mesh can be obtained via the accurate projection process. Moreover, the cutting-edge algorithms can reconstruct frontal faces under extreme pose variations and occlusions. As long as pre-defining the attribution for each vertex, called **Re-segmentation**, we can obtain precise facial segmentation masks by inheriting such attribution during the round-trip process. The round-trip process means performing reconstruction on originals, then projecting back to the 2D plane while keeping the facial landmarks aligned. In the next paragraph, we will elaborate the details of **Re-segmentation**.

Re-segmentation on BFM. We choose the BFM database [3] to fulfil **Re-segmentation** due to its availability and detailed documentation. The geometry of the BFM consists of 53,490 vertexes connected by 160,470 triangles. As illustrated at the top of Fig. 3(a), officials provide the segmentation mask, which divides heads into four rough parts, *i.e.*, eyebrows, nose, mouth, and others, distinguished by colors. Apparently, precise re-segmentation must be obtained manually. Based on the documentation that illustrates the symmetry of vertexes about the perpendicular bisector of the head, we segment vertexes of one side face and directly acquiring the side vertexes, guaranteeing the unity of two side faces. There have three points deserved to be emphasized: (1) Before **Re-segmentation**, we highlight vertexes that correspond to 68 facial landmarks on 2D plane, which can help determine frontal face boundary to exclude ears and neck, meanwhile ensuring accurate segmentation. (2) For eyes and brows, we deliberately enlarge their areas. Take the eye as an example, the eye orbit and the eye bag also carry discriminative features, as well as the eyeball, does. Thus, we ensure they are defined. (3) To achieve complete segmentation on the frontal face, we add two components. They are the regions between brows, the regions between nose and mouth. As illustrated at the bottom of Fig. 3(a), fourteen components are presented with different colors.

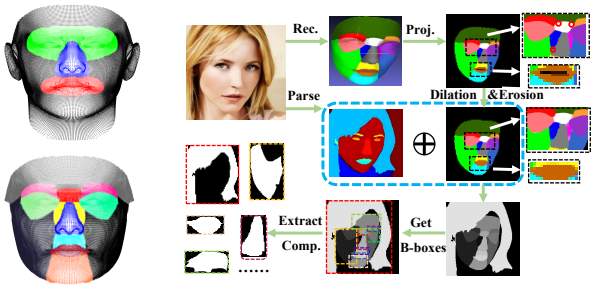
Component Masks and Bounding-boxes Extracting Done the above **Re-segmentation**, by projecting reconstruction results back to 2D plane while keeping 68 landmarks aligned can we obtain masks matching original faces. We refer to the projected mask as the reconstruction mask, denoted by m_{rec}. m_{rec} has pixel values from 0 to 14. These values correspond to 14 components and one non-face background. Now, we need to define the i-th component mask as m^i_{rec} where pixels belonging to the ith component are set to 1 while other pixel are all set to 0. The definition of m^i_{rec} can be formulated as:

$$m^i_{rec} = \begin{cases} 1, & m_{rec} = i, i = 0, 1, 2, \dots 14 \\ 0, & others \end{cases} \quad (1)$$

There exists some hollows in m_{rec}, as obviously shown in Fig. 3(b). The reason is that 3D reconstruction will inevitably cause hollows and become more severe in the projected 2D mask. So we adopt two ordered morphological operations, **Dilation & Erosion**, to eliminate them as much as possible. When projecting to 2D plane, the 3D reconstruction will lead to ignoring the existence of pose variations and occlusions due to its robustness, which causes hallucination in m_{rec}. To ensure the authenticity of the final mask, we adopt the face parsing that judges occlusions as backgrounds to obtain the parsing mask m_{par}. In addition to mitigating the effects of occlusion, m_{par} can not only introduce the hair as our fifteenth component but also take the inner of the mouth into account to handle samples opening the mouth which m_{rec} can not handle (shown in Fig. 3(b)). Concretely, m_{par} pixel values include 0, 1 and 2, where 2 refers to the hair component, 1 corresponds to occlusions, and 0 represents the non-face background. By overlaying m_{par} on m_{rec}, we can obtain the ultimate mask m which takes occlusions into account and derives fifteen component masks. The definition of m is formulated as:

$$m = \begin{cases} 0, & m_{par} \neq 2 \text{ and } m_{rec} = 0 \\ i, & m_{par} \neq 2 \text{ and } m_{rec} = i, i = 1, \cdots 14 \\ 15, & m_{Par} = 2 \end{cases} \quad (2)$$

The definition of m_i is similar to m^i_{rec} (Eqn. 1), except that i ranges from 0 to 15. By multiplying each m_i on feature maps, we obtain all 15 local feature maps. Not original faces but feature maps extracted from the intermediate layer of backbone networks are utilized because feature maps own refined high-level features. We will refer to the feature map as P. So we resize m to be the same as that of P, then multiplying all m_i by P so as to acquire fifteen local feature maps, called p_i where i=1,2,$\cdots$,15. Obviously, p_i may have too many zero values, which will cause redundant information and decrease the distinctiveness. So we compute the minimal outer rectangle for each component,



(a) (b)

Figure 3. (a) The BFM vertex segmentation results, provided by official authorities (top) and by re-segmentation (bottom). We remove the neck and ear parts, and make a finer division for the frontal face to extend to fourteen parts. (b) The complete components acquisition scheme.

i.e. the bounding-box b_i for each p_i. Finally we crop areas on p_i within b_i, obtaining the final local features $\hat{p}_i$, which can be formulated as follow:

$$\hat{p}_i = b_i \cap (m_i * P), , i = 1, 2, \cdots, 15, \tag{3}$$

where $\cap$ refers to cropping areas on p_i within b_i, $*$ refers to Hadamard product. Concrete implementation is illustrated in Fig. 3(b). Some samples obtained via the acquisition scheme are shown in Fig. 4, which validates that our scheme can not only do well in frontal easy samples but also be robust to pose variations and occlusions.

3.2. XFR based on Accurate Facial Compositions

In this section, we will introduce Backbone and LFRM in detail. The overview of AFC-XFR is elaborated in Fig. 2. Suppose that the face dataset $M = (x_1, x_2, \cdots, x_N)$ consists of N samples, we randomly sample to form positive pairs, written as x^a and x^b. At first, we pre-train the Backbone on M, supervised by the Arc-softmax [8] loss, denoted by $\mathcal{L}_{cls}$. During the training stage, we fix the parameter of Backbone and obtain P^a, P^b from its middle Conv layer, V^a_{ins} and V^b_{ins} from the last FC layer. P^a and P^b represent the intermediate feature map for x^a and x^b. V^a_{ins} and V^b_{ins} are **Instance-based representations**. Final local features $\hat{p}^a_i$ and $\hat{p}^b_i$ are obtained from P^a and P^b via the facial components acquisition scheme, according to Eqn.3. Then, $\hat{p}^a_i$ and $\hat{p}^b_i$ are fed into LFRM including six Conv layers and one ROI pooling layer [11] and one FC layer. Concretely, we extract and refine the features by Conv layers, then unify all spatial dimensions into 5×5 by ROI pooling, afterwards map them into new feature space by FC layer. Then we obtain fifteen **Component-based Representations** v^a_i and V^b_i. In addition to them, we also feed P^a or P^b directly and get v^a_{16} or v^b_{16} as the full face representation to provide global information. Then, we concatenate v^a_1,v^a_2,$\cdots$,v^a_{16} to obtain **Fused representation**V^a_{fuse} (similarity for V^b_{fuse}).

Finally, we do two things for training, one is use $\mathcal{L}_{cls}$ to supervise V^a_{fuse} and V^b_{fuse}; another thing is concatenate V^a_{fuse} and V^a_{ins} to obtain V^a_{fin} (similarity for V^b_{fin}), called **Final Representations** .Then we perform L_2 norm on V^a_{fin} and V^b_{fin}, then supervised by $\mathcal{L}_{cls}$. To sum up, $\mathcal{L}_{cls}$ supervision on **Fused Representations** and **Final Representations** completes the training for LFRM. In the next section, we will introduce SARM.

Figure 4. Several samples of component masks. In addition to common frontal face cases, the scheme can perform well when tackling non-frontal and occluded cases. For example, the upper right sample represents the case of the profile face and the bottom right represent the case of occlusion (hat).

3.3. Self-Attention based Reconstruction Module

The final part of AFC-XFR is SARM. The modification compared with the conventional Transformer mainly consists in two aspects: (1) Reducing the repeating times of Encoder and Decoder basic units from 6 to 2; (2) Removing classification heads composed of linear layers and softmax in order to suit the demand for feasibility. SARM plays the role of taking advantage of the explainable insights and thereby helps for learning better component-based features. Its self-attention is the key to adapt the explainability. As we all know, translation between two different corpora is not a simply one-to-one mapping but a comprehensive process that should balance the local mapping between words and global associations between sentences and even more. We argue this advantage benefits from the linguistic fundamental cognition. In other words, the self-attention mechanism can compute and adjust the importance of each word in a weighting and summing way. What makes the self-attention effective is that sentences expressed in two different languages share the same meaning.

These experience and conclusions enlighten us to adopt the Transform architecture in the term of face recognition. Similarly, we deem that different faces of the same ID share consistent identity information despite comparable representation differences for positions, occlusions, illuminations, etc. Specifically, we imitate the setting of translating

sentences, and decompose faces of positive pairs into two sets of local features in the form of embeddings. Here, two sentences are analogous to two faces, and the word embedding sequences are analogous to the facial local feature embedding sequences. The improved recognition performance verifies the rationality of analogy. Meanwhile, it also proves the insight (*facial features are processed holistically*) in reverse.

To be specific, we serialize v_i^a and v_i^b to obtain the embedding sequence E^a,E^b. They are fed into the position encoding (PE) module, Encoder and Decoder respectively to obtain the embedding sequence E^{out}. The loss function is based on the Mean Square Error (MSE) which is formulated by:

$$\mathcal{L}_{MSE} = \sum_{i=1}^{D} \left[\left(E^a - E^{out}\right)^2 + \left(E^b - E^{out}\right)^2 \right], \quad (4)$$

where D is the elements number of embedding sequence. $\mathcal{L}_{MSE}$ and two aforementioned $\mathcal{L}_{MSE}$ form the complete loss function for training AFC-XFR.

4. Experiments

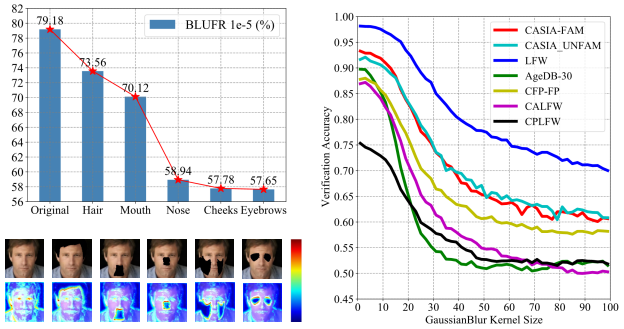
This section is structured as follows. Section 4.1 introduces the datasets and experimental settings. Section 4.2 explores some experimental insights of face recognition and makes detailed analysis. Section 4.3 includes the ablation study which validates the consistent performance improvements among five softmax-based loss functions, meanwhile demonstrates a similar improvement can also be obtained when using a deeper network architecture and a bigger dataset.

4.1. Datasets and Experimental Settings

Training data. We use two public datasets to train our models. In ablation study, we use cleaned CASIA-WebFace [40] which consists of 9879 IDs and 400,943 images totally. In performance experiments, we utilize MS1M-v1c [1] (cleaned version of MS-Celeb-1M [14]) including 72,778 IDs and 3,126,881 images.

Test data. In order to make a thorough evaluation, we adopt LFW [17], BLUFR [19], AgeDB-30 [21], CALFW [48], CFP-FP [29], CPLFW [47], MegaFace [18] datasets. LFW is the most commonly used benchmark for wild face recognition. BLUFR is dedicated to the evaluation with a focus on low false accept rates (FAR), and we report the verification rate at the lowest FAR (1e-5) on BLUFR. AgeDB-30 and CALFW show large age gap sample distributions. CFP-FP and CPLFW focus on cross-pose variants face verification. MegaFace also evaluates the performance of large-scale face recognition with millions of distractors.

Preprocessing. All training and test images are detected by employing FaceBoxes [46]. Then, we align and crop faces to 144×144 by five facial landmarks [10].



(a) (b)

Figure 5. At the bottom of (a), we occlude hair, mouth, nose, cheeks, and eyebrows on the original faces. The first row is occluded images and the second is the corresponding feature responses. Blue indicates the low response value. The top of (a) shows the BLUFR verification rate at FAR = 1e-5 when occluding different components. (b) The curves are verification rates with the increase of kernel size of GaussianBlur on different evaluation sets.

CNN Architecture. To balance performance and time-consuming, we adopt SE-ResNet-18 [16] and deepen it to keep high resolution of feature maps in deeper layers to perform ablation study and interpretability analysis. In performance experiments, we want to validate consistent in a deeper network, we use ResNet-50 [15] architecture.

Training and Evaluation. Four NVIDIA Tesla P40 GPUs are employed for training. For CASIA-WebFace, we first train the Backbone with the batch size of 128 until the convergence of loss. Then we train our framework based on the trained Backbone, where the batch size is 256 and the learning rate starts from 0.1, divided by 10 at 9,12,18 epochs and finishes at 20 iterations. On MS1M-v1c, we also train Backbone as the pre-training model. Then, based on Backbone, we train our framework with batch size 256. Meanwhile, we set the learning rate to 0.1, divide it by 10 at 6,9,12 epochs and finish at 15 epochs. We set the momentum to 0.9 and weight decay to 5e-4. In the evaluation stage, we use the aforementioned **Final Representations** as the face representations. The cosine similarity is utilized as the similarity metric. For strict and precise evaluation, all the overlapping identities between training datasets and test datasets are removed according to the overlapping list [35].

4.2. Explainability Analysis

In this section, we mainly delve into verifying the aforesaid three insights. According to the available local components, we can initially explore their significance by occlusion. As shown in Fig. 5(a), we demonstrate: (1) Without any occlusion, regions around eyebrows have relatively large response values which proves the insight that **the eyebrow plays a critical role in identification**. The dramatic drop of the BLUFR verification rate when occluding eye-

brows makes the insight more convincing. (2) When an occlusion occurs, edges of the occlusion components show large values which is reasonable. More importantly, the overall response values show a uniform decrease which suggests that "**Facial features are processed holistically**". We deem that the occlusion damages the integrity of the face structure, leading to performance decrease. However, such heuristic experiments can not draw solid explainable conclusions. Therefore, we perform three explainability verifying experiments.

Familiarity Analysis. In this experiment, we explore to verify the insight "**When suffering degradation, visual systems are more robust to familiar faces compared with unfamiliar faces**". To achieve it, we choose GaussianBlur as the degradation method and define the familiarity of samples for CNN. Specifically, we define the unfamiliar ID as its samples are removed from the training set, and the familiar ID as only two samples are removed and the remaining samples are used for training. On CASIA-WebFace, we select 3,000 familiar IDs which have more than 22 images per ID to construct a familiar-set "CASIA-FAM", and 3,000 unfamiliar IDs which have the least images among all IDs to build an unfamiliar-set "CASIA-UNFAM". Following the protocol of LFW, we develop two test protocols that contain 6,000 pairs based on "CASIA-FAM" and "CASIA-UNFAM". Then we train AFC-XFR based on the non-overlapping training set. Finally, we calculate the verification rate on LFW, AgeDB-30, CFP-FP, CALFW, CPLFW, "CASIA-FAM" and "CASIA-UNFAM" by gradually increasing the GaussianBlur kernel size.

As shown in Fig. 5(b), we can conclude: firstly, all sets have the same downswing as the degree of blurring increases until the blurring reaches a certain level. It suggests similar to human visual systems, DCNNs based face recognition also weaken due to the degradation. Secondly, except for relatively easy LFW, "CASIA-FAM" and "CASIA-UNFAM" curves are higher than other sets which are reasonable for they are subject to the same data distribution similar to the training set. Thirdly, focusing on "CASIA-FAM" and "CASIA-UNFAM", there are almost no differences between their curves, which contradicts the human visual system that has better robustness when recognizing a familiar face. The DCNN robustness to degradation does not vary much with different degrees of familiarity. The aforesaid insight may be model-dependent.

Similarity Matrix Metric. In order to delve into how each facial component influences the identification, we propose a novel metric called the "Similarity Matrix" to characterize the degree of influence. Specifically, we pick and group from the fifteen facial components to get six masks including noses, mouths, hair, eyebrows, forehead, and cheeks. we define a 7×7 matrix M whose each entry M_{ij} refers to the cosine similarity between paired final representations under one case of occlusion. For example, M_{23} represents the case of one adopting a nose mask and the other adopting a mouth mask in a pair of a sample. M has the following two properties: (1) M_{11} as the baseline standard denotes keeping original pairs without any occlusion. (2) M is the symmetric matrix. Especially the first column and row denote one keeping original and another adopting some mask in a pair.

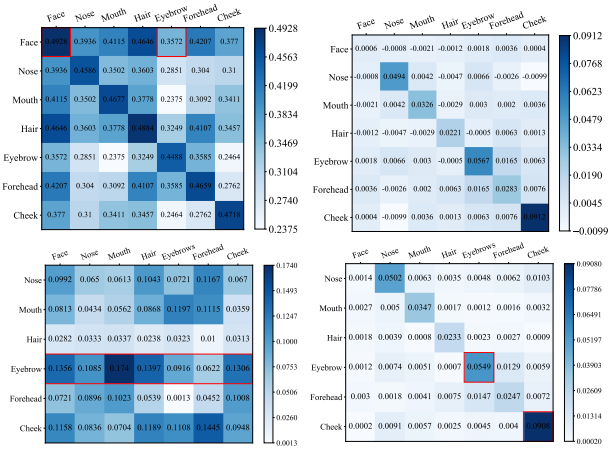
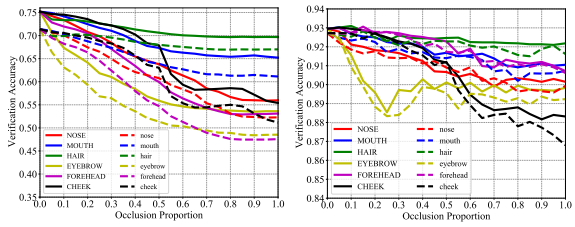


Figure 6. The similarity matrices across the facial components. From top left to bottom right are M_p,M_n,M_{p-dif}, and M_{n-dif}.

Based on CASIA-WebFace, we can obtain three M which includes M_{all} for 6000 pairs, M_p for 3000 positive pairs, and M_n for 3000 negative pairs. For example, we sum and average the M of 3000 positive pairs to obtain the M_p which is more convincing. Moreover, we define M_{dif} to reflect the similarity changes by occluding one face after fixing another of the pair. To be specific, we subtract each row from the first row and take absolute values. Then taking out the first row, we obtain a 6×7 M_{dif}. Therefore, we can also obtain $M_{all-dif}$, M_{p-dif} and M_{n-dif}. The metric results of M_p, M_n, M_{p-dif} and M_{n-dif} are shown in Fig. 6. We can conclude: (1) Focusing on the main diagonal of M_p and M_n, when masking the same components, the similarity of the positive pair decreases, while that of negative pairs is just the opposite. It is consistent with the human visual perception that masking may disturb the confirmation of positive pairs, but decrease the discriminability between negative pairs due to similar masks.

(2) Actually, the confirmation of positive pairs is more meaningful and valuable than the distinction between negative pairs in the real world. So we emphasize on M_p and M_{p-dif}. Observing the first row of M_p, M_{p15} has the biggest gap with M_{p11} which proves eyebrows play an important role in the identification. The fourth row of M_{p-dif}, referring to masking eyebrows in the context of already masking a certain component, has a bigger value than other rows, which can also explain the role of eyebrows.

(3) Now observing M_{n-dif}, except for the cheeks



(a) (b)

Figure 7. Local mask experiment. (a) displays the BLUFR accuracy, where the dashed lines denote the baseline and the solid lines denote AFC-XFR. (b) denotes verification rate on CASIA-WebFace.

$M_{n-dif45}$ which has the biggest area, the similarity of eyebrows $M_{n-dif67}$ changed the most which can also verify the importance of eyebrows.

Local Mask Experiment. We study the influence degree for recognition of these components in a more elaborate way. Concretely, we increase the mask proportion of each component and record performance. Meanwhile, we also compare our AFC-XFR with the baseline.

From Fig. 7, we can not only find AFC-XFR outperforms the baseline, but also prove of all the components except for the forehead and cheeks, the lack of eyebrow will severely affect the recognition.

4.3. Quantity Analysis

We analyze each module used in AFC-XFR and validate its effectiveness for representation learning. Table 1 shows their performance with softmax loss, and some frequently-used margin-based loss functions including Arc-softmax [8], AM-softmax [33], MV AM-softmax[36], and A-softmax [20] loss functions. In Table 1, "Org" denotes the plain training on the backbone. "AFC-XFR" denotes the ultimate training scheme. "TR-MSE" denotes removing SARM, but still conducting supervised learning between embedding sequences with MSE loss. "WO-TR" denotes removing SARM and doing nothing for the embedding sequences. The above employs the modified ResNet-18 network on CASIA-Webface-AFC. Meanwhile, to further explore the ability of AFC-XFR for deeper networks and large datasets, we adopt AFC-XFR on the MS1M-v1c-AFC dataset based on ResNet-50. "Org*" and "AFC-XFR*" denotes the corresponding plain training and ultimate training scheme.

From Table 1, we can conclude: (1) Comparing with "Org", whatever LFRM or SARM, our methods result in a progressive increase, especially in BLUFR and MegaFace benchmarks. Moreover, the ultimate training method "AFC-XFR*" achieves the best performance in most benchmarks, taking the admirable "Arc-softmax" as an example, we gain 2.84% improvements in BLUFR, and 2.48% in MegaFace Id.,2.34% in MegaFace Veri.

Table 1. Performance (%) on LFW, BLUFR, AgeDB-30, CFP-FP, CALFW, CPLFW, and MegaFace. In MegaFace, "Id." refers to face identification rank1 accuracy with 1M distractors, and "Veri." refers to face verification rate at 1e-6 FAR.

Method	LFW	BLUFR	AgeDB	CFP	CALFW	CPLFW	MegaFace	
							Id.	Veri.
softmax								
Org	97.25	59.92	85.38	86.83	83.40	72.57	57.01	59.91
TR-MSE	97.57	62.11	86.25	87.03	83.63	72.53	**61.87**	65.11
WO-TR	97.52	**62.57**	86.33	86.79	83.53	72.50	61.61	**66.74**
AFC-XFR	**97.65**	62.09	**86.45**	**87.09**	**83.78**	**72.78**	61.85	66.08
Org*	99.55	92.69	95.70	93.50	93.47	85.10	90.61	91.66
AFC-XFR*	**99.70**	**92.93**	**95.96**	**93.78**	**93.69**	**85.35**	**90.81**	**91.96**
Arc-softmax								
Org	98.68	79.13	90.67	**89.07**	88.37	77.20	78.06	82.47
TR-MSE	98.45	81.38	91.50	88.93	88.33	76.55	**80.60**	**84.87**
WO-TR	98.57	81.95	91.27	88.94	88.48	**77.43**	80.58	84.69
AFC-XFR	**98.73**	**81.97**	**91.82**	88.97	**88.78**	77.15	80.54	84.81
Org*	99.67	94.60	96.17	93.07	94.52	85.00	93.56	94.70
AFC-XFR*	**99.83**	**94.81**	**96.47**	**93.21**	**94.70**	**85.68**	**93.77**	**95.02**
AM-softmax								
Org	98.42	79.06	**91.33**	90.26	89.10	**78.33**	77.42	82.67
TR-MSE	**98.68**	82.15	91.15	**90.46**	88.98	77.93	81.09	85.14
WO-TR	98.53	81.90	91.27	90.01	88.88	77.95	81.10	**85.80**
AFC-XFR	**98.68**	**82.20**	91.28	89.94	**89.12**	78.05	**81.18**	85.14
Org*	99.60	94.78	96.82	93.14	94.50	86.17	94.36	**95.57**
AFC-XFR*	**99.78**	**95.17**	**96.95**	**93.31**	**94.63**	**86.32**	**94.47**	95.45
MV AM-softmax								
Org	98.70	79.30	90.73	**90.11**	89.08	78.55	78.06	82.47
TR-MSE	98.75	**81.59**	91.17	89.70	89.07	78.45	**81.86**	84.84
WO-TR	98.70	81.40	**91.30**	89.64	**89.40**	78.25	81.63	84.28
AFC-XFR	**98.88**	81.15	91.12	89.77	89.15	**78.62**	81.66	**84.97**
Org*	99.68	95.79	96.88	93.34	94.97	86.08	95.18	95.92
AFC-XFR*	**99.72**	**95.88**	**97.01**	**93.46**	**95.12**	86.08	**95.24**	**95.96**
A-softmax								
Org	97.73	60.94	86.48	87.53	84.32	73.53	60.11	63.71
TR-MSE	97.78	**64.56**	86.98	87.74	84.55	**73.57**	**64.91**	69.04
WO-TR	97.92	64.29	87.20	88.14	**84.80**	73.48	64.74	68.71
AFC-XFR	**97.95**	63.64	**87.35**	**88.20**	84.75	**73.57**	64.86	**69.57**
Org*	99.63	92.56	97.07	**94.63**	94.77	86.42	93.27	94.78
AFC-XFR*	**99.67**	**92.95**	**97.13**	94.56	**94.97**	**86.48**	**93.47**	**94.89**

(2) We can validate the effect of LFRM by comparing "Org" with "WO-TR". Except for the CPLFW and CFP, the "WO-TR" can obtain consistent improvement on other benchmarks.

(3) We can compare "WO-TR" and "AFC-XFR" to study the effect of SARM. Notably, "AFC-XFR" gains great advantages. The comparison indicates the self-attention mechanism can help narrow intra-class variation to obtain more discriminative features.

(4) Comparing "Org*" with "AFC-XFR*", AFC-XFR still achieves the leading accuracy. Although the saturated performance has been obtained based on deep networks, AFC-XFR still achieves consistent improvement in most of the test sets, and also the competitive results on BLUFR and MegaFace.

5. Conclusions

In this paper, we propose an XFR framework to verify three insights about face recognition experimentally. To our knowledge, it is the first time to apply the Transformer structure on the face recognition task to achieve the explainability. Meanwhile, we will release "-AFC" versions of the common face datasets to facilitate future study.

References

[1] http://trillionpairs.deepglint.com/overview.

[2] Irwan Bello, Barret Zoph, Ashish Vaswani, Jonathon Shlens, and Quoc V Le. Attention augmented convolutional networks. In *Proceedings of the IEEE ICCV*, pages 3286–3295, 2019.

[3] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 187–194, 1999.

[4] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. *arXiv:2005.12872*, 2020.

[5] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pages 39–57. IEEE, 2017.

[6] Gregory Castanon and Jeffrey Byrne. Visualizing and quantifying discriminative features for face recognition. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition*, pages 16–23. IEEE, 2018.

[7] Cheng Chi, Fangyun Wei, and Han Hu. Relationnet++: Bridging visual representations for object detection via transformer decoder. *arXiv:2010.15831*, 2020.

[8] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the CVPR*, pages 4690–4699, 2019.

[9] A. Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, M. Dehghani, Matthias Minderer, G. Heigold, S. Gelly, Jakob Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ArXiv*, abs/2010.11929, 2020.

[10] Zhen-Hua Feng, Josef Kittler, Muhammad Awais, Patrik Huber, and Xiao-Jun Wu. Wing loss for robust facial landmark localisation with convolutional neural networks. In *Proceedings of the CVPR*, pages 2235–2245, 2018.

[11] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE ICCV*, pages 1440–1448, 2015.

[12] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[13] Aha D. Gunning D. Darpa's explainable artificial intelligence (xai) program. *AI Magazine*, 40(2), 2019.

[14] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *European conference on computer vision*, pages 87–102. Springer, 2016.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the CVPR*, pages 770–778, 2016.

[16] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the CVPR*, pages 7132–7141, 2018.

[17] Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. 2008.

[18] Ira Kemelmacher-Shlizerman, Steven M Seitz, Daniel Miller, and Evan Brossard. The megaface benchmark: 1 million faces for recognition at scale. In *Proceedings of the CVPR*, pages 4873–4882, 2016.

[19] Shengcai Liao, Zhen Lei, Dong Yi, and Stan Z Li. A benchmark study of large-scale unconstrained face recognition. In *IEEE IJCB*, pages 1–8. IEEE, 2014.

[20] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. In *Proceedings of the CVPR*, pages 212–220, 2017.

[21] Stylianos Moschoglou, Athanasios Papaioannou, Christos Sagonas, Jiankang Deng, Irene Kotsia, and Stefanos Zafeiriou. Agedb: the first manually collected, in-the-wild age database. In *CVPR Workshops*, pages 51–59, 2017.

[22] Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the CVPR*, pages 427–436, 2015.

[23] David Novotny, Diane Larlus, and Andrea Vedaldi. Anchornet: A weakly supervised network to learn geometry-sensitive features for semantic matching. In *Proceedings of the CVPR*, pages 5277–5286, 2017.

[24] Omkar M Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. 2015.

[25] Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Łukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran. Image transformer. *arXiv:1802.05751*, 2018.

[26] Prajit Ramachandran, Niki Parmar, Ashish Vaswani, Irwan Bello, Anselm Levskaya, and Jonathon Shlens. Stand-alone self-attention in vision models. *arXiv:1906.05909*, 2019.

[27] Brandon RichardWebster, So Yon Kwon, Christopher Clarizio, Samuel E Anthony, and Walter J Scheirer. Visual psychophysics for making face recognition algorithms more explainable. In *Proceedings of the ECCV*, pages 252–270, 2018.

[28] Wojciech Samek, Grégoire Montavon, Andrea Vedaldi, Lars Kai Hansen, and Klaus-Robert Müller. *Explainable AI: interpreting, explaining and visualizing deep learning*, volume 11700. Springer Nature, 2019.

[29] Soumyadip Sengupta, Jun-Cheng Chen, Carlos Castillo, Vishal M Patel, Rama Chellappa, and David W Jacobs. Frontal to profile face verification in the wild. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–9. IEEE, 2016.

[30] Pawan Sinha, Benjamin Balas, Yuri Ostrovsky, and Richard Russell. Face recognition by humans: Nineteen results all computer vision researchers should know about. *Proceedings of the IEEE*, 94(11):1948–1962, 2006.

[31] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the CVPR*, pages 1701–1708, 2014.

[32] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia

Polosukhin. Attention is all you need. In *NeurIPS*, pages 5998–6008, 2017.

[33] Feng Wang, Jian Cheng, Weiyang Liu, and Haijun Liu. Additive margin softmax for face verification. *IEEE SPL*, 25(7):926–930, 2018.

[34] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the CVPR*, pages 7794–7803, 2018.

[35] Xiaobo Wang, Shuo Wang, Jun Wang, Hailin Shi, and Tao Mei. Co-mining: Deep face recognition with noisy labels. In *Proceedings of the ICCV*, pages 9358–9367, 2019.

[36] Xiaobo Wang, Shifeng Zhang, Shuo Wang, Tianyu Fu, Hailin Shi, and Tao Mei. Mis-classified vector guided softmax loss for face recognition. *arXiv:1912.00833*, 2019.

[37] Jonathan R Williford, Brandon B May, and Jeffrey Byrne. Explainable face recognition. In *European Conference on Computer Vision*, pages 248–263, 2020.

[38] Bichen Wu, Chenfeng Xu, Xiaoliang Dai, Alvin Wan, Peizhao Zhang, Masayoshi Tomizuka, Kurt Keutzer, and Peter Vajda. Visual transformers: Token-based image representation and processing for computer vision. *arXiv:2006.03677*, 2020.

[39] Tian Xu, Jiayu Zhan, Oliver GB Garrod, Philip HS Torr, Song-Chun Zhu, Robin AA Ince, and Philippe G Schyns. Deeper interpretability of deep networks. *arXiv:1811.07807*, 2018.

[40] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. Learning face representation from scratch. *arXiv:1411.7923*, 2014.

[41] Bangjie Yin, Luan Tran, Haoxiang Li, Xiaohui Shen, and Xiaoming Liu. Towards interpretable face recognition. In *Proceedings of the IEEE ICCV*, pages 9348–9357, 2019.

[42] Timothy Zee, Geeta Gali, and Ifeoma Nwogu. Enhancing human face recognition with an interpretable neural network. In *ICCV Workshops*, 2019.

[43] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *ECCV*, pages 818–833. Springer, 2014.

[44] Matthew D Zeiler, Graham W Taylor, and Rob Fergus. Adaptive deconvolutional networks for mid and high level feature learning. In *2011 ICCV*, pages 2018–2025, 2011.

[45] Jianming Zhang, Sarah Adel Bargal, Zhe Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff. Top-down neural attention by excitation backprop. *IJCV*, 126(10):1084–1102, 2018.

[46] Shifeng Zhang, Xiangyu Zhu, Zhen Lei, Hailin Shi, Xiaobo Wang, and Stan Z Li. Faceboxes: A cpu real-time face detector with high accuracy. In *2017 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–9. IEEE, 2017.

[47] Tianyue Zheng and Weihong Deng. Cross-pose lfw: A database for studying crosspose face recognition in unconstrained environments. *Beijing University of Posts and Telecommunications, Tech. Rep*, pages 18–01, 2018.

[48] Tianyue Zheng, Weihong Deng, and Jiani Hu. Cross-age lfw: A database for studying cross-age face recognition in unconstrained environments. *arXiv:1708.08197*, 2017.

[49] Yaoyao Zhong and Weihong Deng. Deep difference analysis in similar-looking face recognition. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pages 3353–3358. IEEE, 2018.

[50] Yaoyao Zhong and Weihong Deng. Face transformer for recognition. *arXiv preprint arXiv:2103.14803*, 2021.

[51] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv:2010.04159*, 2020.