

# CATFace: Cross-Attribute-Guided Transformer with Self-Attention Distillation for Low-Quality Face Recognition

Niloufar Alipour Talemi<sup>1</sup>, Hossein Kashiani<sup>1</sup>, and Nasser M. Nasrabadi<sup>1</sup>, *Fellow, IEEE*

**Abstract**—Although face recognition (FR) has achieved great success in recent years, it is still challenging to accurately recognize faces in low-quality images due to the obscured facial details. Nevertheless, it is often feasible to make predictions about specific soft biometric (SB) attributes, such as gender, and baldness even in dealing with low-quality images. In this paper, we propose a novel multi-branch neural network that leverages SB attribute information to boost the performance of FR. To this end, we propose a cross-attribute-guided transformer fusion (CATF) module that effectively captures the long-range dependencies and relationships between FR and SB feature representations. The synergy created by the reciprocal flow of information in the dual cross-attention operations of the proposed CATF module enhances the performance of FR. Furthermore, we introduce a novel self-attention distillation framework that effectively highlights crucial facial regions, such as landmarks by aligning low-quality images with those of their high-quality counterparts in the feature space. The proposed self-attention distillation regularizes our network to learn a unified quality-invariant feature representation in unconstrained environments. We conduct extensive experiments on various FR benchmarks varying in quality. Experimental results demonstrate the superiority of our FR method compared to state-of-the-art FR studies.

**Index Terms**—Face recognition, soft biometric attributes, knowledge distillation, self-attention mechanism, feature fusion.

## 1 INTRODUCTION

FACE recognition (FR) has been one of the most popular fields in computer vision due to its wide range of applications in military, public security, and daily life [1]. In the realm of FR, there has been notable progress in recent years with the emergence of advanced network architectures [2], [3]. Alongside these advancements, the field has witnessed the introduction of various designs of loss functions [4], [5], [6], [7], [7], [8], [9] which have played a significant role in enhancing FR performance. Despite all these advancements, it has still been challenging to preserve the high performance of FR methods in unconstrained environments. The majority of FR training datasets [10], [11], [12] consist of high-quality images that differ significantly from real-world environments. This becomes evident when considering images captured by surveillance cameras [13], which present challenging attributes like sensor noise, low resolution, motion blur, and turbulence effect, among others. As a result, when FR models trained on constrained datasets are applied to real-world scenarios, the models' accuracy suffers a significant drop. On the other hand, collecting a large-scale unconstrained face dataset with large variations needs manual labeling, which is time-consuming and costly to provide.

Recently, several alternative approaches have been proposed to fill the gap between the semi-constrained training datasets and unconstrained testing sets. Some of these studies involve super resolution-based techniques [14], [15], [16], which aim to reconstruct high-resolution images from low-resolution ones and then feed them to a FR model. Moreover, with the advent of Generative Adversarial Networks (GANs) [17], [18], [19] several GAN-based frontalization approaches [20], [21] have been proposed to handle faces with extreme poses. However, these approaches primarily address specific image variations, resulting in limited generalizability across diverse conditions. Another concern is that

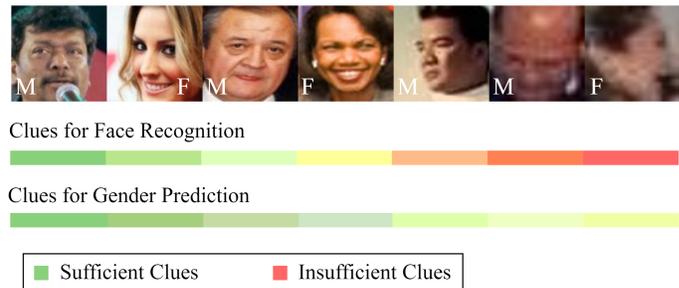


Fig. 1: Examples of face images with different degrees of degradation in various real-world FR benchmarks. In some images, the identity of the person is not easily recognizable due to the lack of some clues that are essential for FR. However, the gender of a person can still be inferred from those images. Therefore, leveraging some SB attributes like gender can enhance FR performance in challenging conditions. Note that M and F stand for male and female, respectively.

during the inference step, performing such preprocessing methods may lead to significantly high computational cost compared to the recognition network itself. Furthermore, despite the significant advancements in GAN models, preserving identity in particular for cases with extreme poses has been still a challenging problem [22], [23].

Looking at the FR task from an alternative perspective, humans inherently analyze facial attributes to discern the identity of a person. This observation can bring up the hypothesis that utilizing facial attributes can improve the performance of FR in challenging cases. Fig. 1 shows that even in the case of low-quality images, it is often feasible to predict certain soft biometric (SB) attributes like

gender, while accurately identifying the exact identity proves to be a challenging task. Therefore, the incorporation of SB information into FR can help the network to recognize the identities more accurately. Inspired by this observation, in this work, we employ facial attributes including gender, and baldness, which remain consistent across different scenarios such as varying illuminations and poses to raise the performance of FR. To this end, we propose a multi-head neural network that not only predicts SB attributes and recognizes identities simultaneously but also employs a novel cross-attribute-guided transformer fusion (CATF) module to effectively integrate feature representation of the SB information into FR features. This module initially enables the synergistic fusion of the SB and FR feature representations using dual cross-attention operations. Subsequently, it conducts feature fusion across global spatial tokens, where each channel is treated as an independent token. By regarding each channel token as an abstract representation of the entire image, the fusion process naturally captures global interactions and representations. Therefore, our proposed fusion module concentrates on the most pertinent areas within both the SB and FR feature representations and enhances the integration of distinctive facial details with SB cues.

To further leverage the advantages of using SB information as an auxiliary modality, we train our proposed multi-branch network using a novel knowledge distillation (KD) based approach. This approach enables our SB prediction branch to exhibit robustness in challenging cases, thereby improving the overall FR accuracy. The idea of utilizing KD, which is a teacher-student-based framework, originated from the observation that although crucial visual details are missing in low-quality images, in many cases humans can still roughly determine an object's regions in such images based on prior knowledge learned from previously viewed corresponding high-quality images [24]. Thus, as in low-quality faces, features from detailed parts of a face may not be captured, our KD approach tries to transfer prior knowledge from the high-quality images to the low-quality ones. Furthermore, to eliminate the necessity of a pre-trained teacher network, we adopt a self-KD method that involves training a single network in a progressive manner to distill its own knowledge [25].

In most self-KD-based approaches, the primary focus lies in minimizing the distance between the feature maps or soft targets of the networks. However, in this work, we adopt a different perspective by leveraging attention values derived from the network's feature representation. When attention is applied to the feature representations, the importance of essential regions, such as face landmarks, is heightened. This leads to the distillation of more significant information that effectively contributes to the FR task. As self-attention modules are the fundamentally important parts of our proposed CATF module which is based on transformers, we emulate the self-attention mechanism in our KD approach. To be more specific, we distill the knowledge from the high-quality self-attention components to their corresponding low-quality counterparts. Furthermore, considering the positive correlation between feature norm and image quality in recent studies [8], [9], our approach centers on aligning only the directional component of attention maps rather than their magnitude. Consequently, we formulate our distillation loss using cosine similarity, which enables us to capture the angular relationship between feature vectors and enhances the discriminative power of the model.

Most previous KD-based FR studies [26], [27] manually create low-resolution images to train the student network while low-resolution is merely one probable characteristic that unconstrained



Fig. 2: Samples of augmented data. The first row shows original images from the CelebA dataset. The second and the third rows depict low-quality versions of the original data which are generated by controllable face synthesis GAN and atmospheric turbulence simulator, respectively.

images may have. Hence, in this work, to fully exploit the advantage of self-KD, we augment the training dataset with diverse properties encountered in real-world scenarios. These properties include atmospheric turbulence, improper illuminations, motion blur, and various styles. By incorporating such variations, we aim to enhance the robustness and generalization ability of the network, enabling it to effectively recognize faces under challenging conditions commonly encountered in everyday life. This comprehensive augmentation approach moves beyond the limited scope of low-resolution images and provides a more realistic and representative training environment for KD-based FR systems. Some examples of the augmented data that we utilize in our KD-based approach are illustrated in Fig. 2.

In summary, the key contributions of this paper are summarized as follows:

- 1) We propose a novel multi-branch neural network to tackle the challenge of FR in low-quality images. By leveraging certain facial attributes from the SB branch, we enhance the performance of the dedicated FR branch.
- 2) We present a cross-attribute-guided transformer fusion (CATF) module which effectively captures and incorporates long-range dependencies, enabling a comprehensive understanding of the intricate relationships between FR and SB feature representations.
- 3) To raise the robustness of our proposed network against low-quality images, we propose a novel KD-based training approach. We prioritize important areas like facial landmarks by distilling self-attention values, outperforming other KD-based methods.
- 4) Extensive experiments with diverse datasets, including manual and real-world low-quality images, strongly support the enhanced performance of FR through employing SB information with the novel KD-based training approach and the proposed CATF module.

## 2 RELATED WORKS

### 2.1 Data Augmentation

One of the most significant challenges of FR methods is low-quality images. Low-quality images can result from several factors such as inherent camera noise, atmospheric turbulence, and

improper illuminations which lead to significant performance degradation. One promising approach to address this challenge is data augmentation, which involves creating additional training data by artificially modifying the existing dataset. By exposing the network to a wide range of image variations during training, it can learn to better cope with different types of distortions that may be present in real-world scenarios. Therefore, in this work, we augment our training data with different styles and attributes to enhance the robustness and generalization ability of our proposed model.

- **Controllable Face Synthesis.** In 2014, Goodfellow et al. introduced GANs [17] to synthesize realistic data samples based on a pair of neural networks, namely a generator and a discriminator. The core concept of GANs involves training the two networks in an adversarial manner, where the generator learns to produce fake samples that are indistinguishable from real ones, while the discriminator is trained to differentiate between real and fake samples. In recent years, many GAN-based face image generation models have been proposed to synthesize face images with desired properties [28], [29]. However, many of these models primarily focus on face editing tasks, such as face aging or transferring diverse expressions and poses from a given face image to a target one. While these approaches are valuable for such tasks, they are not particularly beneficial for enhancing the performance of FR in low-quality scenarios. In this regard, to create realistic useful face images, we use the model proposed in [30] which is a novel face synthesis model that can generate face images similar to the distribution of a target dataset through learning a linear subspace in the style latent space. Therefore, with an unlabeled target dataset including our desired characteristics such as motion blur, inherent sensor noise, and low resolution, we can generate face images containing all such attributes. Employing such realistic low-quality images through our proposed KD approach makes our multi-branch network robust against the different characteristics of the unconstrained scenarios.
- **Atmospheric Turbulence Simulation.** The effects of atmospheric turbulence on long-distance imaging applications, particularly in areas such as surveillance, are substantial. The fluctuation in the refractive index of air caused by atmospheric turbulence leads to variations in the path of light through the atmosphere, resulting in distortions in the captured images [31], [32]. These distortions significantly degrade the image quality, thereby posing challenges in extracting useful information from the affected images. Therefore, we consider the atmospheric turbulence effect in the training process. To generate atmospheric turbulence, we use a Phase-to-Space simulator which is proposed in [31]. This simulator is based on a novel concept called the phase-to-space (P2S) transform, which converts the phase representation of the turbulence to the spatial representation of the image. The P2S transform is implemented by a lightweight neural network that learns the basis functions for the spatially varying convolution from the known turbulence statistics models. By using the P2S transform, the simulation can be significantly accelerated compared to the conventional split-step method, while preserving the essential turbulence statistics.

## 2.2 Multi-task Learning for Face Analysis

Multi-task learning (MTL) is a strategy that simultaneously optimizes several relevant tasks to enhance the generalization performance through an inductive transfer mechanism. The concept of MTL can be traced back to the 1990s [33] which involves leveraging a single neural network to perform multiple related tasks. Following the advent of deep neural networks [34], MTL has been applied in many computer vision tasks such as medical image analysis [35], [36], object detection [37], [38], and facial attribute recognition [39], [40]. Here, we concentrate on MTL approaches for face analysis. Recently, several methods have incorporated the MTL framework for face-related tasks. Levi et al. [41] used the MTL framework to simultaneously perform age and gender prediction from face images. Also, HyperFace [39] is a multi-task learning algorithm that operates on the fused intermediate features to predict facial attributes and to do some other face-related tasks. In All-In-One [40], authors took advantage of MTL for FR in addition to facial attribute prediction which led to considerable improvement in FR for challenging unconstrained environments. However, existing MTL frameworks, do not directly leverage attribute information to enhance the performance of a FR task whereas intrinsically humans analyze facial attributes to recognize identities. In this regard, we propose a new multi-branch neural network that simultaneously performs SB prediction as an auxiliary modality and FR as the main task. To boost the discriminative ability of the FR branch, we integrate SB information with FR feature representation through an attentional module that is capable of learning complex relationships between input features. Moreover, we rely on facial attributes that remain consistent across different images of the same identity. For instance, attributes such as gender and eye shape exhibit consistency across varying illuminations or poses, while others like hair color may vary within different images of the same individual.

## 2.3 Knowledge Distillation

The concept of KD was first introduced by Hinton et al. in 2015 [42]. The basic idea is transferring knowledge from a large neural network (teacher) to a smaller neural network (student) by minimizing the Kullback-Leibler (KL) divergence of soft class probabilities between them. After that, several variants of distillation methods have been suggested to leverage the insights provided by the teacher network more effectively. For instance, many feature-based KD methods have been proposed that focus on distilling intermediate representations from the teacher model into the student network [43], [44]. An alternative approach known as self-distillation involves training a student network using its own knowledge, without the need for a separate teacher network. This approach aims to improve the efficiency, generality, and transferability of the learned knowledge. For instance, Zhang et al. in [45] proposed to first divide the network into several sections and then squeeze the knowledge in the deeper portion of the network into the shallow ones. There are also many augmentation-based strategies for self-distillation approaches [46], [47]. In the field of FR, researchers have also investigated the application of KD methods to improve network performance through the use of augmentation techniques, especially for challenging scenarios [48]. Shin et al. [48] utilized manually created low-resolution images to train the student network. However, our work takes a different approach by utilizing various synthesis methods to generate realistic face images that exhibit the characteristic challenges

observed in real-world scenarios, instead of relying on simple down-sampling. Furthermore, we go beyond typical methods that use network outputs or feature maps of different layers to transfer knowledge from high-quality images to low-quality ones. Instead, we leverage attention distillation to gain enhanced guidance from the self-attention mechanism by effectively identifying the crucial knowledge embedded within high-quality images. In addition to what is distilled, the distance metric for measuring distillation is also a critical factor influencing the model’s performance (see Section 4.3.2). Therefore, in contrast to most KD-based methods [47], [49] that employ general distance metrics like KL divergence or L2-distance, we tailor our distillation loss to our specific application which is FR.

## 2.4 Vision Transformer

Vision transformer (ViT) [50] is a novel neural network architecture that adapts the transformer model, originally developed for natural language processing, to vision processing tasks such as image recognition. ViT relies on the self-attention mechanism as a key component. This enables ViT to effectively attend to different parts of the input image and captures the interactions and long-range dependencies among pixels or patches within visual data. By leveraging self-attention, ViT can recognize complex patterns and features in the input images, thereby achieving state-of-the-art (SoTA) performance in large-scale image classification tasks. Inspired by the success of the seminal ViT, researchers have recognized the potential of the ViT architecture in various computer vision tasks like FR [51], [52], [53]. The potential of the vanilla ViT architecture for FR is explored in [52]. This study finds that although the transformer network encounters challenges when working with smaller databases, it exhibits promising performance when trained on larger datasets. Su et al. [54] introduce the atomic tokens and holistic tokens in the transformer encoder to capture the attentive relationship between facial regions and learn discriminative hybrid tokens to boost FR performance. Unlike the studies in FR that rely on the vanilla transformer architecture as a feature extractor, we exploit the advances in recent ViT studies to tailor the CATF module that can capture both global and local dependencies in face images between the FR and SB tasks. By incorporating the CATF module into our model, we enable a holistic understanding of the relationships between the FR and SB tasks, facilitating effective feature fusion. This would make our model achieve SoTA performance in the FR task which is our main goal.

## 3 PROPOSED METHOD

The proposed architecture, illustrated in Fig. 3, comprises two branches, namely the  $Br_{FR}$  and  $Br_{SB}$ , both sharing a ResNet-based backbone and dedicated convolutional layers. The  $Br_{FR}$  is dedicated to FR, while the  $Br_{SB}$  is intended for SB prediction. To enhance the performance of the network against low-quality images, we employ a self-distillation approach during the simultaneous training of both branches. By leveraging this approach, our network is capable of extracting and distilling valuable information from high-quality samples, thereby enhancing its ability to handle low-quality input images. This methodology showcases a promising direction for addressing the challenges posed by low-quality image inputs in FR and SB prediction tasks. Additionally, as shown in Fig. 4, upon training the multi-branch network, we employ a novel attention mechanism to integrate the

SB information into the FR feature representation. This integration effectively enriches the FR embedding, ultimately improving the model’s ability to identify challenging face images.

### 3.1 Self-Distilled Multi-Branch Network

#### 3.1.1 Multi-Branch Network

For FR, as it is shown in Fig. 3, there exist dedicated convolutional layers in addition to the backbone. The FR branch concludes with a softmax layer, the dimensions of which are determined by the number of classes within the training dataset. Thus, this branch is intrinsically considered for face identification which is basically a classification task based on identities. Conventional softmax loss of a sample  $x_i$  can be expressed as:

$$L(x_i) = -\log \frac{\exp(W_{y_i} \cdot z_i + b_{y_i})}{\sum_j^{N_c} \exp(W_j \cdot z_j + b_j)}, \quad (1)$$

where  $W_j$  represents the  $j$ -th column of the last fully connected layer’s weight,  $W \in \mathbb{R}^{d \times N_c}$ . The face embedding of sample  $x_i$  and its ground truth identity are shown by  $z_i \in \mathbb{R}^d$  and  $y_i$ , respectively.  $N_c$  and  $b_j$  indicate the number of classes and the bias term for the  $j$ -th class, respectively. In the context of training FR models, the features obtained from a simple softmax loss often fail to exhibit sufficient discriminative power. To address this limitation, the prevalent approach is to employ a margin-based softmax loss function. This loss function leads to the minimization of intra-class compactness, ensuring samples within the same identity cluster closely together, while simultaneously maximizing inter-class dispersion, enabling better separability between samples from different identity classes. In margin-based loss functions, for simplicity, the bias term is fixed to 0, and also the inner product of features and weights is considered as  $\|W_j\| \|z_i\| \cos \theta_j$  [55].  $\theta_j$  corresponds to the angle between  $z_i$  and  $W_j$ . Assuming  $\|W_j\|$  to be equal to 1 and  $z_i$  is rescaled with  $s$  during training, margin-based loss functions can be expressed as follows:

$$L_{AdaFace}(x_i) = -\log \frac{\exp(f(\theta_{y_i}, m))}{\exp(f(\theta_{y_i}, m)) + \sum_{j \neq y_i} \exp(s \cos \theta_j)}, \quad (2)$$

where  $m$  is a scalar hyper-parameter referred to as the margin, and  $f$  is a margin function. To achieve better convergence, many margin-based loss functions have been introduced where  $f(\theta_{y_i}, m)$  is the only distinguishing factor among them. AdaFace [9] is one of the recent SoTA margin-based loss functions that emphasizes samples of different difficulties based on their image quality. It approximates the image quality with feature norms. For high norms, it emphasizes samples away from the boundary, and for low norms, it emphasizes samples near the boundary. In this work, to train the FR branch, we utilize AdaFace loss function in which the margin function is defined as follows;

$$f(\theta_{y_i}, m) = \begin{cases} s(\cos(\theta_j + g_{angle}) - g_{add}) & \text{if } j = y_i \\ s \cos \theta_j & \text{if } j \neq y_i \end{cases}, \quad (3)$$

$$g_{angle} = -m \cdot \widehat{\|z_i\|}, \quad g_{add} = m \cdot \widehat{\|z_i\|} + m, \quad (4)$$

$$\widehat{\|z_i\|} = \left[ \left( \frac{\|z_i\| - \mu_z}{\sigma_z} \right) \right]_{-1}^1, \quad (5)$$

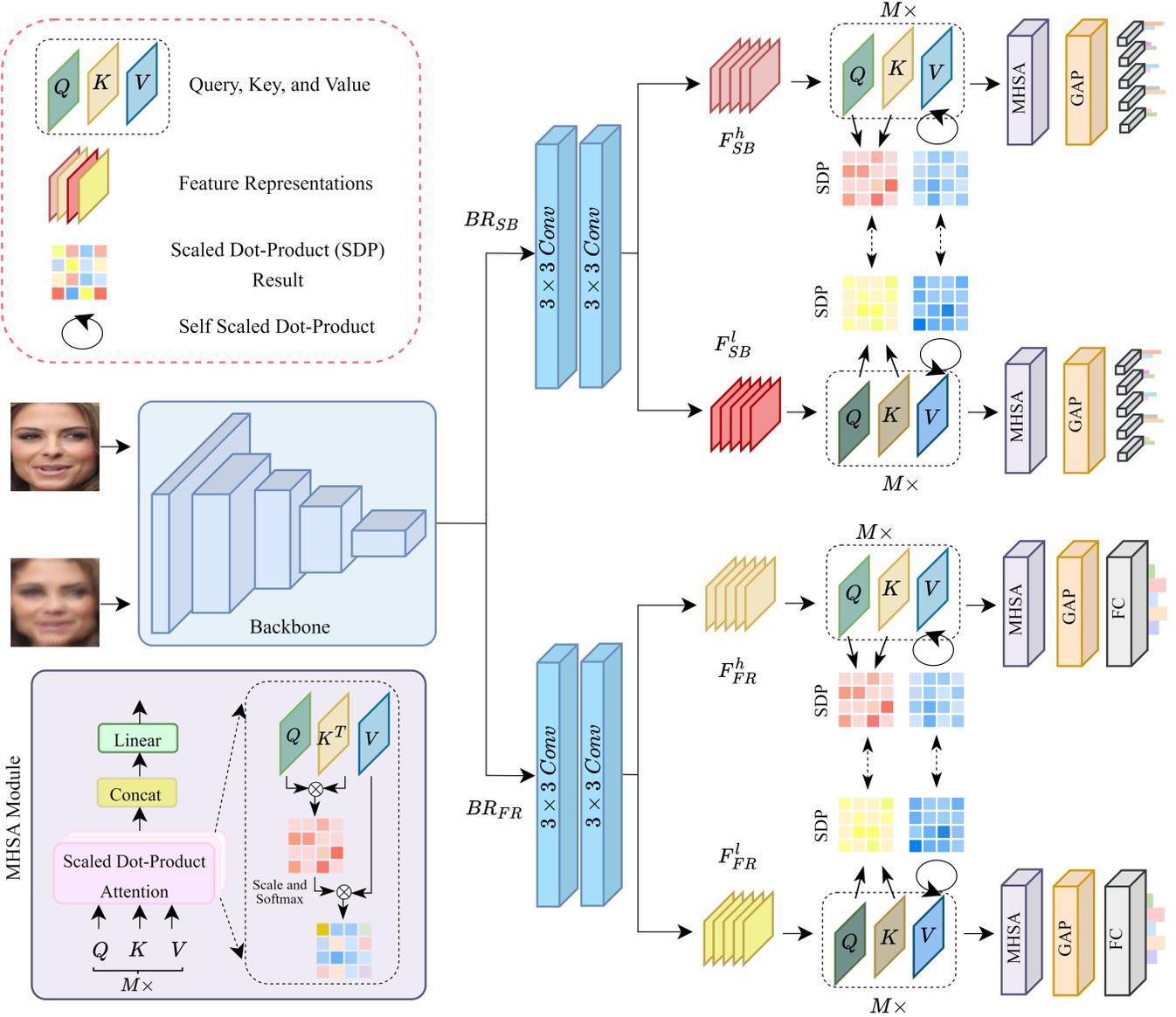


Fig. 3: Multi-branch neural network with self-attention distillation for FR and SB attribute prediction. Note that MHSA stands for multi-head self-attention module. This diagram shows the first step of our two-step training process. The  $BR_{FR}$  and  $BR_{SB}$  branches are jointly trained in the first step of the training process. In the second step, to enrich the FR feature representations, the SB and FR feature representations are fused together through the proposed CATF module (see Fig. 4). It should be noted that the global average pooling (GAP) and the final fully connected (FC) layers are removed from each branch for the second step of the training process.

where  $\|z_i\|$  measures the quality of a sample  $i$ , and  $\|\widehat{z}_i\|$  is the normalized quality using mean ( $\mu_z$ ) and standard deviation ( $\sigma_z$ ) of all  $z_i$  within a batch. It should be noted that over the test time when we are presented with arbitrary pairs of images for comparison (e.g.  $I_1$  and  $I_2$ ), the cosine similarity metric between them ( $\frac{z_1 \cdot z_2}{\|z_1\| \|z_2\|}$ ) determines whether they belong to the same identity or not.

The architecture of the  $BR_{SB}$  is similar to the  $BR_{FR}$ . During the training process, we employ binary classifiers dedicated to predicting different facial attributes, with each classifier equipped with its respective cross-entropy loss function. The total classification loss for the  $BR_{SB}$  is given by:

$$L_{SB} = \sum_{i=1}^n \lambda_{a_i} L_{a_i}, \quad (6)$$

where each  $L_{a_i}$  represents a loss for each individual attribute and  $\lambda_{a_i}$  is the loss-weight corresponding to the attribute  $a_i$ . Also,  $n$  denotes the number of SB attributes in  $BR_{SB}$ . For each attribute,  $L_{a_i}$  is computed as:

$$L_{a_i} = -(a_i \log(p_{a_i}) + (1 - a_i) \log(1 - p_{a_i})), \quad (7)$$

where  $p_{a_i}$  is the probability that the network computes for  $a_i$ .

### 3.1.2 Self-Attention Distillation

To enable our proposed multi-branch network to have a robust performance against low-quality images, we employ a self-distillation mechanism to distill information from high-quality images to their corresponding low-quality ones. In comparison with the most conventional KD-based methods that directly make feature maps close to each other, we focus on keeping the attention maps

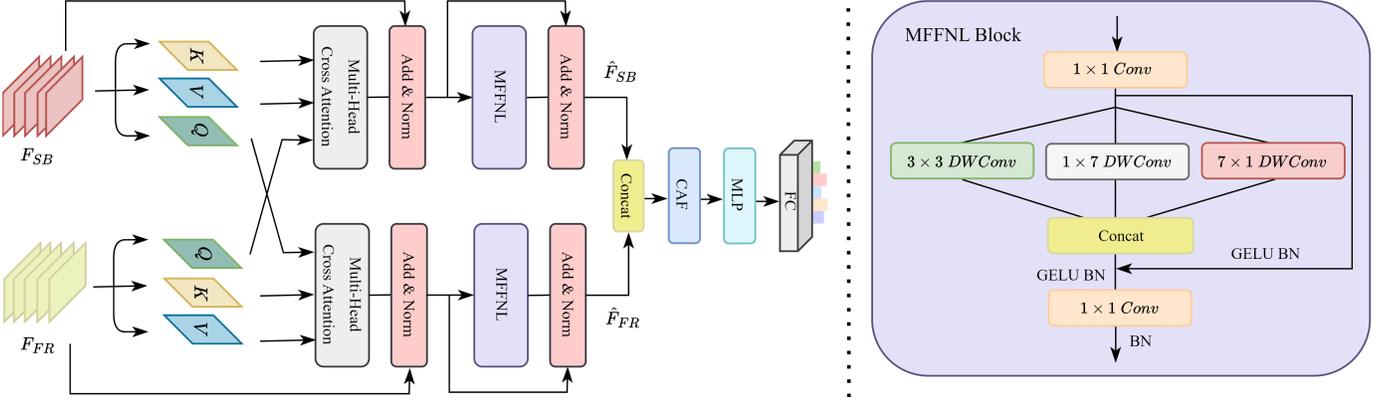


Fig. 4: Proposed cross-attribute-guided transformer fusion (CATF) module for FR. This diagram shows the second step of our two-step training process.

consistent between high-quality and low-quality samples in both the SB and FR branches. The inspiration behind our approach lies in the attention mechanism, which guides feature maps to focus on important regions. As key-points such as eyes and mouth are crucial for both FR and SB prediction, employing attention-based distillation can help the network to distill more informative features.

To identify the most discriminative features, we apply self-attention to the last convolutional layer of both the SB and FR branches. Self-attention can capture global dependencies between different regions of the face image such as face landmarks. Furthermore, in our proposed method, SB and FR feature representations are integrated using a cross-attribute-guided transformer fusion (CATF) module in which self-attention is a principal block. As a result, to provide the best input for the fusion module which includes three key components (i.e., key, query, and value), we leverage the self-attention mechanism to teach the network key information from high-quality samples.

Let  $F_{FR}^h \in \mathbb{R}^{H \times W \times C}$  and  $F_{FR}^l \in \mathbb{R}^{H \times W \times C}$  be the feature maps of the last convolutional layer of the  $BR_{FR}$  for a high-quality sample and its corresponding low-quality one, respectively, where  $H$ ,  $W$ , and  $C$  denote the height, width, and the channel number of each feature map. Similarly, we assume  $F_{SB}^h \in \mathbb{R}^{H \times W \times C}$ , and  $F_{SB}^l \in \mathbb{R}^{H \times W \times C}$  for  $BR_{SB}$ . We first flatten them to  $F_{FR}^l, F_{FR}^h, F_{SB}^l, F_{SB}^h \in \mathbb{R}^{N \times C}$  ( $N = H \times W$ ). Then, based on the self-attention mechanism, each feature map will be linearly projected to three learnable matrices: query matrices ( $Q_{FR}^l, Q_{FR}^h, Q_{SB}^l, Q_{SB}^h \in \mathbb{R}^{N \times C}$ ), key matrices ( $K_{FR}^l, K_{FR}^h, K_{SB}^l, K_{SB}^h \in \mathbb{R}^{N \times C}$ ), and value matrices ( $V_{FR}^l, V_{FR}^h, V_{SB}^l, V_{SB}^h \in \mathbb{R}^{N \times C}$ ). Finally, the attention map will be computed as the dot product of each  $Q$  and its corresponding  $K$ , as follows:

$$A_{FR}^l = \text{Softmax}\left(\frac{Q_{FR}^l (K_{FR}^l)^T}{\sqrt{C}}\right) V_{FR}^l, \quad (8)$$

$$A_{FR}^h = \text{Softmax}\left(\frac{Q_{FR}^h (K_{FR}^h)^T}{\sqrt{C}}\right) V_{FR}^h, \quad (9)$$

By following the same computational process, we can also compute  $A_{SB}^l$  and  $A_{SB}^h$ . We force the network to mimic not only attention maps of high-quality samples but also their corresponding value parameters. To minimize the differences between attention maps and also value parameters of high-quality and

low-quality samples, we employ cosine similarity. Therefore, the distillation loss is computed as:

$$L^{distill} = L_{FR}^{distill} + L_{SB}^{distill}, \quad (10)$$

$$L_{FR}^{distill} = 2 - \langle A_{FR}^l, A_{FR}^h \rangle - \langle V_{FR}^l, V_{FR}^h \rangle, \quad (11)$$

$$L_{SB}^{distill} = 2 - \langle A_{SB}^l, A_{SB}^h \rangle - \langle V_{SB}^l, V_{SB}^h \rangle, \quad (12)$$

where  $\langle \cdot \rangle$  denotes the cosine similarity metric. The total loss for each branch is the weighted sum of the target task's loss and the distillation loss.

### 3.2 Cross-Attribute-Guided Transformer Fusion (CATF).

To selectively focus on the most relevant regions in both the SB and FR feature representations and facilitate the fusion of discriminative facial information with SB cues, we employ a dual cross-attention operations in the CATF module (see Fig. 4). The reciprocal flow of information in the dual cross-attention operations enables a synergistic fusion of the SB and FR feature representations, enhancing the overall performance of FR. The cross-attention operations also effectively capture long-range dependencies, providing a holistic understanding of the relationships between the FR and SB tasks for feature fusion. In addition, we propose a multi-scale feed-forward network with locality (MFFNL), and the channel-wise attentional fusion (CAF) block in the CATF module to further improve the fusion of discriminative facial information with SB cues. Given the feature representations of SB and FR as  $F_{FR} \in \mathbb{R}^{H \times W \times C}$  and  $F_{SB} \in \mathbb{R}^{H \times W \times C}$ , we separately map  $F_{FR}$ , and  $F_{SB}$  to three learnable matrices: query matrices ( $Q_{FR}, Q_{SB} \in \mathbb{R}^{N \times C}$ ), key matrices ( $K_{FR}, K_{SB} \in \mathbb{R}^{N \times C}$ ), and value matrices ( $V_{FR}, V_{SB} \in \mathbb{R}^{N \times C}$ ). To promote effective feature collaboration, we create a cross-attention fusion operation by exchanging the query matrices  $Q_{FR}$  and  $Q_{SB}$  between the two branches as follows:

$$CA_{FR} = \text{Softmax}\left(\frac{Q_{FR} K_{SB}^T}{\sqrt{C}}\right) V_{SB}, \quad (13)$$

$$CA_{SB} = \text{Softmax}\left(\frac{Q_{SB} K_{FR}^T}{\sqrt{C}}\right) V_{FR}, \quad (14)$$

where  $CA_{FR}$  and  $CA_{SB}$  denote the cross-attention operations and  $C$  is the dimension of key matrices ( $K_{FR}, K_{SB} \in \mathbb{R}^{N \times C}$ ).

The single cross-attention operation is performed for each head in parallel to compute the multi-head cross-attention mechanism, denoted by  $MCA_{FR}$  and  $MCA_{SB}$ . Following the concatenation of all head unit outputs along the channel dimension, the resulting tensor is reshaped to match the dimensions of each feature map ( $F_{FR}, F_{SB} \in \mathbb{R}^{H \times W \times C}$ ).

### 3.2.1 Multi-scale Feed-Forward Network with Locality (MFFNL)

The standard transformer encoder includes a feed-forward network with two fully-connected layers for up- and down-projection operations, as well as GELU [56] activation. However, recent studies [57], [58] have shown that the vanilla feed-forward network cannot leverage local context in neighboring pixels, which is essential for an effective FR. To address this shortcoming, we propose a novel multi-scale feed-forward network named MFFNL which is able to learn facial features at different scales. As illustrated in Fig. 4, in our proposed MFFNL block, a multi-scale depth-wise convolution (MDConv) layer is integrated into the vanilla feed-forward network. The MFFNL block consists of two pointwise convolutions for expansion and projection operations and the proposed MDConv layer is positioned in between. The MDConv layer is composed of three parallel streams, each utilizing a distinct depth-wise convolution. The first stream utilizes a  $3 \times 3$  depthwise convolution, while the other two streams employ  $1 \times 7$  and  $7 \times 1$  depthwise convolutions, respectively. Motivated by [59], decomposition is adopted to decompose a  $7 \times 7$  convolution into two  $1 \times 7$  and  $7 \times 1$  convolutions to reduce computational complexity while maintaining the effective receptive field size. These streams are concatenated together to construct the fused feature representation. In addition, a shortcut connection is utilized after the MDConv layer to enhance the gradient propagation capability in the MFFNL block. The computation of the MFFNL block for input  $X$  is represented as:

$$\text{MFFNL}(X_{in}) = \text{PConv}\left(\text{MDConv}\left(\text{PConv}(X_{in})\right)\right), \quad (15)$$

$$\begin{aligned} \text{MDConv}(X_{in}) &= \text{Concat}\left(\text{DConv}_{3 \times 3}(X_{in}), \right. \\ &\quad \left. \text{DConv}_{1 \times 7}(X_{in}), \text{DConv}_{7 \times 1}(X_{in})\right) + X_{in}. \quad (16) \end{aligned}$$

where PConv and DConv denote the pointwise and depthwise convolution layers. After each layer, we use the GELU activation and batch normalization. To sum up, the MDConv layer facilitates multi-scale feature extraction in the MFFNL block, making our CATF module able to capture both short-term and long-term dependencies in FR and SB tasks.

### 3.2.2 Channel-wise Attentional Fusion (CAF)

Once we encode the long- and short-range interactions of the FR and SB features using our cross-attention operations and the MFFNL block, we propose to conduct feature fusion across global channel tokens (see Fig. 5). We first concatenate the outputs of MFFNL blocks, denoted by  $F_{\text{cat}} = \text{Concat}(\hat{F}_{FR}, \hat{F}_{SB})$ , along the channel dimension. Then, we construct channel-wise tokens by transposing the input tokens where the channel dimension determines the token scope and the spatial dimension determines the token feature dimension. In this context, channel-wise tokens can extract global interactions between both the FR and SB tasks. To incorporate attention scores between the channels of these tasks

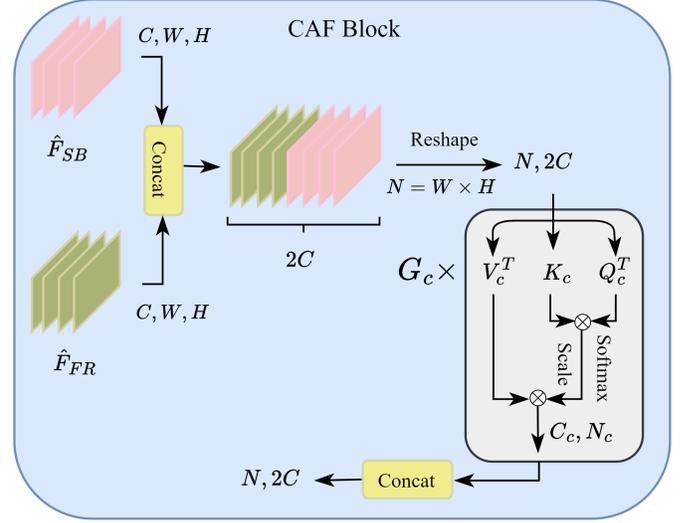


Fig. 5: Proposed channel-wise attentional fusion (CAF) block.

for feature fusion, we apply self-attention to the channel tokens. To achieve this while maintaining computational efficiency, we arrange the channel tokens into  $G_c$  groups with  $C_c$  channels each, where the channel dimension of  $F_{\text{cat}}$  is  $2C = G_c \times C_c$ . The formulation of channel-wise attention that interacts across a group of channels is as follows:

$$\text{CAF}(Q_c, K_c, V_c) = \text{Softmax}\left(\frac{Q_c^T K_c}{\sqrt{C_c}}\right) V_c^T, \quad (17)$$

where  $Q_c, K_c, V_c \in \mathbb{R}^{N_c \times C_c}$  are grouped channel-wise queries, keys, and values, respectively. Note that  $N_c$  stands for the spatial dimension of each channel group. Following the projections in the DMCA module, three projection layers are employed to compute the queries, keys, and values matrices along the channel dimension. Ultimately, we calculate the CAF for all  $G_c$  channel groups and concatenate all of them together to feed the classifier.

## 4 EXPERIMENTS

### 4.1 Implementation Details

#### 4.1.1 Training Datasets

To conduct a fair comparison with other methods, we separately train our model on two datasets: the CelebA [60] and MS1MV2 [61] datasets. CelebA is a large-scale face dataset containing 202,559 face images from more than 10k identities with different poses, backgrounds, and lighting conditions. Each face image is annotated with 40 facial attributes such as gender, the shape of the nose, and the color of the hair. In this work, we rely on identity facial attributes that stay the same in different images of the same person. For instance, gender, the shape of eyes, or being bald remain the same in different situations such as various illuminations or poses while some attributes such as the color of hair and wearing glasses may vary in different images of the same person. In this regard, we utilize five SB attributes which are gender, big nose, chubby, narrow eye, and Bald. In line with SoTA studies, we adhere to the dataset's protocol [60] for both the training and testing sets. The second training set, MS1MV2, is a large-scale dataset of more than 5M face images that make it possible to compare our proposed method with the SoTA methods on benchmark FR datasets. Gender information

is the only attribute provided for this large-scale dataset [62]. Therefore, in this case, our auxiliary modality only includes the information about one attribute.

#### 4.1.2 Test Datasets

To evaluate our proposed model, we utilize the test set of the CelebA dataset. Moreover, when employing the MS1MV2 dataset as the training set, we utilize several widely-used FR benchmarks with high-quality, mixed-quality, and low-quality settings. In high-quality settings, the LFW [63], CFP-FP [64], CPLFW [65], CALFW [66], and AgeDB [67] datasets are utilized. The datasets in high-quality setting exhibit variations in lighting, pose, and age. To investigate the performance of the proposed method on more challenging images, we also test our model on mixed-quality setting with the IJB-B [68], and IJB-C [13] datasets which cover a wide range of face variations and challenges for FR in unconstrained settings. The IJB-B and IJB-C datasets consist of 21.8K and 31.3K images, respectively. The IJB-C dataset, comprising 3,531 identities, is an extension of the IJB-B dataset, covering 1,845 different identities. For both of these datasets, we follow the standard 1:1 verification protocol which is a template-based method. Considering that each template contains multiple frames, we compute the average feature vector for each template. Moreover, to gauge the efficacy of the proposed method in more challenging scenarios, we evaluate our method on low-quality realistic and synthetic FR test sets. For realistic tests, we employ TinyFace [69], a low-resolution in-the-wild dataset, and SCFace [70], a cross-resolution FR dataset captured in uncontrolled indoor environments at three different distances. As for the synthetic tests, we utilize the CelebA dataset to synthesize low-quality data corrupted by the controllable face synthesis GAN and atmospheric turbulence simulator.

#### 4.1.3 Augmentations

As discussed in Section 3, in the training process, we utilize a novel distillation approach to transfer the knowledge learned from the high-quality images to the low-quality ones to boost the model’s performance in challenging scenarios. To create realistic low-quality versions of the training data, we adopt two approaches. In our first approach, we employ a simulator proposed in [31] to generate images corrupted with atmospheric turbulence effect. In this simulator, we can control the strength of the atmospheric turbulence by an aperture diameter,  $D$ , divided by the fied parameter,  $r_0$ . We refer the readers to [31] for detailed information. Fig. 6 shows sample images degraded by different strengths of atmospheric turbulence. It is evident that when turbulence levels are high, facial attributes and landmarks are impacted by considerable deformation and blurring. During the training phase, we randomly corrupt the training data with different ratios of  $D/r_0$  which determines the strength of the atmospheric turbulence effect (between 0.25 to 2).

In addition to the atmospheric turbulence effect, we employ the controllable face synthesis generator introduced in [30] to create low-quality images with unconstrained imaging environment factors including noise, low resolution, and motion blur. The generator is pre-trained on the WiderFace dataset [71], which encompasses a wide array of unconstrained variations, as the target data.

#### 4.1.4 Training details

The scale of the training dataset plays a crucial role in the performance of the FR, as a larger dataset can provide a wider

range of real-world characteristics, leading to better generalization to unseen data. Hence, in the case of training on the CelebA dataset, our backbone is weighted with a pre-trained ResNet-50 [3] on the VGGFace2 dataset [10] that contains more than 3.3M images of about 9k identities. As mentioned before, to gain a better insight into the advantage of utilizing attributes for FR, we have also used MS1MV2 [61], which includes more than 5M face images as a training set. In this case, we employ ResNet-101 as the backbone of our proposed model.

The training process of the proposed method includes two main steps. First, we jointly train our multi-branch network with both classification and distillation losses for each branch. The weight parameters of the total loss function are determined based on the prioritization of FR as our primary objective. We set  $\lambda_{FR} = 3$ ,  $\lambda_{Male} = \lambda_{Bald} = 1$ , and all other weight parameters to 0.5. In the second step of the training phase, to enrich the FR feature representation, we fuse the FR and SB feature representations together through our proposed CATF module. As such, we train this fused branch for the goal of FR with the loss function given in Equation 2.

The model undergoes training using stochastic gradient descent with an initial learning rate of 0.1. In the case of training on the CelebA dataset, the model is trained for 25 epochs, and the scheduling step is set at 3, 7, and 15 epochs. For training on the MS1MV2 dataset, the scheduling step is set at 7, 13, and 18 epochs for a total of 30 epochs.

## 4.2 Comparison with SoTA methods

### 4.2.1 Soft Biometric Prediction

Table 1, presents a comprehensive evaluation of the proposed SB predictor on the two challenging annotated datasets, CelebA [60] and LFWA [63] datasets. These datasets are widely used in the field of facial attribute analysis and serve as benchmarks for evaluating the performance of SB prediction methods. Similar to the SoTA methods, we have followed the same protocol, and the results of the other methods are directly reported from the original papers. In the case of the CelebA dataset, our proposed approach consistently outperformed the existing SB prediction methods, achieving the highest accuracy across all attributes. For the LFW dataset, our method surpassed all other methods for all those attributes except for the narrow eye attribute, where it secured the position of a runner-up among all other methods.

### 4.2.2 Face Recognition

In line with the established methodology of the SoTA FR studies that train the model as a classifier and test it as a verifier [5], [7], [9], we also evaluate our FR model as a verifier. To this end, when employing the CelebA dataset as a training set, a subset of 10,000 pairs is randomly sampled from the CelebA dataset, ensuring that the identities of these pairs are excluded from the training set. Regarding Section 4.1.4, in the first step of the training process, we train our multi-branch network without employing any fusion modules. Thus, we can consider this branch as a baseline to better clarify the effective role of integrating SB into FR. The experimental findings presented in Table 2 provide compelling evidence that our proposed model effectively enhances FR performance through the utilization of SB attributes. We have also performed a comprehensive set of experiments to evaluate the effectiveness of our proposed CATF module and compare it with alternative integration strategies. The experimental findings,



Lower Degradation Degree

Higher Degradation Degree

Fig. 6: Images corrupted by simulated atmospheric turbulence with strengths ranging from 0.25 to 2 (the first image is the original one).

as presented in Table 2, demonstrate that substituting this module with simple operations like addition or concatenation leads to even inferior performance compared to the baseline approach, particularly for specific false acceptance rates (FARs). Moreover, results prove that among recent feature integration studies [72], [73], [74], our proposed module establishes the most effective integration approach for enhancing the performance of FR feature representation. Furthermore, we extend our explorations to investigate the impact of the number of attributes utilized for FR, and the last rows of Table 2 reveal that incorporating more identity facial attributes contributes to improved accuracy.

To obtain a better understanding of the benefits associated with employing SB attributes for FR and also have a fair comparison with recent SoTA methods, we have additionally trained our model on the MS1MV2 dataset [5]. In this case, we adopt ResNet-101 as the backbone and conduct evaluations on nine widely recognized FR benchmarks. As shown in Table 3, results demonstrate that the observed enhancements for the mixed-quality and low-quality datasets are more notable in comparison with the improvements in the case of high-quality datasets. This can be attributed to the accuracy saturation in high-quality datasets such as LFW and CFP-FP benchmarks. As high-quality images inherently contain a wealth of important facial information, the distillation of knowledge from such images to their corresponding low-quality counterparts becomes less noticeable. Similarly, the same scenario holds true for the impact of utilizing SB attributes to help the FR branch. As a result, the marginal gains achieved in the high-quality scenario do not reflect the full potential of the proposed method. Instead, the efficacy of the proposed method becomes particularly evident when dealing with mixed-quality and low-quality images, as these cases greatly benefit from the supplementary knowledge transferred from the higher-quality images. It is worth noting that due to the availability of only one attribute for this training set, we focused on a single SB attribute (gender) to evaluate the effectiveness of the proposed method while considering multiple attributes could lead to even greater improvements in especially low-quality cases.

To further demonstrate the effectiveness of our proposed approach, we expanded our experiments to include a realistic cross-resolution dataset. Table 5 presents a comparison of our proposed method with the recent SoTA methods on the SCFace dataset [70]. Some approaches, such as RPCL [75] and RI [76], optimized their methods through fine-tuning on the SCFace training set. For models such as FAN [77] and TRM [78], which reported performance with and without fine-tuning on this dataset, it is evident that fine-tuning significantly enhances the performance of the FR model. As indicated in Table 5, our model surpasses all non-fine-tuned methods and even rivals the performance of models that are fine-tuned on the SCFace dataset. This result underscores

TABLE 1: Performance comparison in terms of accuracy (%) between the proposed SB predictor and the SoTA methods.

Data	Methods	Bald	Big Nose	Chubby	Male	Narrow Eye
CelebA	Z. Liu et al. [79]	98.00	78.00	91.00	98.00	81.00
	Moon [80]	98.77	84.00	95.44	98.10	86.52
	HyperFace [39]	-	-	-	97.00	-
	All-In-One [40]	-	-	-	99.00	-
	MCFA [81]	99.00	84.00	96.00	98.00	87.00
	DMM [82]	99.03	84.78	95.86	98.29	87.73
	Ours	<b>99.11</b>	<b>85.41</b>	<b>96.13</b>	<b>99.19</b>	87.69
LFWA	Z. Liu et al. [79]	88.00	81.00	73.00	94.00	81.00
	HyperFace [39]	-	-	-	94.00	-
	All-In-One [40]	-	-	-	93.12	-
	MCFA [81]	91.00	81.00	74.00	93.00	78.00
	DMM [82]	91.96	83.67	77.66	94.14	83.67
	Ours	<b>92.19</b>	<b>84.49</b>	<b>77.71</b>	<b>95.36</b>	<b>83.81</b>

TABLE 2: Performance comparison between the proposed method (CATFace), the baseline, and other SoTA feature integration methods. Results are based on TAR@FAR, in which TAR and FAR stand for True Acceptance Rate, and False Acceptance Rate, respectively. Also, M, B, and NE stand for male, bald, and narrow eyes, respectively.

Methods	CelebA				
	$10^{-5}$	$10^{-4}$	$10^{-3}$	$10^{-2}$	$10^{-1}$
Baseline (without SB)	89.23	90.51	92.08	93.26	94.48
Concatenation	88.95	90.06	92.06	93.31	94.50
Addition	88.83	89.96	91.74	93.29	94.39
SENET [72]	89.98	90.96	92.84	94.39	95.65
Cross-Attention [74]	90.68	92.12	93.30	94.50	95.68
FFM [73]	90.73	92.59	93.49	94.52	95.79
CATFace	<b>91.10</b>	<b>92.91</b>	<b>93.78</b>	<b>94.83</b>	<b>96.18</b>
CATFace (B)	89.54	90.89	92.67	93.90	94.81
CATFace (B & M)	89.97	91.03	92.83	94.13	94.97
CATFace (B & M & NE)	90.23	91.37	93.08	94.29	95.13

the generalization capability of our proposed model to handle unseen cross-resolution face images.

### 4.3 Ablation and Analysis

#### 4.3.1 Effect of Self-Attention Distillation

To better clarify the effect of self-attention distillation on our proposed model, we investigate its impact on the FR and SB branches separately. Table 4 and Table 6 demonstrate the effect of self-attention distillation on SB prediction and FR, respectively. In the training phase, to generate distorted versions of both CelebA and LFWA datasets, within each batch, we randomly distort 50 percent of images by atmospheric turbulence [31] and the rest are augmented by a GAN-based controllable face synthesis

TABLE 3: Performance comparison of our proposed method (CATFace) with recent SoTA FR methods. TAR is reported at FAR = 0.01% (All these methods are trained on the MS1MV2 dataset).

Methods	Venue	High Quality (Verification Accuracy)					Mixed Quality (TAR)		Low Quality TinyFace	
		LFW	CFP-FP	CPLFW	AgeDB	CALFW	IJB-B	IJB-C	Rank-1	Rank-5
CosFace [7]	CVPR18	99.81	98.12	92.28	98.11	95.76	94.80	96.37	-	-
ArcFace [5]	CVPR19	99.83	98.27	92.08	98.28	95.45	94.25	96.03	-	-
MV-Softmax [83]	AAAI20	99.80	98.28	92.83	97.95	96.10	93.60	95.20	-	-
URL [84]	CVPR20	99.78	98.64	-	-	-	-	96.60	63.89	68.67
SCF-ArcFace [85]	CVPR21	99.82	98.40	93.16	98.30	96.12	94.74	96.09	-	-
MagFace [8]	CVPR21	99.83	98.46	92.87	98.17	96.15	94.51	95.97	-	-
MIND [86]	LSP21	-	-	-	-	-	-	-	66.82	-
ElasticFace [87]	CVPRW22	99.80	<b>98.73</b>	93.23	98.28	96.18	95.43	96.65	-	-
LS [88]	FG23	99.50	-	-	-	-	-	-	66.30	-
AdaFace [9]	CVPR22	99.82	98.49	93.53	98.05	96.08	95.67	96.89	68.21	71.54
CATFace <sup>1</sup>	-	99.83	98.57	<b>93.71</b>	<b>98.14</b>	<b>96.17</b>	<b>95.82</b>	<b>97.07</b>	<b>68.52</b>	<b>71.92</b>
CATFace <sup>2</sup>	-	<b>99.84</b>	98.68	<b>93.84</b>	<b>98.33</b>	<b>96.32</b>	<b>96.13</b>	<b>97.43</b>	<b>68.95</b>	<b>72.31</b>

<sup>1</sup>This is our proposed FR method trained with the self-distillation approach without employing SB attributes.

<sup>2</sup>This is our proposed FR method trained with both the self-distillation approach and the proposed CATF module to employ SB attributes.

TABLE 4: Ablation of our self-attention distillation approach on the SB branch.

Test Data	Approach			Bald	Big Nose	Chubby	Male	Narrow Eye	
	Aug	Distill							
		Feat	CBAM	SA					
CelebA					99.10	84.84	96.09	99.16	87.56
	✓				99.10	84.95	96.00	99.10	87.30
	✓	✓			99.11	84.91	96.05	99.14	87.57
	✓		✓		99.12	85.16	96.10	99.16	87.63
	✓			✓	99.11	<b>85.41</b>	<b>96.13</b>	<b>99.19</b>	<b>87.69</b>
Distorted CelebA	✓				96.53	80.73	93.84	97.68	84.79
	✓				97.70	82.01	94.24	97.93	85.32
	✓	✓			98.09	82.49	94.30	97.96	85.68
	✓		✓		98.20	83.05	94.54	98.41	85.93
	✓			✓	<b>98.31</b>	<b>83.11</b>	<b>94.69</b>	<b>98.74</b>	<b>86.10</b>
LFWA	✓				90.83	82.73	76.67	93.10	82.96
	✓				90.96	82.97	76.83	93.13	82.95
	✓	✓			91.23	82.98	76.91	93.45	83.02
	✓		✓		91.87	83.07	77.39	94.09	83.85
	✓			✓	<b>92.19</b>	<b>84.49</b>	<b>77.71</b>	<b>95.36</b>	<b>83.81</b>
Distorted LFWA	✓				88.00	78.12	72.22	90.62	77.38
	✓				88.91	79.35	73.10	90.99	78.45
	✓	✓			89.15	79.43	73.93	91.75	78.90
	✓		✓		89.88	80.91	74.50	92.90	79.25
	✓			✓	<b>90.26</b>	<b>81.51</b>	<b>74.66</b>	<b>93.02</b>	<b>80.62</b>

method [30]. For each set of test data, we have considered five different models to predict SB attributes. The first model is the SB branch of a simple multi-branch network trained without any data augmentation and distillation approaches (the first row of Table 4). The second model is trained with augmented data, in addition to the original data without employing any distillation approach. The third model utilizes a general feature-based distillation method during the training phase. The fourth and last models use attention-based self-distillation approaches which are based on CBAM [89] and self-attention, respectively. The results in Table 4 indicate that while data augmentation can improve the accuracy of predicting certain SB attributes, the incorporation of KD proves to be more effective in fully utilizing the potential of the synthesized low-quality images. Furthermore, our experimental findings verify the superiority of the self-attention mechanism in producing informative feature maps in comparison with the CBAM method. Similarly, for the FR task, we have considered five different models (see Table 6). The first model corresponds to the FR branch of a simple multi-branch network trained without data augmentation or distillation approaches. The other four models

TABLE 5: Performance comparison of our proposed method (CATFace) with recent SoTA FR methods on the SCFace dataset.

Methods	Fine-Tuned	Distance			Avg.
		4m	2.6m	1m	
FAN [77]	✓	77.50	95.00	98.30	90.30
ArcFace [5], [90]	✓	80.50	98.00	99.50	92.70
RPCL [75]	✓	90.40	98.00	98.00	95.47
TRM [78]	✓	91.25	99.50	99.50	96.75
DDL [91]	✓	93.20	99.20	98.50	97.00
RI [76]	✓	97.07	99.23	99.80	98.70
FAN [77]	-	62.00	90.00	94.80	82.30
ArcFace [5], [90]	-	58.90	98.30	99.50	85.50
DCR [92]	-	73.30	93.50	98.00	88.27
TRM [78]	-	79.25	97.00	97.75	91.33
CATFace	-	<b>90.64</b>	<b>98.85</b>	<b>99.61</b>	<b>96.37</b>

are also similar to what was mentioned for the SB branch. Results obtained from these experiments for the FR task further reinforce the effectiveness of our proposed KD approach over alternative methods in KD.

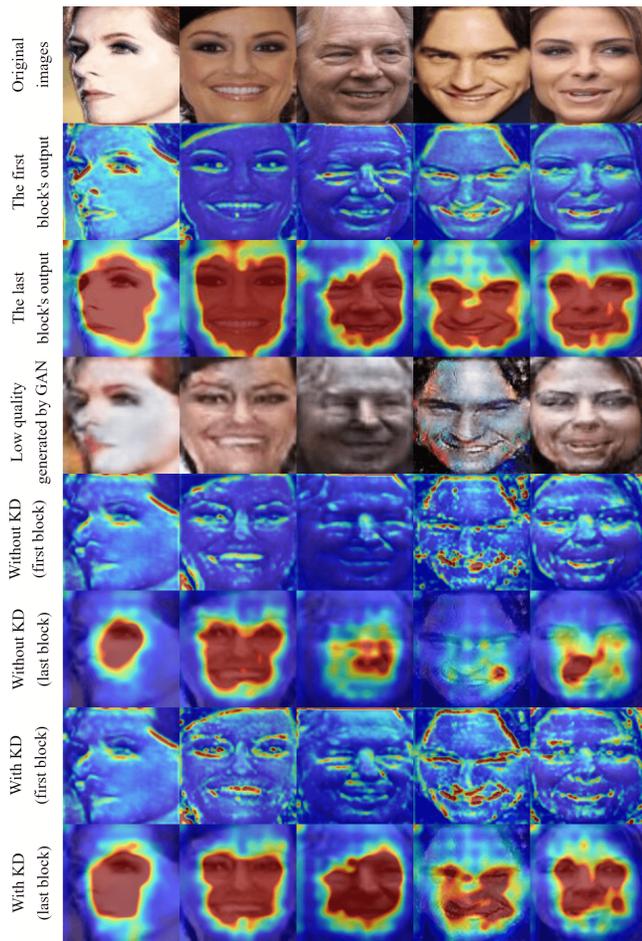


Fig. 7: The visualization of the feature maps of the first and the last block of the FR branch. The second and the third rows are related to the original images. The fourth row shows the low-quality versions of the original images generated by the GAN. The fifth and the sixth rows are corresponding to the low-quality images without using the KD approach while the last two rows depict the feature maps of the low-quality images when the network is utilizing the proposed KD approach.

TABLE 6: Ablation of our self-attention distillation approach on the FR branch (TAR is reported at FAR = 0.01%).

Aug	Approach			CelebA	Distorted CelebA
	Distill				
	Feat	CBAM	SA		
✓				92.64	89.16
✓	✓			92.83	90.36
✓		✓		93.05	90.73
✓			✓	93.18	91.27
✓	✓		✓	<b>93.26</b>	<b>91.43</b>

#### 4.3.2 Effect of Cosine Similarity metric on Distillation

Table 7 compares four different approaches, each varying in attention map creation and distillation metrics. The first two rows utilize the L2-distance metric to distill attention maps from high-quality samples to their corresponding low-quality counterparts. The subsequent rows adopt a similar approach but employ the cosine similarity metric for attention distillation. Results conspicuously verify that utilizing cosine similarity, whether with CBAM or self-attention methods, outperforms using L2-distance

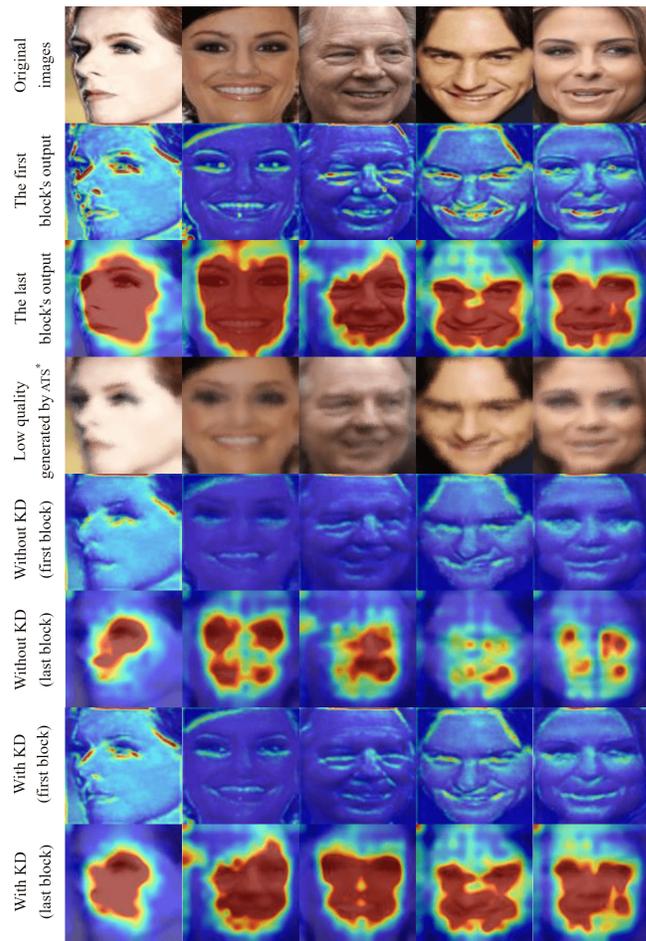


Fig. 8: The visualization of the feature maps of the first and the last block of the FR branch. The second and the third rows are related to the original images. The fourth row shows the low-quality versions of the original images which are disturbed by atmospheric turbulence effects. The fifth and the sixth rows are corresponding to the low-quality images without using the KD approach while the last two rows depict the feature maps of the low-quality images when the network is utilizing the proposed KD approach.

\*ATS stands for atmospheric turbulence simulator.

for distillation. It is demonstrated that feature norm is positively correlated with image quality [8], [9]. Thus, to effectively utilize the information extracted from the high-quality samples, we need to distill only the directional component of the attention maps while the L2-distance tries to align both the norm and angle components of the attention maps. As shown in Equations 11 and 12, our distillation loss minimizes exclusively the angle between the attention maps of the high-quality and low-quality samples which enables the network to focus on richer information.

#### 4.3.3 Effect of Soft Biometric Attributes

As explained in Section 4.2.2, to explore the contribution of SB attributes to our proposed model, we conduct several experiments such as exploring the impact of the number of attributes employed for FR or clarifying the crucial role of the proposed CATF module in the integration stage. All the experimental results verify the benefits of serving SB information as auxiliary data in the FR task (see Tabel 2). In this section, we perform further experiments to

TABLE 7: Ablation of our distillation metric on the FR branch (TAR is reported at FAR = 0.01%).

Approach				CelebA	Distorted CelebA
Distance Metric		Distill			
L2-Distance	Cosine Sim	CBAM	SA		
✓		✓		93.07	90.77
✓			✓	93.10	90.89
	✓	✓		93.18	91.27
	✓		✓	<b>93.26</b>	<b>91.43</b>

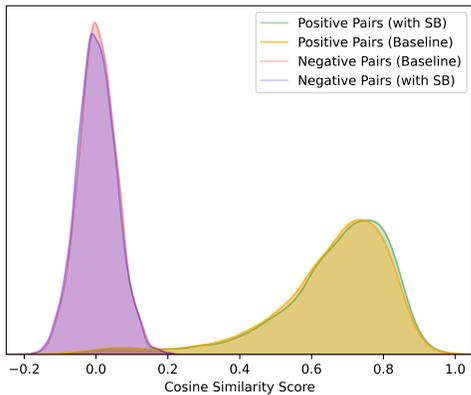


Fig. 9: Comparison of cosine similarity distribution between the baseline model without employing SB information and the proposed model employing SB information. The positive and negative pairs are sampled from the CelebA dataset.

assess the contribution of each component of our proposed CATF module, namely MFFNL and CAF blocks, to the FR task. Table 8 demonstrates that optimal performance is attained when both components are utilized in conjunction. Furthermore, the CAF block appears to have a more significant impact on the overall performance of the CATF module compared to the MFFNL block.

To gain deeper insights into the role of SB attributes, we visualize the distributions of the similarity scores on 10,000 pairs of the CelebA dataset both with and without the utilization of SB attributes. As depicted in Fig. 9, the peak values of the cosine similarity score distribution for both the positive and negative pairs are shifted. To be more specific, the peak value of the cosine similarity score distribution is shifted rightward for the positive pairs and leftward for the negative pairs. These shifts indicate that leveraging SB attributes leads to better separation between the similarity scores of the positive and negative pairs which implies a reduction in false positive and false negative errors.

#### 4.3.4 Visualization

The features of the first and last convolutional block of the FR branch are visualized in Fig. 7 and Fig. 8, through an attention map introduced by [93]. We compare the output feature representations between the original data and its low-quality version generated by the controllable GAN and the atmospheric turbulence simulator. To generate these maps, we normalize values within a range of 0 to 1, making them more visually interpretable. The attention maps provide valuable insights into the network’s focus during FR tasks. Regarding these attention maps, in the case of high-quality images (the original data), the network focuses on critical facial features such as the eyes, nose, and lips, which play a pivotal role in ensuring accurate recognition (rows 2 and 3, Fig. 7 and Fig. 8). However, when it comes to the low-quality images, the

TABLE 8: Ablation of our proposed CATF module on the MFFNL and CAF blocks (results are based on TAR@FAR).

Approach		CelebA				
MFFNL	CAF	$10^{-5}$	$10^{-4}$	$10^{-3}$	$10^{-2}$	$10^{-1}$
		90.68	92.12	93.30	94.50	95.68
✓		90.81	92.76	93.55	94.73	95.90
	✓	90.97	92.83	93.72	94.78	96.04
✓	✓	<b>91.10</b>	<b>92.91</b>	<b>93.78</b>	<b>94.83</b>	<b>96.18</b>

network’s attention to intricate details such as facial landmarks is compromised due to the absence of information (rows 5 and 6, Fig. 7 and Fig. 8). This is where our attention-based KD approach comes into play and enhances the ability of the model to concentrate on detailed facial features (the last two rows, Fig. 7 and Fig. 8). It achieves this by matching the attention maps of the low-quality images to their corresponding high-quality counterparts.

## 5 CONCLUSION

This paper addresses the poor performance of FR models with regard to low-quality images. Inspired by the fact that humans intrinsically analyze facial attributes to recognize identities, we utilize SB attributes as auxiliary information to improve the performance of FR. We propose a novel feature-level fusion module to effectively integrate SB information into the FR feature representations. We also incorporate a self-distillation technique during the simultaneous training of both the SB and FR branches which empowers our network to extract and distill effective information from high-quality samples, thereby strengthening its capacity to handle low-quality input images.

## REFERENCES

- [1] R. Khallaf and M. Khallaf, “Classification and analysis of deep learning applications in construction: A systematic literature review,” *Autom. Constr.*, vol. 129, p. 103760, 2021.
- [2] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2015, pp. 1–9.
- [3] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 770–778.
- [4] F. Wang, X. Xiang, J. Cheng, and A. L. Yuille, “NormFace: L2 hypersphere embedding for face verification,” in *Proc. 29th ACM Int. Conf. Multimedia*, 2017, pp. 1041–1049.
- [5] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, “ArcFace: Additive angular margin loss for deep face recognition,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 4690–4699.
- [6] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, “SphereFace: Deep hypersphere embedding for face recognition,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 212–220.
- [7] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu, “CosFace: Large margin cosine loss for deep face recognition,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 5265–5274.
- [8] Q. Meng, S. Zhao, Z. Huang, and F. Zhou, “MagFace: A universal representation for face recognition and quality assessment,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021, pp. 14 225–14 234.
- [9] M. Kim, A. K. Jain, and X. Liu, “AdaFace: Quality adaptive margin for face recognition,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 18 750–18 759.
- [10] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, “VggFace2: A dataset for recognising faces across pose and age,” in *Proc. 13th IEEE Int. Conf. Automat. Face Gesture Recognit. (FG)*, 2018, pp. 67–74.

- [11] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Learning face representation from scratch," *arXiv preprint arXiv:1411.7923*, 2014.
- [12] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, "MS-Celeb-1M: A dataset and benchmark for large-scale face recognition," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 87–102.
- [13] B. Maze, J. Adams, J. A. Duncan, N. Kalka, T. Miller, C. Otto, A. K. Jain, W. T. Niggel, J. Anderson, J. Cheney *et al.*, "IARPA Janus benchmark-c: Face dataset and protocol," in *Proc. Int. Conf. Biometrics (ICB)*, 2018, pp. 158–165.
- [14] L. Tran, X. Yin, and X. Liu, "Disentangled representation learning GAN for pose-invariant face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 1415–1424.
- [15] N. Singh, S. S. Rathore, and S. Kumar, "Towards a super-resolution based approach for improved face recognition in low resolution environment," *Multimed. Tools Appl.*, vol. 81, no. 27, pp. 38 887–38 919, 2022.
- [16] J. Chen, J. Chen, Z. Wang, C. Liang, and C.-W. Lin, "Identity-aware face super-resolution for low-resolution face recognition," *IEEE Signal Process. Lett.*, vol. 27, pp. 645–649, 2020.
- [17] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Proc. Adv. Neural Inf. Process. Syst.*, pp. 2672–2680, 2014.
- [18] M.-Y. Liu and O. Tuzel, "Coupled generative adversarial networks," in *Proc. Adv. Neural Inf. Process. Syst.*, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, Eds., vol. 29, 2016.
- [19] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, June 2019.
- [20] X. Tu, J. Zhao, Q. Liu, W. Ai, G. Guo, Z. Li, W. Liu, and J. Feng, "Joint face image restoration and frontalization for recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 3, pp. 1285–1298, 2021.
- [21] X. Yin, X. Yu, K. Sohn, X. Liu, and M. Chandraker, "Towards large-pose face frontalization in the wild," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2017.
- [22] Y. Yin, J. P. Robinson, S. Jiang, Y. Bai, C. Qin, and Y. Fu, "SuperFront: From low-resolution to high-resolution frontal face synthesis," in *Proc. 29th ACM Int. Conf. Multimedia*, 2021, pp. 1609–1617.
- [23] J. Cao, Y. Hu, H. Zhang, R. He, and Z. Sun, "Towards high fidelity face frontalization in the wild," *Int. J. Comput. Vis.*, vol. 128, pp. 1485–1504, 2020.
- [24] Z. Jin, J.-Y. Yang, Z.-S. Hu, and Z. Lou, "Face recognition based on the uncorrelated discriminant transformation," *Pattern Recognit.*, vol. 34, no. 7, pp. 1405–1416, 2001.
- [25] M. Ji, S. Shin, S. Hwang, G. Park, and I.-C. Moon, "Refine myself by teaching myself: Feature refinement via self-knowledge distillation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021, pp. 10 664–10 673.
- [26] S. Ge, S. Zhao, C. Li, Y. Zhang, and J. Li, "Efficient low-resolution face recognition via bridge distillation," *IEEE Trans. Image Process.*, vol. 29, pp. 6898–6908, 2020.
- [27] S. Ge, S. Zhao, C. Li, and J. Li, "Low-resolution face recognition in the wild via selective knowledge distillation," *IEEE Trans. Image Process.*, vol. 28, no. 4, pp. 2051–2062, 2018.
- [28] D. Roich, R. Mokady, A. H. Bermano, and D. Cohen-Or, "Pivotal tuning for latent-based editing of real images," *ACM Trans. Graph.*, vol. 42, no. 1, pp. 1–13, 2022.
- [29] T. Xiao, J. Hong, and J. Ma, "ELEGANT: Exchanging latent encodings with gan for transferring multiple face attributes," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 168–184.
- [30] F. Liu, M. Kim, A. Jain, and X. Liu, "Controllable and guided face synthesis for unconstrained face recognition," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Springer, 2022, pp. 701–719.
- [31] Z. Mao, N. Chittim, and S. H. Chan, "Accelerating atmospheric turbulence simulation via learned phase-to-space transform," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2021, pp. 14 759–14 768.
- [32] W. Robbins and T. E. Boulton, "On the effect of atmospheric turbulence in the feature space of deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 1618–1626.
- [33] R. Caruana, "Multitask learning," *Mach. Learn.*, vol. 28, pp. 41–75, 1997.
- [34] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Proc. Adv. Neural Inf. Process. Syst.*, vol. 25, 2012.
- [35] M.-T. Tran, S.-H. Kim, H.-J. Yang, and G.-S. Lee, "Multi-task learning for medical image inpainting based on organ boundary awareness," *Appl. Sci.*, vol. 11, no. 9, 2021.
- [36] S. Chen, G. Bortsova, A. García-Uceda Juárez, G. Van Tulder, and M. De Bruijne, "Multi-task attention-based semi-supervised learning for medical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2019, pp. 457–465.
- [37] A. Khattar, S. Hegde, and R. Hebbalaguppe, "Cross-domain multi-task learning for object detection and saliency estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021, pp. 3639–3648.
- [38] W. Zhang, K. Wang, Y. Wang, L. Yan, and F.-Y. Wang, "A loss-balanced multi-task model for simultaneous detection and segmentation," *Neurocomputing*, vol. 428, pp. 65–78, 2021.
- [39] R. Ranjan, V. M. Patel, and R. Chellappa, "HyperFace: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 1, pp. 121–135, 2017.
- [40] R. Ranjan, S. Sankaranarayanan, C. D. Castillo, and R. Chellappa, "An all-in-one convolutional neural network for face analysis," in *Proc. 12th IEEE Int. Conf. Automat. Face Gesture Recognit. (FG)*, 2017, pp. 17–24.
- [41] G. Levi and T. Hassner, "Age and gender classification using convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, 2015, pp. 34–42.
- [42] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.
- [43] J. Yim, D. Joo, J. Bae, and J. Kim, "A gift from knowledge distillation: Fast optimization, network minimization and transfer learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 4133–4141.
- [44] W. Park, D. Kim, Y. Lu, and M. Cho, "Relational knowledge distillation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 3967–3976.
- [45] L. Zhang, J. Song, A. Gao, J. Chen, C. Bao, and K. Ma, "Be your own teacher: Improve the performance of convolutional neural networks via self distillation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2019, pp. 3713–3722.
- [46] T.-B. Xu and C.-L. Liu, "Data-distortion guided self-distillation for deep neural networks," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, no. 01, 2019, pp. 5565–5572.
- [47] S. Yun, J. Park, K. Lee, and J. Shin, "Regularizing class-wise predictions via self-knowledge distillation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 13 876–13 885.
- [48] S. Shin, Y. Yu, and K. Lee, "Enhancing low-resolution face recognition with feature similarity knowledge distillation," *arXiv preprint arXiv:2303.04681*, 2023.
- [49] Y. Liu, J. Cao, B. Li, W. Hu, J. Ding, and L. Li, "Cross-architecture knowledge distillation," in *Proc. Asian Conf. Comput. Vis. (ACCV)*, 2022, pp. 3396–3411.
- [50] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent.*
- [51] M. Luo, H. Wu, H. Huang, W. He, and R. He, "Memory-modulated transformer network for heterogeneous face recognition," *IEEE Trans. Inf. Forensics Secur.*, vol. 17, pp. 2095–2109, 2022.
- [52] Y. Zhong and W. Deng, "Face transformer for recognition," *arXiv preprint arXiv:2103.14803*, 2021.
- [53] A. George, A. Mohammadi, and S. Marcel, "Prepended domain transformer: Heterogeneous face recognition without bells and whistles," *IEEE Trans. Inf. Forensics Secur.*, vol. 18, pp. 133–146, 2022.
- [54] W. Su, Y. Wang, K. Li, P. Gao, and Y. Qiao, "Hybrid token transformer for deep face recognition," *Pattern Recognit.*, vol. 139, p. 109443, 2023.
- [55] W. Liu, Y. Wen, Z. Yu, and M. Yang, "Large-margin softmax loss for convolutional neural networks," *arXiv preprint arXiv:1612.02295*, 2016.
- [56] D. Hendrycks and K. Gimpel, "Gaussian error linear units (GELUs)," *arXiv preprint arXiv:1606.08415*, 2016.
- [57] K. Yuan, S. Guo, Z. Liu, A. Zhou, F. Yu, and W. Wu, "Incorporating convolution designs into visual transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2021, pp. 579–588.
- [58] Z. Wang, X. Cun, J. Bao, W. Zhou, J. Liu, and H. Li, "Uformer: A general u-shaped transformer for image restoration," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 17 683–17 693.
- [59] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 2818–2826.
- [60] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, December 2015.
- [61] Z. Zhu, G. Huang, J. Deng, Y. Ye, J. Huang, X. Chen, J. Zhu, T. Yang, J. Lu, D. Du *et al.*, "WebFace260M: A benchmark unveiling the power of million-scale deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021, pp. 10 492–10 502.

- [62] Z. Huang, J. Zhang, and H. Shan, "When age-invariant face recognition meets face age synthesis: A multi-task learning framework," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021, pp. 7282–7291.
- [63] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," in *Proc. Workshop Faces Real-Life Images, Detect., Alignment, Recognit.*, 2008.
- [64] S. Sengupta, J.-C. Chen, C. Castillo, V. M. Patel, R. Chellappa, and D. W. Jacobs, "Frontal to profile face verification in the wild," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, 2016, pp. 1–9.
- [65] T. Zheng and W. Deng, "Cross-Pose LFW: A database for studying cross-pose face recognition in unconstrained environments," Beijing University of Posts and Telecommunications, Tech. Rep. 18-01, 2018.
- [66] T. Zheng, W. Deng, and J. Hu, "Cross-Age LFW: A database for studying cross-age face recognition in unconstrained environments," *arXiv preprint arXiv:1708.08197*, 2017.
- [67] S. Moschoglou, A. Papaioannou, C. Sagonas, J. Deng, I. Kotsia, and S. Zafeiriou, "AgeDB: The first manually collected, in-the-wild age database," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, 2017, pp. 1997–2005.
- [68] C. Whitelam, E. Taborsky, A. Blanton, B. Maze, J. Adams, T. Miller, N. Kalka, A. K. Jain, J. A. Duncan, K. Allen *et al.*, "IARPA Janus benchmark-b face dataset," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, 2017, pp. 90–98.
- [69] Z. Cheng, X. Zhu, and S. Gong, "Low-resolution face recognition," in *Proc. Asian Conf. Comput. Vis. (ACCV)*, 2019, pp. 605–621.
- [70] M. Grgic, K. Delac, and S. Grgic, "SCFace—surveillance cameras face database," *Multimed. Tools Appl.*, vol. 51, pp. 863–879, 2011.
- [71] S. Yang, P. Luo, C.-C. Loy, and X. Tang, "WIDER FACE: A face detection benchmark," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 5525–5533.
- [72] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 7132–7141.
- [73] J. Zhang, H. Liu, K. Yang, X. Hu, R. Liu, and R. Stiefelhagen, "CMX: Cross-modal fusion for rgb-x semantic segmentation with transformers," *arXiv preprint arXiv:2203.04838*, 2022.
- [74] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *Proc. Annu. Meet. Assoc. Comput. Linguist.*, vol. 2019, 2019, p. 6558.
- [75] P. Li, S. Tu, and L. Xu, "Deep rival penalized competitive learning for low-resolution face recognition," *Neural Networks*, vol. 148, pp. 183–193, 2022.
- [76] J. C. L. Chai, T.-S. Ng, C.-Y. Low, J. Park, and A. B. J. Teoh, "Recognizability embedding enhancement for very low-resolution face recognition and quality estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp. 9957–9967.
- [77] X. Yin, Y. Tai, Y. Huang, and X. Liu, "FAN: Feature adaptation network for surveillance face recognition and normalization," in *Proc. Asian Conf. Comput. Vis. (ACCV)*, 2020.
- [78] H. Wang, S. Wang, and L. Fang, "Two-stage multi-scale resolution-adaptive network for low-resolution face recognition," in *Proc. 29th ACM Int. Conf. Multimedia*, 2022, pp. 4053–4062.
- [79] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2015, pp. 3730–3738.
- [80] E. M. Rudd, M. Günther, and T. E. Boulton, "MOON: A mixed objective optimization network for the recognition of facial attributes," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 19–35.
- [81] N. Zhuang, Y. Yan, S. Chen, and H. Wang, "Multi-task learning of cascaded CNN for facial attribute classification," in *Proc. Int. Conf. Pattern Recognit. (ICPR)*, 2018, pp. 2069–2074.
- [82] L. Mao, Y. Yan, J.-H. Xue, and H. Wang, "Deep multi-task multi-label CNN for effective facial attribute classification," *IEEE Trans. Affect. Comput.*, vol. 13, no. 2, pp. 818–828, 2020.
- [83] X. Wang, S. Zhang, S. Wang, T. Fu, H. Shi, and T. Mei, "Mis-classified vector guided softmax loss for face recognition," in *Proc. AAAI Conf. Artif. Intell.*, no. 07, 2020, pp. 12 241–12 248.
- [84] Y. Shi, X. Yu, K. Sohn, M. Chandraker, and A. K. Jain, "Towards universal representation learning for deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 6817–6826.
- [85] S. Li, J. Xu, X. Xu, P. Shen, S. Li, and B. Hooi, "Spherical confidence learning for face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021, pp. 15 624–15 632.
- [86] C.-Y. Low, A. B.-J. Teoh, and J. Park, "MIND-Net: A deep mutual information distillation network for realistic low-resolution face recognition," *IEEE Signal Process. Lett.*, vol. 28, pp. 354–358, 2021.
- [87] F. Boutros, N. Damer, F. Kirchbuchner, and A. Kuijper, "ElasticFace: Elastic margin loss for deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 1578–1587.
- [88] H. Wang and S. Wang, "Low-resolution face recognition enhanced by high-resolution facial images," in *Proc. 12th IEEE Int. Conf. Automat. Face Gesture Recognit. (FG)*. IEEE, 2023, pp. 1–8.
- [89] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 3–19.
- [90] R. K. Soni and N. Nain, "Synthetic data approach for unconstrained low-resolution face recognition in surveillance applications," in *Proc. of the Indian Conf. on Comput. Vis., Graph. and Image Process. (ICVGIP)*, 2022, pp. 1–6.
- [91] Y. Huang, P. Shen, Y. Tai, S. Li, X. Liu, J. Li, F. Huang, and R. Ji, "Improving face recognition from hard samples via distribution distillation loss," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Springer, 2020, pp. 138–154.
- [92] Z. Lu, X. Jiang, and A. Kot, "Deep coupled resnet for low-resolution face recognition," *IEEE Signal Process. Lett.*, vol. 25, no. 4, pp. 526–530, 2018.
- [93] S. Zagoruyko and N. Komodakis, "Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer," *arXiv preprint arXiv:1612.03928*, 2016.



**Niloufar Alipour Talemi** received the B.Sc. degree in electrical engineering from Shahid Beheshti University, Tehran, Iran, and the M.Sc. degree in electrical engineering digital electronic systems from University of Guilan, Rasht, Iran. She is currently pursuing the Ph.D. degree with West Virginia University, USA. Her current research interests include computer vision, pattern recognition, deep learning, machine learning, and their applications in biometrics.



**Hossein Kashiani** received the M.Sc. degree in electrical engineering from Iran University of Science and Technology, Tehran, Iran. Currently, he is pursuing the Ph.D. degree at West Virginia University, USA. He has authored over 15 publications, including journals and peer-reviewed conferences. His current research interests encompass computer vision, deep learning, machine learning, and their applications in biometrics. Furthermore, he has served as a reviewer for IEEE Transactions on Circuits and Systems for Video Technology, Pattern Recognition, Expert Systems With Applications, Knowledge-Based Systems, and other related journals and conferences.



**Nasser M. Nasrabadi** (Fellow, IEEE) received the B.Sc. (Eng.) and Ph.D. degrees in electrical engineering from the Imperial College of Science and Technology, University of London, London, U.K., in 1980 and 1984, respectively. In 1984, he was with IBM, U.K., as a Senior Programmer. From 1985 to 1986, he was with Philips Research Laboratory, New York, NY, USA, as a member of the Technical Staff. From 1986 to 1991, he was an Assistant Professor with the Department of Electrical Engineering, Worcester Polytechnic Institute, Worcester, MA, USA. From 1991 to 1996, he was an Associate Professor with the Department of Electrical and Computer Engineering, State University of New York at Buffalo, Buffalo, NY, USA. From 1996 to 2015, he was a Senior Research Scientist with the U.S. Army Research Laboratory. Since 2015, he has been a Professor with the Lane Department of Computer Science and Electrical Engineering. His current research interests are in image processing, computer vision, biometrics, statistical machine learning theory, sparsity, robotics, and neural networks applications to image processing. He has served as an Associate Editor for the IEEE Transactions on Image Processing, IEEE Transactions on Circuits and Systems for Video Technology, and the IEEE Transactions on Neural Networks. He is a Fellow of ARL and SPIE.