



# Towards High Fidelity Face Frontalization in the Wild

Jie Cao<sup>1,2,3,4</sup> · Yibo Hu<sup>1,2,3,4</sup> · Hongwen Zhang<sup>1,2,3,4</sup> · Ran He<sup>1,2,3,4</sup> · Zhenan Sun<sup>1,2,3,4</sup>

Received: 29 September 2018 / Accepted: 5 September 2019 / Published online: 12 October 2019  
© Springer Science+Business Media, LLC, part of Springer Nature 2020

## Abstract

Face frontalization refers to the process of synthesizing the frontal view of a face from a given profile. Due to self-occlusion and appearance distortion in the wild, it is extremely challenging to recover faithful high-resolution results meanwhile preserve texture details. This paper proposes a high fidelity pose in-variant model (HF-PIM) to produce photographic and identity-preserving results. HF-PIM frontalizes the profiles through a novel texture fusion warping procedure and leverages a dense correspondence field to bind the 2D and 3D surface spaces. We decompose the prerequisite of warping into dense correspondence field estimation and facial texture map recovering, which are both well addressed by deep networks. Different from those reconstruction methods relying on 3D data, we also propose adversarial residual dictionary learning to supervise facial texture map recovering with only monocular images. Furthermore, a multi-perception guided loss is proposed to address the practical misalignment between the ground truth frontal and profile faces, allowing HF-PIM to effectively utilize multiple images during training. Quantitative and qualitative evaluations on five controlled and uncontrolled databases show that the proposed method not only boosts the performance of pose-invariant face recognition but also improves the visual quality of high-resolution frontalization appearances.

**Keywords** Face frontalization · Realistic face generation · Pose-invariant face recognition

---

Communicated by Xavier Alameda-Pineda, Elisa Ricci, Albert Ali Salah, Nicu Sebe, Shuicheng Yan.

---

✉ Zhenan Sun  
znsun@nlpr.ia.ac.cn

Jie Cao  
jie.cao@cripac.ia.ac.cn

Yibo Hu  
yibo.hu@cripac.ia.ac.cn

Hongwen Zhang  
hongwen.zhang@cripac.ia.ac.cn

Ran He  
rhe@nlpr.ia.ac.cn

- <sup>1</sup> Center for Research on Intelligent Perception and Computing, CASIA, Beijing, China
- <sup>2</sup> National Laboratory of Pattern Recognition, CASIA, Beijing, China
- <sup>3</sup> School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
- <sup>4</sup> Center for Excellence in Brain Science and Intelligence Technology, CAS, Beijing, China

## 1 Introduction

Face frontalization refers to predicting the frontal view image from a given profile. It is an effective preprocessing method for pose-invariant face recognition. Frontalized profile faces can be directly used by general face recognition methods without retraining the recognition models. Recent studies have shown that frontalization is a promising approach to address long-standing problems caused by pose variation in the face recognition system. Additionally, generating photographic frontal faces are beneficial for a series of face-related tasks, including face reconstruction, face attribute analysis, facial animation, etc.

Due to the appealing prospect in theories and applications, research interest has been lasting for years. In the early stage, most traditional face frontalization methods (Dovgard and Basri 2004; Hassner 2013; Hassner et al. 2015; Ferrari et al. 2016; Zhu et al. 2015) are 3D-based. These methods mainly leverage theories of monocular face reconstruction to recover 3D faces and then render frontal view images. The well-known 3D Morphable Model (3DMM) (Blaiz and Vetter 1999) has been widely employed to express facial shape and appearance information. Recently, great breakthroughs

have been made by the methods based on Generative Adversarial Networks (GAN) (Goodfellow et al. 2014). These methods frontalize faces from the perspective of 2D image-to-image translation and build deep networks with novel architectures. The visual realism has been improved significantly, for instance, in Multi-PIE (Gross et al. 2010), some synthesized results (Huang et al. 2017; Zhao et al. 2018) from small pose profiles are so photographic that it is difficult for human observers to distinguish them from the real ones. Moreover, frontalized results have been proven to be effective for tackling the pose discrepancy in face recognition. Through the “recognition via generation” framework, current methods (Zhao et al. 2017, 2018) achieve state-of-the-art pose-invariant face recognition performance on multiple datasets, including Multi-PIE, IJB-A (Klare et al. 2015) and CFP (Sengupta et al. 2016).

Even though significant improvements have been made, there are still some ongoing issues for in-the-wild face frontalization. For traditional 3D-based approaches, due to the shortage of 3D data and the limited representation power of existing 3D models, their performances are commonly less competitive compared with GAN-based methods albeit some improvements (Cole et al. 2017; Tran and Liu 2018) have been made. However, GAN-based methods heavily rely on minimizing the pixel-wise loss to deal with the noisy data for in-the-wild settings. As discussed in many other image restoration tasks (Huang et al. 2017; Johnson et al. 2016), the consequence is that the outputs lack variations and tend to keep close to the statistical meaning of the training data. The results will be over-smoothed with little high-level texture information. Hence, current frontalization results are less appealing in high-resolution, and the output size is often limited to  $128 \times 128$ .

To address the above issues, this paper proposes a High Fidelity Pose Invariant Model (HF-PIM) that combines the advantages of 3D and GAN based methods. In HF-PIM, we frontalize the profiles via a novel texture fusion warping procedure. Inspired by recent progress in 3D face analysis (Güler et al. 2017, 2018), we introduce a dense correspondence field to bind the 2D and 3D surface spaces. Thus, the prerequisite of our warping procedure is decomposed into two well-constrained problems: dense correspondence field estimation and facial texture map recovering. We build an end-to-end neural network to simultaneously address the two problems and benefit from its greater representation power than traditional 3D-based methods. Furthermore, we propose Adversarial Residual Dictionary Learning (ARDL) to get rid of the reliance on 3D data. Thanks to the 3D-based deep framework and the capacity of ARDL for fine-grained texture representation (Dana 2017), high-resolution results with faithful texture details can be obtained. Considering that the profile and frontal face pairs obtained in unconstrained environments have large discrepancies in expressions, illu-

minations, backgrounds, etc., (i.e., misalignment), we extend our model to multi-perception guided HF-PIM to address this problem. In this modification, we propose a multi-perception guided loss to utilize supervision information from multiple possible resources. We make extensive comparisons with state-of-the-art methods on the Multi-PIE, IJB-A, LFW (Huang et al. 2007), and CFP datasets. We also frontalize  $256 \times 256$  images from CelebA-HQ (Karras et al. 2018) to push forward the advance in high-resolution face frontalization. Quantitative and qualitative results demonstrate our method dramatically improves pose-invariant face recognition and produces photographic high-resolution results, potentially benefitting many real-world applications.

A preliminary version (Cao et al. 2018) of this work has been accepted by the Thirty-second Conference on Neural Information Processing Systems (NeurIPS 2018). We extend it in four ways: (1) we emphasize the proposed fusion warping that can be easily integrated into an end-to-end neural network to address the problem of background synthesis effectively. (2) We extend the perceptual loss (Johnson et al. 2016) to the multi-perception guided loss and propose multi-perception guided HF-PIM to address the misalignment between the ground truth frontal and profile faces obtained in unconstrained environments. (3) We conduct experiments on CFP and add numerous evaluation metrics, including FID (Heusel et al. 2017), ROC curves and more visualization results, to demonstrate the effectiveness of our model. (4) We conduct ablation study to reveal how the components of our model work and make discussions about limitations.

To summarize, our main contributions are listed as follows:

- A novel High Fidelity Pose Invariant Model (HF-PIM) is proposed to produce identity-preserving and realistic frontalized face images with a higher resolution.
- Through dense correspondence field estimation and facial texture map recovering, our warping procedure can frontalize profile images with large poses and preserve abundant latent 3D shape information.
- Without the need of 3D data, we propose ARDL to supervise the process of facial texture map recovering, effectively compensating the texture representation capacity of the 3D-based framework.
- Towards overcoming the misalignment between the ground truth frontal and profile faces, we introduce a multi-perception guided loss to utilize supervision information from multiple possible resources.
- A unified end-to-end neural network is built to integrate all algorithmic components, which makes the training process elegant and flexible.
- Extensive experimental results on five face frontalization databases demonstrate that the proposed method not only boosts pose-invariant face recognition in the wild, but

also dramatically improves the visual quality of high-resolution images.

The remainder of this paper is organized as follows: Sect. 2 provides a brief review of related works. Section 3 presents the details of our HF-PIM for face frontalization. Section 4 describes the experimental results, Sect. 5 discusses the limitations and future work, and Sect. 6 summarizes the conclusions.

## 2 Related Works

### 2.1 Generative Adversarial Network

Introduced by Goodfellow et al. (2014), Generative Adversarial Network (GAN) plays a min-max game to improve both discriminator and generator. The generator tries to map a given input distribution to a target data distribution whereas the discriminator tries to distinguish the data produced by the generator from the real one. With the constraints of the min-max game, GAN can encourage the generated images to be close to the real image manifolds. Recently, deep convolutional generative adversarial network (DCGAN) (Radford et al. 2016) has demonstrated the superior performance of image generation. Info-GAN (Chen et al. 2016) applies information regularization to optimization. Furthermore, Wasserstein GAN (Arjovsky et al. 2017) improves the learning stability of GAN and provides solutions for debugging and hyperparameter searching for GAN. These successful theoretical analyses of GAN show the effectiveness and possibility of photorealistic face image generation and synthesis.

### 2.2 Face Frontalization

Face frontalization is an extremely challenging synthesis problem due to its ill-posed nature. Recent methods addressing this problem can be divided into three categories: 3D/2D local texture warping (Hassner et al. 2015; Zhu et al. 2015), statistic modeling (Sagonas et al. 2015) and deep learning based methods (Cole et al. 2017; Kan et al. 2014; Li et al. 2019; Yang et al. 2015; Yim et al. 2015). Hassner et al. (2015) employ a single unmodified 3D reference surface to produce the frontal view face. A joint frontal view reconstruction and landmark localization are optimized by the minimization of the nuclear norm in Sagonas et al. (2015).

With the development of deep learning, Kan et al. (2014) propose SPAE for face frontalization by employing auto-encoders. Yim et al. (2015) introduce multi-task learning for frontal view synthesis. Yang et al. (2015) employ a recurrent transformation unit to synthesize discrete 3D views. Moreover, Cole et al. (2017) decompose faces into a sparse set of

landmarks and aligned texture maps, and then combine them by a differentiable image warping operation.

GAN has dominated the field of face frontalization since it is firstly used by DR-GAN (Tran et al. 2017). Later, TP-GAN (Huang et al. 2017) is proposed with a two-pathway structure and perceptual supervision. CAPG-GAN (Hu et al. 2018) introduces pose guidance through inserting conditional information carried by five-point heatmaps. PIM (Zhao et al. 2018) proposes a “learning to learn” strategy for high-quality and identity-preserving face frontalization. CR-GAN (Tian et al. 2018) introduces a generation sideway to maintain the completeness of the learned embedding space and utilizes both labeled and unlabeled data to further enrich the embedding space for realistic generations. All those methods treat face frontalization as a 2D image-to-image translation problem without considering the intrinsic 3D properties of human face. They indeed perform well in the situation where training data is sufficient and captured well controlled. However, in-the-wild setting often leads to inferior performance, as we discussed in Sect. 1.

The attempt to combine prior knowledge of 3D face has been made by FF-GAN (Yin et al. 2017), 3D-PIM (Zhao et al. 2018) and UV-GAN (Deng et al. 2018). Their and our methods are all 3D-based but there are many differences. In FF-GAN, a CNN is trained to regress the 3DMM coefficients of the input. Those coefficients are integrated as a supplement of low-frequency information. 3D-PIM incorporates a simulator with the aid of a 3DMM to obtain prior information to accelerate the training process and reduce the amount of required training data. In contrast, we do not employ 3DMM to present shape or texture information. We introduce a novel dense correspondence field and frontalize the profiles through warping. UV-GAN leverages an out-of-the-box method to project a 2D face to a 3D surface space. Their network can be regarded as a 2D image-to-image translation model in the facial texture space. In contrast, once the training procedure is finished, our model can estimate the latent 3D information from the profiles without the need for any additional out-of-the-box methods.

### 2.3 Pose-invariant Face Recognition

Different from frontalization methods, conventional approaches aim to achieve pose-invariant face recognition by leveraging robust local descriptors or metric learning. Among them, patch-based approaches are very effective for eliminating relatively small pose discrepancy. For instance, StackFlow (Ashraf et al. 2008) discovers viewpoint-induced spatial deformities at the patch level and warps the non-frontal face image to the frontal one progressively. Without the need of accurate face alignment, Lucey and Chen (2008) propose “patch-whole” algorithm that decomposes the gallery image into an ensemble of statistically independent patches.

However, in recent years, most of these methods are exceeded by deep learning approaches. With the support of big data, a deep network is trained in a pose-agnostic manner or several ones designed for a specific pose are fused jointly to tackle pose variance. Taigman et al. (2014) propose the first recognition pipeline that includes face frontalization and deep identity representation extraction. Subsequently, Masi et al. (2016) employ multiple pose-specific models and utilize multiple identity representations with the aid of 3D modeling. Similarly, the idea of utilizing an ensemble of pose-specific CNN features to reduce the sensitivity of the recognition system to pose variations is proposed by AbdAlmageed et al. (2016). To further improve the robustness of the pose-invariant representations extracted from multiple feature extractors, Li et al. (2016) propose a novel data cleaning strategy. Consensus-Driven Propagation (CDP) (Yan et al. 2018) is proposed to carefully aggregate multi-view information by constructing a relational graph in a bottom-up manner. Inspired by the human visual system, Han et al. (2018) propose the contrastive convolution network, which focuses on those contrastive characteristics between the two faces to be compared. DREAM block is proposed by Cao et al. (2018) to bridge the discrepancy between frontal and profile faces. However, most of the mentioned methods require well-annotated data but collecting faces covering all poses is expensive and even impractical. In contrast, by employing profiling methods to synthesize the ground truth, our HF-PIM can utilize both paired or unpaired faces during the training process.

Achieving pose-invariant face recognition by data augmentation is another promising research direction. Masi et al. (2016) render unseen viewpoints by 3D modeling to better capture intra-subject appearance variations and then train a deep network to achieve face recognition across poses. Later on, inspired by Shrivastava et al. (2017), DA-GAN (Zhao et al. 2017, 2018), which acts as a 2D face image refiner, is proposed to boost pose-invariant face recognition. In brief, the refiner improves the quality of data augmented by conventional methods. The training process benefits from those refined data and the performance is boosted. Thus, DA-GAN is a method for augmenting training data. Note that UV-GAN mentioned above can be used to benefit face recognition in the same manner with DA-GAN, so it is also a data augmentation method. In contrast, our HF-PIM is trained to directly rotate the given profile to the frontal face, which can be directly used for face recognition/verification.

### 3 High Fidelity Pose Invariant Model

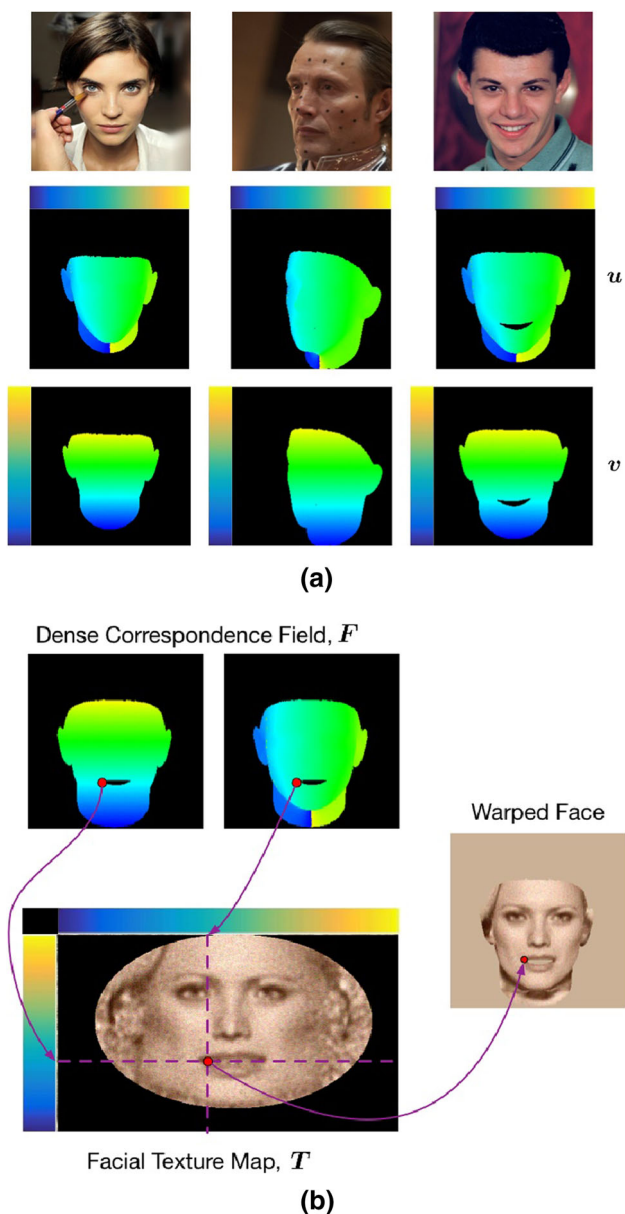
The goal of face frontalization is to model the mapping from profile face  $X$  to the corresponding frontal face  $Y$  ( $X, Y \in \mathbb{R}^{N \times N \times 3}$ ). As a reminder, we use  $I_{ij}$  to denote the pixel

value of the coordinate  $(i, j)$  in an image  $I$ . To learn the mapping, we employ image pairs  $(X, Y)$  for model training and expect the produced frontalized result  $\hat{Y}$  to be as close to the ground truth as possible. Inspired by recent progress in 3D face analysis (Güler et al. 2017, 2018), we propose a brand-new framework which frontalizes the given profile face through recovering geometry and texture information without explicitly building a 3D face model. Concretely, the facial texture map and a novel dense correspondence field are leveraged to produce  $Y$  through warping. The facial texture map  $T$  lies in UV space—a space in which the manifold of the face is flattened into a contiguous 2D atlas.  $T$  represents the surface of the 3D face. The dense correspondence field  $F = (u; v)(u, v \in \mathbb{R}^{N \times N})$  is specified by the following statement: assume that the coordinate of a point in  $T$  is  $(u_{ij}, v_{ij})$ , after warping, the corresponding coordinate in  $Y$  is  $(i, j)$ . Figure 1 provides an intuitionistic illustration. Formally, given  $T$  and  $F$  with respect to  $Y$ ,  $T$  is warped into  $Y$  through the following formulation:

$$Y_{ij} = \text{warp}(i, j; F, T) = T_{u_{ij}, v_{ij}}. \quad (1)$$

Equation 1 depicts how  $F$  establishes the correspondence between a 2D face image and its facial texture map. For every pixel in a 2D face, its value is represented by the corresponding pixels in the facial texture map, and  $F$  specifies the location of the corresponding pixels. Since our proposed  $F$  builds a point-to-point correspondence, we refer to it as dense correspondence field.

Our proposed warping procedure inherits the virtue of morphable model construction: geometry and texture are well disentangled by dense correspondence. However, there are many limitations for traditional construction methods, including heavily relying on 3D data, neglect of image background, etc. To overcome those limitations, we design an end-to-end neural network to provide  $F$  and  $T$  for Eq. 1. Concretely, we introduce a fully convolutional network  $C$  to estimate the dense correspondence field  $F$ , a transformative autoencoder  $E_t - D_t$  to recover the facial texture map  $T$ , and a deep backend neural network  $R$  with fusion warping. As shown in Fig. 2a, our correspondence field estimator  $C$  takes the profile as the input and produces the predicted dense correspondence field of the frontal view face. Our transformative autoencoder,  $E_t - D_t$ , also takes the profile and produces the recovered facial texture map and the facial texture feature map. Finally, the frontalized face is synthesized by fusion warping: the warping layer combines the predicted dense correspondence field and the facial texture map to produce the frontalized facial part; Then, the backend neural network  $R$  produces the complete frontalized face by fusing the frontalized facial part and the facial texture feature map. In the following, we describe how to estimate the



**Fig. 1** **a** Visual examples of the dense correspondence field  $F = (u; v)$ . RGB color images are on the first row. Corresponding  $u$  and  $v$  are on the second and the third row respectively. Color indicates the values of the pixels in  $F$ . **b** A visual illustration about the warping procedure proposed in Eq. 1. Those red dots and purple lines indicate the relationships between the facial texture map, dense correspondence field and the RGB color image (Color figure online)

dense correspondence field  $F$  of the frontal view in Sect. 3.1. Then, the recovering procedure of the facial texture map  $T$  via ARDL is illustrated in Sect. 3.2. The backend network and the fusion warping are introduced in Sect. 3.3. Furthermore, HF-PIM is extended to the multi-perception guided version in Sect. 3.4 to deal with the practical data misalignment problem. The overall training method is summarized in Sect. 3.5.

### 3.1 Dense Correspondence Field Estimation

To obtain the ground truth dense correspondence field  $F$  of monocular frontal face images for training, we employ a face reconstruction method for 3D shape information estimation. Concretely, we employ BFM (Paysan et al. 2009) as the 3D face model. Through the model fitting method proposed by Zhu et al. (2016), we get estimated shape parameters containing coordinates of vertices. To build  $F$ , we follow Cao et al. (2019) and map those vertices to UV space via the cylindrical unwrapping described in Booth and Zafeiriou (2014). We cull those non-visible vertices via z-buffering.

To infer the dense correspondence field of the frontal view from the profile image, we build a transformative autoencoder,  $C$ , with U-Net Ronneberger et al. (2015) architecture. U-Net architecture has been widely employed in segmentation tasks for dense prediction. Considering that predicting the segmentation masks and predicting the correspondence fields can be both regarded as the process of extracting meaningful geometrical representations and making dense predictions, we use U-Net architecture to build the correspondence field autoencoder  $C$ . Given the input profile,  $C$  first encodes it into pose-invariant shape representations and then recover dense correspondence field of the frontal view. Those shortcuts in U-Net guarantee the preservation of spatial information in the output. To supervise  $C$  during training, we minimize the pixel-wise error between the estimated map and the ground truth  $F$ , namely:

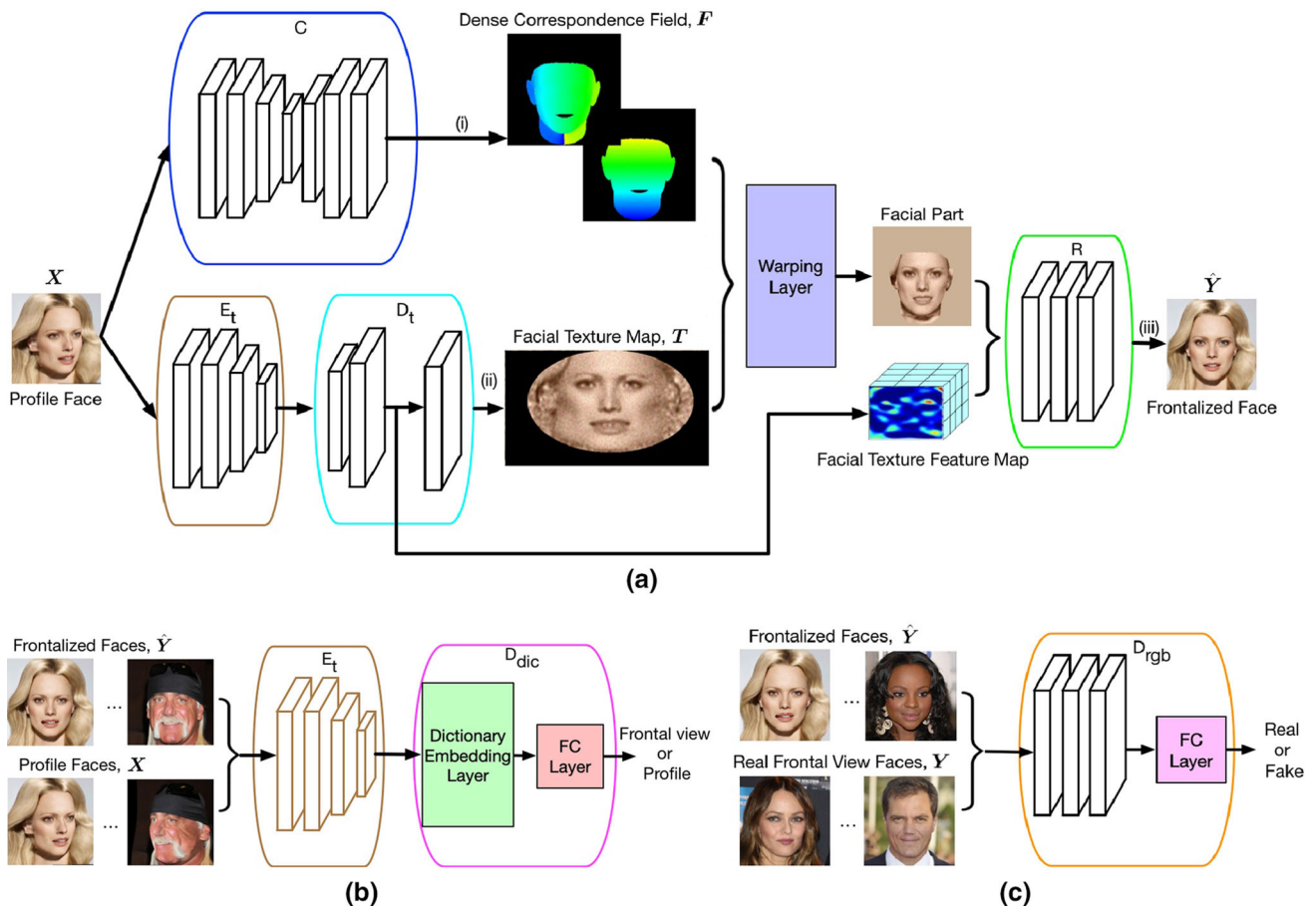
$$L_{corr} = \|C(X) - F\|_1, \tag{2}$$

where  $\|\cdot\|_1$  denotes calculating the mean of the element-wise absolute value summation of a matrix.

### 3.2 Facial Texture Map Recovery

We employ a transformative autoencoder consisting of the encoder  $E_t$  and the decoder  $D_t$  for facial texture map recovering. However, the ground truth facial texture map  $T$  of monocular face image captured in the wild is absent. To sidestep the demand for the ground truth  $T$ , we introduce Adversarial Residual Dictionary Learning (ARDL), as illustrated in Fig. 2b. During the training procedure, only  $Y$  is required instead of  $T$ .

We first describe the dictionary layer in our network. In the field of fine-grained visual recognition, recent advances (Cimpoi et al. 2015; Dana 2017; Sun et al. 2018) have demonstrated the superiority of dictionary embeddings. Inspired by the fact that dictionary embeddings are effective for representing the texture details of different sub-species, we integrate a dictionary embedding layer into our network. Given a set of facial texture feature embeddings  $B = \{b_1, \dots, b_n\}$ , our dictionary layer encodes



**Fig. 2** **a** The framework of our HF-PIM to frontalize face images. The procedures of dense correspondence field estimation (i), facial texture map recovering (ii) and frontal view warping (iii) are discussed in Sects. 3.1, 3.2 and 3.3, respectively. **b** The discriminator employed

for ARDL, which is discussed in Sect. 3.2. “FC Layer” denotes a fully connected layer. **c** The discriminator employed for ordinary adversarial learning, which is discussed in Sect. 3.4

the embeddings into a fixed length dictionary representation  $E = \{e_1, \dots, e_m\}$ . Concretely,  $E$  is calculated as:

$$e_k = \sum_{i=1}^n e_{ik} = \sum_{i=1}^n w_{ik}(\mathbf{b}_i - \mathbf{c}_k) = \sum_{i=1}^n w_{ik} \mathbf{r}_{ik}, \quad (3)$$

where  $C = \{\mathbf{c}_1, \dots, \mathbf{c}_m\}$  denotes the learnable codebook. Our dictionary layer takes the residual as the input, and the residual vector  $\mathbf{r}_{ik} = \mathbf{b}_i - \mathbf{c}_k$ .  $w_{ik}$  is the corresponding weight for  $\mathbf{r}_{ik}$ . Inspired by Van Gemert et al. (2008) that assigns a descriptor to each codeword, we set the weight as:

$$w_{ik} = \frac{\exp(-s_k \|\mathbf{b}_i - \mathbf{c}_k\|^2)}{\sum_{j=1}^m \exp(-s_j \|\mathbf{b}_i - \mathbf{c}_j\|^2)} = \frac{\exp(-s_k \|\mathbf{r}_{ik}\|^2)}{\sum_{j=1}^m \exp(-s_j \|\mathbf{r}_{ij}\|^2)}, \quad (4)$$

where  $s = (s_1, \dots, s_m)$  is the smoothing factor, which is also learnable. In summary, we optimize the codebook and the

smoothing factor to make the dictionary layer map the feature embeddings to the dictionary representation. The mapping is denoted as  $D_{dic}(\cdot)$  in the following parts. In our network, we use the output of the encoder  $E_t$  as the input of the dictionary layer.

We then introduce how we combine dictionary representation with adversarial learning. Our inspiration comes from such an observation: when the identity label is fixed, for  $X$  across different poses, the recovered texture map  $T$  should be invariant. To this end,  $E_t$  should eliminate those discrepancies caused by different views and encode the input into pose-invariant facial texture representation. We introduce the adversarial learning mechanism to supervise  $E_t$  by making  $D_{dic}$  as its rival. Formally, the adversarial loss introduced by ARDL is formulated as:

$$L_{adv} = \mathbb{E}_{\mathbf{X} \sim p_{data}} [\log D_{dic}(E_t(\mathbf{X}))]. \quad (5)$$

Accordingly,  $D_{dic}$  is optimized to minimize:

$$L_{dic} = \mathbb{E}_{\mathbf{X}, \mathbf{Y} \sim p_{data}} [\log D_{dic}(E_t(\mathbf{Y})) + \log(1 - D_{dic}(E_t(\mathbf{X})))] \tag{6}$$

where we add a fully connected (FC) layer upon  $D_{dic}$  to make binary predictions standing for real and fake. Through optimizing Eqs. 5 and 6 alternatively,  $E_t$  manages to make the encodings of the profile and the frontal view as similar as possible. In the meantime,  $D_{dic}$  tries to find the clues standing for pose information, which provides the adversarial supervision information for  $E_t$ . This procedure is illustrated in Fig. 2b.

### 3.3 Fusion Warping

In this subsection, we propose fusion warping that augments the procedure in Eq. 1 to produce the non-facial parts simultaneously with the facial one. As indicated by Fig. 1b, ordinary warping only produces the facial part of the image. The values of pixels standing for non-facial parts are undefined, so post process is necessary to complete the missing regions. In our fusion warping, a backend neural network  $R$  with several convolution layers is employed to integrate the warping result with the second last convolution layers in  $D_t$ . The final output of  $D_t$ , i.e., the predicted facial texture map, remains to be warped into the facial region through Eq. 1. In the meantime, the output of the second last layer of  $D_t$ , which can be regarded as the facial texture feature map (denoted as  $\mathbf{T}^*$ ), is fed into  $R$  along with the warped facial part. The reconstruction loss with fusion warping is formulated as follows:

$$L_{rec} = \|R(warp(\mathbf{F}, \mathbf{T}) \odot \mathbf{T}^*) - \mathbf{Y}\|_1, \tag{7}$$

where  $\odot$  denotes the concatenation operation on the feature channel. In Eq. 7, the dense correspondence field  $\mathbf{F}$  and the facial texture map  $\mathbf{T}$  are the predicted facial geometric and texture representations, respectively. Whereas the concatenated  $\mathbf{T}^*$  provides the information of the non-facial part, which stands for the background in the frontalized image. Our backend network  $R$  integrates these representations and produces the final results. Hence, our network is trained to produce realistic frontalized faces as well as the backgrounds without the need of post process.

### 3.4 Multi-perception Guided Loss

In this subsection, we introduce the multi-perception guided loss to deal with the practical data misalignment problem. Recall that employing the perceptual loss introduced by a fixed neural network is originally proposed by Johnson et al. (2016) for image style transfer. TP-GAN Huang et al. (2017)

first use the perceptual loss in face frontalization. The perceptual loss focuses on the similarity in feature-level and preserve the identity information of the input, which greatly benefits pose-invariant face recognition. In this paper, the perceptual loss is formulated as:

$$L_p = \|\phi(R(warp(\mathbf{F}, \mathbf{T}) \odot \mathbf{T}^*)) - \phi(\mathbf{Y})\|_2^2, \tag{8}$$

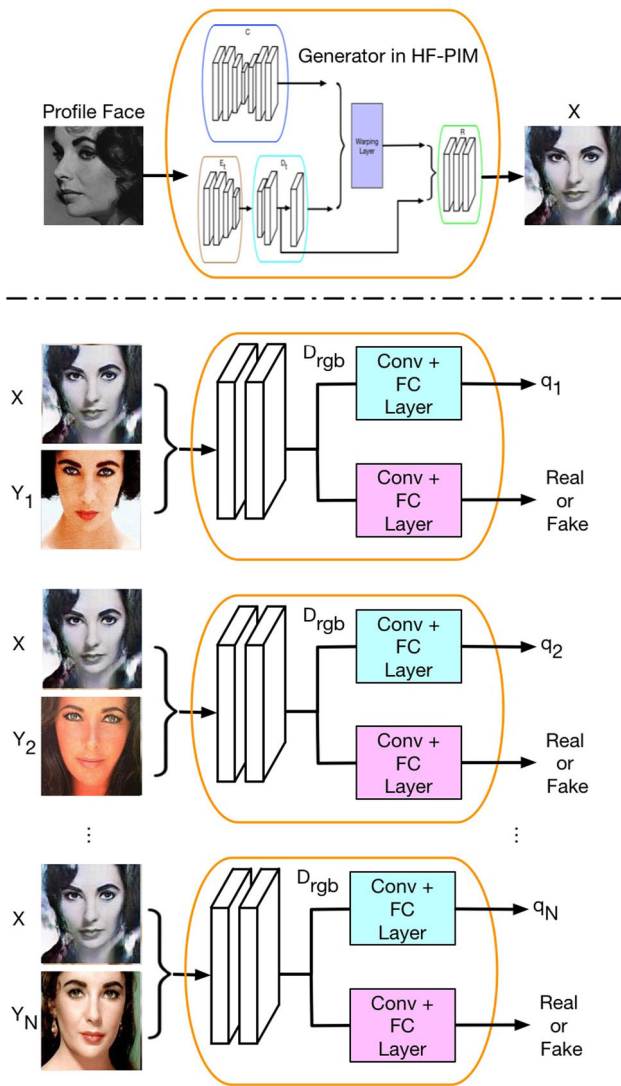
where  $\phi(\cdot)$  denotes the extracted identity representation obtained by the second last fully connected layer within the identity preserving network and  $\|\cdot\|_2$  denotes the vector 2-norm. In our experiment, we employ Light CNN (Wu et al. 2018) as our identity preserving network.

Furthermore, we extend the original perceptual loss to multi-perception guided loss. The intuition comes from such an observation: when employing the original perceptual loss, we assume that the identity label of each face pair is different across the entire training set. However, in most unconstrained datasets (e.g., the CFP dataset), multiple frontal faces can be provided for each person. But those images are not optimal: their expressions, illuminations, backgrounds, and other factors may vary from each other and the given profiles. This condition is referred to as the misalignment problem between the ground truth frontal and profile faces. On the one hand, it is desirable to utilize the multiple images for guidance. On the other hand, the noise introduced by them must be reduced to yield better performance. To this end, we propose multi-perception guided HF-PIM by replacing the original perceptual loss with the multi-perception guided loss, as illustrated in Fig. 3. Formally, we formulate the modified perceptual loss as:

$$L_p = \left\| \phi(R(warp(\mathbf{F}, \mathbf{T}) \odot \mathbf{T}^*)) - \sum_{i=1}^N d_i \phi(\mathbf{Y}_i) \right\|_2^2. \tag{9}$$

As described in Eq. 3, our multi-perception guided HF-PIM is trained to minimize the distance between the frontalized images and the fused representation of the  $N$  ground truth ones in feature-level.  $d_i$  denotes the learnable weighting factor for each ground truth frontal view face. Since  $\mathbf{d}$  is learnable, our network is trained to adjust the contribution of each frontal face to the fused representation. In our experiments, we set  $N = 4$  for the sake of the convenience of network training.

Our discriminator  $D_{rgb}$  is modified to evaluate the coefficient  $\mathbf{d}$  jointly with predicting the reality of frontalized faces. Specifically,  $D_{rgb}$  takes paired images as input. The real pair consists of two real faces, and the fake pair consists of a real face as well as a frontalized face. The two images are drawn from the same identity and concatenated together as the input. At the end of the second last convolution layer in  $D_{rgb}$ , we add another branch which consists of a convolution layer and a fully connected one to output the weight prediction  $\mathbf{q}$  for a



**Fig. 3** An illustration about the modification in multi-perception guided HF-PIM. Given the frontalized face and a set of real frontal faces of the same identity as the ground truth,  $D_{rgb}$  outputs the weight prediction (denoted as  $q$ ) to calculate the weighting factor  $d$  for obtaining the fused identity representation

given image pair. Sigmoid activation is applied to scale  $q$  to the range of  $[0, 1]$ . Then the weighting factor is calculated as  $d_i = \frac{q_i}{\sum_{j=1}^N q_j}$ . Inspired by Tran et al. (2018), we also apply dropout on  $q$ . This simple modification makes the network takes perceptual guidance of multiple images varying from 1 to  $N$ .

Besides the minor modifications in the network structure, the data sampling scheme is also altered accordingly to ensure  $N$  images of the same identity are sampled for each training batch. Compared with the original version, no modification is required for our multi-perception guided HF-PIM during the testing phase because the discriminators are discarded and only the generator is employed to frontalize the profiles.

### 3.5 Overall Training Method

We also introduce the adversarial loss in the RGB color image space following those GAN-based methods (Tran et al. 2017; Huang et al. 2017; Hu et al. 2018; Zhao et al. 2018; Yin et al. 2017). A CNN named  $D_{rgb}$  is employed to give adversarial supervision in color space, as shown in Fig. 2c. Note that our method can be easily extended to those advanced versions (Arjovsky et al. 2017; Mao et al. 2017) of GAN. But in this paper, we simply use the original form of adversarial loss function (Goodfellow et al. 2014) to prove that the effectiveness comes from our own contributions.

In summary, all the involved algorithmic components in our network are differentiable. Hence, the parameters can be optimized in an end-to-end manner via gradient backpropagation. The whole training process is described in Algorithm 1. Note that  $\lambda_{rec}$ ,  $\lambda_{corr}$ ,  $\lambda_{adv}$ ,  $\lambda_p$ , and  $\lambda_g$  are pre-defined weights for the corresponding loss terms.

#### Algorithm 1 Training algorithm of HF-PIM

- 1: **Input:** profile  $X$ , the ground truth frontal view  $Y$  with ground truth dense correspondence field  $F$ , maximum iteration  $iter$  and the identity preserving network Wu et al. (2018).
- 2: **Output:** the frontalized result  $\hat{Y}$
- 3: Initializing  $C$ ,  $E_t$ ,  $D_t$ ,  $R$ ,  $D_{dic}$ ,  $D_{rgb}$
- 4:  $i \leftarrow 0$
- 5: **while**  $i < iter$  **do**
- 6:   Sampling training data
- 7:   Model forward propagation
- 8:   Calculating  $L_{rec}$ ,  $L_{corr}$ ,  $L_{dic}$ ,  $L_{adv}$  and  $L_p$
- 9:   Calculating the adversarial losses in the RGB color image space, i.e.,  $L_g$  (for the generator) and  $L_d$  (for the discriminator)
- 10:    $L \leftarrow \lambda_{rec}L_{rec} + \lambda_{corr}L_{corr} + \lambda_{adv}L_{adv} + \lambda_pL_p + \lambda_gL_g$
- 11:   Optimize  $C$ ,  $R$ ,  $E_t$ ,  $D_t$  by minimizing  $L$
- 12:   Optimize  $D_{dic}$  by minimizing  $L_{dic}$
- 13:   Optimize  $D_{rgb}$  by minimizing  $L_d$
- 14:    $i \leftarrow i + 1$
- 15: **end while**

## 4 Experiments

### 4.1 Datasets

To demonstrate the superiority of our method in both controlled and unconstrained environments and produce high-resolution face frontalization results, we conduct our experiment on the following five datasets:

*Multi-PIE* (Gross et al. 2010) is established for studying on PIE (pose, illumination, and expression) invariant face recognition. 4 sessions, 20 illumination conditions, 15 poses and 6 expressions of 337 subjects were captured in controlled environments. We employ Setting-2 proposed by Huang et al. (2017) to split the Multi-PIE dataset. Concretely, we use the



face images with the neutral expression under 20 illuminations and 13 poses within  $\pm 90^\circ$ . The two poses from the two additional cameras (08\_1 and 19\_1) located above the subject are not considered. The first 200 subjects are used for training and the rest 137 ones for testing. Each testing identity has one gallery image from his/her first appearance. Hence, there are 72,000 and 137 images in the probe and gallery sets, respectively.

*LFW* (Huang et al. 2007) is a benchmark database for face recognition. It contains 13,233 images of 5,749 people and has been widely used to evaluate synthesis or verification performance of various methods under unconstrained environments. Since the face images in LFW are collected from the web and contain various pose, expression and illumination variations, it is extremely challenging to synthesize a photorealistic frontal face. In our experiment, LFW is only used for testing. Following the verification protocol (Huang et al. 2007), face images are divided into 10 folds that contain different identities and 600 face pairs.

*IJB-A* (Klare et al. 2015) is one of the most challenging in-the-wild face recognition datasets at present. It has 5,396 images and 20,412 video frames of 500 subjects with large pose variations captured from in-the-wild environments to avoid the near frontal bias. In this paper, we follow the testing protocols in Klare et al. (2015) for face recognition and verification.

*CFP* (Sengupta et al. 2016) is a face dataset established specifically for frontal to profile face verification problem. The yaw angle of most profiles in this dataset are nearly  $90^\circ$ . It consists of 500 subjects and there are 10 frontal and 4 profile images per subject. The experimental protocols provided by CFP divide the data into 10 splits with a pairwise disjoint set of individuals in each split. There are 50 individuals, 350 same and 350 not-same pairs per split. In this paper, we focus on the Frontal-Profile protocols to demonstrate the effectiveness of our model.

*CelebA-HQ* (Karras et al. 2018) established by Karras et al. (2018) is a high-resolution subset of CelebA (Liu et al. 2015). CelebA (Liu et al. 2015) is a large-scale face attributes dataset. Contained images cover large pose variations and background clutter. CelebA-HQ consists of 30,000 faces chosen from CelebA. In Karras et al. (2018), a series of processing methods have been applied to improve the resolution and image quality of the selected faces.

## 4.2 Experimental Settings

*Implementation Details.* In our experiment, face images are normalized to  $128 \times 128$  and  $256 \times 256$  for regular and high-resolution face frontalization, respectively. Image intensities are linearly scaled to the range of  $[-1, 1]$ . We use the landmarks of the centers of eyes and mouth to normalize face images by the method proposed by Wu et al. (2018). To test

on different datasets, we train several different HF-PIM by using different training data. To test on the Multi-PIE, we train our network on the training set of Multi-PIE. To test on LFW and IJB-A, the training data consists of the training set of CelebA-HQ and Multi-PIE. Note that the images in CelebA-HQ are downsampled to  $128 \times 128$  in this case. To test on CFP, we fine-tune the network employed for testing on LFW and IJB-A on the training split of CFP. To generate  $256 \times 256$  results, we train our HF-PIM only on the training set of CelebA-HQ. We adapt the model architecture in Zhu et al. (2017) to build our networks. We determine the weights for various loss terms in Algorithm 1 by cross-validation on Multi-PIE. Our empirical rule is that adjusting the weight to make all the losses have the same order of magnitude, and we simply increase/decrease the weight by a factor of 10. In our experiments, we set  $\lambda_{rec}$  to 10 and all the other weights to 1. We use Adam optimizer with a learning rate of  $1e-4$  and  $\beta_1 = 0.5$ ,  $\beta_2 = 0.99$ . Our proposed method is implemented based on the deep learning library Pytorch (Paszke et al. 2017). Two NVIDIA Titan X GPUs with 12GB GDDR5X RAM are employed for the training and testing processes.

*Evaluation Metrics* Evaluating the face recognition performances by “recognition via generation” is the most common method to measure the quality of the frontalized faces, which means profiles are frontalized first, and then the performance is evaluated on those processed face images. Besides, the visual quality is compared, as most GAN-based methods do. The Fréchet Inception Distance (FID) (Heusel et al. 2017) has been recently proposed to measure the performance of image generation tasks. In our experiment, FID is also employed to measure the performances quantitatively.

*Generating Paired Training Data* For the Multi-PIE dataset, subjects have images across different poses captured simultaneously. Those images are perfect for training. However, the faces in CelebA-HQ are all captured in the wild, so paired data is not available. To address this issue, we employ HPEN (Zhu et al. 2015) to rotate the frontal view faces to profiles. To meet the requirement of HPEN, we first detect the facial landmarks by applying the official implementation<sup>1</sup> of Bulat and Tzimiropoulos (2017). In this process, we discard 29 samples which the algorithm failed to detect the landmarks. Then, we employ the “Pose Adaptive 3DMM fitting” part of HPEN to fit the BFM (Paysan et al. 2009) face model and get the fitted parameters, including pose, shape and expression. We calculate the yaw angles from the pose parameters and divide the subjects by yaw angles: if the yaw angle is smaller than  $\pm 5^\circ$ , we treat this face as the frontal view one and send it to the training set; if the yaw angle is larger than  $\pm 15^\circ$ , we regard this face as a profile and send it to the testing set; otherwise, we discard this sample. Hence, we have 19,203 and 5,998 in the training and the testing sets of

<sup>1</sup> <https://github.com/ladrianb/face-alignment>.

**Table 1** Comparisons of rank-1 recognition rates (%) across views on Multi-PIE

Method	$\pm 15^\circ$	$\pm 30^\circ$	$\pm 45^\circ$	$\pm 60^\circ$	$\pm 75^\circ$	$\pm 90^\circ$
DR-GAN Tran et al. (2017)	94.9	91.1	87.2	84.6	–	–
FF-GAN Yin et al. (2017)	94.6	92.5	89.7	85.2	77.2	61.2
Light CNN Wu et al. (2018)	98.6	97.4	92.1	62.1	24.2	5.5
TP-GAN Huang et al. (2017)	98.7	98.1	95.4	87.7	77.4	64.6
CAPG-GAN Hu et al. (2018)	99.8	99.6	97.3	90.3	83.1	66.1
3D-PIM Zhao et al. (2018)	99.6	99.5	98.8	98.4	95.2	86.7
PIM Zhao et al. (2018)	99.3	99.0	98.5	98.1	95.0	86.5
HF-PIM (Ours)	99.99	99.98	99.88	99.14	96.40	92.32

CelebA-HQ, respectively. Finally, we employ the rotation part of HPEN to produce the corresponding profiles of the frontal faces in the training set. Concretely, we profile these faces to  $\pm 15^\circ, \pm 30^\circ, \dots, \pm 75^\circ, \pm 90^\circ$  (just like Multi-PIE). We only alter the pose parameters, so other attributes, e.g., expressions, illuminations, etc., remain unchanged.

### 4.3 Frontalization Results in Controlled Situations

In this subsection, we systematically compare our method with DR-GAN (Tran et al. 2017), TP-GAN (Huang et al. 2017), FF-GAN (Yin et al. 2017), CAPG-GAN (Hu et al. 2018), PIM (Zhao et al. 2018) and 3D-PIM (Zhao et al. 2018) on the Multi-PIE dataset. Those profiles with extreme poses ( $75^\circ$  and  $90^\circ$ ) are very challenging cases.

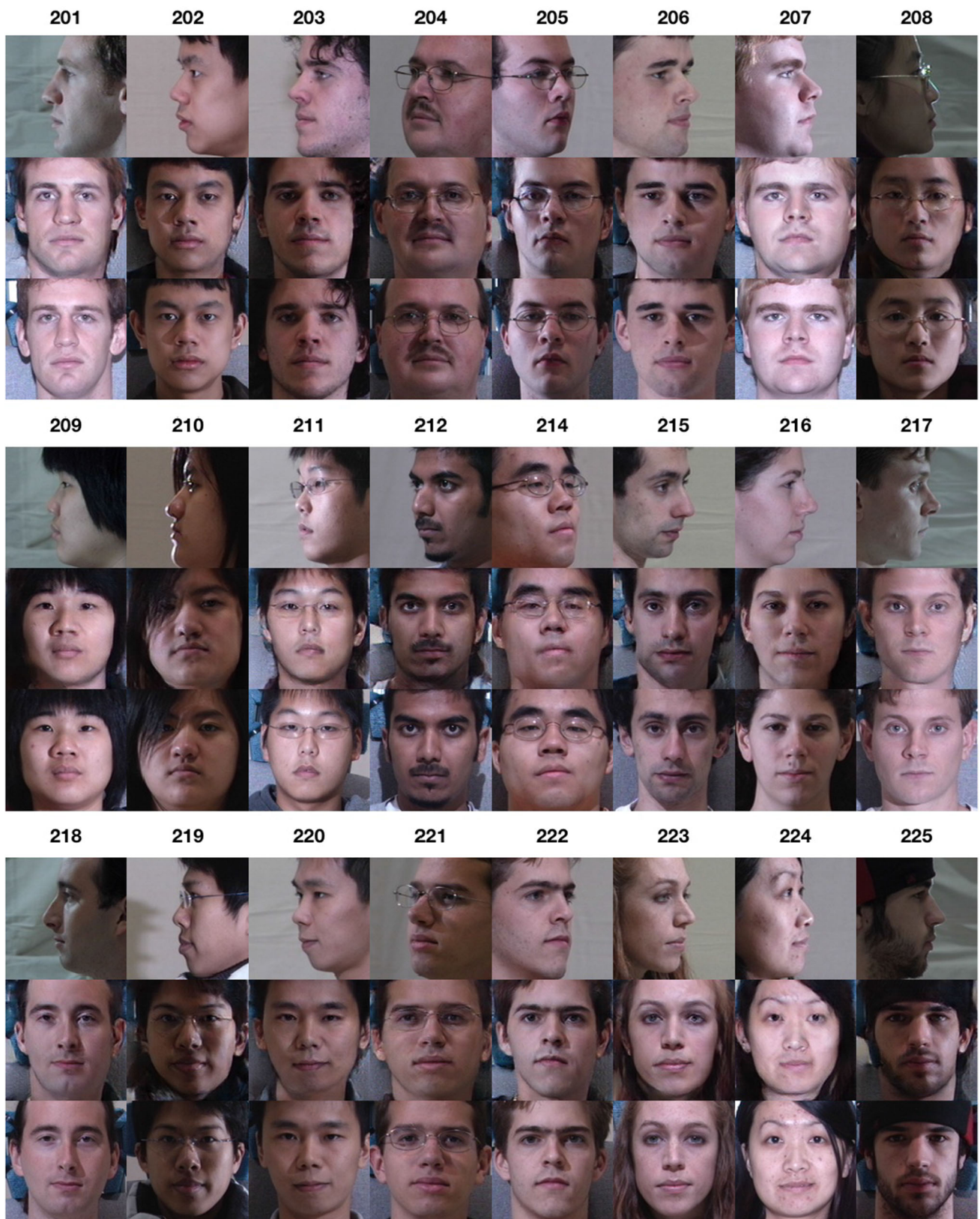
To demonstrate the superiority of our model on identity preserving, we test the face recognition performance on Multi-PIE. Remind that our performance is evaluated by the “recognition via generation” framework. Concretely, profiles are first frontalized by our model and then used directly for recognition. After the frontalization preprocessing, Light CNN (Wu et al. 2018) is employed as the feature extractor. We compute the cosine distance of the extracted feature vectors for face recognition. The manners for evaluating TP-GAN, FF-GAN, PIM, CAPG-GAN, and 3D-PIM are the same as our model. Light CNN is used for these methods except FF-GAN (their feature extractor is not publicly available). DR-GAN is evaluated in a different manner: the feature vectors are directly extracted by their model. Thus, no extra feature extractor is needed for DR-GAN. Besides frontalization methods, the performance of Light CNN is also included as the baseline. The results are reported across different poses in Table 1. For those poses less than  $60^\circ$ , the performances of most methods are quite good, whereas our method performs obviously better. We infer that the performance has almost saturated in this case. For those extreme poses, our method still achieves state-of-the-art recognition performance. Note that we employ the original HF-PIM rather than the multi-perception guided one. The reason is that for the Multi-PIE dataset, images across different poses are captured simulta-

neously in a totally controlled situation. Thus, for a given profile, the corresponding frontal face is perfectly matched. There is no need to involve multi-perception guidance in such a case.

Visual inspections on Multi-PIE are shown in Fig. 4. The ID numbers of the subjects in the figure are consecutive (from 201 to 225). We observe that HF-PIM can not only well preserve the overall facial structure but also recover the unseen ears and cheeks in an identity consistent way. For most subjects, the frontalized results preserve both the visual realism and the characteristics of identities very well. These results also suggest that given enough training data and a proper network structure, it is feasible to synthesize a photorealistic frontal face image from a large pose.

FID is adopted to reveal the performance of general image generation tasks. Lower FID score indicates that the Wasserstein distance between two distributions is smaller. In our experiment, we follow Miyato and Koyama (2018) and compute FID between the frontal faces and the frontalized ones. In Miyato and Koyama (2018), Inception V3 model (Szegedy et al. 2015) is employed to extract feature vectors from images for calculating FID. However, we argue that Inception models are trained for image classification and may not effectively reflect the quality of faces. Thus, we also employ Light CNN as the feature extractor. The results are reported across different yaw angles in Table 2. FID between the profiles and the real frontal faces across different angles are also reported as a reference. We can see that employing different feature extractors, Inception V3 or Light CNN, leads to similar observations: our frontalized faces have lower FID scores than the original profiles, indicating that faces frontalized by our HF-PIM become more similar with the frontal ones in feature level.

The time complexity is also an important factor in face recognition applications. Hence, we make comparisons of the inference time with both the traditional 3D-based methods (LFW-HPEN Zhu et al. 2015 and LFW-3D Hassner et al. 2015) and GAN-based methods (DR-GAN Tran et al. 2017 and CAPG-GAN Hu et al. 2018), as shown in Table 3. We employ a single NVIDIA Titan XP GPU and a single CPU



**Fig. 4** Visualization results on Multi-PIE. The numbers stand for the ID of the subjects below. For each subject, the input image is on the top, the frontalized result is in the middle and the ground truth on the bottom. The illuminations of the inputs are randomly collected

**Table 2** FID across different views on Multi-PIE

Feature Extractor	$\pm 15^\circ$	$\pm 30^\circ$	$\pm 45^\circ$	$\pm 60^\circ$	$\pm 75^\circ$	$\pm 90^\circ$
Light CNN Wu et al. (2018)	6.920/ <b>6.681</b>	7.826/ <b>7.467</b>	10.495/ <b>8.196</b>	13.982/ <b>9.600</b>	19.571/ <b>11.669</b>	33.885/ <b>14.663</b>
Inception V3 Szegedy et al. (2015)	16.583/ <b>15.698</b>	45.269/ <b>33.368</b>	68.367/ <b>46.529</b>	77.426/ <b>64.120</b>	96.710/ <b>82.192</b>	154.055/ <b>129.166</b>

Numbers in italics denote the distance between the profile set and the frontal face set. Numbers in bold type denote the distance between the frontalized face set and the real frontal face set

**Table 3** Comparisons of the inference time (ms) on frontalizing a  $128 \times 128$  image

Setting	LFW-HPEN Zhu et al. (2015)	LFW-3D Hassner et al. (2015)	DR-GAN Tran et al. (2017)	CAPG-GAN Hu et al. (2018)	Ours
CPU	$1.01 \times 10^3$	$1.37 \times 10^2$	–	–	$4.57 \times 10^2$
GPU	–	–	14.95	29.67	5.77

“–” means the result is not reported

**Table 4** Face recognition/verification performance (%) comparisons on IJB-A

Method	Verification		Recognition	
	FAR=0.01	FAR=0.001	Rank-1	Rank-5
DR-GAN Tran et al. (2017)	77.4 $\pm$ 2.7	53.9 $\pm$ 4.3	85.5 $\pm$ 1.5	94.7 $\pm$ 1.1
FF-GAN Yin et al. (2017)	85.2 $\pm$ 1.0	66.3 $\pm$ 3.3	90.2 $\pm$ 0.6	95.4 $\pm$ 0.5
PIM Zhao et al. (2018)	93.3 $\pm$ 1.1	87.5 $\pm$ 1.8	94.4 $\pm$ 1.1	–
Light CNN Wu et al. (2018)	91.5 $\pm$ 1.0	84.3 $\pm$ 2.4	93.0 $\pm$ 1.0	–
HF-PIM (Ours)	95.3 $\pm$ 0.7	89.9 $\pm$ 1.3	96.4 $\pm$ 0.5	98.1 $\pm$ 0.2

The results are averaged over 10 testing splits. “–” means the result is not reported

**Table 5** Face verification performance (%) comparisons on LFW

Method	ACC	AUC
TP-GAN Huang et al. (2017)	96.13	99.42
FF-GAN Yin et al. (2017)	96.42	99.45
CAPG-GAN Hu et al. (2018)	99.37	99.90
Light CNN Wu et al. (2018)	99.39	99.87
HF-PIM (Ours)	<b>99.41</b>	<b>99.92</b>

Bold values indicate the best performance among all the compared methods

(i7-4790) for GPU and CPU settings in the table, respectively. We can see that our HF-PIM on GPU is the fastest among all the compared methods. Even in the CPU setting, our HF-PIM still takes less than 1 second.

#### 4.4 Frontalization Results in the Wild

Extending face frontalization to in-the-wild setting is a very challenging problem of significant importance. In this paper, we evaluate our method on three widely used in-the-wild datasets, including IJB-A, LFW, and CFP. Note that to make fair comparisons with previous methods, we only use the training set of CelebA-HQ and Multi-PIE to test on IJB-A and LFW. To test on CFP, we fine-tune the network employed for testing on LFW and IJB-A on the training set of CFP

for each split and introduce multi-perception guidance. This model is denoted as HF-PIM (mpg). Quantitative results are summarized in Tables 4, 5, and 6 for IJB-A, LFW, and CFP, respectively. The ROC curves are shown in Figs. 5, 6, and 7. Additionally, we also compare our results with DA-GAN Zhao et al. (2017) in Table 7. Since all the methods compared in Table 4 do not use the training data of IJB-A, which is different from DA-GAN, we do not include DA-GAN in that table. To compare with DA-GAN, we follow their training protocol and finetune our baseline model, i.e., Light CNN Wu et al. (2018), on the training data of each split.

We can see that the face frontalization methods only marginally improve the performance on LFW because most faces in this dataset are (near) frontal view. Besides, the baseline model Light CNN has already achieved relatively high performance. But our method still outperforms existing frontalization methods in this case.

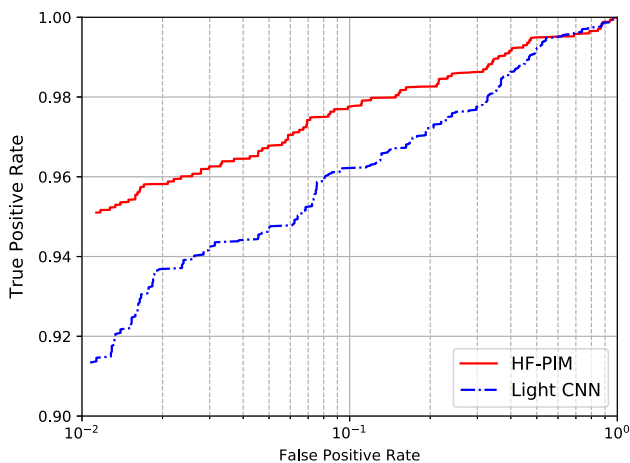
When tested on IJB-A that contains lots of images with large poses, our method shows a significant improvement for face verification and recognition. The visualization comparison<sup>2</sup>, which is shown in Fig. 8, also proves our superiority for preserving the identity information and the texture details.

<sup>2</sup> Visualization results produced by other methods are released by their authors. Different methods usually report visual examples of different identities. We try our best to find those identities reported by most methods.

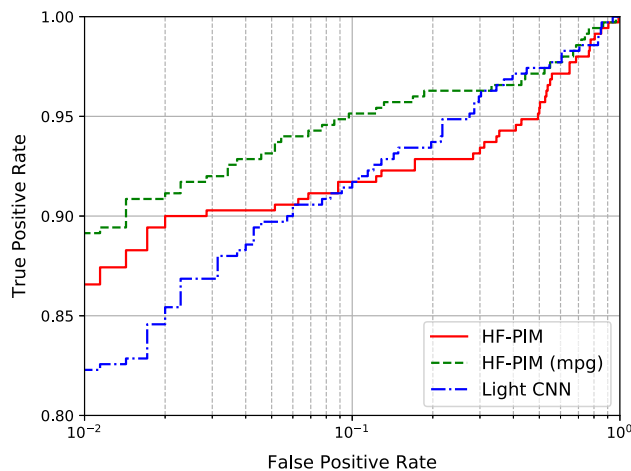
**Table 6** Face recognition performance (%) comparisons on CFP

Method	Frontal-profile		
	ACC	EER	AUC
DR-GAN Tran et al. (2017)	93.41 ± 1.17	6.45 ± 0.16	97.96 ± 0.06
DR-GAN (color) Tran et al. (2018)	93.64 ± 1.51	6.22 ± 0.32	98.18 ± 0.38
PIM Zhao et al. (2018)	93.10 ± 1.01	7.69 ± 1.29	97.65 ± 0.62
Light CNN Wu et al. (2018)	92.47 ± 1.44	8.71 ± 1.80	97.77 ± 0.76
HF-PIM	94.00 ± 0.97	8.86 ± 1.74	97.41 ± 0.65
HF-PIM (mpg)	<b>94.71 ± 0.83</b>	5.86 ± 1.13	98.21 ± 0.55
Human	94.57 ± 1.10	<b>5.02 ± 1.07</b>	<b>98.92 ± 0.46</b>

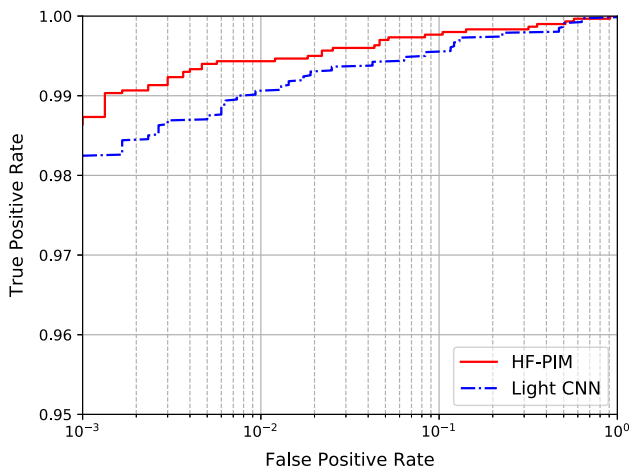
The results are averaged over 10 testing splits. “-” means the result is not reported



**Fig. 5** ROC curves on the IJB-A verification protocol



**Fig. 7** ROC curves on the CFP verification protocol



**Fig. 6** ROC curves on the LFW verification protocol

Thanks to the 3D-based framework and the powerful adversarial residual dictionary learning, our HF-PIM produces results with very high fidelity. For other methods, they indeed produce reasonable images but redundant manipulations can be observed. For instance, DR-GAN makes the eyes of the subject in the second row in IJB-A open; TP-GAN and CAPG-GAN tend to change the skin color and background.

When compared with DA-GAN, we can see despite that our baseline model is inferior, our method still achieves a comparable performance. We argue that a completely fair comparison with DA-GAN is impossible because our model and DA-GAN are designed to boost face recognition in entirely different manners. Besides, DA-GAN does not focus on generating photorealistic face images.

The results on the CFP dataset also demonstrate the effectiveness of our model. It is notable that the training set for the original HF-PIM consists of no profiles in 90° captured under unconstrained situation but nearly all the profiles in CFP are 90°. Thus, the limitations of HF-PIM mainly result from the lack of data, which is overcome by the multi-perception guided version. Although the original HF-PIM does not show obvious advantages over the baseline Light CNN in Table 6 and Fig. 7, HF-PIM (mpg) dramatically improves the verification performance, even comparable with human experts. Since face frontalization on CFP is a very challenging problem, few methods have reported their visualization results. Thus, we mainly make comparisons with DR-GAN Tran et al. (2018) in Fig. 9. We can see that generating convincing frontalization results even for extreme poses in the wild is possible. Our HF-PIM can faithfully preserve the visual

**Table 7** Comparisons between our HF-PIM and DA-GAN (Zhao et al. 2017) on IJB-A. The results are averaged over 10 testing splits

Method	Verification		Recognition	
	FAR=0.01	FAR=0.001	Rank-1	Rank-5
Baseline	96.3 ± 0.7	92.0 ± 0.6	96.6 ± 0.6	98.7 ± 0.3
DA-GAN Zhao et al. (2017)	<b>97.6 ± 0.7</b>	93.0 ± 0.5	97.1 ± 0.7	<b>98.9 ± 0.3</b>
Baseline	93.95 ± 1.04	91.79 ± 0.74	95.29 ± 0.81	97.60 ± 0.65
HF-PIM	96.53 ± 0.72	<b>93.49 ± 0.66</b>	<b>97.26 ± 0.46</b>	98.17 ± 0.23

The baseline models of HF-PIM and DA-GAN are different  
 Bold values indicate the best performance among all the compared methods



**Fig. 8** Visualization comparisons of face frontalization results. The samples on the left are drawn from LFW and the right side are from IJB-A



**Fig. 9** Visualization comparisons of the frontalization results on CFP. **a** The input profiles. **b** The faces frontalized by our Multi-perception Guided HF-PIM. **c** The faces frontalized by our HF-PIM without the multi-perception guided loss. **d** The faces synthesized by DR-GAN (Tran et al. 2018)

**Fig. 10** High-resolution frontalized results on the testing set of CelebA-HQ. From left to right, the first column is the input profile images, and the second column is our frontalized results. The results of CAPG-GAN and TP-GAN are shown in the third and the fourth columns, respectively



properties except pose. The expression, illumination and resolution of our frontalization results are very consistent with the inputs. In contrast, DR-GAN tends to produce results with similar hues, which indicates the heavy reliance on the pixel-wise loss.

#### 4.5 High-Resolution Face Frontalization

Generating high-resolution results has great importance on extending the application of face frontalization. However, due to its difficulty, few methods consider producing images with a size larger than  $128 \times 128$ . To further demonstrate our superiority, frontalized  $256 \times 256$  results on CelebA-HQ are proposed in this paper. Some samples are shown in Figs. 10 and 11. We also make comparisons with TP-GAN and CAPG-GAN. Note that since results on CelebA-HQ have not been reported by previous methods, we contact the authors to get their model and produce  $128 \times 128$  results through carefully following their instructions. The images in CelebA-HQ contain rich textures that are difficult for the generator to reproduce faithfully. Even in such a challenging

situation, HF-PIM is still able to produce plausible results. By contrast, the results of Hu et al. (2018) and Huang et al. (2017) look less appealing.

#### 4.6 Ablation Study

The superiority of our novel 3D-based deep framework has been verified by the comparisons with other methods. In this section, we propose two groups of variants: one group is designed to verify the effectiveness of the reconstruction loss ( $L_{rec}$ ) as well as the adversarial loss ( $L_g$ ), and whether it is necessary to employ symmetric loss like TP-GAN Huang et al. (2017); The other group is designed to verify the effectiveness of our proposed ARDL.

In the first group, let us denote our HF-PIM as “with  $L_{rec}$ , with  $L_g$ , without  $L_{sym}$ ”. We investigate the following variants:

- V1 (“without  $L_{rec}$ , with  $L_g$ , without  $L_{sym}$ ”): We directly remove  $L_{rec}$  from the total loss function.

**Fig. 11** More frontalization results on CelebA-HQ. The first and the third rows are the input images. The second and the fourth rows are corresponding frontalized images



- V2 (“with  $L_{rec}$ , without  $L_g$ , without  $L_{sym}$ ”): We remove  $L_g$ , i.e., the adversarial loss measured in the RGB color space, from the loss function. Accordingly,  $D_{rgb}$  is also excluded from the training process.
- V3 (“with  $L_{rec}$ , with  $L_g$ , with  $L_{sym}$ ”): The symmetric loss originally used in TP-GAN is added to the total loss function.

In the second group, recall that our proposed ADRL subtly integrates both the deep dictionary representation and the adversarial learning into HF-PIM. Hence, we denote our HF-PIM as “with DL, with  $L_{adv}$ ” in the second group. DL means dictionary learning, and  $L_{adv}$  is the adversarial loss measured in the embedding space, as described in Eq. 5. Specifically, the following variants are investigated:

- V4 (“without DL, with  $L_{adv}$ ”): We remove the dictionary learning in ADLR. In this case, the dictionary embedding layer is replaced by a fully connected layer.
- V5 (“with DL, without  $L_{adv}$ ”): We remove the adversarial learning in ADLR. In this case,  $D_{dic}$  is removed and

$L_{adv}$  is obtained through calculating the vector 2-norm of  $(E_t(Y) - E_t(X))$ .

- V6 (“without DL, without  $L_{adv}$ ”): We directly remove ARDL from our network. In this case,  $D_{dic}$  is removed and  $L_{adv}$  is excluded from the training process.

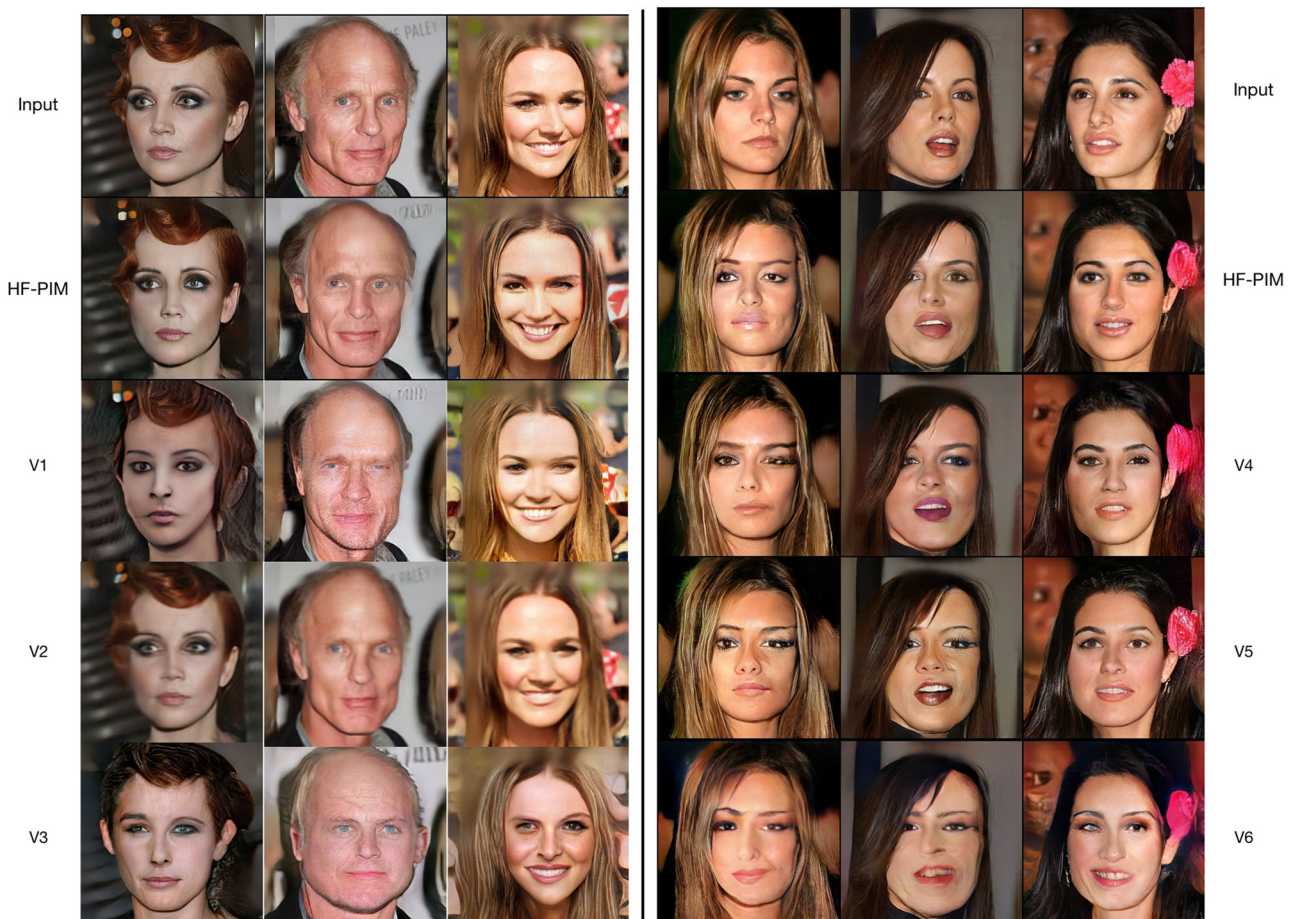
We summarize the mentioned variants and report quantitative results in Table 8. Furthermore, we report the visualization results in Fig. 12. Through the comparisons between our HF-PIM and its variants, the effectiveness of our model is demonstrated. The effectiveness of the adversarial loss and the reconstruction loss is verified by the inferior performances of V1 and V2, respectively. We can see that very symmetric faces can be obtained by V3. However, the visual quality and the recognition performances drop markedly in this case. The reason is that human faces are not strictly symmetric (hair, accessories, expressions, etc.). Thus, the symmetric loss is not included in our HF-PIM. The effectiveness of our ARDL can be demonstrated by V4, V5, and V6. Both quantitative and qualitative results indicate that the dictionary embedding layer is more powerful than ordinary fully



**Table 8** (a) A summary of the variants involved in our ablation study. (b) Face recognition performance (%) comparisons for our HF-PIM and its variations on the IJB-A dataset

Model	$L_{rec}$	$L_g$	$L_{sym}$	Model	DL	$L_{adv}$
(a)						
HF-PIM	✓	✓	–	HF-PIM	✓	✓
V1	–	✓	–	V4	–	✓
V2	✓	–	–	V5	✓	–
V3	✓	✓	✓	V6	–	–
Model	Verification		Recognition			
	FAR=0.01	FAR=0.001	Rank-1	Rank-5		
(b)						
V1	92.1±1.2	85.3±2.0	91.9±1.4	93.2±1.1		
V2	94.4±0.8	88.6±1.7	94.0±1.1	95.8±0.9		
V3	93.9±1.0	87.3±2.0	93.4±1.2	94.1±0.8		
V4	95.0±0.7	89.1±1.5	95.6±0.8	97.0±0.4		
V5	94.7±0.8	89.1±1.6	95.2±0.9	96.9±0.4		
V6	94.1±1.0	87.7±1.9	94.3±1.1	96.1±0.8		
HF-PIM	<b>95.3±0.7</b>	<b>89.9±1.3</b>	<b>96.4±0.5</b>	<b>98.1±0.2</b>		

Bold values indicate the best performance among all the compared methods  
 ✓/– mean that the variant includes/excludes the corresponding loss. The results are averaged over 10 testing splits



**Fig. 12** Qualitative comparisons on synthetic results between HF-PIM and its variants

connected one on representing fine-grained texture information. Besides, we can see that adversarial training criterion on texture space brings more improvements.

## 5 Limitations and Discussion

Although our method can achieve compelling results and preserve the identity information across many different datasets, there is still room for improvement, especially for high-resolution face frontalization. We can see that the visual qualities of the high-resolution face frontalization results look less convincing. Sometimes our model tends to produce more warped bulged regions than the normal side (e.g., the fourth subject in the last row in Fig. 11). We argue that high-resolution face frontalization is much challenging but few datasets provide paired face images at  $256 \times 256$ . So, we employ the profiling method in Zhu et al. (2015) to overcome the lack of data. However, these synthetic training data is not perfect, making the high-resolution results less appealing than  $128 \times 128$  ones. Given training data with higher quality, it is predictable that our model indeed produces realistic results.

Existing methods (Huang et al. 2017; Tran et al. 2017; Zhao et al. 2018; Yin et al. 2017; Hu et al. 2018) measure the performance of face recognition to reflect the quality of frontalized results. This measurement cannot be applied to those datasets without identity labels (e.g., CelebA-HQ) and neglects texture information that is not sensitive to identity. However, those textures also play an import role on the visual quality and should be preserved faithfully. For face attribute analysis, data augmentation and many other applications, recovering high-resolution frontal faces with detailed texture information has great potential for making progress. Therefore, finding new applications for face frontalization and putting forward new metrics need further research.

## 6 Conclusion

This paper has proposed High Fidelity Pose Invariant Model (HF-PIM) to produce realistic and identity-preserving frontalization results with a higher resolution. HF-PIM combines the advantages of 3D and GAN based methods and frontalizes profile images via a novel facial texture fusion warping procedure. By leveraging a novel dense correspondence field, the prerequisite of warping is decomposed into dense correspondence field estimation and facial texture recovering, which are well addressed by a unified end-to-end neural network. We also have introduced Adversarial Residual Dictionary Learning (ARDL) to supervise facial texture map recovering without the need of 3D data. Furthermore, the multi-perception guided loss is proposed to overcome

the practical data misalignment problem in unconstrained environments. Exhaustive experiments have shown that our proposed method can preserve more identity information as well as texture details, which makes the high-resolution results far more realistic.

**Acknowledgements** This work is funded by the National Key Research and Development Program of China (Grant Nos. 2016YFB1001001, 2017YFC0821602), the National Natural Science Foundation of China (Grant Nos. 61622310, 61427811, U1836217), and Beijing Natural Science Foundation (Grant No. JQ18017).

## References

- AbdAlmageed, W., Wu, Y., Rawls, S., Harel, S., Hassner, T., Masi, I., Choi, J., Lekust, J., Kim, J., Natarajan, P., et al. (2016). Face recognition using deep multi-pose representations. In *IEEE winter conference on applications of computer vision (WACV)* (pp. 1–9).
- Arjovsky, M., Chintala, S., & Bottou, L. (2017). Wasserstein GAN. In *International conference on machine learning (ICML)* (pp. 214–223).
- Ashraf, A. B., Lucey, S., & Chen, T. (2008). Learning patch correspondences for improved viewpoint invariant face recognition. In *IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 1–8).
- Blanz, V., & Vetter, T. (1999). A morphable model for the synthesis of 3D faces. In *Annual conference on computer graphics and interactive techniques (SIGGRAPH)* (pp. 187–194).
- Booth, J., & Zafeiriou, S. (2014). Optimal UV spaces for facial morphable model construction. In *IEEE international conference on image processing (ICIP)* (pp. 4672–4676).
- Bulat, A., & Tzimiropoulos, G. (2017). How far are we from solving the 2D & 3D face alignment problem? (and a dataset of 230,000 3D facial landmarks). In *IEEE international conference on computer vision (ICCV)* (pp. 1021–1030).
- Cao, J., Hu, Y., Yu, B., He, R., & Sun, Z. (2019). 3D aided duet GANs for multi-view face image synthesis. *IEEE Transactions on Information Forensics and Security (TIFS)*, 14(8), 2028–2042.
- Cao, J., Hu, Y., Zhang, H., He, R., & Sun, Z. (2018). Learning a high fidelity pose invariant model for high-resolution face frontalization. In *Conference on neural information processing systems (NeurIPS)*.
- Cao, K., Rong, Y., Li, C., Tang, X., & Loy, C.C. (2018). Pose-robust face recognition via deep residual equivariant mapping. In *IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 5187–5196).
- Chen, X., Duan, Y., Houthoofd, R., Schulman, J., Sutskever, I., & Abbeel, P. (2016). InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets. In *Conference on neural information processing systems (NeurIPS)* (pp. 2172–2180).
- Cimpoi, M., Maji, S., & Vedaldi, A. (2015). Deep filter banks for texture recognition and segmentation. In *IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 3828–3836).
- Cole, F., Belanger, D., Krishnan, D., Sarna, A., Mosseri, I., & Freeman, W.T. (2017). Synthesizing normalized faces from facial identity features. In *IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 3386–3395).
- Dana, H.Z.J.X.K. (2017). Deep TEN: Texture encoding network. In *IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 2896–2905).

- Deng, J., Cheng, S., Xue, N., Zhou, Y., & Zafeiriou, S. (2018). UV-GAN: Adversarial facial UV map completion for pose-invariant face recognition. In *IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 7093–7102).
- Dovgird, R., & Basri, R. (2004). Statistical symmetric shape from shading for 3D structure recovery of faces. In *European conference on computer vision (ECCV)* (pp. 99–113).
- Ferrari, C., Lisanti, G., Berretti, S., & Del Bimbo, A. (2016). Effective 3D based frontalization for unconstrained face recognition. In *International conference on pattern recognition (ICPR)* (pp. 1047–1052).
- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. In *Conference on neural information processing systems (NeurIPS)* (pp. 2672–2680).
- Gross, R., Matthews, I., Cohn, J., Kanade, T., & Baker, S. (2010). Multi-PIE. *Image and Vision Computing (IVC)*, 28(5), 807–813.
- Güler, R. A., Neverova, N., & Kokkinos, I. (2018). DensePose: Dense human pose estimation in the wild. In *IEEE conference on computer vision and pattern recognition (CVPR)*.
- Güler, R. A., Trigeorgis, G., Antonakos, E., Snape, P., Zafeiriou, S., & Kokkinos, I. (2017). Densereg: Fully convolutional dense shape regression in-the-wild. In *IEEE Conference on computer vision and pattern recognition (CVPR)* (vol. 2, p. 5).
- Han, C., Shan, S., Kan, M., Wu, S., & Chen, X. (2018). Face recognition with contrastive convolution. In *European conference on computer vision (ECCV)* (pp. 118–134).
- Hassner, T. (2013). Viewing real-world faces in 3D. In *IEEE international conference on computer vision (ICCV)* (pp. 3607–3614).
- Hassner, T., Harel, S., Paz, E., & Enbar, R. (2015). Effective face frontalization in unconstrained images. In *IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 4295–4304).
- Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Klambauer, G., & Hochreiter, S. (2017). GANs trained by a two time-scale update rule converge to a Nash equilibrium. In *Conference on neural information processing systems (NeurIPS)* (pp. 6629–6640).
- Hu, Y., Wu, X., Yu, B., He, R., & Sun, Z. (2018). Pose-guided photorealistic face rotation. In *IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 8398–8406).
- Huang, G. B., Ramesh, M., Berg, T., & Learned-Miller, E. (2007). Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Tech. rep., University of Massachusetts, Amherst.
- Huang, H., He, R., Sun, Z., & Tan, T. (2017). Wavelet-SRnet: A wavelet-based CNN for multi-scale face super resolution. In *IEEE international conference on computer vision (ICCV)* (pp. 1689–1697).
- Huang, R., Zhang, S., Li, T., & He, R. (2017). Beyond face rotation: Global and local perception GAN for photorealistic and identity preserving frontal view synthesis. In *IEEE international conference on computer vision (ICCV)* (pp. 2458–2467).
- Johnson, J., Alahi, A., & Fei-Fei, L. (2016). Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision (ECCV)* (pp. 694–711).
- Kan, M., Shan, S., Chang, H., & Chen, X. (2014). Stacked progressive auto-encoders (SPAEC) for face recognition across poses. In *IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 1883–1890).
- Karras, T., Aila, T., Laine, S., & Lehtinen, J. (2018). Progressive growing of GANs for improved quality, stability, and variation. In *International conference on learning representations (ICLR)*.
- Klare, B.F., Jain, A.K., Klein, B., Taborsky, E., Blanton, A., Cheney, J., Allen, K., Grother, P., Mah, A., & Burge, M. (2015). Pushing the frontiers of unconstrained face detection and recognition: IARPA janus benchmark a. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1931–1939).
- Li, J., Zhao, J., Zhao, F., Liu, H., Li, J., Shen, S., Feng, J., & Sim, T. (2016). Robust face recognition with deep multi-view representation learning. In *ACM international conference on multimedia (ACM-MM)* (pp. 1068–1072).
- Li, P., Wu, X., Hu, Y., He, R., & Sun, Z. (2019). M2FPA: A multi-yaw multi-pitch high-quality database and benchmark for facial pose analysis. *IEEE international conference on computer vision (ICCV)*.
- Liu, Z., Luo, P., Wang, X., & Tang, X. (2015). Deep learning face attributes in the wild. In *IEEE international conference on computer vision (ICCV)* (pp. 3730–3738).
- Lucey, S., & Chen, T. (2008). A viewpoint invariant, sparsely registered, patch based, face verifier. *International Journal of Computer Vision (IJCV)*, 80(1), 58–71.
- Mao, X., Li, Q., Xie, H., Lau, R. Y., Wang, Z., & Smolley, S. P. (2017). Least squares generative adversarial networks. In *IEEE international conference on computer vision (ICCV)* (pp. 2813–2821).
- Masi, I., Rawls, S., Medioni, G., & Natarajan, P. (2016). Pose-aware face recognition in the wild. In *IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 4838–4846).
- Masi, I., Trn, A. T., Hassner, T., Leksut, J. T., & Medioni, G. (2016). Do we really need to collect millions of faces for effective face recognition? In *European conference on computer vision (ECCV)* (pp. 579–596).
- Miyato, T., & Koyama, M. (2018). cGANs with projection discriminator. In *International conference on learning representations (ICLR)*.
- Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., & Lerer, A. (2017). Automatic differentiation in pytorch. In *Conference on neural information processing systems (NeurIPS-W)*.
- Paysan, P., Knothe, R., Amberg, B., Romdhani, S., & Vetter, T. (2009). A 3D face model for pose and illumination invariant face recognition. In *IEEE international conference on advanced video and signal based surveillance (AVSS)* (pp. 296–301).
- Radford, A., Metz, L., & Chintala, S. (2016). Unsupervised representation learning with deep convolutional generative adversarial networks. In *International conference on learning representations (ICLR)*.
- Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In *International conference on medical image computing and computer assisted intervention (MICCAI)* (pp. 234–241).
- Sagonas, C., Panagakis, Y., Zafeiriou, S., & Pantic, M. (2015). Robust statistical face frontalization. In *IEEE international conference on computer vision (ICCV)* (pp. 3871–3879).
- Sengupta, S., Chen, J. C., Castillo, C., Patel, V. M., Chellappa, R., & Jacobs, D. W. (2016). Frontal to profile face verification in the wild. In *IEEE winter conference on applications of computer vision (WACV)* (pp. 1–9).
- Shrivastava, A., Pfister, T., Tuzel, O., Susskind, J., Wang, W., & Webb, R. (2017). Learning from simulated and unsupervised images through adversarial training. In *IEEE conference on computer vision and pattern recognition (CVPR)* (vol. 2, p. 5).
- Sun, X., Nasrabadi, N. M., & Tran, T. D. (2018). Supervised deep sparse coding networks. In *IEEE international conference on image processing (ICIP)* (pp. 346–350).
- Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2015). Going deeper with convolutions. In *IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 1–9).
- Taigman, Y., Yang, M., Ranzato, M., & Wolf, L. (2014). Deepface: Closing the gap to human-level performance in face verification. In *IEEE international conference on computer vision (ICCV)* (pp. 1701–1708).

- Tian, Y., Peng, X., Zhao, L., Zhang, S., & Metaxas, D. N. (2018). CR-GAN: Learning complete representations for multi-view generation. In *International joint conference on artificial intelligence (IJCAI)*.
- Tran, L., & Liu, X. (2018). Nonlinear 3D face morphable model. In *IEEE conference on computer vision and pattern recognition (CVPR)*.
- Tran, L., Yin, X., & Liu, X. (2017). Disentangled representation learning GAN for pose-invariant face recognition. In *IEEE conference on computer vision and pattern recognition (CVPR)* (vol. 3, p. 7).
- Tran, L.Q., Yin, X., & Liu, X. (2018). Representation learning by rotating your faces. In *IEEE transactions on pattern analysis and machine intelligence (TPAMI)*.
- Van Gemert, J. C., Geusebroek, J. M., Veenman, C. J., & Smeulders, A. W. (2008). Kernel codebooks for scene categorization. In *European conference on computer vision (ECCV)* (pp. 696–709).
- Wu, X., He, R., Sun, Z., & Tan, T. (2018). A light cnn for deep face representation with noisy labels. *IEEE Transactions on Information Forensics and Security*, 13(11), 2884–2896.
- Yan, J., Lin, D., & Loy, C. C. (2018). Consensus-driven propagation in massive unlabeled data for face recognition. In *European conference on computer vision (ECCV)*.
- Yang, J., Reed, S.E., Yang, M.H., & Lee, H. (2015). Weakly-supervised disentangling with recurrent transformations for 3D view synthesis. In *Conference on neural information processing systems (NeurIPS)* (pp. 1099–1107).
- Yim, J., Jung, H., Yoo, B., Choi, C., Park, D., & Kim, J. (2015). Rotating your face using multi-task deep neural network. In *IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 676–684).
- Yin, X., Yu, X., Sohn, K., Liu, X., & Chandraker, M. (2017). Towards large-pose face frontalization in the wild. In *IEEE international conference on computer vision (ICCV)* (pp. 1–10).
- Zhao, J., Cheng, Y., Xu, Y., Xiong, L., Li, J., Zhao, F., Jayashree, K., Pranata, S., Shen, S., Xing, J., et al. (2018). Towards pose invariant face recognition in the wild. In *IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 2207–2216).
- Zhao, J., Xiong, L., Cheng, Y., Cheng, Y., Li, J., Zhou, L., Xu, Y., Karlekar, J., Pranata, S., Shen, S., et al. (2018). 3D-aided deep pose-invariant face recognition. In *International joint conference on artificial intelligence (IJCAI)*.
- Zhao, J., Xiong, L., Jayashree, P. K., Li, J., Zhao, F., Wang, Z., Pranata, P. S., Shen, P. S., Yan, S., & Feng, J. (2017). Dual-agent GANs for photorealistic and identity preserving profile face synthesis. In *Conference on neural information processing systems (NeurIPS)* (pp. 66–76).
- Zhao, J., Xiong, L., Li, J., Xing, J., Yan, S., & Feng, J. (2018). 3D-aided dual-agent gans for unconstrained face recognition. *IEEE transactions on pattern analysis and machine intelligence (TPAMI)*.
- Zhu, J. Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *IEEE international conference on computer vision (ICCV)* (pp. 2242–2251).
- Zhu, X., Lei, Z., Liu, X., Shi, H., & Li, S. Z. (2016). Face alignment across large poses: A 3D solution. In *IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 146–155).
- Zhu, X., Lei, Z., Yan, J., Yi, D., & Li, S. Z. (2015). High-fidelity pose and expression normalization for face recognition in the wild. In *IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 787–796).

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.