Joint Face Image Restoration and Frontalization for Recognition

Xiaoguang Tu, Jian Zhao, *Member, IEEE*, Qiankun Liu, Wenjie Ai, Guodong Guo, *Senior Member, IEEE*, Zhifeng Li, *Senior Member, IEEE*, Wei Liu, *Senior Member, IEEE*, and Jiashi Feng, *Member, IEEE*

***Abstract*— In real-world scenarios, many factors may harm face recognition performance, *e.g.*, large pose, bad illumination, low resolution, blur and noise. To address these challenges, previous efforts usually first restore the low-quality faces to high-quality ones and then perform face recognition. However, most of these methods are stage-wise, which is sub-optimal and deviates from the reality. In this paper, we address all these challenges jointly for unconstrained face recognition. We propose an Multi-Degradation Face Restoration (MDFR) model to restore frontalized high-quality faces from the given low-quality ones under arbitrary facial poses, with three distinct novelties. First, MDFR is a well-designed encoder-decoder architecture which extracts feature representation from an input face image with arbitrary low-quality factors and restores it to a high-quality counterpart. Second, MDFR introduces a pose residual learning strategy along with a 3D-based Pose Normalization Module (PNM), which can perceive the pose gap between the input initial pose and its real-frontal pose to guide the face frontalization. Finally, MDFR can generate frontalized high-quality face images by a single unified network, showing a strong capability of preserving face identity. Qualitative and quantitative experiments on both controlled and in-the-wild benchmarks demonstrate the superiority of MDFR over state-of-the-art methods on both face frontalization and face restoration.**

***Index Terms*— 3D based face normalization, multi-degradation face restoration, unconstrained face recognition.**

Manuscript received January 2, 2021; revised March 27, 2021; accepted April 29, 2021. Date of publication May 10, 2021; date of current version March 9, 2022. This work was supported in part by the Open Fund Project of Key Laboratory of Flight Technology and Flight Safety of CAFUC under Grant FZ2020KF10; in part by the National Science Foundation of China under Grant 62006244; in part by the Project of Comprehensive Reform of Electronic Information Engineering Specialty for Civil Aircraft Maintenance under Grant 14002600100017J172; in part by the Project of Civil Aviation Flight University of China under Grant J2018-56, Grant CJ2019-01, and Grant J2020-060; in part by the National Key Research and Development Program of China under Grant 2018AAA0103203; and in part by the Sichuan University Failure Mechanics & Engineering Disaster Prevention and Mitigation Key Laboratory of Sichuan Province Open Foundation under Grant 2020FMSCU02. This article was recommended by Associate Editor H. Shi. *(Corresponding author: Jian Zhao.)*

Xiaoguang Tu is with the Aviation Engineering Institute, Civil Aviation Flight University of China, Guanghan 618307, China (e-mail: xguangtu@outlook.com).

Jian Zhao is with the Institute of North Electronic Equipment, Beijing 100864, China (e-mail: zhaojian90@u.nus.edu).

Qiankun Liu is with Pensees Ptd Ltd., Singapore 117583 (e-mail: allen.liu@pensees.ai).

Wenjie Ai is with the School of Information and Communication Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China (e-mail: 201821011405@std.uestc.edu.cn).

Guodong Guo is with the Institute of Deep Learning, Baidu Research, Beijing 100080, China (e-mail: guodong.guo@mail.wvu.edu).

Zhifeng Li and Wei Liu are with the Tencent AI Lab, Shenzhen 518000, China (e-mail: michaelzfli@tencent.com; wl2223@columbia.edu).

Jiashi Feng is with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore 119260 (e-mail: elefjia@nus.edu.sg).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCSVT.2021.3078517.

Digital Object Identifier 10.1109/TCSVT.2021.3078517

I. Introduction

UNCONSTRAINED face recognition [1]–[5] is an important task in computer vision. In real-world applications, the enrolled faces in the gallery are usually frontal high-quality photos, while the probe faces may show large poses, bad illumination, low resolution, blur and noise, which may fail face recognition systems, as shown in Fig. 1. For this reason, the unconstrained face recognition is hardly considered to be solved.

Existing methods typically tackle the above challenges individually. For example, some works [6]–[9] synthesize a frontalized face from the profile face to achieve face recognition across poses. Though impressive face recognition accuracy has been achieved on lab-environment datasets such as Multi-PIE [10], their performance drops dramatically on benchmarks containing real-world samples such as IJB-C [11] with considerable low resolution and blurred data. Apart from the large pose, another challenge that may fail face recognition is the low quality of probe images, including bad illumination, low resolution, blur and noise. To improve the robustness of face recognition models against various image quality degradation factors, techniques like super-resolution [12], [13], [60], [61], [71], illumination normalization [14], [15], [69], deblurring [16], [17] and denoising [18] have been proposed. However, these methods merely focus on a single low-quality factor and are less effective for tackling the cases involving multiple challenging factors. For instance, super-resolution [12], [19], [20] methods are fragile for blurred faces.

Inspecting previous works on face frontalization [6], [7], [9], [21], [70] or face restoration [12], [13], [22], [23] from low-quality images, we observe a major problem: they are generally limited to just one of the contaminating factors and easily fail on multi-factor cases. Building a unified model to solve the large pose and low-quality problems at one shot is intuitively a straightforward solution, which however is not easy, considering the complex architecture design as well as the deep entanglement of different contaminating factors. To be more specific, a face restoration model takes the raw image as input while a face frontalization model needs additional input channels to encode the facial pose information, which means their architectures are not compatible. More

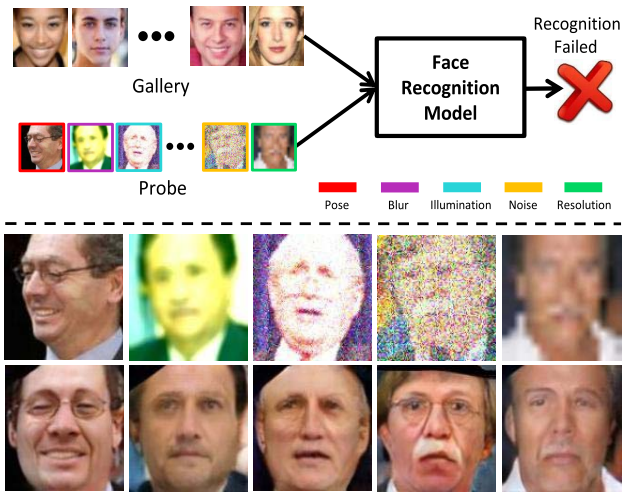


Fig. 1. Top: Illustration of challenging unconstrained face recognition. The presence of large poses and low quality fails the face recognition system. Each color in the legend indicates one influencing factor. Bottom: Example face images recovered by MDFR. *Row* 1: Input profile and low-quality face images (some images contain multiple contaminating factors). *Row* 2: Recovered frontalized high-quality faces.

importantly, the degraded facial details of input low-quality face images may fail at facial landmark detection and thus hinder face frontalization that heavily relies on landmark information. It is thus very crucial to obtain reliable facial landmarks from low-quality faces.

To tackle these challenges, we develop an **M**ulti-**D**egradation **F**ace **R**estoration (MDFR) model which is mainly driven by two task-specific generators during training, one for face restoration with multiple low-quality factors and the other for face frontalization. However, even when the two separate generators perform well on each specific task, there is a domain gap between their features in the identity metric space, which makes their identity representations inconsistent and hence affects the final face recognition. This identity inconsistency may in turn affect the results of face generation if these tasks are separately performed. To remove such a domain gap, we further propose an Task-Integrated (TI) training scheme to merge the learnings of these two tasks into a single one, enabling all contamination factors to be tackled by a unified network. Moreover, the TI training ensures face images to be frontalized from a single profile face image, *without requesting any priors such as input facial landmarks and target frontal landmarks*. Please note, although the proposed MDFR is able to jointly address face frontalization and face restoration from multiple degradation factors, it can also perform each task separately, such as face frontalization from high-quality inputs, face super-resolution, face deblurring and denoising and so on.

Structurally, our MDFR consists of two main components, *i.e.,* a dual-agent generator and a dual-agent prior-guided discriminator. The dual-agent generator learns to synthesize frontalized high-quality faces from the degraded inputs via two task-specific agents: a **F**ace **R**estoration sub-**N**et (FRN) and a **F**ace **F**rontalization sub-**N**et (FFN). FRN learns to recover facial details from low-quality images while FFN learns to rotate faces by leveraging the given target facial poses. The dual-agent discriminator consists of a **P**ose **C**onditioned **D**iscriminator (PCD) and an **I**dentity **C**onditioned **D**iscriminator (ICD), which are used to criticize the generated face images by referring to prior knowledge, making the outputs satisfy the input requirements. The well-designed dual-agent generator and dual-agent discriminator work together to achieve high-fidelity and identity-preserving frontal face generations from the low-quality inputs. The proposed training scheme is a two-stage training strategy, which contains the separate training and TI training. The 3D-based **P**ose **N**ormalization **M**odule (PNM) is used to guide real-frontal face generation during TI training, which merges face frontalization and restoration into a single unified network, so that the tasks could be blended into the same identity representation space and boost each other to learn more powerful representations for recognition. Our contributions are summarized as follows:

- We propose a novel **M**ulti-**D**egradation **F**ace **R**estoration (MDFR) model which recovers frontalized high-quality face images from given face images with arbitrary poses and multiple low-quality factors.
- We formulate face frontalization by pose residual learning and propose a 3D-based **P**ose **N**ormalization **M**odule (PNM), which normalizes 2D facial landmarks to real-frontal for guiding face frontalization learning.
- We develop an effective TI training strategy to merge face restoration and frontalizaiton into a unified network, which further enhances the output quality and improves the face recognition performance.
- Our method shows the ability to synthesize photorealistic frontal faces from low-quality faces with arbitrary poses and achieves remarkable face recognition performance under unconstrained environments.

II. Related Work

A. Face Frontalization

Earlier methods address face frontalization through 2D/3D local texture warping [6], [24] or statistical modeling [25]. In [26], Kan *et al.* use Stacked Progressive Auto-Encoders to rotate a profile face to frontal. Later, Hassner *et al.* [6] apply a single and unmodified 3D surface to approximate the shape of all the input faces. In [25], Sagonas *et al.* propose joint frontal face reconstruction and landmark detection by solving a constrained low-rank minimization problem. Such methods show effectiveness on face frontalization, but they tend to suffer a great performance degradation for profile and near-profile[1] faces due to severe texture loss and artifacts.

With the advent of GANs [27] in the field of computer vision, several GAN-based approaches [8], [21], [28], [29] have been proposed to synthesize frontal face images from the profile counterparts. In [7], Tran *et al.* propose the DR-GAN for frontal face generation for the first time. Then, TP-GAN [21] is proposed with a two-pathway structure and perceptual supervision. It leverages a well pre-trained face recognition model to guide an identity preserving inference of frontal views from profiles. PIM [8] aims to generate high-quality results through adding regularization terms to

[1]Faces with yaw angle greater than 60°.

learn more robust face representations. In [30], Tian *et al.* introduce a generation sideway to maintain the completeness of the learned embedding space and utilize both labelled and unlabeled data to further enrich the embedding space for realistic generations.

All these methods treat face frontalization as a direct 2D image-to-image translation problem and use frontal faces as ground-truths. However, we argue that such frontal faces are **not real ground-truths**, say near-frontal, as they may differ from the real world frontal poses at pixel level due to the variations in the collecting process. *If such pseudo ground-truths are used directly for training, the model can hardly converge well*. Even though CAPG-GAN [9] can alleviate this issue by offering a target pose, it still needs external assistance to obtain the target frontal pose. In this paper, we formulate face frontalization by pose residual learning and introduce a **P**ose **N**ormalization **M**odule (PNM) based on 3D face morphable model to offer real-frontal poses for face frontalization. PNM projects facial landmarks into a standard 3D space and automatically rotates them to real-frontal, serving as a pose target for refining the outputs towards frontal restoration. Once the training is done, our model can generate real-frontal face images from a single input **without requesting any target poses**.

B. Face Restoration

To restore high-quality face images from low-quality counterparts, methods such as super-resolution [19], [31], [32], denoising [33], [34], deblurring [35]–[37] and illumination normalization [38]–[40] have been proposed. For example, Kim *et al.* [31] utilize a very deep convolutional network by cascading many small filters to extract contextual information for super-resolution. Lai *et al.* [32] propose the LPSR Network to restore high-resolution images based on cascaded CNNs. In [19], [62], [63], the researchers recover high-resolution images with the aid of facial attributes, *i.e.*, facial landmark, parsing segmentation information and geometry prior estimation. Besides, some works [64]–[66] focus on architecture design to improve the super-resolution performance. For image deblurring or denoising, earlier methods [33]–[35] mainly exploit frequency-domain knowledge to restore band-pass frequency components. In [36], Svoboda *et al.* tackle this problem using custom CNN for the first time. Later, Xu *et al.* [37] perform joint deblurring and super-resolution for face and text images with GANs to learn a category-specific prior to solve this problem. In [16], Shen *et al.* exploit global and local semantic cues and incorporate perceptual and adversarial losses to restore photorealistic face images with finer details. Illumination mainly changes the weight of pixel values in a face image, which also leads to degraded recognition performance in extreme conditions. Prior methods addressing this problem are mainly based on holistic normalization [38], [39] or invariant feature extraction [14], [40]. Methods of the first category redistribute the intensities of the original image in a more normalized way, which is less prone to lighting changes; invariant feature extraction methods extract illumination-invariant features, such as the high-frequency and gradient-based components.

Though the mentioned methods are effective for face image enhancement, their performances usually drop when encountering cross degradation factors. For instance, using the deblurring to enhance low-resolution images would not help, and sometimes leads to worse performance because of overfitting to blurring factors.

III. Multi-Degradation Face Restoration

An overview of the proposed MDFR is shown in Fig. 2. As can be seen, our model consists of a Dual-Agent Generator, a Dual-Agent Prior-Guided Discriminator and a Pose Normalization Module. We now present each component in details.

A. Dual-Agent Generator

The dual-agent generator contains a **F**ace **R**estoration sub-**N**et (FRN) and a **F**ace **F**rontali-zation sub-**N**et (FFN), each consisting of an encoder to map the input into an embedding space, and a decoder to recover the embedding code to the target face, with the same architecture but taking in different inputs, as shown in Fig. 3.

FRN takes as input a low-quality face I_l and outputs a high-quality counterpart $\hat{I}_h$:

$$\hat{I}_h = G_1(I_l) = F_1(E_1(I_l)), \tag{1}$$

where E_1 and F_1 are the encoder and decoder of FRN, respectively.

FFN takes in three inputs, including a high-quality face image I_h, I_h's corresponding facial landmarks L_p and a target facial landmarks L_t. During the FFN separate training, L_t is the corresponding landmarks of the target near-frontal face image, while during the TI training,[2] L_t is the real-frontal facial landmarks L_f that normalized by PNM. We use 18 facial landmarks to indicate a facial pose and encode them as Gaussian heatmaps to represent L_p and L_t. Different from previous work [9] that uses the target pose to directly guide face generation, we feed L_p and L_t into the encoder and perform subtraction to obtain the pose residual. The intuition is that learning only the differences between poses avoids the redundant pose-irrelevant information, such as static backgrounds, which remains unchanged during the transformation. Therefore, the rotated face $\hat{I}_{ht}$ can be generated from I_h, conditioned on the pose residual:

$$\hat{I}_{ht} = G_2(I_h) = F_2(E_2(I_h) \oplus [E_2(L_p) - E_2(L_t)]), \tag{2}$$

where E_2 and F_2 are the encoder and decoder of FFN, respectively, and $\oplus$ denotes concatenation. To make the decoder easier reuse features of different spatial positions and facilitate feature propagation, we add dense connections in the decoder. The outputs of each block are connected to the first convolutional layers located in all subsequent blocks in the decoder. As the blocks have different feature resolutions, we upsample feature maps with lower resolutions when we use them as inputs into higher resolution layers.

[2]For the details of separate training and task-integrated training, please refer to Sec. 3.4.

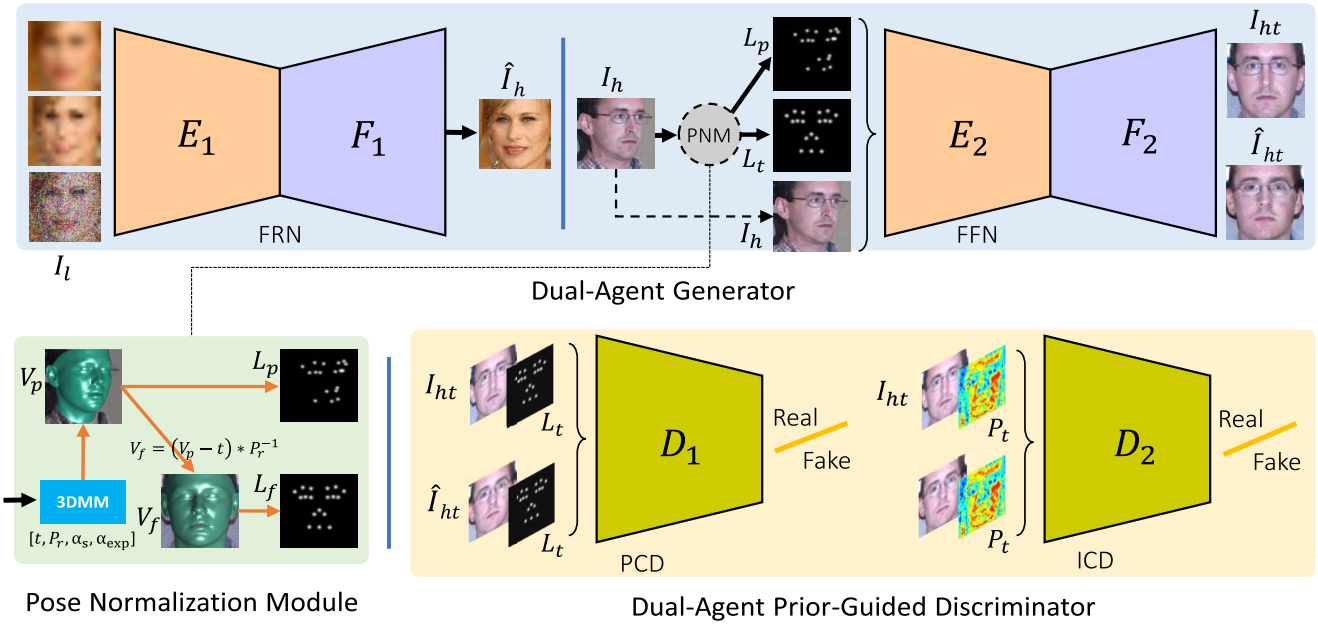


Fig. 2. Overview of MDFR. MDFR consists of two main components, *i.e.,* the dual-agent generator which includes a **F**ace **R**estoration sub-**N**et (FRN) and a **F**ace **F**rontalization sub-**N**et (FFN), and the dual-agent prior-guided discriminator which contains a **P**ose **C**onditioned **D**iscriminator (PCD) and an **I**dentity **C**onditioned **D**iscriminator (ICD). During the task-integrated training, the **P**ose **N**ormalization **M**odule (PNM) is proposed to offer real-frontal facial landmarks L_f, serving as the target facial landmarks L_t for guiding face generation.

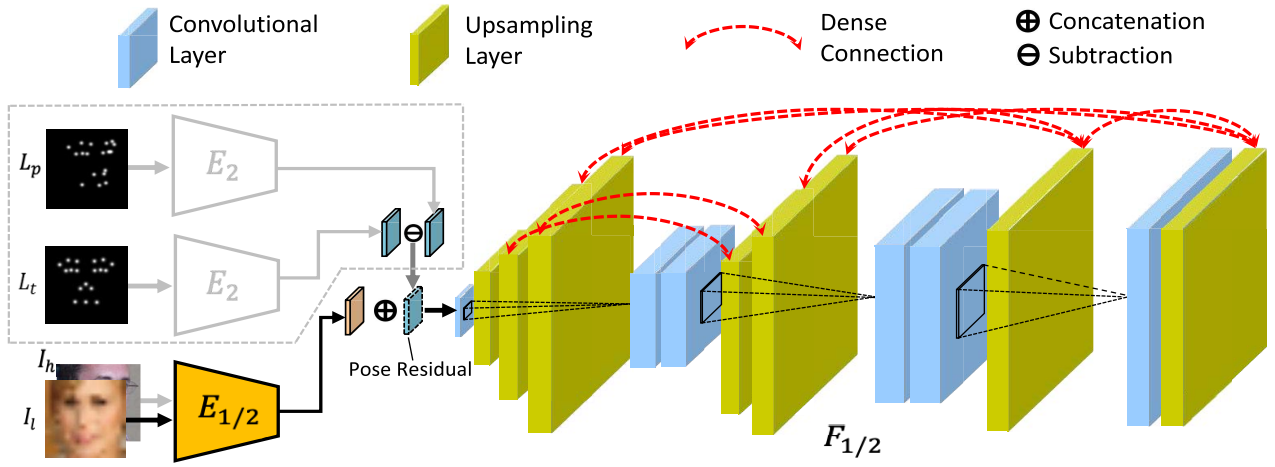


Fig. 3. Architecture of FRN/FFN. FRN and FFN share the same architecture. Decoder F_1 of FRN takes as input the representations of the input image encoded by E_1, while F_2 of FFN takes as input the concatenated representation of the input image and the pose residual between profile and frontal facial landmarks.

B. Pose Normalization Module

We devise a **P**ose **N**ormalization **M**odule (PNM) to perform pose normalization. Note that PNM is only used during the TI training. PNM offers a real-frontal pose with uniform face scale to guide face frontalization. Based on the 3D morphable model [41], the 3D vertices of a 2D face image can be expressed as a linear combination over a set of PCA bases as follows:

$$\boldsymbol{S} = \overline{\boldsymbol{S}} + \boldsymbol{A}_{\mathrm{s}}\boldsymbol{\alpha}_{\mathrm{s}} + \boldsymbol{A}_{\mathrm{exp}}\boldsymbol{\alpha}_{\mathrm{exp}}, \tag{3}$$

where $\overline{\boldsymbol{S}} \in \mathbb{R}^{3\times N}$ is the mean shape, $\boldsymbol{A}_{\mathrm{s}} \in \mathbb{R}^{3\times N}$ is the shape principle basis trained on the 3D face scans, $\boldsymbol{\alpha}_{\mathrm{s}} \in \mathbb{R}^{40}$ is the shape representation coefficient, $\boldsymbol{A}_{\mathrm{exp}} \in \mathbb{R}^{3\times N}$ is the expression principle basis, $\boldsymbol{\alpha}_{\mathrm{exp}} \in \mathbb{R}^{10}$ is the corresponding expression coefficient, and N is the number of vertices.

The 3D face vertices $\boldsymbol{S}$ can be projected onto a 2D image plane with scale orthographic projection to generate the 2D profile face from a specified viewpoint:

$$\boldsymbol{V}_p = f * \boldsymbol{Pr} * \boldsymbol{\Pi} * \boldsymbol{S} + \boldsymbol{t}, \tag{4}$$

where $\boldsymbol{V}_p$ denotes the 2D coordinates of the 3D vertices projected onto the 2D plane, f is the scale factor, $\boldsymbol{\Pi}$ is a fixed orthographic projection matrix, $\boldsymbol{Pr}$ is the rotation matrix, and $\boldsymbol{t}$ is the translation vector. As $\boldsymbol{Pr}$ and $\boldsymbol{t}$ indicate the rotation and offset variance, when removed from Eqn. (4), the normative frontal coordinates of a face image with arbitrary poses can

be obtained by

$$\boldsymbol{V}_f = f * \boldsymbol{\Pi} * \boldsymbol{S} = (\boldsymbol{V}_p - \boldsymbol{t}) * Pr^{-1}. \tag{5}$$

Here $\boldsymbol{V}_p$ and $\boldsymbol{V}_f$ store the profile and real-frontal dense 2D coordinates (x, y) for a given face image in a standard 3D space, with z coordinates removed. We use the state-of-the-art 3D face reconstruction method 2DASL [42] for 3DMM parameters regression and obtain dense coordinates (over 50, 000 points) from a given 2D face image. The 18 common key-points are sampled from V_p and V_f, respectively, to generate the Gaussian heatmaps L_p and L_f.

C. Dual-Agent Prior-Guided Discriminator

The discriminative loss for face super-resolution is firstly proposed in the work URDGN [67]. Following this idea, we propose to condition the discriminator with two kinds of additional prior knowledge, *i.e.,* the target facial landmarks and the frontal face identity feature map, letting the generated image approach the real one not only in terms of the target pose but also in terms of identity representation.

Our prior-guided discriminator is initialized with a VGG-11 [43] backbone. The first discriminator **P**ose **C**onditioned **D**iscriminator (PCD) takes the target pose L_t as condition and pairs with FFN's output $\hat{I}_{ht}$ (or the target high-quality face image I_{ht}), *i.e.,* $[\hat{I}_{ht}, L_t]$ vs. $[I_{ht}, L_t]$. The second discriminator **I**dentity **C**onditioned **D**iscriminator (ICD) takes the target face identity feature P_t as condition and pairs with $\hat{I}_{ht}$ or I_{ht}, *i.e.*, $[\hat{I}_{ht}, P_t]$ vs. $[I_{ht}, P_t]$. Following such conditions, F_2 would generate images $\hat{I}_{ht}$ which approach I_{ht}'s appearance and meanwhile satisfy the frontal-pose requirement. Specifically, PCD and ICD can not only distinguish real/fake of the outputs, but also learn the distinctions of facial pose and identity representation between the fake and real images.

D. Overall Training

Our overall training includes two phases: separate training and TI training. We now describe each training process in details. Algorithm 1 describes the whole process of the proposed training strategy.

1) Separate Training: We first train FRN and FFN separately. **FRN S**eparate (FRN-S) training restores high-quality face images from low-quality ones and **FFN S**eparate (FFN-S) training rotates profile faces to the target pose.

a) FRN-S Training: The identity-preserving loss $\mathcal{L}_{\text{id}}$ is employed during FRN-S training to preserve the identity information of the generated face image. We use a pre-trained face recognition model R_{id} to extract identity features and fix the parameters during training. $\mathcal{L}_{\text{id}}$ is defined as

$$\mathcal{L}_{\text{id}}(X, Y) = \left\| \frac{R_{\text{id}}(X)}{||R_{\text{id}}(X)||_2} - \frac{R_{\text{id}}(Y)}{||R_{\text{id}}(Y)||_2} \right\|_2^2, \tag{6}$$

where X is the input and Y is the output of FRN.

The restoration loss for FRN-S training is defined as

$$\mathcal{L}_{\text{r}}(I_h, \hat{I}_h) = ||I_h - \hat{I}_h||_2^2, \tag{7}$$

Algorithm 1 Overall Training

Separate training:

Phase 1: Train face restoration net FRN by $\mathcal{L}_{\text{FRN}}$, image pair $\{I_l, I_h\}$. I_l is the input low-quality face, I_h is I_l's high-quality face taken as the ground-truth.

Phase 2: Train face frontalization net FFN by $\mathcal{L}_{\text{FFN}}$, image pair $\{I_h, I_{ht}\}$ and landmark pair $\{L_h, L_t\}$. I_h is the high-quality face, I_{ht} is I_h's target face image (near frontal) taken as ground-truth, $\{L_h, L_t\}$ are their corresponding facial landmarks.

Task-Integrated training:

Fixing FFN's parameters.

Taking high-quality face I_h as image input of FFN, I_h's real frontal landmark L_f (normalized by PNM) as landmark input of FFN, generating high-quality and frontal face I_{hf}. Train face restoration net FRN by $\mathcal{L}_{\text{TI}}$, image pair $\{I_l, I_{hf}\}$. I_l is the low-quality face, I_{hf} is the output of FFN when taking I_h as input. I_h is I_l's corresponding high-quality face image. I_{hf} is used as the ground-truth for FRN training.

1: **while** not converge **do**
2: Choose one minibatch of N low-quality image I_l^i, $i = 1, \ldots, N$.
3: FRN outputs one minibatch of N restored images $\hat{I}_{hf}^i$ from the low-quality images I_l^i;
FFN outputs one minibatch of N images I_{hf}^i from I_h^i; from the restored images $G_1(I_l^i)$;
4: Update FRN by descending its stochastic gradient: $\nabla_{\theta_{FRN}} \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}_{\text{TI}}$.
5: **end while**

where $\hat{I}_h$ is the restored face image by FRN-S and I_h is the high-quality face image. The overall loss function for FRN-S is

$$\mathcal{L}_{\text{FRN}}(I_h, \hat{I}_h) = \mathcal{L}_{\text{r}}(I_h, \hat{I}_h) + \lambda_1 \mathcal{L}_{\text{id}}(I_h, \hat{I}_h), \tag{8}$$

where λ_1 is a weighting parameter balancing different losses.

b) FFN-S Training: The FFN-S training is supervised by four losses, *i.e.*, identity-preserving loss $\mathcal{L}_{\text{id}}$, frontalization loss $\mathcal{L}_{\text{f}}$, and conditional adversarial losses $\mathcal{L}_{\text{pcd}}$ and $\mathcal{L}_{\text{icd}}$. $\mathcal{L}_{\text{id}}$ for FFN-S is the same as that for FRN-S. For $\mathcal{L}_{\text{f}}$, we penalize the pixel-wise Euclidean distance between the rotated image and its corresponding ground-truth,

$$\mathcal{L}_{\text{f}}(I_{ht}, \hat{I}_{ht}) = ||I_{ht} - \hat{I}_{ht}||_2^2, \tag{9}$$

where I_{ht} is the target near-frontal face image and $\hat{I}_{ht}$ is the output of FFN.

The decoder of FFN takes as input the latent code of the profile face and the pose residual between the profile facial landmark heatmap L_p and the target facial landmark heatmap L_t, which provides strong guidance for face rotation. The dual-agent discriminator is used to refine $\hat{I}_{ht}$ according to the given prior knowledge. For PCD, D_1 takes L_t as condition and pairs with $\hat{I}_{ht}$ and I_{ht} as input. For ICD, D_2 takes the identity feature map P_t as condition and pairs with $\hat{I}_{ht}$ and I_{ht} as input. The condition adversarial losses $\mathcal{L}_{\text{pcd}}$ and $\mathcal{L}_{\text{icd}}$ can thus

be defined as

$$\mathcal{L}_{\text{pcd}} = \mathbb{E}_{I_p \in \mathcal{I}}[\log(D_1([L_t, I_{ht}])) + \log(1 - D_1([L_t, \hat{I}_{ht}]))],$$
$$\mathcal{L}_{\text{icd}} = \mathbb{E}_{I_p \in \mathcal{I}}[\log(D_2([P_t, I_{ht}])) + \log(1 - D_2([P_t, \hat{I}_{ht}]))]. \tag{10}$$

The overall loss function is a weighted sum of the above losses. The parameters of generator (θ_G), PCD (θ_P) and ICD (θ_I) are trained alternatively to optimize the following min-max problem:

$$\min_{\theta_G} \max_{\theta_P, \theta_I} \mathcal{L}_{\text{FFN}} = \mathcal{L}_{\text{f}}(I_{ht}, \hat{I}_{ht}) + \lambda_2 \mathcal{L}_{\text{id}}(I_{ht}, \hat{I}_{ht}) + \lambda_3(\mathcal{L}_{\text{pcd}} + \mathcal{L}_{\text{icd}}), \tag{11}$$

where λ_2 and λ_3 are weighting parameters trading off different losses.

2) Task-Integrated (TI) Training: After FRN and FFN are pre-trained, we perform the TI training, which we term as FRN Task-Integrated (FRN-TI) training. We use the output of FFN as ground-truth to train FRN. FRN-TI behaves somewhat like the distillation [44] process by transferring knowledge from the teacher model to the student. During FFN-S training, FFN learns to generate the target face according to the given target facial landmarks. During TI training, we use the real-frontal facial landmarks L_f that normalized by PNM to guide face generation for FFN, so the normalized pose representation is embedded into the outputs as well as the feature maps of FFN, which can serve as ground-truth for face frontalization by FRN. We thus formulate the training of FRN-TI at both image and feature levels. Specifically, we use FFN's output as well as its deep feature maps as the ground-truth to guide the learning of FRN. After the training is completed, PNM and FFN could be removed and the high-quality frontal face can be generated by merely using FRN without given the target facial landmarks. We train FRN-TI end-to-end. During the FRN-TI training, the parameters of FFN are fixed and only FRN is optimized.

We use the last dense block of FRN and FFN to perform feature level supervision via a **F**eature **A**lignment (FA) loss. As FRN and FFN have the same architecture, the FA loss $\mathcal{L}_{\text{FA}}$ can be easily defined as the mean squared error between their feature maps:

$$\mathcal{L}_{\text{FA}} = \frac{1}{N} \left\| \sum_{i=0}^{N} (B^i_{FFN}(G_1(I_l)) - B^i_{FRN}(I_l)) \right\|_2^2, \tag{12}$$

where I_l is an arbitrary low-quality face image, $B_{FRN}(\cdot)$ and $B_{FFN}(\cdot)$ are the feature representations from the last block of the decoders F_1 and F_2, respectively, N is the number of feature maps, and i is the i-th feature map.

Then, the overall loss function for TI training is

$$\mathcal{L}_{\text{TI}} = \mathcal{L}_{\text{r}}(\hat{I}_{ht}, G_1(I_l)) + \lambda_4 \mathcal{L}_{\text{id}}(\hat{I}_{ht}, G_1(I_l)) + \lambda_5 \mathcal{L}_{\text{FA}}, \tag{13}$$

where $\hat{I}_{ht}$ is the output of FFN, and λ_4 and λ_5 are weighting parameters among different losses. During FRN-TI training, FFN generates aligned images and features to guide face generation in FRN. After the task-integrated training is completed, FRN-TI is capable of generating frontalized high-quality faces by itself.

IV. Experiments

Implementation The size of face images is fixed as 128×128; constraint factors λ_1, λ_2, λ_3, λ_4, and λ_5 are fixed as 10^4, 10^4, 10^4, 0.1 and 1, respectively; batch size is set as 8; initial learning rate lr for FRN, FFN, PCD and ICD is 10^{-4}, 10^{-4}, 10^{-3} and 10^{-3}, respectively. We use 2DASL [42] for 3D face reconstruction. We initialize the face recognition network with ResNet-50 [49] to extract face identity features, which is pre-trained on CASIA-Webface [45] dataset using the AAM loss [50].

A. Datasets

a) CASIA-Webface: The CASIA-Webface [45] dataset contains 494,414 face images from 10,575 identities detected from the Internet. We use it for face recognition model training and FRN-S training. To generate the low-quality face images, we use color warp to randomly change the pixel's RGB value; Gaussian, uniform and average filters to randomly perform blurring; Gamma adjust to randomly change images' brightness level; bicubic to perform down-sampling and Gaussian noise to contaminate the images. The mean and standard deviation for Gaussian filter is a random number ranged from [0.1, 0.2] and [0.1, 0.2], respectively. The mean and standard deviation for color wrap is a random number ranged from [0.1, 0.2] and [0.1, 0.2], respectively. The low and high bound for Gamma adjust is a random number ranged from [0.1, 0.3] and [1, 3], respectively. The mean and standard deviation for Gaussian noise is a random number ranged from [0.1, 0.5] and [0.1, 0.5], respectively.

b) Multi-PIE: The CMU Multi-PIE [45] is the largest multi-view face recognition benchmark and is collected in four sessions. Following previous face frontalization works [8], [9], we conduct experiments under two settings: **Setting-1** only uses the images in session one, which contains 250 identities. The images with 11 poses within 90° of the first 150 identities are used for training. For testing, one frontal view with neutral expression and illumination is used as the gallery image for each of the remaining 100 identities and other images are used as probes. **Setting-2** uses the images with neutral expression from all four sessions, which contains 337 identities. The images with 11 poses within 90o of the first 200 identities are used for training. For testing, one frontal view with neural illumination is used as the gallery image for each of the remaining 137 identities and other images are used as probes. The training subsets of Multi-PIE are used for FFN-S training and FRN-TI training. The testing subsets are used for face frontalization evaluation.

c) LFW: The Labelled Faces in the Wild (LFW) [46] dataset contains 13,233 high-quality face images of 5,749 identities. The images are obtained by trawling the Internet followed by face centering, scaling and cropping based on bounding boxes provided by an automatic face locator. We use LFW for face frontalization and restoration testing.

d) IJB-C: The IARPA Janus Benchmark (IJB-C) [11] contains both still images and video frames from "in-the-wild" environment, which is believed to be the most unconstrained face dataset to date. We use it for face restoration testing.

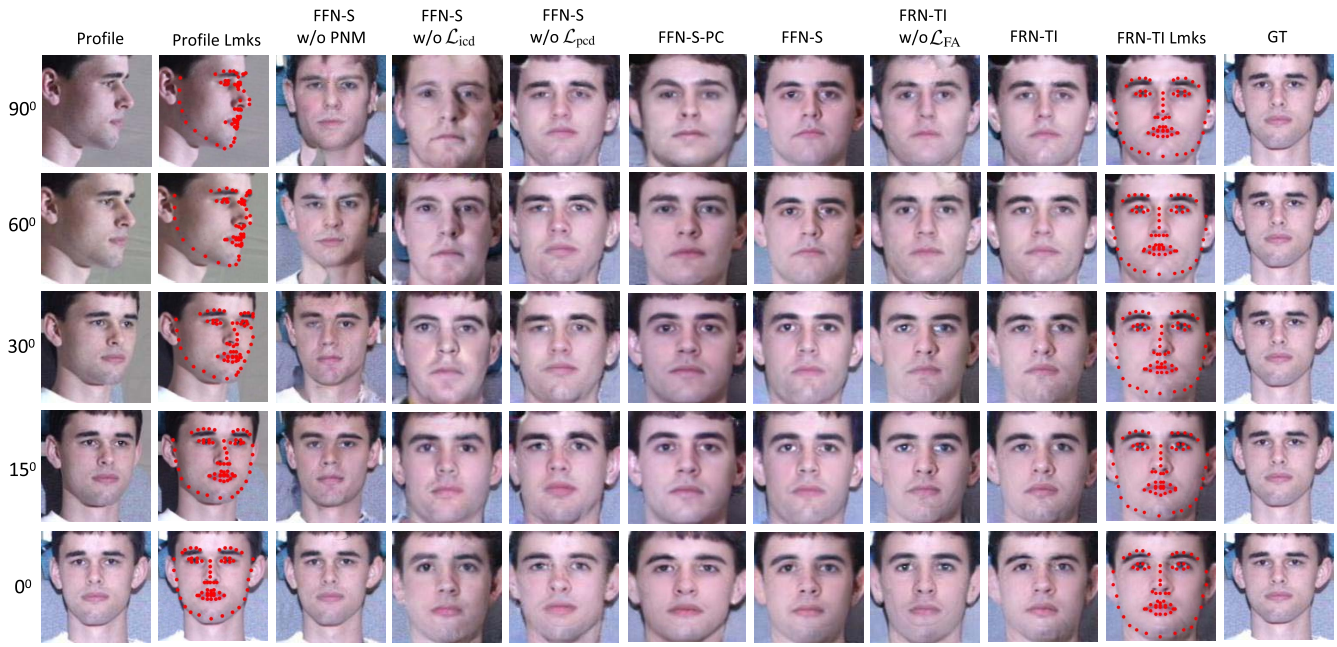


Fig. 4. Component analysis. Outputs of FFN-S and FRN-TI and their variants. *Col.* 2 and *Col.* 10 are the landmark detection results by FAN [68] for the profile images and FRN-TI outputs, respectively.

e) CelebA: CelebFaces Attributes Dataset (CelebA) is a large-scale face attribute dataset with more than 200K celebrity images, each with 40 attribute annotations. The images in this dataset cover large pose variations and background clutter. We use this dataset for model testing.

B. Evaluation on Face Frontalization

We first verify MDFR's effectiveness on face frontalization. During inference, we feed FFN-S and FRN-TI with the high-quality profile images, thus FRN-TI can be viewed as solely focusing on face frontalization.

1) Component Analysis: We investigate different architectures and loss combinations of FFN-S and FRN-TI to see their respective roles in face generation. We compare the results of FFN-S with three variants: w/o Pose Normalization Module (PNM), w/o $\mathcal{L}_{pcd}$ and w/o $\mathcal{L}_{icd}$, in each case. For FFN-S w/o PNM, we remove the facial landmark heatmaps from both FFN's encoder and PCD. We also compare the results with FRN-TI by using/removing $\mathcal{L}_{FA}$ to investigate $\mathcal{L}_{FA}$'s effectiveness when performing FRN-TI training. To evaluate the effectiveness of the pose residual learning module, we compare FFN-S with the modified one, which directly concatenates the pose features together. The modified pose learning strategy is named FFN-S-PC (Pose Concatenating).

The visualized results of all variants are shown in Fig. 4, where we observe all the variants perform well within a pose range of $\pm 30°$. However, when the pose is larger than 30°, FFN-S w/o PNM, $\mathcal{L}_{icd}$ and $\mathcal{L}_{pcd}$ all arise artifacts to some extent. PNM offers real-frontal landmarks to guide face frontalization. If removed, the model is unable to locate face regions correctly, making the recovered face distorted in part of face regions (see *Col.* 3). As $\mathcal{L}_{icd}$ helps refine the generated face images according to the given identity information, if removed, the output's identities are totally changed in larger poses (*Col. 4*). Although the identity-preserving loss $\mathcal{L}_{id}$ draws the input and output closer in the identity metric space, it is difficult to affect the appearance at image level. However, $\mathcal{L}_{icd}$ is able to directly refine the output to its original appearance by distinguishing real/fake between the concatenation pair of face image and identity feature map. For FFN-S w/o $\mathcal{L}_{pcd}$, the outputs (*Col.* 5) are not real-frontal but slightly rightward. This small angle drift may make the outputs involve some artifacts for large pose frontalization. FRN-TI achieves nearly the same visual results with FFN-S, which verifies the effectiveness of our proposed TI training. However, if $\mathcal{L}_{FA}$ is removed during FRN-TI training, the visual performance drops slightly, *e.g.,* the eyes that look strange in *Col.* 8. For FFN-S-PC (*Col.* 6), it achieves comparable results with FFN-S within small pose changes ($\leq$ 45°). However for the poses larger than 45°, especial for the pose of 90°, the result of FFN-S-PC is not as clear as that of FFN-S. The reason should be that the concatenation of pose heatmaps cannot well eliminate the background, which may influence the generation process, causing the degradation of the generated results.

The averaged rank-1 recognition rates of all variations are compared on **Setting-1** in Tab. I. The results on the original profile images serve as our baseline (*i.e.*, b). The results of all experiments are based on Resnet-50 [49] for face recognition. By comparing the results of *row* 2-8, we observe FFN-S and FRN-TI achieve the top-2 recognition rate among all the variants, which confirms the visual results in Fig. 4. For poses with 45^o, FFN-S w/o PNM, $\mathcal{L}_{icd}$, $\mathcal{L}_{pcd}$ and FRN-TI w/o $\mathcal{L}_{fa}$ even perform worse than the baseline, while FFN-S and FRN-TI achieve better than the baseline across all views. It seems the larger the head pose, the greater improvements

TABLE I

COMPONENT ANALYSIS. RANK-1 RECOGNITION RATES (%) UNDER MULTI-PIE [12] SETTING-1. B DENOTES THE PERFORMANCE OF RESNET-50 ON ORIGINAL PROFILE IMAGES

1	Methods	$\pm 90^o$	$\pm 75^o$	$\pm 60^o$	$\pm 45^o$	$\pm 30^o$	$\pm 15^o$
2	b	18.80	63.82	92.21	98.30	99.23	99.40
3	FFN-S w/o PNM	63.27	81.32	91.05	97.12	98.40	98.54
4	FFN-S w/o $\mathcal{L}_{\text{icd}}$	43.53	66.34	76.50	84.21	92.62	93.56
4	FFN-S w/o $\mathcal{L}_{\text{pcd}}$	63.70	82.04	91.41	97.42	98.58	98.54
6	FFN-S	80.11	89.27	94.94	99.06	99.51	99.92
7	FRN-TI w/o $\mathcal{L}_{\text{fa}}$	75.47	86.83	93.17	98.12	98.92	99.68
8	FRN-TI	79.83	88.52	94.20	98.78	99.36	99.90

TABLE II

RANK-1 RECOGNITION RATES (%) ACROSS DIFFERENT VIEWS ON MULTI-PIE SETTING-2. B DENOTES THE PERFORMANCE OF RESNET-50 ON ORIGINAL PROFILE IMAGES. "–" MEANS THE RESULT IS NOT REPORTED

1	Methods	$\pm 90°$	$\pm 75°$	$\pm 60°$	$\pm 45°$	$\pm 30°$	$\pm 15°$
2	b	15.50	55.10	85.92	97.13	98.41	98.62
3	MVP[51]	-	-	60.10	72.90	83.70	92.80
4	CPF[52]	-	-	61.90	79.90	88.50	95.00
5	DR-GAN [7]	-	-	83.20	86.20	90.10	94.00
6	TP-GAN [21]	64.64	77.43	87.72	95.38	98.06	98.68
7	CAPG-GAN [9]	66.05	83.05	90.63	97.33	99.56	99.82
8	FFN-S	70.20	85.31	91.81	98.05	99.82	99.83
9	FRN-TI	69.61	84.93	91.04	97.76	99.73	99.82

are achieved by FFN-S and FRN-TI. For 90^o yaw angle, FFN-S achieves 80.11% recognition rate, 61.31% higher than b, while FRN-TI achieves 61.03% higher recognition rate than the baseline.

2) Comparison With State-of-the-Arts: Tab. II shows recognition rate comparisons of FFN-S and FRN-TI with other methods on Multi-PIE **Setting-2**. FFN-S achieves the best recognition performance across all poses among all the compared methods, followed by FRN-TI. For poses within $\pm 45°$, FFN-S and FRN-TI achieve comparable recognition results with CAPG-GAN [9]. However, for poses larger than 45°, FFN-S and FRN-TI achieve much better recognition performance than CAPG-GAN. In particular, FFN-S and FRN-TI outperform CAPG-GAN by 4.15% and 3.56% under pose $\pm 90°$, respectively.

To further validate our model's generalizability to in-the-wild face images, we qualitatively compare the visual frontalization results of FFN-S and FRN-TI with Hassner *et al.* [6], DR-GAN [7], TP-GAN [21] and CAPG-GAN [9] on LFW datasets in Fig. 5. It is quite obvious that all the methods perform well on small pose cases. See *Rows* 1 & 2. However, for large pose cases, *i.e., Rows* 3, 4 and 5, Hassner *et al.*, DR-GAN, TP-GAN and CAPG-GAN distort the output faces and involve artifacts to some extent, and the face identities are also changed. FFN-S (*Col.* 6) and FRN-TI (*Col.* 7) still faithfully recover high-fidelity frontalized face images with finer local details and global face shapes while well preserving the identities.

Different from DR-GAN, TP-GAN and CAPG-GAN that need to provide the initial input or target facial landmarks to help face frontalization, FRN-TI does not need any facial landmark information in inference. During FRN-TI training, the normalized pose is embedded into FFN's outputs as well as the last blocks' feature maps to guide the learning of FRN, enabling FRN to spontaneously perceive the input and target poses and then map face regions to frontal.

C. Evaluation on Face Restoration

In unconstrained environment, the face image captured by a camera can be a high-quality one, which presents no degraded factors. A desired face restoration model is expected to preserve the original details for the high-quality inputs. Therefore, we first investigate the case when the inputs are high-quality face images. The visualization results of FRN-S and FRN-TI are shown in Fig. 6. FRN-S means the face restoration net is trained separately, while FRN-TI means the restoration net is trained by the task-integrated strategy. As can be seen, the outputs of FRN-S (*Row* 2) are almost the same as with the original ones, meaning FRN separately training is able to well preserve the fine details of the original high-quality face images, without any degradation of image quality. The results of FRN-TI (*Row* 3) further confirm the effectiveness of TI training on face frontalization.

Then we investigate the case when the inputs are low-quality face images. We verify the effectiveness of FRN-S and FRN-TI on face restoration using two datasets, LFW and IJB-C. We also compare with FRN-FFN where face restoration and frontalization are performed separately one by one. For LFW, we generate the low-quality face images using the same methods on CASIA-Webface, denoted as LFW-Lq dataset. IJB-C contains many low-quality samples collected from real-world applications. We use NIMA [53] to select low-quality ones for evaluation, denoted as IJBC-Lq dataset. We empirically set the confidence value τ of NIMA as 4.9 and 13, 729 image pairs are selected. As few existing methods address all the low-quality modalities at one shot,

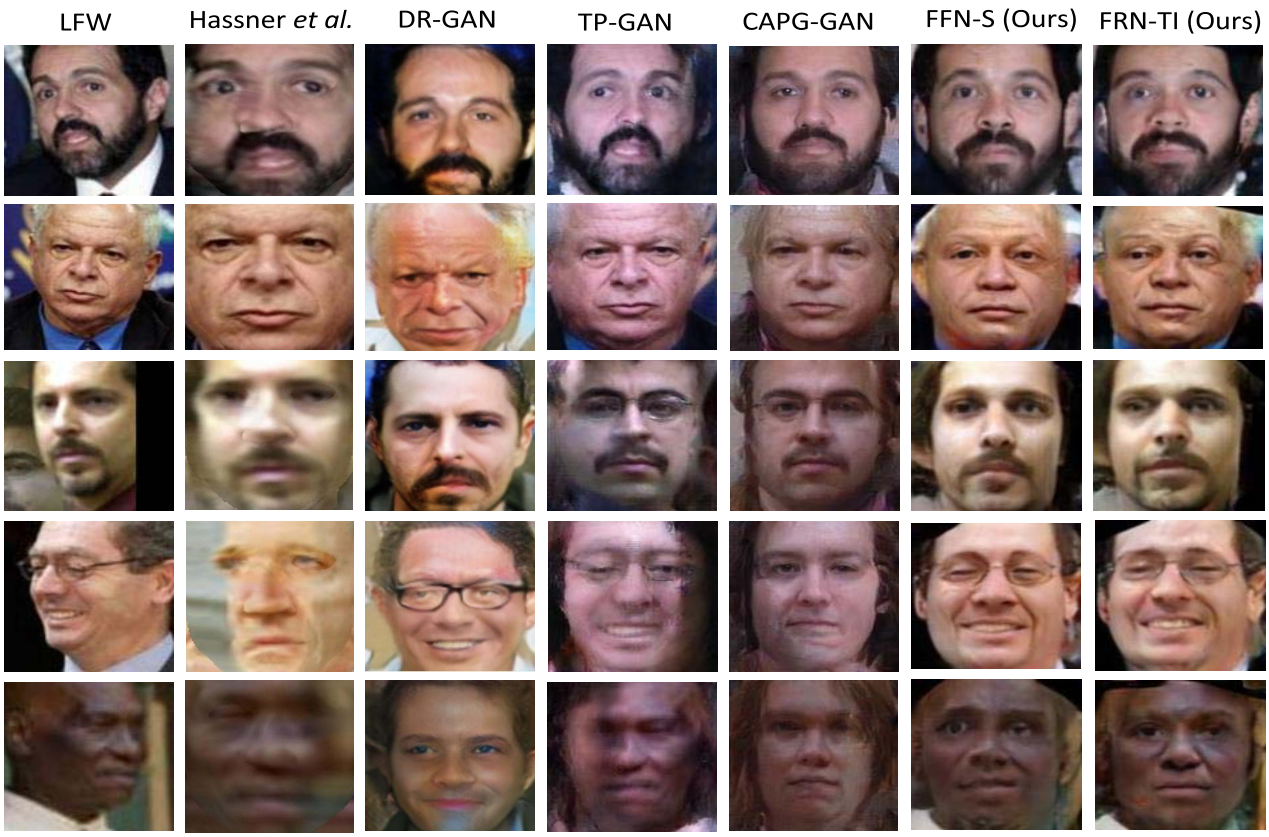


Fig. 5. The comparison results between our model and other popular face frontalization methods on LFW. The inputs to FFN are high-quality face images with arbitrary facial poses.

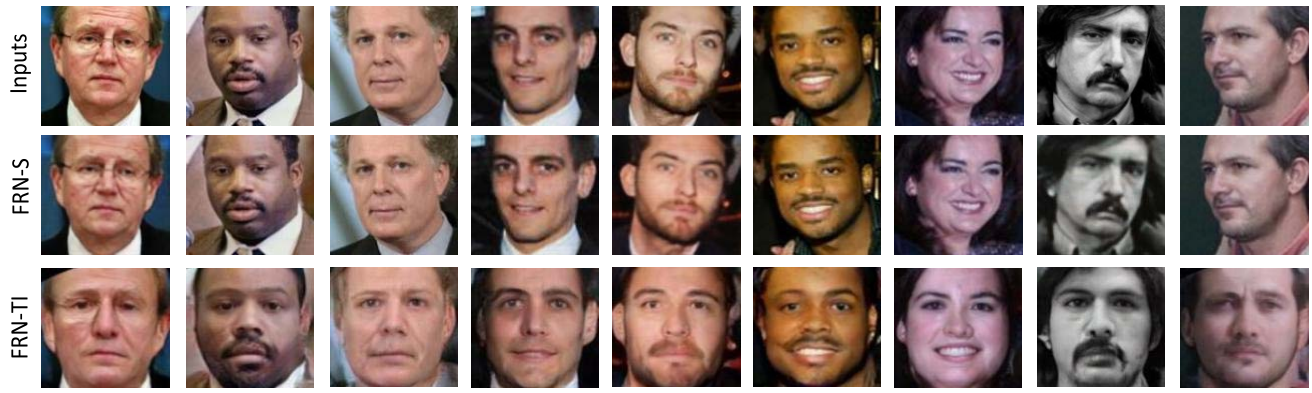


Fig. 6. The results of FRN-S and FRN-TI on LFW, respectively. The images of the first row are the high-quality inputs, the images of the second row are the outputs of FRN-S, the images of the third row are the outputs of FRN-TI.

we compare our results with state-of-the-art methods that focus on each modality separately. Specifically, the super-resolution methods ESRGAN [13] and SI [12], the denoising methods FFDNet [22] and PD-Denoising [23] as well as the deblurring method Deblur-GAN [54] are compared with FRN-S, FRN-FFN and FRN-TI. All the methods are trained on CASIA-Webface using the same low-quality processing methods.

We first compare the visual results of FRN-S, FRN-FFN and FRN-TI with other methods on LFW-Lq. For super-resolution, we up-sample the low-resolution inputs to 128×128 as the inputs of our models. The comparison results are shown in Fig. 7. By comparing *Rows* 3 & 5, we observe ESRGAN, SI and FRN-S are all effective to hallucinate high-resolution images from the low-resolution inputs. ESRGAN and SI achieve better visual results than FRN-S. This is because FRN-S only uses ℓ_2 pixel-wise loss, while ESRGAN and SI use other losses like adversarial or perceptual loss to recover more facial details. When training FFN, we use the adversarial loss for supervision, thus the outputs of FRN-FFN and FRN-TI are much sharper with more high-frequency details preserved than FRN-S. Moreover, ESRGAN and SI cannot address

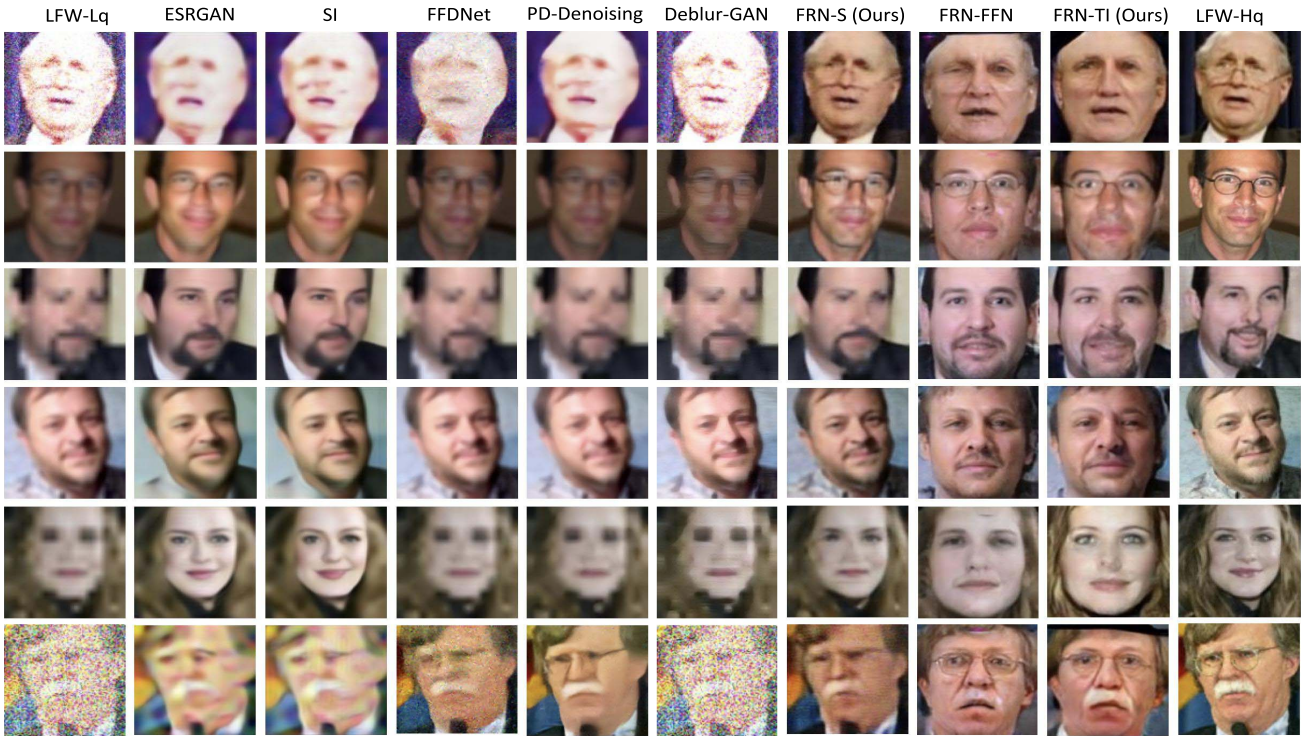


Fig. 7. Comparison of face restoration on LFW-Lq. The low-quality factors contain low resolution (*Rows* 3 & 5), bad illumination (*Rows* 1 & 2), image noise (*Rows* 1 & 6), and image blur (*Rows* 2 & 4). The last column shows the high-quality face images of LFW.

other low-quality cases, such as bad illumination (*Rows* 1 & 2), noise (*Rows* 1 & 6) and image blur (*Rows* 2 & 4), while FRN-S, FRN-FFN and FRN-TI are effective for all low-quality factors. Similar experimental results can also be obtained from denoising methods FFDNet and PD-Denoising, and also deblurring method Deblur-GAN. FFDNet and PD-Denoising can only address the problem of image noise (*Rows* 1 & 6), but they fail on others such as low resolution, bad illumination and image blur. For Deblur-GAN, it is only effective on blurred images (*Rows* 2 & 4) and fails on others. FRN-FFN and FRN-TI recover more facial details than FRN-S and meanwhile frontalize the profile face to its frontal pose. Compared with FRN-IT, FRN-FFN loses more facial details and slightly changes the face identity in some cases (*Rows* 2 & *5*). The reason behind this may be the identity representation inconsistency between the outputs of FRN-S and FFN-S. To investigate this challenge, we visualize the learned identity features in Fig. 8 and observe there exists a large margin between FRN-S domain and FFN-S domain. This identity inconsistency presents a gap when performing face frontalization from the outputs of FRN-S, which makes the model difficult to converge and affects the identity preservation.

From the results of Fig. 6 and Fig. 7, we can conclude that the proposed FRN is able to output high-quality face images, no matter the inputs are degenerated face images or high-quality ones. Since we overlay different degradation factors such as low-resolution, noise, blur and bad illumination during data processing, our restoration task is more challenging than the traditional ones that focus on only one degradation factor.

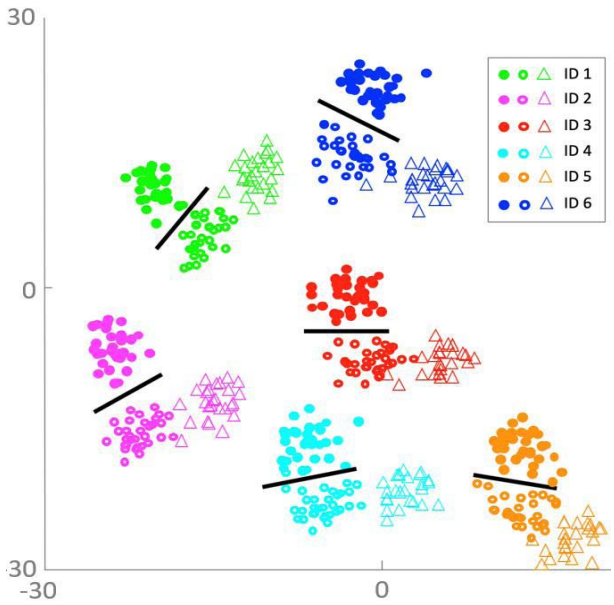


Fig. 8. Distribution of identity features from FRN-S domain (dot), FFN-S domain (circle) and FRN-TI domain (triangle). The identities are randomly selected from the training set. We observe a large gap between FRN-S and FFN-S domains in the identity metric space.

However, there are two limitations to the proposed FRN. The first one is that, our training data is driven by artificial degradation method, which may reduce the efficacy when addressing unseen data from real world. The second one is that, our method only uses the high-quality images to guide

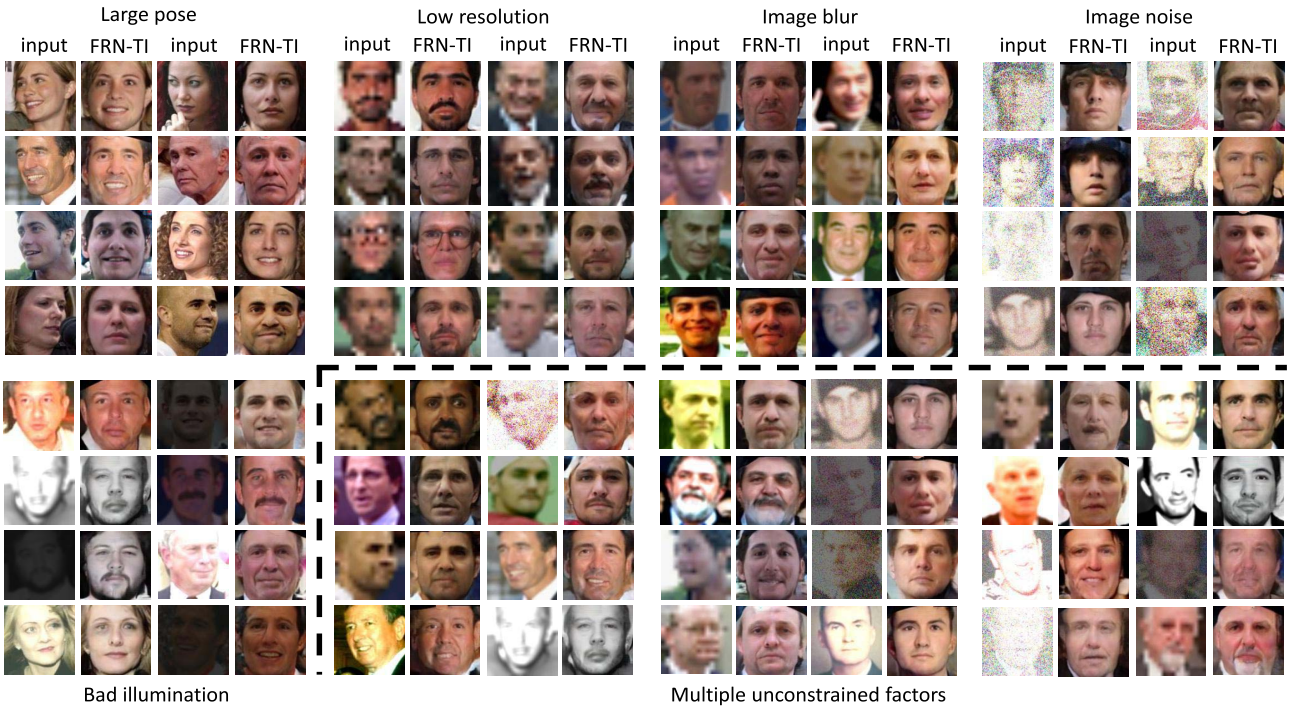


Fig. 9. Case study in extremely environments. Face images that present very large pose, very low resolution, image blur, image noise, bad illumination and multiple unconstrained factors are studied. The images are selected from LFW.

feature learning without considering the entangled influence of difference degradation factor in feature space, which may reduce the representation ability for the encoded features. Perhaps the disentangled representation learning is a better solution, which can remove the influence of other factors, and choose the best one for face restoration.

We follow the literatures [58], [59] for a qualitative comparisons with other state-of-the-art methods. For a fair comparison, we only consider the low-resolution factor during the retraining of our model. Since the works [58], [59] also focus on joint face frontalization and restoration and have reported their results, we copy the reported results from the work [59] and compare with ours on the dataset Multi-PIE and CelebA [57], using the average Peak Signal-to-Noise (PSNR) and the Structural SIMilarity (SSIM) scores. The qualitative comparison results are reported in Tab. III. As can be seen, the proposed MDFR achieves the best performance in both metrics on Multi-PIE and CelebA datasets. Specifically, MDFR outperforms VividGAN by 0.967 dB and 0.882 dB in PSNR on the datasets Multi-PIE and CelebA, respectively. For the metric SSIM, MDFR outperforms VividGAN by 0.022 and 0.014 on Multi-PIE and CelebA, respectively. The results indicate that the proposed MDFR achieves more authentic results than the other comparison methods.

We then report face recognition performance of FRN-S, FRN-FFN and FRN-TI and compare with other state-of-the-art face restoration methods on LFW-Lq and IJBC-Lq. Specifically, we use face restoration methods to recover high-quality images from low-quality ones and then use Resnet-50 to extract identity features for face recognition. The recognition results are reported in Tab. IV, where b is the result on

TABLE III
QUANTITATIVE COMPARISON RESULTS WITH OTHER METHODS

1		Multi-PIE		CelebA	
2		PSNR	SSIM	PSNR	SSIM
3	TANN [58]	24.426	0.831	25.690	0.870
4	VividGAN [59]	26.289	0.876	26.965	0.893
5	MDFR (ours)	**27.256**	**0.898**	**27.847**	**0.907**

low-quality images without restoration used for face recognition directly. From Tab. IV, we find FRN-TI achieves the best performance on both LFW-Lq and IJBC-Lq, followed by FRN-FFN and FRN-S. Compared with b, FRN-TI improves face recognition rate by 23.61% and 9.93% on LFW-Lq and IJBC-Lq, respectively; FRN-FFN improves by 22.29% and 7.88%, respectively; FRN-S improves by 21.89% and 6.17%, respectively. By comparing FRN-S and FRN-TI, FRN-TI's outputs contain more facial details which can be used for recognition. In addition, the enhanced faces are normalized to frontal, further boosting the performance for FRN-TI than FRN-S. As FRN-FFN performs high-quality face frontalization in two separate phases, the identity representation gap between these two tasks hinders them from linking well to each other, hence decreasing the final recognition results. As other compared methods can only address a certain low-quality factor, they report poor results on LFW-Lq. Even-though they are effective for a specific low-quality aspect, they may fail on others. In addition, no significant recognition improvements are observed using the compared methods on IJBC-Lq, which contains real world low-quality samples. Some of the methods, *e.g.,* ESRGAN and FFDNet even

TABLE IV

QUANTITATIVE COMPARISONS ON IDENTITY RECOGNIZABILITY. B DENOTES THE RESULTS ON LOW-QUALITY IMAGES WITHOUT RESTORATION. *Row* 3 REPORTS THE FACE VERIFICATION ACC (%) ON LFW-LQ; *Row* 4 REPORTS THE FACE VERIFICATION ACC (%) ON IJBC-LQ @FFR = 0.1

Methods	b	ESRGAN	SI	FFDNet	PD-Denoising	Deblur-GAN	FRN-S	FRN-FFN	FRN-TI
Input Size	128×128	12×14	12×14	128×128	128×128	128×128	128×128	128×128	128×128
LFW-Lq Acc.	71.62	78.21	83.50	75.21	78.35	76.32	93.51	93.91	**95.23**
IJBC-Lq Acc.	64.25	63.25	67.23	64.21	64.26	65.12	70.42	72.13	**74.18**

TABLE V

FACE VERIFICATION RESULTS (%) OF DIFFERENT METHODS ON LFW, AGEDB-30 [47] AND CFP-FP [48]. MDFR SERVES AS PRE-PROCESSING OF FACE IMAGES FOR EACH OF THE METHODS

Methods	SphereFace	SphereFace + MDFR	CosFace	CosFace + MDFR	ArcFace	ArcFace + MDFR
LFW [46]	99.42	99.48	99.51	99.54	99.53	**99.56**
AgeDB-30 [47]	91.70	91.81	94.56	94.68	95.15	**95.20**
CFP-FP [48]	94.38	94.51	95.44	95.59	95.56	**95.62**

slightly decrease the recognition performance compared with the baseline.

We then use our MDFR (FRN-TI) as face image preprocessing and use the face recognition methods, Sphereface [55], CosFace [56] and ArcFace [50] to perform face verification. The comparison results are shown in Tab. V. It is clear to see, when MDFR is used to pre-process the original face images, the recognition performance for each of the method is improved. Comparing with the state-of-the-art face recognition method ArcFace, if MDFR is used the recognition can be improved by 0.03%, 0.05% and 0.06% on LFW, AgeDB-30 and CFP-FP, respectively. Based on the observation, our ARFM can serve as a plug-and-play module to any state-of-the-art methods for high-performance unconstrained face recognition.

For more visualization results of FRN-TI, please refer to Fig. 9, where we have shown case study results in extremely environments, *i.e.*, very large pose (1^{st} panel), very low resolution (2^{nd} panel), image blur (3^{rd} panel), image noise (4^{th} panel), bad illumination (5^{th} panel) and multiple contamination factors ($6^{th} - 8^{th}$ panels) that contain at least two of the unconstrained factors. It is clear to see that our FRN-TI is able to generate photorealistic and identity-preserving frontal view faces from profile face images, no matter the profile images contain only one unconstrained factor or multiple unconstrained factors. In particular, when the input face images present low quality factors such as low resolution, noise, blur or bad illumination that may fail facial landmark detection, our model still can recover the high-quality and frontal view counterpart, which is more advanced than many of the landmark guided face frontalization methods.

V. CONCLUSION

In this paper, we propose a novel **M**ulti-**D**egradation **F**ace **R**estoration (MDFR) model. MDFR contains a dual-agent generator and a dual-agent prior-guided discriminator which cooperate with each other to learn frontalized high-quality faces from face images with multiple low-quality factors and arbitrary facial poses. The Pose Normalization Module (PNM) based on a 3D morphable model is proposed to normalize facial landmarks to real-frontal, serving as a unified criterion to guide the learning of MDFR. The Task-Integrated (TI) training is further developed to merge face restoration and frontalization into a unified network. When the TI training is done, MDFR is able to restore frontal and high-quality face images from low-quality ones with arbitrary facial pose without requesting any prior input landmarks. We demonstrate that the proposed unified framework outputs more visually realistic face images with more discriminative features preserved for face recognition than performing face restoration and frontalization separately. Comprehensive experiments on both controlled and "in-the-wil" face benchmarks illustrate the superiority of our method compared with other state-of-the-art face frontalization and face restoration methods.

REFERENCES

[1] B. F. Klare *et al.*, "Pushing the frontiers of unconstrained face detection and recognition: IARPA janus benchmark a," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1931–1939.

[2] J. Zhao, L. Xiong, J. Li, J. Xing, S. Yan, and J. Feng, "3D-aided dual-agent GANs for unconstrained face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 10, pp. 2380–2394, Oct. 2019.

[3] J. Zhao *et al.*, "Multi-prototype networks for unconstrained set-based face recognition," 2019, *arXiv:1902.04755*. [Online]. Available: http://arxiv.org/abs/1902.04755

[4] I. Masi, A. T. Trân, T. Hassner, G. Sahin, and G. Medioni, "Face-specific data augmentation for unconstrained face recognition," *Int. J. Comput. Vis.*, vol. 127, nos. 6–7, pp. 642–667, Jun. 2019.

[5] J. Zhao, J. Xing, L. Xiong, S. Yan, and J. Feng, "Recognizing profile faces by imagining frontal view," *Int. J. Comput. Vis.*, vol. 128, no. 2, pp. 460–478, Feb. 2020.

[6] T. Hassner, S. Harel, E. Paz, and R. Enbar, "Effective face frontalization in unconstrained images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 4295–4304.

[7] L. Tran, X. Yin, and X. Liu, "Disentangled representation learning GAN for pose-invariant face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1415–1424.

[8] J. Zhao *et al.*, "Towards pose invariant face recognition in the wild," in *CVPR*, Jun. 2018, pp. 2207–2216.

[9] Y. Hu, X. Wu, B. Yu, R. He, and Z. Sun, "Pose-guided photorealistic face rotation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8398–8406.

[10] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-PIE," *Image Vis. Comput.*, vol. 28, no. 5, pp. 807–813, May 2010.

[11] B. Maze *et al.*, "IARPA Janus benchmark-C: Face dataset and protocol," in *Proc. Int. Conf. Biometrics (ICB)*, Feb. 2018, pp. 158–165.

[12] K. Zhang *et al.*, "Super-identity convolutional neural network for face hallucination," in *Proc. ECCV*, 2018, pp. 183–198.

[13] X. Wang *et al.*, "ESRGAN: Enhanced super-resolution generative adversarial networks," in *Proc. ECCV*, 2018, pp. 63–79.

[14] N. P. Ramaiah, E. P. Ijjina, and C. K. Mohan, "Illumination invariant face recognition using convolutional neural networks," in *Proc. IEEE Int. Conf. Signal Process., Informat., Commun. Energy Syst. (SPICES)*, Feb. 2015, pp. 1–4.

[15] X. Tu, J. Gao, M. Xie, J. Qi, and Z. Ma, "Illumination normalization based on correction of large-scale components for face recognition," *Neurocomputing*, vol. 266, pp. 465–476, Nov. 2017.

[16] Z. Shen, W.-S. Lai, T. Xu, J. Kautz, and M.-H. Yang, "Deep semantic face deblurring," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8260–8269.

[17] I. Vasiljevic, A. Chakrabarti, and G. Shakhnarovich, "Examining the impact of blur on recognition by convolutional networks," 2016, *arXiv:1611.05760*. [Online]. Available: http://arxiv.org/abs/1611.05760

[18] F. Wang *et al.*, "The devil of face recognition is in the noise," in *Proc. ECCV*, 2018, pp. 765–780.

[19] Y. Chen, Y. Tai, X. Liu, C. Shen, and J. Yang, "FSRNet: End-to-end learning face super-resolution with facial priors," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2492–2501.

[20] H. Kong, J. Zhao, X. Tu, J. Xing, S. Shen, and J. Feng, "Cross-resolution face recognition via prior-aided face hallucination and residual knowledge distillation," 2019, *arXiv:1905.10777*. [Online]. Available: http://arxiv.org/abs/1905.10777

[21] R. Huang, S. Zhang, T. Li, and R. He, "Beyond face rotation: Global and local perception GAN for photorealistic and identity preserving frontal view synthesis," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2439–2448.

[22] K. Zhang, W. Zuo, and L. Zhang, "FFDNet: Toward a fast and flexible solution for CNN-based image denoising," *IEEE Trans. Image Process.*, vol. 27, no. 9, pp. 4608–4622, Sep. 2018.

[23] Y. Zhou *et al.*, "When AWGN-based denoiser meets real noises," 2019, *arXiv:1904.03485*. [Online]. Available: http://arxiv.org/abs/1904.03485

[24] X. Zhu, Z. Lei, J. Yan, D. Yi, and S. Z. Li, "High-fidelity pose and expression normalization for face recognition in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 787–796.

[25] C. Sagonas, Y. Panagakis, S. Zafeiriou, and M. Pantic, "Robust statistical face frontalization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3871–3879.

[26] M. Kan, S. Shan, H. Chang, and X. Chen, "Stacked progressive auto-encoders (SPAE) for face recognition across poses," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1883–1890.

[27] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. NIPS*, 2014, pp. 2672–2680.

[28] X. Yin, X. Yu, K. Sohn, X. Liu, and M. Chandraker, "Towards large-pose face frontalization in the wild," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3990–3999.

[29] J. Zhao *et al.*, "Dual-agent gans for photorealistic and identity preserving profile face synthesis," in *Proc. NIPS*, 2017, pp. 66–76.

[30] Y. Tian, X. Peng, L. Zhao, S. Zhang, and D. N. Metaxas, "CR-GAN: Learning complete representations for multi-view generation," 2018, *arXiv:1806.11191*. [Online]. Available: http://arxiv.org/abs/1806.11191

[31] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1646–1654.

[32] W.-S. Lai, J.-B. Huang, N. Ahuja, and M.-H. Yang, "Deep Laplacian pyramid networks for fast and accurate super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 624–632.

[33] A. Mosleh, J. P. Langlois, and P. Green, "Image deconvolution ringing artifact detection and removal via PSF frequency analysis," in *Proc. ECCV*. Berlin, Germany: Springer, 2014, pp. 247–262.

[34] H. Yue, X. Sun, J. Yang, and F. Wu, "CID: Combined image denoising in spatial and frequency domains using Web images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 2933–2940.

[35] S. Anwar, C. P. Huynh, and F. Porikli, "Class-specific image deblurring," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 495–503.

[36] P. Svoboda, M. Hradis, L. Marsik, and P. Zemcik, "CNN for license plate motion deblurring," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2016, pp. 3832–3836.

[37] X. Xu, D. Sun, J. Pan, Y. Zhang, H. Pfister, and M.-H. Yang, "Learning to super-resolve blurry face and text images," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 251–260.

[38] V. P. Vishwakarma, "Illumination normalization using fuzzy filter in DCT domain for face recognition," *Int. J. Mach. Learn. Cybern.*, vol. 6, no. 1, pp. 17–34, Feb. 2015.

[39] Y.-F. Yu, D.-Q. Dai, C.-X. Ren, and K.-K. Huang, "Discriminative multi-layer illumination-robust feature extraction for face recognition," *Pattern Recognit.*, vol. 67, pp. 201–212, Jul. 2017.

[40] B. Wang, W. Li, W. Yang, and Q. Liao, "Illumination normalization based on Weber's law with application to face recognition," *IEEE Signal Process. Lett.*, vol. 18, no. 8, pp. 462–465, Aug. 2011.

[41] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li, "Face alignment across large poses: A 3D solution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 146–155.

[42] X. Tu *et al.*, "3D face reconstruction from a single image assisted by 2D face images in the wild," 2020, *arXiv:1903.09359*. [Online]. Available: https://arxiv.org/abs/1903.09359

[43] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556

[44] S. Gupta, J. Hoffman, and J. Malik, "Cross modal distillation for supervision transfer," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2827–2836.

[45] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Learning face representation from scratch," 2014, *arXiv:1411.7923*. [Online]. Available: http://arxiv.org/abs/1411.7923

[46] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database forstudying face recognition in unconstrained environments," Tech. Rep., 2008.

[47] 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 Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 51–59.

[48] 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)*, Mar. 2016, pp. 1–9.

[49] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[50] 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)*, Jun. 2019, pp. 4690–4699.

[51] Z. Zhu, P. Luo, X. Wang, and X. Tang, "Multi-view perceptron: A deep model for learning face identity and view representations," in *Proc. NIPS*, 2014, pp. 217–225.

[52] J. Yim, H. Jung, B. Yoo, C. Choi, D. Park, and J. Kim, "Rotating your face using multi-task deep neural network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 676–684.

[53] H. Talebi and P. Milanfar, "NIMA: Neural image assessment," *IEEE Trans. Image Process.*, vol. 27, no. 8, pp. 3998–4011, Aug. 2018.

[54] O. Kupyn, V. Budzan, M. Mykhailych, D. Mishkin, and J. Matas, "DeblurGAN: Blind motion deblurring using conditional adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8183–8192.

[55] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "SphereFace: Deep hypersphere embedding for face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 212–220.

[56] H. Wang *et al.*, "CosFace: Large margin cosine loss for deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5265–5274.

[57] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3730–3738.

[58] X. Yu, F. Shiri, B. Ghanem, and F. Porikli, "Can we see more? Joint frontalization and hallucination of unaligned tiny faces," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 9, pp. 2148–2164, Sep. 2020.

[59] Y. Zhang, I. W. Tsang, J. Li, P. Liu, X. Lu, and X. Yu, "Face hallucination with finishing touches," *IEEE Trans. Image Process.*, vol. 30, pp. 1728–1743, 2021.

[60] Y. Zhang, I. W. Tsang, Y. Luo, C.-H. Hu, X. Lu, and X. Yu, "Copy and paste GAN: Face hallucination from shaded thumbnails," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 7355–7364.

[61] X. Yu and F. Porikli, "Hallucinating very low-resolution unaligned and noisy face images by transformative discriminative autoencoders," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3760–3768.

[62] X. Yu, B. Fernando, B. Ghanem, F. Porikli, and R. Hartley, "Face super-resolution guided by facial component heatmaps," in *Proc. ECCV*, 2018, pp. 217–233.

[63] X. Yu, B. Fernando, R. Hartley, and F. Porikli, "Super-resolving very low-resolution face images with supplementary attributes," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 908–917.

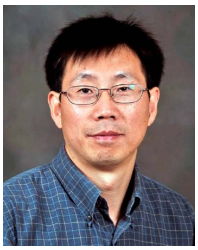
[64] X. Yu, F. Porikli, B. Fernando, and R. Hartley, "Hallucinating unaligned face images by multiscale transformative discriminative networks," *Int. J. Comput. Vis.*, vol. 128, no. 2, pp. 500–526, Feb. 2020.
[65] Y. Zhang, I. Tsang, Y. Luo, C. Hu, X. Lu, and X. Yu, "Recursive copy and paste GAN: Face hallucination from shaded thumbnails," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Feb. 23, 2021, doi: 10.1109/TPAMI.2021.3061312.
[66] X. Yu and F. Porikli, "Imagining the unimaginable faces by deconvolutional networks," *IEEE Trans. Image Process.*, vol. 27, no. 6, pp. 2747–2761, Jun. 2018.
[67] X. Yu and F. Porikli, "Ultra-resolving face images by discriminative generative networks," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2016, pp. 318–333.
[68] A. Bulat and G. Tzimiropoulos, "How far are we from solving the 2D & 3D face alignment problem? (And a dataset of 230,000 3D facial landmarks)," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 1021–1030.
[69] S. Du and R. K. Ward, "Adaptive region-based image enhancement method for robust face recognition under variable illumination conditions," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 20, no. 9, pp. 1165–1175, Sep. 2010.
[70] G.-S. Hsu, H.-C. Shie, C.-H. Hsieh, and J.-S. Chan, "Fast landmark localization with 3D component reconstruction and CNN for cross-pose recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 11, pp. 3194–3207, Nov. 2018.
[71] G. Gao, Y. Yu, J. Yang, G.-J. Qi, and M. Yang, "Hierarchical deep CNN feature set-based representation learning for robust cross-resolution face recognition," *IEEE Trans. Circuits Syst. Video Technol.*, early access, Dec. 3, 2020, doi: 10.1109/TCSVT.2020.3042178.

Xiaoguang Tu received the Ph.D. degree from the University of Electronic Science and Technology of China (UESTC) in 2020. From 2018 to 2020, he was a Visiting Scholar with the Learning and Vision Lab, National University of Singapore (NUS) under the supervision of Dr. Jiashi Feng. He is currently a Lecturer with the Aviation Engineering Institute, Civil Aviation Flight University of China. His research interests include convex optimization, computer vision, and deep learning.

Jian Zhao (Member, IEEE) received the bachelor's degree from Beihang University in 2012, the master's degree from the National University of Defense Technology in 2014, and the Ph.D. degree from the National University of Singapore in 2019. He is currently an Assistant Professor with the Institute of North Electronic Equipment, Beijing, China. His main research interests include deep learning, pattern recognition, computer vision, and multimedia analysis. He has published over 40 cutting-edge articles. He has received the Young Talent Support Project from the China Association for Science and Technology, and the Beijing Young Talent Support Project from Beijing Association for Science and Technology, the Lee Hwee Kuan Award (Gold Award) on PREMIA 2019, the Best Student Paper Award on ACM MM 2018, and the top-3 awards several times on worldwide competitions. He is the SAC of VALSE, and the Committee Member of CSIG-BVD. He has served as the Invited Reviewer of NSFC, T-PAMI, IJCV, NeurIPS (one of the top 30% highest-scoring reviewers of NeurIPS 2018), and CVPR.

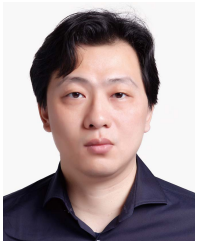
Qiankun Liu is currently an AI Scientist with Pensees Pte. Ltd., Singapore. His research interests include generative adversarial networks, optical flow estimation, and face recognition.

Wenjie Ai is currently pursuing the master's degree with the School of Information and Communication Engineering, University of Electronic Science and Technology of China (UESTC). His research interests include computer vision and deep learning, in particular, super resolution and deblur.

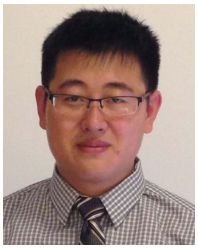


Guodong Guo (Senior Member, IEEE) received the Ph.D. degree in computer science from the University of Wisconsin, Madison, WI, USA. He is currently the Deputy Head of the Institute of Deep Learning, Baidu Research, and also an Associate Professor with the Department of Computer Science and Electrical Engineering, West Virginia University (WVU), USA. His research interests include computer vision, biometrics, machine learning, and multimedia. He received the North Carolina State Award for Excellence in Innovation in 2008, Outstanding Researcher at CEMR, WVU, from 2017 to 2018 and from 2013 to 2014, and New Researcher of the Year at CEMR, WVU, from 2010 to 2011. He was selected the People's Hero of the Week by BSJB under Minority Media and Telecommunications Council (MMTC) in 2013. Two of his articles were selected as The Best of FG'13 and The Best of FG'15, respectively.

Zhifeng Li (Senior Member, IEEE) received the Ph.D. degree from The Chinese University of Hong Kong in 2006. He was a Post-Doctoral Fellow with The Chinese University of Hong Kong and Michigan State University for several years. Before joining Tencent AI Lab, he was a Full Professor with the Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences. He is currently a Top-Tier Principal Researcher with Tencent AI Lab. His research interests include deep learning, computer vision and pattern recognition, and face detection and recognition. He is a fellow of the British Computer Society (FBCS). He is also serving on the Editorial Boards of *Neurocomputing* and IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY.



Wei Liu (Senior Member, IEEE) was a Research Staff Member of the IBM T. J. Watson Research Center, Yorktown Heights, NY, USA, from 2012 to 2015. He is currently a Distinguished Scientist of Tencent, China, and the Director of the Computer Vision Center, Tencent AI Lab. He has long been devoted to research and development in the fields of machine learning, computer vision, pattern recognition, information retrieval, and big data. He is a fellow of the International Association for Pattern Recognition (IAPR) and an Elected Member of the International Statistical Institute (ISI). He also serves on the editorial boards of IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, and *Pattern Recognition*.



Jiashi Feng (Member, IEEE) received the B.E. degree from the University of Science and Technology of China, Hefei, China, in 2007, and the Ph.D. degree from the National University of Singapore, Singapore, in 2014. From 2014 to 2015, he was a Post-Doctoral Researcher with the University of California. He is currently an Assistant Professor with the Department of Electrical and Computer Engineering, National University of Singapore. His current research interests include machine learning and computer vision techniques for large-scale data analysis.