DotFAN: A Domain-Transferred Face Augmentation Net

Hao-Chiang Shao, *Member, IEEE*, Kang-Yu Liu, Weng-Tai Su, *Member, IEEE*, Chia-Wen Lin, *Fellow, IEEE*, and Jiwen Lu, *Senior Member, IEEE*

***Abstract*— The performance of a convolutional neural network (CNN) based face recognition model largely relies on the richness of labeled training data. However, it is expensive to collect a training set with large variations of a face identity under different poses and illumination changes, so the diversity of within-class face images becomes a critical issue in practice. In this paper, we propose a 3D model-assisted domain-transferred face augmentation network (DotFAN) that can generate a series of variants of an input face based on the knowledge distilled from existing rich face datasets of other domains. Extending from StarGAN's architecture, DotFAN integrates with two additional subnetworks, i.e., face expert model (FEM) and face shape regressor (FSR), for latent facial code control. While FSR aims to extract face attributes, FEM is designed to capture a face identity. With their aid, DotFAN can separately learn facial feature codes and effectively generate face images of various facial attributes while keeping the identity of augmented faces unaltered. Experiments show that DotFAN is beneficial for augmenting small face datasets to improve their within-class diversity so that a better face recognition model can be learned from the augmented dataset.**

***Index Terms*— Face augmentation, convolutional neural networks, generative adversarial networks, domain knowledge transfer, generative model.**

I. Introduction

FACE recognition is one of the most considerable research topics in the field of computer vision. Benefiting from meticulously-designed CNN architectures and loss functions [1]–[3], the performance of face recognition models have been significantly advanced. The performance of a CNN-based face recognition model largely relies on the richness of labeled training data. However, collecting a training set with large variations of a face identity under different poses and illumination changes is very expensive, making the diversity of within-class face images a critical issue in practice. This is a considerable problem in developing a surveillance system for small/medium-size real-world applications. In such cases, each identity usually has only a few face samples (we call it **Few-Face learning problem**), making the data processing strategy equally important as the face recognition algorithm.

A face recognition model may fail, if the training set is too anemic to well train the model. To avoid this circumstance, our idea is to distill the knowledge within a rich data domain and then transfer the distilled knowledge to enrich an incomprehensive set of training samples in a target domain via domain-transferred augmentation. Specifically, we aim to train a composite network, which learns a *attribute-decomposed representation* of faces from rich face datasets, so that this network can generate face variants—each being associated with a different pose angle, a different facial expression, or a shading pattern due to a different illumination condition—of each face subject in an anemic dataset for the data augmentation purpose. Hence, we propose in this paper a **Domain-transferred Face Augmentation Net (DotFAN)**, that aims to learn the distributions of the faces of distinct identities in the feature space from rich training data so that it can augment face data, including frontalized neutral faces, during inference by transferring the knowledge it learned, as its design concept illustrated in Fig. 1.

The proposed DotFAN is a face augmentation approach through which any identity class—no matter a minority class or not—can be enriched by synthesizing face samples based on the knowledge learned from rich face datasets of other domains via domain transfer. To this end, DotFAN first learns a facial representation from rich datasets to decompose facial information into essential facial attribute codes that are vital for identity identification and face manipulation. Then, exploiting this attribute-decomposed facial representation, DotFAN can generate synthetic face samples neighboring to the input faces in the sample space so that the diversity of each face-identify class can be significantly enhanced. As a result, the performance of a face recognition model trained on the enriched dataset can be improved as well.

Utilizing two auxiliary subnetworks, namely a data-driven face-expert model (FEM) [4], [5] and a model-assisted face shape regressor (FSR), DotFAN operates in a model-assisted

Manuscript received July 8, 2020; revised June 2, 2021 and August 21, 2021; accepted October 6, 2021. Date of publication October 20, 2021; date of current version October 26, 2021. This work was supported in part by the Ministry of Science and Technology, Taiwan, under Grant MOST 109-2634-F-007-013 and Grant MOST 110-2634-F-007-015. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Julian Fierrez. *(Corresponding author: Chia-Wen Lin.)*

Hao-Chiang Shao is with the Department of Statistics and Information Science, Fu Jen Catholic University, New Taipei City 242062, Taiwan (e-mail: shao.haochiang@gmail.com).

Kang-Yu Liu is with Realtek Semiconductor Corporation, Hsinchu 300092, Taiwan (e-mail: asdfghj49888@gmail.com).

Weng-Tai Su is with the Department of Electrical Engineering, National Tsing Hua University, Hsinchu 300044, Taiwan (e-mail: wengtai2008@hotmail.com).

Chia-Wen Lin is with the Department of Electrical Engineering and the Institute of Communications Engineering, National Tsing Hua University, Hsinchu 300044, Taiwan, and also with the Electronic and Optoelectronic System Research Laboratories, Industrial Technology Research Institute, Hsinchu 310401, Taiwan (e-mail: cwlin@ee.nthu.edu.tw).

Jiwen Lu is with the Department of Automation, Tsinghua University, Beijing 100084, China (e-mail: lujiwen@tsinghua.edu.cn).

Digital Object Identifier 10.1109/TIP.2021.3120313

1941-0042 © 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

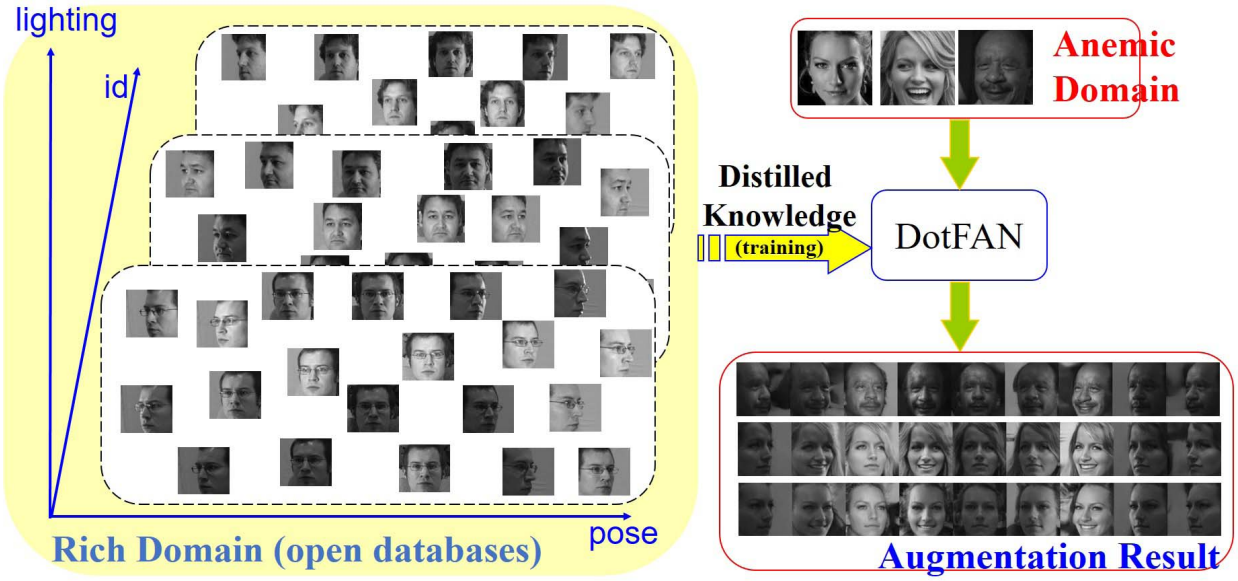


Fig. 1. DotFAN aims to enrich an anemic domain via identity-preserving face generation based on the knowledge, i.e., separated facial representation, distilled from data in a rich domain.

data-driven fashion. FEM is a purely data-driven subnetwork pretrained on a domain rich in face identities, whereas FSR is driven by a 3D face model and pretrained on another domain with rich poses and expressions. Hence, FEM ensures that the synthesized variants of an input face are of the same identity as the input, while FSR collaborating with illumination code enables the model to generate faces with various poses, lighting (shading) conditions, and different expressions. In addition, we use a 3D face model (e.g., 3DMM [6], [7]) to characterize face attributes related to pose and expression with only hundreds of parameters. Thereby, the size of FSR, and its training set of faces with labelled poses and expressions as well, is largely reduced, making it realizable with a light CNN with a significantly reduced number of parameters. Furthermore, the loss terms related to FEM and FSR act as regularizers during the training stage. This design prevents DotFAN from common issues in data-driven approaches, e.g., overfitting due to small training dataset.

Moreover, DotFAN is distinguishable from previous face augmentation and face synthesis methods. As for face augmentation, for example, Masi *et al.* [8] proposed their face-specific data augmentation method designed for maximizing the appearance variation of training images during training on-the-fly. To achieve this goal, the method focuses on two face-specific appearance variations, namely, pose and shape. It exploits face pose estimation, texture mapping, ray casting, precomputed projection matrix and 3D shape models to render new facial views. Although the method proposed in [8] well exploits graphics models to assist face recognition, it does not take into account facial appearance variations due to expressions and lighting conditions, thereby reducing its effectiveness for tackling the **Few-Face** problem in real-world surveillance systems. In addition, current face synthesis methods, including FaceID-GAN and the method proposed in [9], are not so suitable for the **Few-Face** problem, neither. Although the method in [9] elegantly incorporates 3DMM model to synthesize photorealistic faces, their method was primarily designed for generating faces of new identities. Moreover, FaceID-GAN regards its face-expert model as an additional discriminator that needs to be trained jointly with its generator and discriminator in an adversarial training manner. In FaceID-GAN's 3-player game strategy, its face-expert model assists its discriminator rather than its generator, and accordingly FaceID-GAN guarantees only the upper-bound of identity-dissimilarity. This design may prevent FaceID-GAN's face expert model from pretraining and impede the whole training speeds. Because FaceID-GAN cannot be pretrained on a rich-domain data, this fact makes it difficult to transfer knowledge from a rich dataset to another in an on-line learning manner.

On the contrary, DotFAN, a GAN-based domain-transferred face augmentation network, utilizes a concatenation of attribute-decomposed facial features to synthesize faces with appearance variations in poses, expressions and shadows. Also, DotFAN regards its FEM as a regularizer to guarantee that the identity information is not altered by the generator. As a result, its FEM can be pretrained on a rich dataset and play a role of an inspector in charge of overseeing identity-preservability. This design not only carries out the identity-preserving face generation task, but also stabilizes and speeds up the training process by not intervening the competition between generator and discriminator. In sum, DotFAN has following four primary contributions.

- We are the first to propose a domain-transferred face augmentation scheme. The proposed scheme can effectively transfer the knowledge distilled from a rich domain to an anemic domain, while preserving the identity of augmented faces in the target domain.
- DotFAN provides a learning-based universal solution for the **Few-Face** problem. Specifically, i) when a face

recognizer is re-trainable, DotFAN enriches the **Few-Face Set** by data augmentation, and then the recognizer can be re-trained on the enriched set to improve its performance; and, ii) if the face recognizer is pretained on an incomprehensive dataset (e.g., with mainly frontal faces and/or neutral illumination) and is NOT re-trainable, DotFAN can assist the recognizer by frontalizing/neutralizing a to-be-recognized face.

- Through a concatenation of facial attribute codes learned separately from existing face datasets, DotFAN offers a unique unified framework that can incorporate prominent face attributes (e.g. id-information, pose, illumination, shape, and expression codes derived by different subnetworks) for face recognition and can be easily extended to other face related tasks.
- DotFAN well beats the state-of-the-arts by a significant gain margin in face recognition application with small-size training data available. This makes it a powerful tool for low-shot learning applications.

II. Related Work

Recently, various algorithms have been proposed to address the issue of small sample size with dramatic variations in facial attributes in face recognition [10]–[13]. This section reviews works on GAN-based image-to-image translation, face generation, and face frontalization/rotation techniques related to face augmentation.

A. GAN-Based Image-to-Image Translation

GAN and its variants have been widely adopted in a variety of fields, including image super-resolution, image synthesis, image style transfer, and domain adaptation. DCGAN [14] incorporates deep CNNs into GAN for unsupervised representation learning. DCGAN enables arithmetic operations in the feature space so that face synthesis can be controlled by manipulating attribute codes. The concept of generating images with a given condition has been adopted in succeeding works, such as Pix2pix [15] and CycleGAN [16]. Pix2pix requires pair-wise training data to derive the translation relationship between two domains, whereas CycleGAN relaxes such limitation and exploits unpaired training inputs to achieve domain-to-domain translation. After CycleGAN, StarGAN [10] addresses the multi-domain image-to-image translation issue. With the aids of a multi-task learning setting and a design of domain classification loss, StarGAN's discriminator minimizes only the classification error associated to a known label. As a result, the domain classifier in the discriminator can guide the generator to learn the differences among multiple domains. Recently, an attribute-guided face generation method based on a conditional CycleGAN was proposed in [11]. This method synthesizes a high-resolution face based on an low-resolution reference face and an attribute code extracted from another high-resolution face. Consequently, by regarding faces of the same identity as one sub-domain of faces, we deem that face augmentation can be formulated as a multi-domain image-to-image translation problem that can be solved with the aid of attribute-guided face generation strategy.

Additionally, although DotFAN is skeletally an extension of StarGAN, DotFAN is conceptually different from StarGAN. Specifically, DotFAN is a framework specialized for domain-knowledge-transferred face augmentation, whereas StarGAN, as well as other GAN-based face synthesizers, does not have a proper network structure and a suitable loss function design that supports the concept of domain knowledge transfer particularly to face synthesis.

B. Face Frontalization and Rotation

We regard the identity-preserving face synthesis task as an inverse problem of the face frontalization technique used to synthesize a frontal face from a face image with arbitrary pose variation. Typical face frontalization and rotation methods synthesize a 2D face via 3D surface model manipulation, including pose angle control and facial expression control, such as FFGAN [17], FaceID-GAN [7], ExpNet [18], FacePoseNet [19], Rotate-and-render [20], and FFWM [21]. Still, some designs utilize specialized sub-networks or loss terms to reach the goal. For example, based on TPGAN [22], the pose invariant module (PIM) proposed in [23] contains an identity-preserving frontalization sub-network and a face recognition sub-network; the CNN proposed in [24] establishes a dense correspondence between paired non-frontal and frontal faces; and, the face normalization model (FNM) proposed in [5] involves a face-expert network, a pixel-wise loss, and a face attention discriminators to generate a faces with canonical-view and neutral expression. Finally, some methods approached this issue by means of disentangled representations [25], [26]. For example, DR-GAN [25] utilizes an encoder-decoder structure to learn a disentangled representation for face rotation, whereas CAPG-GAN [26] adopts a two-discriminator framework to learn simultaneously pose and identity information. Therefore, we integrate the encoder-decoder framework with DotFAN to learn a *attribute-decomposition representation* for face augmentation.

C. Data Augmentation for Face Recognition

To facilitate face recognition, there are several face normalization and data augmentation methods. Face normalization methods aim to align face images by removing the volatility resulting from illumination variations, changes of facial expressions, and different pose angles [5], whereas the data augmentation method attempts to increase the richness of face images, often in aspects of pose angle and illumination conditions, for the training routine. To deal with illumination variations, conventional approaches utilized either physical models, e.g. Retinex theory [27], or 3D reconstruction strategy to remove/correct the shadow on a 2D image [28], [29]. Moreover, to mitigate the influence brought by pose angles, two categories of methods were proposed, namely pose-invariant face recognition methods and face rotation methods. While the former category focuses on learning pose-invariant features from a large-scale dataset [30]–[32], the latter category, including face frontalization techniques, aims to learn the relationship between rotation angle and resulting face image via a generative model [7], [17], [22], [23], [25], [26], [33], [34]. Because face rotation methods are designed

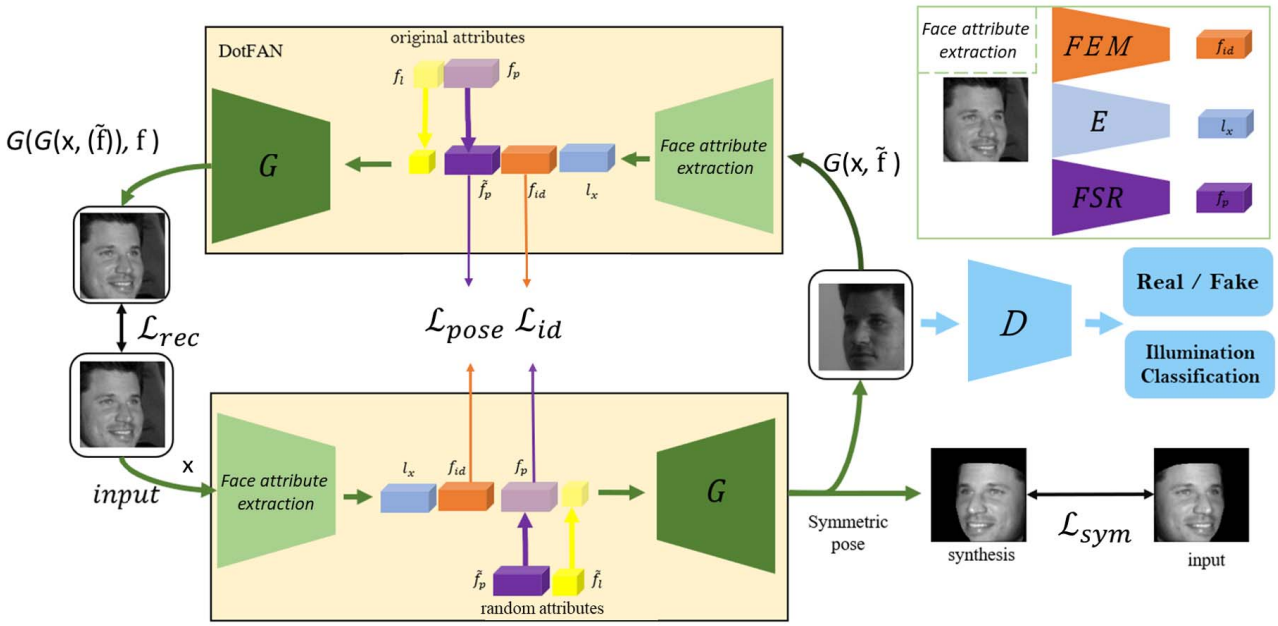


Fig. 2. Data flow of DotFAN's training process. FEM and FSR are independently pre-trained subnetworks, whereas E, G, and D are trained as a whole. $\tilde{f}_p$ and $\tilde{f}_l$ denote respectively a pose code and an illumination code randomly given in the training routine; and, f_l is the ground-truth illumination code provided by the training set. For inference, the data flow begins from $\mathbf{x}$ and ends at $y = G(\mathbf{x}, \tilde{\mathbf{f}})$. Note that $\tilde{\mathbf{f}} = [l_x, f_{id}, \tilde{f}_p, \tilde{f}_l]$, and $\mathbf{f} = [E(G(\mathbf{x}, \tilde{\mathbf{f}})), \Phi_{fem}(G(\mathbf{x}, \tilde{\mathbf{f}})), f_p, f_l]$.

to increase the diversity of the view-points of face image data, they are also beneficial for augmentation tasks. Still, face swapping methods provide another direction for face augmentation tasks. For example, FaceShifter [35] was designed to swap the face region of one image to one another photo-realistically by exploiting the GAN framework with an identity-preserving loss. FaceShifter feeds face identity feature jointly with attribute codes into its generator in a multi-level fashion to produce swapped faces with high-fidelity, and it adopts the cosine similarity to guarantee the (directional) similarity of its id-feature. Nevertheless, FaceShifter takes paired input faces during deployment and guarantees only the visual fidelity of the swapped face via the cosine similarity rather than the consistency of extracted face identity feature via L2-norm, and hence it may not be applied to face augmentation tasks directly.

Based on these meticulous designs, DotFAN is implemented as an extension of StarGAN, involving an encoder-decoder framework and two sub-networks for learning attribute codes separately, and triggered by several loss terms, including reconstruction loss and domain classification loss, as will be elaborated later.

III. Proposed Method

DotFAN is a framework to synthesize face images of one domain based on the knowledge, i.e., attribute-decomposed facial representation, learned from others. Given an input face $\mathbf{x}$, the generator G of DotFAN is trained to synthesize a face $G(\mathbf{x}, \mathbf{f})$ based on an input attribute code $\mathbf{f}$ comprising i) a general latent code $l_{\mathbf{x}} = E(x)$ extracted from $\mathbf{x}$ by the general facial encoder, ii) an identity code f_{id} indicating the face identity, iii) an attribute code f_p describing facial attributes including pose angle and facial expressions, and iv) an illumination code f_l. Through this design, a face image can be embedded via a concatenation of these attribute codes, i.e., $\mathbf{f} = [l_{\mathbf{x}}, f_{id}, f_p, f_l]$. Fig. 2 depicts the flow-diagram of DotFAN, and each component will be elaborated in following subsections.

A. Attribute-Decomposed Facial Representation

To obtain a decomposed representation, the attribute code $\mathbf{f}$ used by DotFAN for generating face variants is derived collaboratively by a general facial encoder E, a face-expert sub-network FEM, a shape-regression sub-network FSR, and an illumination code f_l. FEM and FSR are two well pre-trained sub-networks. FEM learns to extract identity-aware features from faces (of each identity) with various head poses and facial expressions, whereas FSR aims to learn pose features based on a 3D model. The illumination code is a 14×1 one-hot vector specifying 1 label-free case (corresponding to data from CASIA [36]) and 13 illumination conditions (associated with selected Multi-PIE dataset [37]).

1) Face-Expert Model (FEM): FEM Φ_{fem}, architecturally a ResNet-50 trained via ArcFace loss [2], enables DotFAN to extract and then transplant the face identity from an input source to synthesized face images. Though conventionally face identity extraction is considered as a classification problem and optimized by using a cross-entropy loss, recent methods, e.g., CosFace [3] and ArcFace, proposed adopting angular information instead. ArcFace maps face features onto a unit hyper-sphere and adjust between-class distances by using a pre-defined margin value so that a more discriminative feature representation can be obtained. Using the ArcFace loss, FEM ensures not merely a fast training speed for learning

face identity but also the efficiency in optimizing the whole DotFAN network.

2) Face Shape Regressor (FSR): The FSR Φ_{fsr}, architecturally a MobileNet [38], aims to extract face attributes including face shape, pose, and expression. Based on a widely used 3D Morphable Model (3DMM [6]), we designed our FSR as a model-assisted CNN rather than a fully data-driven network, which is complex and must be trained on a large variety of labeled face samples for characterizing face attributes because of the lack in prior knowledge. Moreover, because 3DMM can characterize the face attributes using only hundreds of parameters, the model size of FSR can be significantly reduced. To train FSR, firstly, we follow HPEN's strategy [39] to prepare ground-truth 3DMM parameters $\Theta_{\mathbf{x}}$ of an arbitrary face $\mathbf{x}$ from CASIA dataset [6]. Then, we train FSR via Weighted Parameter Distance Cost (WPDC) [40] defined in (1), with a modified importance matrix, as shown in (2).

$$\mathcal{L}_{wpdc} = \big(\Phi_{fsr}(\mathbf{x}) - \Theta_{\mathbf{x}}\big)^t \mathbf{W} \big(\Phi_{fsr}(\mathbf{x}) - \Theta_{\mathbf{x}}\big) \tag{1}$$

$$\mathbf{W} = (w_R, w_T, w_{shape}, w_{exp}), \tag{2}$$

where w_R, $w_{t_{3d}}$, w_{shape}, and w_{exp} are distance-based weighting coefficients for the $\Theta_{\mathbf{x}}$ derived by 3DMM, and $\Theta_{\mathbf{x}}$ is a vector consisting of a 9×1 vectorized rotation matrix R, a 3×1 translation vector T, a 199×1 vector α_{shape}, and a 29×1 α_{exp}. Note that the facial attribute code $f_p = \Phi_{fsr}(\mathbf{x})$ extracted by FSR is a 240×1 vector mimicking $\Theta_{\mathbf{x}}$. While training DotFAN, we keep α_{shape}'s counterpart—representing facial shape—in f_p unchanged, and we replace f_p's other code segments corresponding to translation T, rotation R, and expression α_{exp} by arbitrary values.

3) General Facial Encoder E and Illumination Code f_l: E is used to capture other features, which cannot be represented by shape and identity codes, on a face. f_l is a one-hot vector specifying the lighting condition, based on which our model synthesizes a face. Note that because CASIA has no shadow labels, for f_l of a face from CASIA, its former 13 entries are set to be 0's and its 14−th entry $f_l^{casia} = 1$; this means to skip shading and to generate a face with the same illumination setting and the same shadow as the input.

B. Generator

The generator G takes an attribute code $\mathbf{f} = [l_{\mathbf{x}}, f_{id}, f_p, f_l]$ as its input to synthesize a face $G(\mathbf{x}, \mathbf{f})$. Described below are loss terms composing the loss function of our generator.

1) Reconstruction Loss: In our design, we exploit a reconstruction loss to retain face contents after performing two transformations dual to each other. That is,

$$\mathcal{L}_{rec} = \|G\big(G(\mathbf{x}, \tilde{\mathbf{f}}), \mathbf{f}\big) - \mathbf{x}\|_2^2 / N, \tag{3}$$

where N is the number of pixels, $G(\mathbf{x}, \tilde{\mathbf{f}})$ is a synthetic face derived according to an input attribute code $\tilde{\mathbf{f}}$. This loss guarantees our generator can learn the transformation relationship between any two dual attribute codes.

2) Pose-Symmetric Loss: Based on a common assumption that a human face is symmetrical, a face with an x° pose angle and a face with a $-x^\circ$ angle should be symmetric about the 0° axis. Consequently, we design a pose-symmetric loss based on which DotFAN can learn to generate $\pm x^\circ$ faces from either training sample. This pose-symmetric loss is evaluated with the aid of a face-mask $M(\cdot)$, which is defined as a function of 3DMM parameters predicted by FSR and makes this loss term focus on the face region by filtering out the background, as described below:

$$\mathcal{L}_{sym} = \|M(\hat{\mathbf{f}}^-) \cdot \big(G(\mathbf{x}, \hat{\mathbf{f}}^-) - \hat{\mathbf{x}}^-\big)\|_2^2 / N. \tag{4}$$

Here, $\hat{\mathbf{f}}^- = [l_{\mathbf{x}}, f_{id}, \hat{f}_p^-, f_l]$, in which $\hat{f}_p^- = \Phi_{fsr}(\hat{\mathbf{x}}^-)$ with $\hat{\mathbf{x}}^-$ denoting the horizontally-flipped version of $\mathbf{x}$, and the other three attribute codes are extracted from $\mathbf{x}$. In sum, this term measures the L_2-norm of the difference between a synthetic face and the horizontally-flipped version of $\mathbf{x}$ within a region-of-interest defined by a mask M.

3) Identity-Preserving Loss: We adopt the following identity-preserving loss to ensure that the identity code of a synthesized face $G(\mathbf{x}, \tilde{\mathbf{f}})$ is identical to that of input face $\mathbf{x}$. That is,

$$\mathcal{L}_{id} = \|\Phi_{fem}(\mathbf{x}) - \Phi_{fem}\big(G(\mathbf{x}, \tilde{\mathbf{f}})\big)\|_2^2 / N_1, \tag{5}$$

where N_1 denotes the length of $\Phi_{fem}(\mathbf{x})$.

4) Pose-Consistency Loss: This term guarantees that the pose and expression feature extracted from a synthetic face is consistent with $\tilde{f}_p$ used to generate the synthetic face. That is,

$$\mathcal{L}_{pose} = \|\tilde{f}_p - \Phi_{fsr}\big(G(\mathbf{x}, \tilde{\mathbf{f}})\big)\|_2^2 / N_2, \tag{6}$$

where N_2 denotes the length of $\tilde{f}_p$.

C. Discriminator

By regarding faces of the same identity as one sub-domain of faces, the task of augmenting faces of different identities becomes a multi-domain image-to-image translation problem addressed in StarGAN [10]. Hence, we exploit an adversarial loss to make augmented faces photo-realistic. To this end, we use the domain classification loss to verify if $G(\mathbf{x}, \tilde{\mathbf{f}})$ is properly classified to a target domain label f_l, which specifies the illumination condition of $G(\mathbf{x}, \tilde{\mathbf{f}})$. In addition, in order to stabilize the training process, we adopted the loss design used in WGAN-GP [41]. Consequently, these two loss terms can be expressed as follows:

$$\begin{aligned} \mathcal{L}_{adv}^D &= D_{src}(G(\mathbf{x}, \tilde{\mathbf{f}})) - D_{src}(\mathbf{x}) \\ &\quad + \lambda_{gp} \cdot \big(\|\nabla_{\hat{x}} D_{src}(\hat{x})\|_2 - 1\big)^2 \\ \mathcal{L}_{adv}^G &= -D_{src}(G(\mathbf{x}, \tilde{\mathbf{f}})), \end{aligned} \tag{7}$$

where λ_{gp} is a trade-off factor for the gradient penalty, $\hat{x}$ is uniformly sampled from the linear interpolation between $\mathbf{x}$ and synthesized $G(\mathbf{x}, \tilde{\mathbf{f}})$, and D_{src} reflects a distribution over sources given by the discriminator; and,

$$\begin{aligned} \mathcal{L}_{cls}^D &= -\log D_{cls}(f_l | \mathbf{x}) \\ \mathcal{L}_{cls}^G &= -\log D_{cls}(\tilde{f}_l | G(\mathbf{x}, \tilde{\mathbf{f}})), \end{aligned} \tag{8}$$

where f_l is the ground-truth illumination code of $\mathbf{x}$, and $\tilde{f_l}$ is the illumination code embedded in $\tilde{\mathbf{f}}$.

In sum, the discriminator aims to produce probability distributions over both source and domain labels, i.e., $D: \mathbf{x} \rightarrow \{D_{src}(\mathbf{x}),\ D_{cls}(\mathbf{x})\}$. Empirically, $\lambda_{gp} = 10$.

D. Full Objective Function

In order to optimize the generator and alleviate the training difficulty, we pretrained FSR and FEM with corresponding labels. Therefore, while training the generator and the discriminator, no additional label is needed. The full objective functions of DotFAN can be expressed as:

$$\begin{aligned}\mathcal{L}_G &= \alpha\mathcal{L}^G_{adv} + \beta\mathcal{L}^G_{cls} + \gamma\,\mathcal{L}_{id} + \zeta\mathcal{L}_{pose} \\ &\quad +\eta\mathcal{L}_{sym} + \xi\mathcal{L}_{rec} \\ \mathcal{L}_D &= \mathcal{L}^D_{adv} + \mathcal{L}^D_{cls}. \end{aligned} \tag{9}$$

where, the two loss terms in $\mathcal{L}_D$ are equal-weighted, and the weighting factors in $\mathcal{L}_G$ are $\alpha = 1$, $\beta = 1$, $\gamma = 8$, $\zeta = 6$, $\eta = 5$, and $\xi = 5$, empirically. Note that the alternative training of generator and discriminator was performed with ratio 1 : 1.

IV. Experiments

A. Datasets

DotFAN is trained jointly on **CMU Multi-PIE** [37] and **CASIA** [36]. Multi-PIE contains more than 750, 000 images of 337 identities, each with 20 different sorts of illumination and 15 different poses. We select images of pose angles ranging in between ±45° and illumination codes from 0 to 12 to form our first training set, containing totally 84, 000 faces. From this training set, DotFAN learns the representative features for a wide range of pose angles, illumination conditions, and resulting shadows. Our second dataset is the whole CASIA set that contains 494, 414 images of 10, 575 identifies, each having about 50 images of different poses and expressions. Since CASIA contains a rich collection of face identities, it helps DotFAN learn features for representing identities.

We exploit CelebA to simulate the data augmentation process. CelebA contains 202, 599 face images collected from 10, 177 identities with 40 kinds of diverse binary facial attributes. We randomly select a fixed number of face images of each identity from CelebA to form our simulation set, called "**sub-CelebA**" and conducted data augmentation experiments on both CelebA and sub-CelebA by using DotFAN.

Moreover, to evaluate the performance of DotFAN on face synthesis, four additional datasets are used: **LFW** [42], **IJB-A** [43], **SurveilFace-1**, and **SurveilFace-2**. LFW has 13, 233 images of 5, 749 identities; IJB-A has 25, 808 images of 500 identities; SurveilFace-1 has 1, 050 images of 73 identities; and SurveilFace-2 contains 1, 709 images of 78 identities. Because faces in two SurveilFace datasets are taken in uncontrolled real working environments, as demonstrated in Fig. 3, they are contaminated by strong backlight, motion blurs, extreme shadow conditions, or influences from various viewpoints. Hence, they mimic the real-world conditions and thus are suitable for evaluating the face augmentation performance. The two SurveilFace sets are private data provided by a video surveillance provider. We will make them publicly available after removing personal labels. Finally, as for DotFAN's capability of face frontalization, we follow the general experiment design to evaluate its performance on LFW and IJB-A.

Fig. 3. Image samples of SurveilFace dataset. Here we show four extreme conditions: (a) strong-backlight, (b) motion-blur, (c) extreme shadow, and (d) unconstrained viewpoint.

In this paper, we demonstrate all face images in grayscale because of two reasons. First, two **SurveilFace** datasets are all grayscale, as is a general case in surveillance applications. Second, DotFAN was trained partially on Multi-PIE in which images have reddish color-drift, so the same color-drift may occur on faces generated by DotFAN. Because such color-drift does not alter the id-features, we do not demonstrate color faces to avoid misunderstanding.

B. Implementation Details

Before training, we align the face images in the Multi-PIE and CASIA by MTCNN [44]. Structurally, our FEM is obtained by Resnet-50 pretained on MS-Celeb-1M [45], and FSR is implemented by a MobileNet [38] pretained on CASIA. To train DotFAN, each input face is resized to 112×112. Both generator and discriminator exploit Adam optimizer [46] with $\beta_1 = 0.5$ and $\beta_2 = 0.999$. The total number of training iterations is 420, 000 with a batch-size of 28, and the number of training epochs is 12. The learning rate is initially set to be 10^{-4} and begins to decay after the 6-th training epoch. Table I shows the network structures of DotFAN's encoder (E), generator (G), and discriminator (D).

C. Face Augmentation

Because DotFAN is a face augmentation network, the experiments in this subsection are designed to show how face recognition accuracy can be improved with DotFAN-augmented training data. We adopt MobileFaceNet[1] as our face recognition model rather than other state-of-the-arts (SOTAs) because it is suitable to be deployed on mobile/embedded devices

[1]We use the third-party open-source implementation in "https://github.com/Xiaoccer/MobileFaceNet_Pytorch".

TABLE I

ARCHITECTURES OF DOT-FAN'S ENCODER (E), GENERATOR (G), AND DISCRIMINATOR (D)

(a) **Encoder (E)**	
Layers	**Output Size**
Input	(1×H×W)
CONV-(N16, K7x7, S1, P3), Batch, ReLU	(16×H×W)
CONV-(N32, K4x4, S2, P1), Batch, ReLU	(32×H/2×W/2)
Residual Block: CONV-(N32, K3x3, S1, P1), Batch, ReLU	(32×H/2×W/2)
Residual Block: CONV-(N32, K3x3, S1, P1), Batch, ReLU	(32×H/2×W/2)
CONV-(N64, K4x4, S1, P1), Batch, ReLU	(64×H/4×W/4)
Residual Block: CONV-(N64, K3x3, S2, P1), Batch, ReLU	(64×H/4×W/4)
Residual Block: CONV-(N64, K3x3, S1, P1), Batch, ReLU	(64×H/4×W/4)
CONV-(N128, K4x4, S2, P1), Batch, ReLU	(128×H/8×W/8)
Residual Block: CONV-(N128, K3x3, S1, P1), Batch, ReLU	(128×H/8×W/8)
Residual Block: CONV-(N128, K3x3, S1, P1), Batch, ReLU	(128×H/8×W/8)
CONV-(N256, K4x4, S2, P1), Batch, ReLU	(256×H/16×W/16)
Residual Block: CONV-(N256, K3x3, S1, P1), Batch, ReLU	(256×H/16×W/16)
Residual Block: CONV-(N256, K3x3, S1, P1), Batch, ReLU	(256×H/16×W/16)
AvgPool(K H/16×W/16, S1)	(256×1×1)

(b) **Generator (G)**	
Layers	**Output Size**
Input	$((256 + F))$
FC-((256+F), 2560)	(2560)
FC-(2560, 256×H/16×W/16), Reshape	(256×H/16×W/16)
Residual Block: CONV-(N256, K3x3, S1, P1), Batch, ReLU	(256×H/16×W/16)
Residual Block: CONV-(N256, K3x3, S1, P1), Batch, ReLU	(256×H/16×W/16)
TRANSPOSECONV-(N128, K4x4, S2, P1),	(128×H/8×W/8)
Residual Block: CONV-(N128, K3x3, S1, P1), Batch, ReLU	(128×H/8×W/8)
Residual Block: CONV-(N128, K3x3, S1, P1), Batch, ReLU	(128×H/8×W/8)
TRANSPOSECONV-(N64, K4x4, S2, P1),	(64×H/4×W/4)
Residual Block: CONV-(N64, K3x3, S1, P1), Batch, ReLU	(64×H/4×W/4)
Residual Block: CONV-(N64, K3x3, S1, P1), Batch, ReLU	(64×H/4×W/4)
TRANSPOSECONV-(N32, K4x4, S2, P1),	(32×H/2×W/2)
Residual Block: CONV-(N32, K3x3, S1, P1), Batch, ReLU	(32×H/2×W/2)
Residual Block: CONV-(N32, K3x3, S1, P1), Batch, ReLU	(32×H/2×W/2)
TRANSPOSECONV-(N16, K4x4, S2, P1)	(16×H×W)
Residual Block: CONV-(N16, K3x3, S1, P1), Batch, ReLU	(16×H×W)
CONV-(N16, K7x7, S1, P3), Tanh	(1×H×W)

(c) **Discriminator (D)**	
Layers	**Output Size**
Input	(1×H×W)
CONV-(N64, K4x4, S2, P1), Leaky ReLU	(64×H/2×W/2)
CONV-(N128, K4x4, S2, P1), Leaky ReLU	(128×H/4×W/4)
CONV-(N256, K4x4, S2, P1), Leaky ReLU	(256×H/8×W/8)
CONV-(N512, K4x4, S2, P1), Leaky ReLU	(512×H/16×W/16)
CONV-(N1024, K4x4, S2, P1), Leaky ReLU	(1024×H/32×W/32)
CONV-(N2048, K4x4, S2, P1), Leaky ReLU	(2048×H/64×W/64)
CONV-(N1, K3x3, S1, P1), AvgPool(K H/64×W/64, S1)(D_{src})	(1)
CONV-(N14, K H/64×W/64, S1, P0)(D_{cls})	(14×1×1)

(less than 1M parameters) for small/medium-size real-world applications. Fig. 4 shows some augmented faces with shadows assigned with four different illumination codes. Note that all synthesized faces presented in this paper are produced by the same DotFAN model without manually data-dependent modifications.

To evaluate the effectiveness of face augmentation with DotFAN, we perform data augmentation on the same dataset by using DotFAN, FaceID-GAN, and StarGAN first. We then train MobileFaceNet [47] on the different datasets augmented by the three augmentation models to obtain the corresponding face recognition models. Consequently, we compare the accuracy of these MobileFaceNet models, each trained on an augmented dataset, on LFW, SurveilFace-1 and SurvilFace-2, respectively. StarGAN used in this experiment is trained on Multi-PIE that is rich in illumination conditions; meanwhile, FaceID-GAN is trained on CASIA to learn pose and expression representations.

Table II summarizes the results of this experiment set. We interpret the results focusing on Sub-experiment(a). In Sub-experiment(a) of Table II, we randomly select 3 faces of each identity from CelebA to form the **RAW** training set, namely **Sub-CelebA(3)**, leading to about $30,000$ training samples in raw **Sub-CelebA(3)**. The MobileFaceNet trained on raw Sub-CelebA(3) achieves a face verification accuracy of 83.1% on LFW, a true accept rate (TAR) of 20.5% at false accept rate (FAR) $= 0.1\%$ on SurveilFace-1, and a TAR of 18.0% at FAR $= 0.1\%$ on SurveilFace-2. After giving each face in raw Sub-CelebA(3) a random facial attribute $\tilde{f}_p$ and a random illumination code $\tilde{f}_l$ to generate a new face and thus to double the size of the training set via DotFAN, the verification accuracy on LFW becomes 93.6%, and the TAR values on SurveilFace datasets are all nearly doubled, as shown in the row named **DotFAN 1x**. This shows DotFAN is effective in face augmentation and outperforms StarGAN and FaceID-GAN significantly. Furthermore, when we augment about $90,000$ additional faces to quadruple the size of training set, i.e., **DotFAN 3x**, we have only a minor improvement in verification accuracy compared to **DotFAN 1x**. This fact reflects that the marginal benefit a model can extract from the data diminishes as the number of samples increases when there is information overlap among data, similar to that reported in [48]. Consequently, Table II and Fig. 5 reveal the following remarkable points.

- First, by integrating attribute controls on pose angle, illuminating condition, and facial expression with an identity-preserving design, DotFAN outperforms StarGAN and FaceID-GAN in domain-transferred face augmentation tasks.
- Second, DotFAN's results obey *the law of diminishing marginal utility* in Economics[2] [49], as demonstrated in all **(DotFAN 1x, DotFAN 3x)** data pairs. Take LFW-experiment in Table II(a) for example. An additional

[2]This law primarily says that the marginal utility of each homogeneous unit decreases as the supply of units increases, and vice versa.

Fig. 4. Face augmentation examples (CelebA) containing augmented faces with 4 illumination conditions and 7 poses.

TABLE II

PERFORMANCE COMPARISON OF FACE RECOGNITION MODELS TRAINED ON THE DATASETS AUGMENTED BY FIVE DIFFERENT AUGMENTATION MODELS, WHERE OUR DOTFAN IS TRAINED ON CASIA, MULTI-PIE, AND MS-CELEB-1M, AND **SUB-CELEBA**(x) DENOTES A SUBSET FORMED BY RANDOMLY SELECTING x IMAGES OF EACH FACE SUBJECT FROM CELEBA

Method	LFW		SurveilFace-1			SurveilFace-2			IJB-A		
	ACC	AUC	@FAR= 0.001	@FAR= 0.01	AUC	@FAR= 0.001	@FAR= 0.01	AUC	@FAR= 0.001	@FAR= 0.01	AUC
(a) **Sub-CelebA(3)** (totally 30, 120 images)											
RAW	83.1	90.2	20.5	34.4	83.2	18.0	33.3	84.8	24.6	44.0	89.1
StarGAN	85.9	92.5	25.1	39.6	87.5	27.4	46.7	91.4	29.5	49.4	90.3
FaceID-GAN	92.5	97.6	34.6	53.5	92.8	32.3	54.0	94.3	11.4	25.5	82.0
FFWM	83.2	91.2	23.3	36.3	85.7	19.1	35.7	87.1	28.8	48.6	90.6
Rotate-and-Render	84.3	91.6	25.5	42.7	88.6	24.3	41.8	89.5	38.3	58.5	93.3
DotFAN 1x	93.6	98.1	35.7	56.2	93.6	34.7	57.8	95.0	48.5	69.0	95.8
DotFAN 3x	94.7	98.7	36.8	58.3	94.6	36.5	60.8	95.6	–	–	–
(b) **Sub-CelebA(8)** (totally 75, 796 images)											
RAW	94.0	98.5	37.8	58.7	94.4	38.3	61.0	95.2	50.8	71.1	96.1
StarGAN	94.3	98.5	42.6	60.7	94.9	42.8	65.6	95.8	50.8	69.8	95.6
FaceID-GAN	96.5	99.3	48.1	65.6	96.0	45.7	67.9	96.8	44.7	65.9	95.1
FFWM	93.2	98.1	40.1	56.2	94.3	37.8	58.6	94.9	56.7	77.0	97.1
Rotate-and-Render	94.8	98.7	38.6	58.7	94.9	36.0	56.7	94.9	59.1	77.7	97.2
DotFAN 1x	97.3	99.5	53.2	71.2	97.0	49.1	72.2	97.2	63.8	82.5	97.8
DotFAN 3x	97.2	99.5	53.2	68.9	96.9	47.3	70.0	97.1	–	–	–
(c) **Sub-CelebA(13)** (totally 116, 659 images)											
RAW	96.3	99.1	47.4	67.8	96.2	43.5	67.0	96.5	60.2	77.9	97.2
StarGAN	96.7	99.3	48.3	68.1	96.7	46.3	70.0	96.7	57.8	77.3	97.0
FaceID-GAN	97.2	99.5	53.3	71.3	97.0	50.2	72.3	97.4	53.1	73.5	96.5
FFWM	96.0	99.3	44.7	68.0	96.5	38.8	64.6	96.6	62.4	80.5	97.7
Rotate-and-Render	96.1	99.2	43.2	63.7	96.3	37.7	60.6	95.7	58.1	81.0	97.7
DotFAN 1x	97.6	99.6	56.2	75.1	97.7	50.4	73.9	97.7	68.4	85.0	97.9
DotFAN 3x	97.5	99.7	56.7	75.5	97.7	53.9	72.2	97.8	–	–	–
(d) **CelebA (full CelebA dataset,** 202, 599 **images)**											
RAW	97.6	99.6	53.5	73.8	97.7	48.7	73.0	97.5	67.0	83.1	97.8
StarGAN	97.7	99.6	55.0	74.2	97.7	53.0	73.8	97.6	68.2	84.0	97.8
FaceID-GAN	98.0	99.7	57.6	76.4	98.1	54.1	76.5	98.0	54.1	73.8	96.4
FFWM	97.3	99.6	50.6	72.7	97.4	43.3	66.9	97.0	68.9	85.3	98.1
Rotate-and-Render	97.5	99.6	50.9	72.4	97.4	44.7	68.3	97.1	61.2	82.5	97.9
DotFAN 1x	98.3	99.8	62.4	80.9	98.4	57.1	76.7	98.1	73.2	88.0	98.2
DotFAN 3x	98.4	99.7	61.4	78.9	98.2	54.7	77.8	98.0	–	–	–

one-unit consumption of training data (1x-augmentation) brings an accuracy improvement, i.e., marginal utility, of $93.6\% - 83.1\% = 10.5\%$; when two more additional units (3x-augmentation) are given, the improvement of accuracy is only $94.7\% - 93.6\% = 1.1\%$. Therefore, a **1x** augmentation is effective to enrich a small dataset, whereas the performance improvement with **DotFAN 3x** augmentation is already saturated as excessive face augmentation does not add much to the richness of face data.

- Third, although the improvement in verification accuracy decreases as the size of raw training set increases, DotFAN achieves a significant performance gain on augmenting a small-size face training set, as demonstrated in all (RAW, DotFAN 1x) data pairs.
- Table II shows that DotFAN 3x is always not as good as DotFAN 1x, and there is a gap between Sub-CelebA (3) DotFAN 1x and Sub-CelebA (13) RAW. Being a feature-space augmentation scheme, DotFAN fixes the identity feature f_{id} extracted from an input face and manipulates attribute codes, e.g. including l_x, f_p, and f_l, to create synthetic features $\mathbf{f} = [l_x, f_{id}, f_p, f_l]$ that distribute around the fixed identity feature of the source face for face augmentation. That is, while real faces are those capable of spanning the whole sample space, generated faces are those locally dense around source data points.

TABLE III

PERFORMANCE EVALUATION ON DOMAIN KNOWLEDGE TRANSFER, WHERE DOTFAN-CASIA (I.E., DOTFAN TRAINED ON CASIA ONLY) IS APPLIED TO AUGMENT THE CASIA DATASET FOR DATA AUGMENTATION

Method	LFW		SurveilFace-1			SurveilFace-2			IJB-A		
	ACC	AUC	@FAR= 0.001	@FAR= 0.01	AUC	@FAR= 0.001	@FAR= 0.01	AUC	@FAR= 0.001	@FAR= 0.01	AUC
(a) **CASIA** (totally 494, 414 images), tested via **MobileFaceNet**											
RAW	99.0	99.8	60.9	80.0	98.4	54.9	77.8	98.3	71.5	89.5	98.7
StarGAN	99.05	99.8	62.0	81.3	98.6	52.0	76.6	98.3	71.1	89.9	98.7
FaceID-GAN	98.1	99.8	55.3	77.2	97.9	51.0	71.2	97.6	54.1	73.8	96.4
FFWM	99.0	99.8	59.3	80.4	98.5	52.2	76.7	98.4	72.6	91.1	98.8
Rotate-and-Render	98.9	99.8	57.9	78.4	98.5	47.1	74.7	97.9	47.3	87.3	98.3
DotFAN-casia 1x	99.05	99.8	62.7	82.7	98.7	61.6	80.5	98.5	71.0	89.9	98.6
(b) **CASIA** (totally 494, 414 images), tested via **ResNet-101**											
RAW	97.2	99.6	50.9	73.6	97.8	42.0	69.5	97.3	69.8	87.5	98.2
StarGAN	98.2	99.7	52.1	74.7	97.7	45.6	71.0	97.7	71.7	86.6	98.3
FaceID-GAN	98.1	99.8	55.3	77.2	97.9	51.0	71.2	97.6	56.7	85.9	98.4
FFWM	97.9	99.7	49.5	71.0	97.3	43.1	68.7	97.2	68.1	88.9	98.7
Rotate-and-Render	98.2	99.8	61.5	79.1	98.6	49.7	74.0	97.8	54.2	87.3	98.2
DotFAN-casia 1x	98.3	99.8	55.6	77.6	98.3	48.8	74.6	98.1	75.1	89.7	98.7

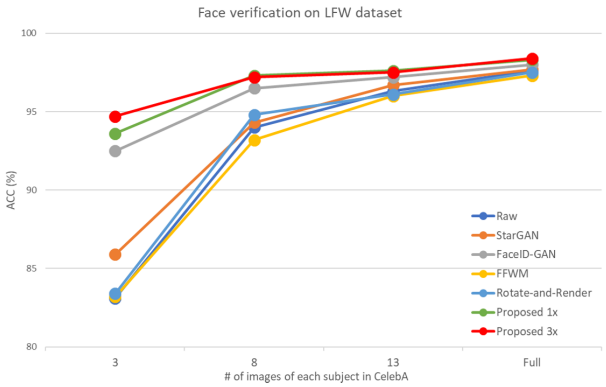


Fig. 5. Comparison of face verification accuracy on LFW trained on different augmented dataset. The horizontal spacing highlights the size of raw training dataset sampled from CelebA.

Fig. 6. Ablation study on loss terms. (a) Full loss. (b) w/o $\mathcal{L}_{id}$, (c) w/o $\mathcal{L}_{cls}$, (d) w/o $\mathcal{L}_{rec}$, (e) w/o $\mathcal{L}_{pose}$, and (f) w/o $\mathcal{L}_{sym}$.

Hence, generated faces may deteriorate the models, and real images are still more effective than generated images for performance improvement.

D. Ablation Study

We then verify the effect brought by each loss term. Fig. 6 depicts the faces generated by using different combinations of loss terms. The top-most row shows faces generated with the full generator loss $\mathcal{L}_G$ in (9), whereas the remaining rows respectively show synthetic results derived without one certain loss term.

As shown in Fig. 6(b), without $\mathcal{L}_{id}$, DotFAN fails to preserve the identity information although other facial attributes can be successfully retained. By contrast, without $\mathcal{L}_{cls}$, DotFAN cannot control the illumination condition, and the resulting faces all share the same shade (see Fig. 6(c)). These two rows evidence that $\mathcal{L}_{cls}$ and $\mathcal{L}_{id}$ are indispensable in DotFAN design. Moreover, Fig. 6(d) shows some unrealistic faces, e.g., a rectangular-shaped ear in the frontalized face; accordingly, $\mathcal{L}_{rec}$ is important for photo-realistic synthesis. Finally, Figs. 6(e)–(f) show that $\mathcal{L}_{pose}$ and $\mathcal{L}_{sym}$ are complementary to each other. As long as either of them functions, DotFAN can generate faces of different face angles. However, because $\mathcal{L}_{sym}$ is designed to learn only the mapping relationship between $+x^\circ$ face and $-x^\circ$ face by ignoring the background outside the face region, artifacts may occur in the background region if $\mathcal{L}_{sym}$ works solely (see Fig. 6(e)). Finally, Fig. 7 illustrate two other visual examples of our ablation study.

E. Verification on Domain Knowledge Transfer

This subsection clarifies the doubt on how effective DotFAN can be when training datasets and augmentation target are highly overlapped. Because DotFAN is trained on three datasets, i.e., CASIA, Multi-PIE, and MS-Celeb-1M, we conduct a face augmentation experiment on CASIA to simulate such situation for clearing up this doubt.

Table III demonstrates the experimental results obtained based on classifiers trained on the CASIA dataset augmented by DotFAN-casia and the compared methods. Because the raw CASIA dataset (494, 414 faces) is much richer than the full CelebA (202, 599 faces), MobileFaceNet trained on raw CASIA performs slightly better than those trained on **1x-augmented** CelebA comprising 202, 599 source faces and

Fig. 7. Two other examples of ablation study on loss terms. For each example, from top to bottom: i) Full loss, ii) w/o $\mathcal{L}_{id}$, iii) w/o $\mathcal{L}_{cls}$, iv) w/o $\mathcal{L}_{rec}$, v) w/o $\mathcal{L}_{pose}$, and vi) w/o $\mathcal{L}_{sym}$.

202, 599 augmented faces. As shown in Table III, DotFAN-casia beats FaceID-GAN and Rotate-and-Render and performs competitively with StarGAN and FFWM. This demonstrates DotFAN is still effective under this situation that there are overlaps between the source-domain and target-domain face data, although the improvement here is less significant compared with those shown in Table II. In sum, Tables II and III jointly evidence that DotFAN effectively augments data via domain knowledge transfer. For each experiment set shown in Tables II and III, the data in **RAW** rows are used as the comparison baseline, and MobileFaceNet and ResNet-101 are trained with the same training configurations to guarantee the fairness of evaluation.

Still, we can draw the following concluding remarks from the IJB-A experiment sets shown in Tables II and III. First, DotFAN (learning knowledge from multiple source domains) well beats DotFAN-casia (learning knowledge from one single domain). Although the augmented **Sub-CelebA(13)** contains significantly fewer face samples ($116,659 \times 2$) than that of the 1x-augmented **CASIA** ($494,414 \times 2$), Table II(c) still shows a better recognition accuracy on IJB-A than Table III(a). This fact implies that DotFAN can augment reliable and diverse face samples more effectively when it is trained on data of multiple domains thoroughly. Second, we can notice that in Table III all face GAN models, except for StaGAN that is not specialized for face image translation, result in augmented face datasets that deteriorate the recognition performance of both MobileFaceNet and ResNet-101 on IJB-A. This is probably due to the various illumination conditions and different kinds of shadows within the IJB-A dataset. Hence, only StarGAN, a general most multi-domain image translation model, can render an input into faces with different lighting/shading conditions while keeping other image contents almost unaltered. Third, however, even though in Table III DotFAN, FaceID-GAN, FFWM, and Rotate-and-Render models do not perform well on IJB-A, DotFAN still outperforms other three and cause least degeneration of face recognizer's ability due to the contribution of DotFAN's FEM module. This fact implies that DotFAN is the most reliable identity-preserving face synthesis model when i) the training dataset and ii) the to-be-augmented dataset are not rich enough simultaneously.

F. Attribute-Decomposed Facial Representation

Fig. 8 show several synthesized faces to demonstrate the capability of DotFAN's attribute-decomposed facial representation. This experiment demonstrates that we can control the face synthesis result by editing our face attribute code. In Fig. 8, each row shows a sequence of faces. Each sequence is obtained by modifying a code segment through linear combination. For example, the aces in the first row are synthesized by editing the face identity. Specifically, letting the attribute codes of the **input** and the **target** images be respectively $\mathbf{f}_L = [l_x^l, f_{id}^l, f_p^l, f_l^l]$ and $\mathbf{f}_R = [l_x^r, f_{id}^r, f_p^r, f_l^r]$, the interpolated faces in the first row are generated by using an edited code $\tilde{\mathbf{f}} = [l_x^l, \tilde{f}_{id}, f_p^l, f_l^l]$ with $\tilde{f}_{id} = \alpha f_{id}^l + (1-\alpha) f_{id}^r$. The first row illustrates that by fixing l_x^l, f_p^l, and f_l^l, the face identity varies smoothly with α while the other attributes keep unchanged. The second and third rows show the face interpolation results of controlled pose code $\tilde{f}_p$. Because both pose information and expression information are encoded into f_p, these two sequences demonstrate that we can control the face synthesis by simply editing only a small segment of f_p. Finally, the fourth row shows the faces synthesized according to edited general feature $\tilde{l}_x$. In this sequence, the bangs and the eye-shadow (glasses) vary smoothly, but the identify information, the lighting condition, and the pose/expression remain unchanged.

G. Face Synthesis

Finally, we verify the efficacy of DotFAN through i) face frontalization and ii) face rotation results. This experiment set is designed for i) simulating the situation in which the

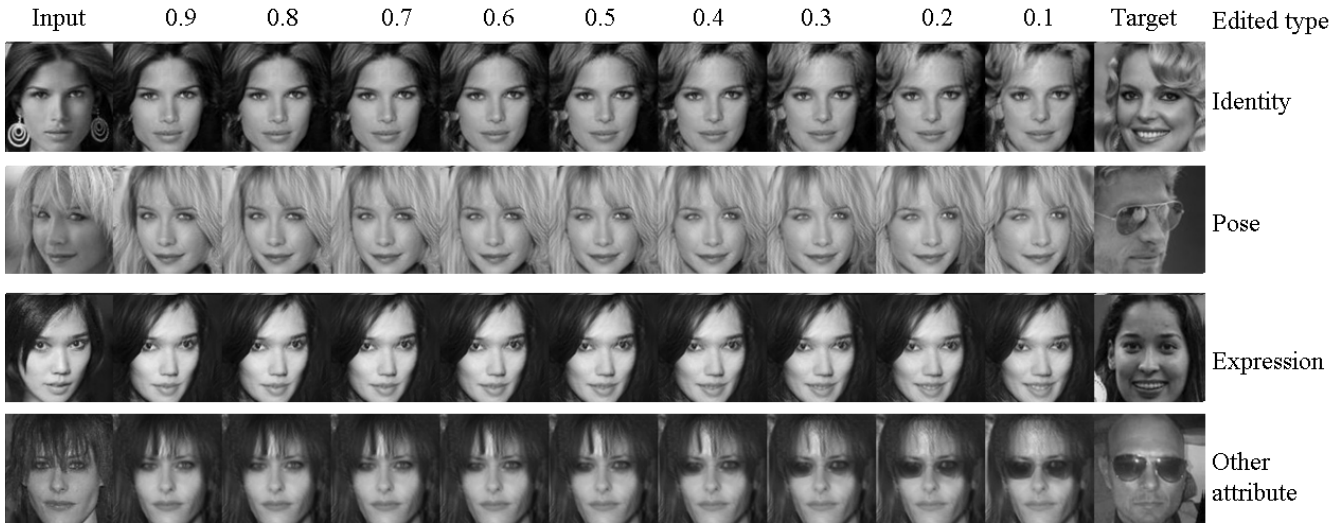


Fig. 8. Attribute-decomposed facial representation. For each row, the morphing sequence is generated by editing an attribute code segment. For example, in the first row, we edited the identity information by $\tilde{f}_{id} = \alpha f_{id}^{input} + (1-\alpha) f_{id}^{target}$ with α ranging from 0.9 to 0.1. Other three rows show sequences obtained by editing pose, expression, and general facial features.

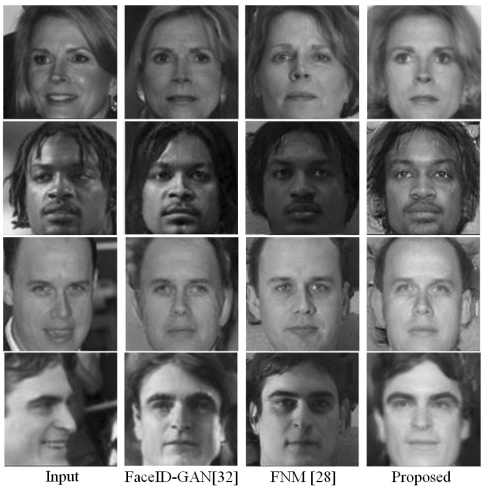


Fig. 9. Face frontalization results (LFW) obtained by different methods.

face recognizer is pre-trained but is not re-trainable and ii) proving that DotFAN can assist the recognizer by frontalizing/neutralizing a to-be-recognized face.

1) Face Frontalization: First, we verify if the identity information extracted from a frontalized face, produced by DotFAN, is of the same class as the identity of a given source face. Following [7], we measure the performance by using a face recognition model trained on MS-Celeb-1M. Next, we conduct frontalization experiments on LFW and CASIA, as the examples demonstrated in Fig. 9 and Fig. 10. Particularly, Fig. 10 demonstrates that DotFAN, being an identity-preserving face augmentation method based on domain knowledge transfer, can retain the identity feature best after face frontalization, even for cases of faces with extreme pose angles. Also, Fig. 10 shows that StarGAN can generate faces with diverse illumination conditions while keeping other image contents unaltered, as we already discussed in Sec. IV-E.

Fig. 10. Face frontalization results (CASIA) obtained by different methods. From left to right: i) input, ii) StarGAN [10], iii) FaceID-GAN [7], iv) Rotate-and-Render [20], v) FFWM [21], and vi) DotFAN.

Table IV shows the comparison of face verification results of the frontalized faces. This experiment set validates that i) compared with other methods, DotFAN achieves comparable visual quality in face frontalization, ii) shadows can be effectively removed by DotFAN, and iii) both DotFAN and DotFAN-casia (i.e., a DotFAN trained only on CASIA dataset) outperform the other methods in terms of verification accuracy, especially in the experiment on IJB-A shown in Table IV(b), where DotFAN reports a much better TAR, i.e., 89.3% on FAR@0.001 and 93.7% on FAR@0.01, than existing approaches. In addition, DotFAN is designed for general face image augmentation, and face frontalization is only a special case of face augmentation. Hence, face frontailization is just an added value of DotFAN that is not particularly optimized for this purpose (i.e., no frontalization-based cost function is used to train DotFAN). Therefore, it is reasonable that when DotFAN's training set is not rich enough, its frontalization performance (i.e., experiment values of DotFAN-casia) might be slightly inferior to FFWM that was particularly designed for face frontalization, as shown in Table IV(a).

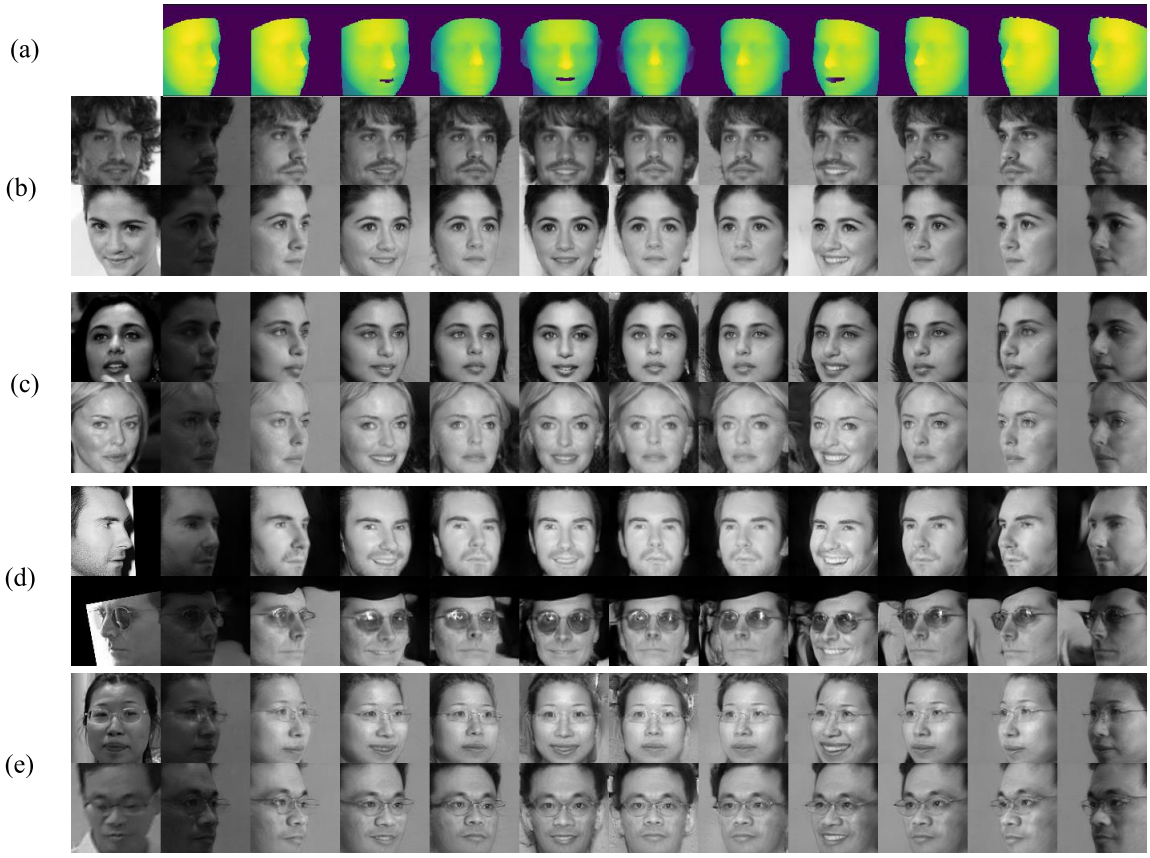


Fig. 11. Synthesized faces for face samples from different datasets generated by DotFAN. The left-most column shows the inputs with random attributes (e.g., poses, expressions, and motion blurs). The top-most row illustrates 3D templates with specific poses and expressions. To guarantee the identity information of each synthetic face is observable, columns 3–11 show shadow-free results, and columns 2 and 12 show faces with shadows. (a) 3D templates, (b) CelebA, (c) LFW, (d) CFP, and (e) SurveilFace.

TABLE IV

FACE VERIFICATION COMPARISON. (a) VERIFICATION ACCURACY ON LFW. (b) TRUE-ACCEPT-RATE (TAR) OF VERIFICATIONS ON IJB-A. NOTE THAT WHILE DOTFAN HAS AN FEM TRAINED ON MS-CELEB-1M IN OUR DESIGN, THE FEM OF DOTFAN-CASIA WAS TRAINED ON CASIA DATASET

(a)

Method	Verification Accuracy
HPEN [39]	96.25±0.76
FF-GAN [17]	96.42±0.89
FaceID-GAN [7]	97.01±0.83
Rotate-and-Render [20]	98.40±0.61
FFWM [21]	98.67±0.53
DotFAN-casia	**98.55±0.52**
DotFAN	**99.18±0.39**

(b)

Method	FAR@0.001	FAR@0.01
PAM [30]	55.2±3.2	73.3±1.8
DCNN [50]	-	78.7±4.3
DR-GAN [25]	53.9±4.3	77.4±2.7
FF-GAN [17]	66.3±3.3	85.2±1.0
FaceID-GAN [7]	69.2±2.7	87.6±1.1
Rotate-and-Render [20]	82.48±2.5	91.87±0.7
FFWM [21]	83.48±3.4	92.53±1.6
DotFAN-casia	**82.3±2.4**	**90.5±0.7**
DotFAN	**89.3±1.0**	**93.7±0.5**

Note that the numerical values shown in Table IV are those reported in FaceID-GAN paper. FaceID-GAN derives the verification values by measuring the cosine distance between the two id-features, extracted by FaceID-GAN's classification module (i.e., module-C) that was trained on the training data and further used to generate synthetic faces, of an real image and that of a synthesized image. Hence, the accuracy values reported in FaceID-GAN form a theoretical upper-bound. On the contrary, DotFAN is a face augmentation network, and therefore we verify DotFAN by using a face recognizer, e.g. MobileFaceNet, different from DotFAN's FEM under the assumption that users may use an arbitrary face recognition network to recognize faces synthesized by DotFAN. Consequently, we use a different face recognition network to impartially claim DotFAN's capability of "identity-preservation". This experiment fairly shows DotFAN performs better than previous methods.

2) Face Rotation: Fig. 11 demonstrates DotFAN's capability in synthesizing faces of given attributes, including pose angles, facial expressions, and shadows, while retaining the associated identities. The source faces presented in the left-most column in Fig. 11 come from four datasets, i.e., CelebA, LFW, CFP [51], and SurveilFace. CelebA and LFW are two widely-adopted face datasets; CFP contains images with extreme pose angles, e.g., ±90°; and, SurveilFace contains faces of variant illumination conditions and faces affected by motion-blurs. This experiment shows that DotFAN can stably synthesize visually-pleasing face images based on 3DMM parameters describing 3D templates.

V. Conclusion

We proposed a Domain-transferred Face Augmentation net (DotFAN) for generating a series of variants of an input face image based on the knowledge of attribute-decomposed face representation distilled from rich datasets. DotFAN is designed in StarGAN's style with two extra subnetworks to learn separately the facial attribute codes and produce a normalized face so that it can effectively generate face images of various facial attributes while preserving identity of synthetic images. Moreover, we proposed a pose-symmetric loss through which DotFAN can synthesize a pair of pose-symmetric face images directly at once. Extensive experiments demonstrate the effectiveness of DotFAN in augmenting small-size face datasets and improving their within-subject diversity. As a result, a better face recognition model can be learned from an enriched training set derived by DotFAN.

References

[1] 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.

[2] 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.

[3] 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.

[4] F. Cole, D. Belanger, D. Krishnan, A. Sarna, I. Mosseri, and W. T. Freeman, "Synthesizing normalized faces from facial identity features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3703–3712.

[5] Y. Qian, W. Deng, and J. Hu, "Unsupervised face normalization with extreme pose and expression in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 9851–9858.

[6] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3D faces," in *Proc. 26th Annu. Conf. Comput. Graph. Interact. Techn. (SIGGRAPH)*, 1999, pp. 187–194.

[7] Y. Shen, P. Luo, P. Luo, J. Yan, X. Wang, and X. Tang, "FaceID-GAN: Learning a symmetry three-player GAN for identity-preserving face synthesis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 821–830.

[8] I. Masi, A. T. Tran, 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.

[9] B. Gecer, B. Bhattarai, J. Kittler, and T.-K. Kim, "Semi-supervised adversarial learning to generate photorealistic face images of new identities from 3D morphable model," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 217–234.

[10] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "StarGAN: Unified generative adversarial networks for multi-domain Image-to-Image translation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8789–8797.

[11] Y. Lu, Y.-W. Tai, and C.-K. Tang, "Attribute-guided face generation using conditional CycleGAN," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 282–297.

[12] T. Li *et al.*, "BeautyGAN: Instance-level facial makeup transfer with deep generative adversarial network," in *Proc. 26th ACM Int. Conf. Multimedia*, 2018, pp. 645–653.

[13] W. Shen and R. Liu, "Learning residual images for face attribute manipulation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4030–4038.

[14] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," 2015, *arXiv:1511.06434*. [Online]. Available: http://arxiv.org/abs/1511.06434

[15] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1125–1134.

[16] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2223–2232.

[17] 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.

[18] F.-J. Chang, A. Tuan Tran, T. Hassner, I. Masi, R. Nevatia, and G. Medioni, "ExpNet: Landmark-free, deep, 3D facial expressions," in *Proc. 13th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2018, pp. 122–129.

[19] F.-J. Chang, A. T. Tran, T. Hassner, I. Masi, R. Nevatia, and G. Medioni, "FacePoseNet: Making a case for landmark-free face alignment," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2017, pp. 1599–1608.

[20] H. Zhou, J. Liu, Z. Liu, Y. Liu, and X. Wang, "Rotate-and-render: Unsupervised photorealistic face rotation from single-view images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5911–5920.

[21] Y. Wei, M. Liu, H. Wang, R. Zhu, G. Hu, and W. Zuo, "Learning flow-based feature warping for face frontalization with illumination inconsistent supervision," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 558–574.

[22] 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.

[23] J. Zhao *et al.*, "Towards pose invariant face recognition in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2207–2216.

[24] Z. Zhang, X. Chen, B. Wang, G. Hu, W. Zuo, and E. R. Hancock, "Face frontalization using an appearance-flow-based convolutional neural network," *IEEE Trans. Image Process.*, vol. 28, no. 5, pp. 2187–2199, May 2019.

[25] 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.

[26] 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.

[27] E. H. Land and J. J. McCann, "Lightness and Retinex theory," *J. Opt. Soc. Amer.*, vol. 61, no. 1, pp. 1–11, 1971.

[28] G. D. Finlayson, S. D. Hordley, and M. S. Drew, "Removing shadows from images," in *Proc. Eur. Conf. Comput. Vis.*, 2002, pp. 823–836.

[29] Y. Wang *et al.*, "Face relighting from a single image under arbitrary unknown lighting conditions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 11, pp. 1968–1984, Oct. 2009.

[30] I. Masi, S. Rawls, G. Medioni, and P. Natarajan, "Pose-aware face recognition in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4838–4846.

[31] K. Cao, Y. Rong, C. Li, X. Tang, and C. C. Loy, "Pose-robust face recognition via deep residual equivariant mapping," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5187–5196.

[32] I. Masi, F.-J. Chang, J. Choi, and S. Harel, "Learning pose-aware models for pose-invariant face recognition in the wild," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 379–393, Feb. 2019.

[33] 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.

[34] J. Zhao *et al.*, "Dual-agent GANs for photorealistic and identity preserving profile face synthesis," in *Proc. Adv. Neural. Inf. Process. Syst.*, 2017, pp. 66–76.

[35] L. Li, J. Bao, H. Yang, D. Chen, and F. Wen, "FaceShifter: Towards high fidelity and occlusion aware face swapping," 2019, *arXiv:1912.13457*. [Online]. Available: http://arxiv.org/abs/1912.13457

[36] 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

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

[38] A. G. Howard *et al.*, "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*. [Online]. Available: http://arxiv.org/abs/1704.04861

[39] 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.

[40] 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.

[41] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of Wasserstein GANs," in *Proc. Adv. Neural Inf. Proc. Syst.*, 2017, pp. 5767–5777.
[42] 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," Univ. Massachusetts, Amherst, MA, USA, Tech. Rep. 07-49, Oct. 2007.
[43] 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.
[44] J. Dai, K. He, and J. Sun, "Instance-aware semantic segmentation via multi-task network cascades," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3150–3158.
[45] 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.*, 2016, pp. 87–102.
[46] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1–15.
[47] S. Chen, Y. Liu, X. Gao, and Z. Han, "MobileFaceNets: Efficient CNNs for accurate real-time face verification on mobile devices," in *Proc. Chin. Conf. Biometric Recognit.*, 2018, pp. 428–438.
[48] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, and S. Belongie, "Class-balanced loss based on effective number of samples," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 9268–9277.
[49] N. G. Mankiw, *Principles of Economics*. Boston, MA, USA: Cengage Learning, 2020.
[50] J.-C. Chen, V. M. Patel, and R. Chellappa, "Unconstrained face verification using deep CNN features," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2016, pp. 1–9.
[51] 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.

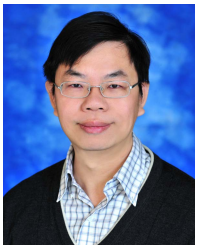
Hao-Chiang Shao (Member, IEEE) received the Ph.D. degree in electrical engineering from the National Tsing Hua University, Taiwan, in 2012. From 2012 to 2017, he was a Postdoctoral Researcher with the Institute of Information Science, Academia Sinica, involved in a series of Drosophila brain research projects and an Research and Development Engineer with the Computational Intelligence Technology Center, Industrial Technology Research Institute, Taiwan, from 2017 to 2018, taking charges of DNN-based automated optical inspection (AOI) projects. He has been an Assistant Professor with the Department of Statistics and Information Science, Fu Jen Catholic University, Taiwan, since 2018. His research interests include 2D+Z image atlasing, 3D mesh processing, big industrial image data analysis, and machine learning.

Kang-Yu Liu received the B.S. degree from the National Chung Cheng University in 2017 and the M.S. degree from the National Tsing Hua University in 2020, both in electrical engineering.

He has been working with Realtek Semiconductor Corporation as an AI Algorithm Engineer since 2020. His research interests lie in computer vision, machine learning, and visual analytics.

Weng-Tai Su (Member, IEEE) received the B.S. degree in electrical engineering from the National Yunlin University of Science and Technology, Yunlin County, Taiwan, in 2012, and the M.S. degree in electrical engineering from the National Tsing Hua University (NTHU), Hsinchu, Taiwan, in 2014, where he is currently pursuing the Ph.D. degree with the Department of Electrical Engineering.

His research interests mainly lie in machine learning, image and video processing, and computer vision.



Chia-Wen Lin (Fellow, IEEE) received the Ph.D. degree in electrical engineering from the National Tsing Hua University (NTHU), Hsinchu, Taiwan, in 2000.

He was with the Department of Computer Science and Information Engineering, National Chung Cheng University, Taiwan, from 2000 to 2007. He is currently a Professor with the Department of Electrical Engineering and the Institute of Communications Engineering, NTHU, and the Research and Development Director of the Electronic and Optoelectronic System Research Laboratories, Industrial Technology Research Institute, Hsinchu. He is also the Deputy Director of the NTHU AI Research Center. Prior to joining academia, he worked with the Information and Communications Research Laboratories, Industrial Technology Research Institute, from 1992 to 2000. His research interests include image and video processing, computer vision, and video networking.

Dr. Lin served as a Distinguished Lecturer of IEEE Circuits and Systems Society from 2018 to 2019, a Steering Committee Member of IEEE TRANSACTIONS ON MULTIMEDIA from 2014 to 2015, and the Chair of the Multimedia Systems and Applications Technical Committee of the IEEE Circuits and Systems Society from 2013 to 2015. His articles received the Best Paper Award of IEEE VCIP 2015, the Top 10% Paper Awards of IEEE MMSP 2013, and the Young Investigator Award of VCIP 2005. He received the Outstanding Electrical Professor Award presented by the Chinese Institute of Electrical Engineering in 2019 and the Young Investigator Award presented by the Ministry of Science and Technology, Taiwan, in 2006. He has been serving as the President of the Chinese Image Processing and Pattern Recognition Association, Taiwan, since 2019. He has served as the Technical Program Co-Chair of IEEE ICME 2010, the General Co-Chair of IEEE VCIP 2018, and the Technical Program Co-Chair of IEEE ICIP 2019. He is currently the Chair of the Steering Committee of IEEE ICME. He has served as an Associate Editor for IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, IEEE TRANSACTIONS ON MULTIMEDIA, and IEEE MULTIMEDIA.

Jiwen Lu (Senior Member, IEEE) received the B.E. degree in mechanical engineering and the M.E. degree in electrical engineering from the Xi'an University of Technology, Xi'an, China, in 2003 and 2006, respectively, and the Ph.D. degree in electrical engineering from Nanyang Technological University, Singapore, in 2012. He is currently an Associate Professor with the Department of Automation, Tsinghua University, Beijing, China. His current research interests include computer vision and pattern recognition. He is a member of the Image, Video and Multidimensional Signal Processing Technical Committee, the Multimedia Signal Processing Technical Committee, and the Information Forensics and Security Technical Committee of the IEEE Signal Processing Society and also a member of the Multimedia Systems and Applications Technical Committee and the Visual Signal Processing and Communications Technical Committee of the IEEE Circuits and Systems Society. He is a fellow of IAPR. He was a recipient of the National Outstanding Youth Foundation of China Award. He serves/has served the General Co-Chair of ICME'2022 and the Program Co-Chair of FG'2023, VCIP'2022, AVSS'2021, and ICME'2020. He also serves as the Co-Editor-in-Chief of *Pattern Recognition Letters* and an Associate Editor for the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, the IEEE TRANSACTIONS ON BIOMETRICS, BEHAVIOR, AND IDENTITY SCIENCE, and *Pattern Recognition*.