Facial UV map completion for pose-invariant face recognition: a novel adversarial approach based on coupled attention residual UNets

In Seop Na[1†], Chung Tran[2†], Dung Nguyen[3] and Sang Dinh[2*] ○iD

*Correspondence:
sangdv@soict.hust.edu.vn
†In Seop Na and Chung Tran
contributed equally to this
work
[2] Hanoi University of Science
and Technology, 1 Dai Co
Viet, Hanoi, Vietnam
Full list of author information
is available at the end of the
article

Abstract

Pose-invariant face recognition refers to the problem of identifying or verifying a person by analyzing face images captured from different poses. This problem is challenging due to the large variation of pose, illumination and facial expression. A promising approach to deal with pose variation is to fulfill incomplete UV maps extracted from in-the-wild faces, then attach the completed UV map to a fitted 3D mesh and finally generate different 2D faces of arbitrary poses. The synthesized faces increase the pose variation for training deep face recognition models and reduce the pose discrepancy during the testing phase. In this paper, we propose a novel generative model called Attention ResCUNet-GAN to improve the UV map completion. We enhance the original UV-GAN by using a couple of U-Nets. Particularly, the skip connections within each U-Net are boosted by attention gates. Meanwhile, the features from two U-Nets are fused with trainable scalar weights. The experiments on the popular benchmarks, including Multi-PIE, LFW, CPLWF and CFP datasets, show that the proposed method yields superior performance compared to other existing methods.

Keywords: Generative adversarial networks, Pose-invariant face recognition, Deep learning, AI

Introduction

Face recognition has gained much attention for decades [1–3]. Contrary to other popular biometrics, face recognition can be applied to uncooperative subjects in a non-instructive manner. While (near)-frontal face recognition has gradually matured, face recognition in the wild is still challenging due to different unconstrained factors. In fact, the performance of a face recognition system heavily depends on the pose of input faces. Recent studies show that the performance of face verification with the same view, such as frontal–frontal or profile–profile, is really good. However, the performance dramatically degrades when verifying faces in different views like frontal-profile [4].

Pose-invariant face recognition refers to the problem of identifying or verifying a person by analyzing face images captured from different poses. In recent years, numerous pose-invariant face recognition methods have been proposed. In [5–13], the authors train deep neural networks on large-scale datasets to ease the effect of pose variation,

Na *et al. Hum. Cent. Comput. Inf. Sci.*   (2020) 10:45

Page 2 of 17

which leads to significant improvements in the performance of face recognition. In [14], Masi et al. propose a method to enrich the pose variation in the training dataset by rotating faces across 3D space. Beyond, in [15], Sagonas et al. propose a novel method to jointly learn both frontal view reconstruction and landmark localization by solving a constrained optimization problem. Kan et al. [16] introduce stacked progressive auto-encoders (SPAE), which can learn pose-robust features through a complicated deep neural network to transform profile faces to frontal ones. In [17], Hassner et al. introduce a straightforward approach to generate frontal faces from a simple 3D shape. Peng et al. [18] propose a new reconstruction loss for disentangled learning that encourages identity features of the same subject to be clustered together despite the pose variation.

Recently, generative adversarial networks (GANs) [19] have proved to be powerful to mimic data distribution. GANs have been successfully applied to many computer vision tasks such as image inpainting [18, 20, 21], style transfer [22, 23], image synthesis [24, 25], super-resolution [26] and so on. These successful applications have motivated researchers to apply GANs to pose-invariant feature disentanglement [4, 27], face completion [28] and face frontalization [4, 29–32]. In [28], Wang et al. propose a recurrent generative adversarial network (RGAN), which consists of a CompletionNet and a DiscriminationNet, for completing face and recovering the missing region automatically. Dual et al. [32] propose a boosting GAN (BoostGAN) for face deocclusion and frontalization. BoostGAN can generate photorealistic frontal faces with identity preservation from occluded but profile ones. TP-GAN [29] uses a two-pathway GAN that simultaneously learns global structures and local information for photorealistic frontal view synthesis. Zhao et al. [33] propose a unified deep architecture containing a face frontalization module and a discriminative learning module, which can be jointly learned in an end-to-end fashion. Zhang et al. [34] propose a geometry guided GAN to generate facial images with arbitrary expressions and poses conditioned on a set of facial landmarks. They embed a classifier into the GAN to facilitate image synthesis and perform facial expression recognition. In [27], Tran et al. propose DR-GAN that can take one or multiple input images and produce one unified identity representation along with synthesized identity-preserved faces of various target poses. However, all methods mentioned above usually require a large amount of paired faces across different poses for training, which is overdemanding in real-world applications.

In [35], Deng et al. propose an adversarial UV map completion framework called UV-GAN to solve pose-invariant face recognition without the need of extensive pose coverage in the training dataset. The authors in [35] first fit a 3DMM [36] to 2D profile face and get an incomplete UV map, which is then fulfilled by a straightforward pix2pix [37, 38]. The generator architecture in pix2pix follows the general shape of U-Net [39] to add skip connections between encoder and decoder subnetworks in order to enhance the transfer of low-level information between input and output. One weakness of the original UV-GAN is the plain architecture of the generator, which is shown to be worse than residual networks [40]. Another weakness is that one U-Net block seems to be not enough to mix well low-level information in the encoder with high-level semantic features in the decoder. In [41], Deng et al. use UV-GAN with similar architecture as in [35] to extract side information as well as subspaces, and combine UV-GAN with robust PCA for the face recognition task. He et al. [42] introduce a framework for heterogeneous

Na *et al. Hum. Cent. Comput. Inf. Sci.*   (2020) 10:45

Page 3 of 17

face synthesis from near-infrared (NIR) to visible domain. The framework consists of two adversarial generators to estimate a UV map and a facial texture map from an input NIR face, and then generate a corresponding frontal visible face. Nevertheless, both generators in this framework are based on the general U-Net structure [23, 39]. Some efforts [43, 44] stack multiple U-Nets together, but skip connections are utilized only inside each single U-Net. Ibtehaz et al. [45] propose residual paths with additional convolutional layers in skip connections to reduce the semantic gap between encoder and decoder features. In [46], Oktay et al. introduce attention gates to implicitly learn to suppress irrelevant regions in an input image while highlighting salient features useful for a specific task. In [47], Tang et al. introduce coupled U-Nets architecture, where coupling connections are utilized to improve the information flow across U-Nets.

In this paper, we propose a new generative model architecture called Attention ResCUNet-GAN, where the generator is coupled U-Nets, and the backbone of each encoder is enhanced by residual network architecture. We use attention gates for skip connections within each U-Net to suppress irrelevant low-level information from encoders. We also use skip connections across two U-Nets to limit gradient vanishing and promote feature reuse. The experiments on the popular benchmarks demonstrate that our Attention ResCUNet-GAN yields considerably better results than the original UV-GAN model.

The rest of this paper is organized as follows. Details of our proposed method are presented in "Our proposed method" section. "Experiments and evaluation" section presents our experimental results on the Multi-PIE dataset. Finally, the conclusion is made in "Conclusions and future work" section.
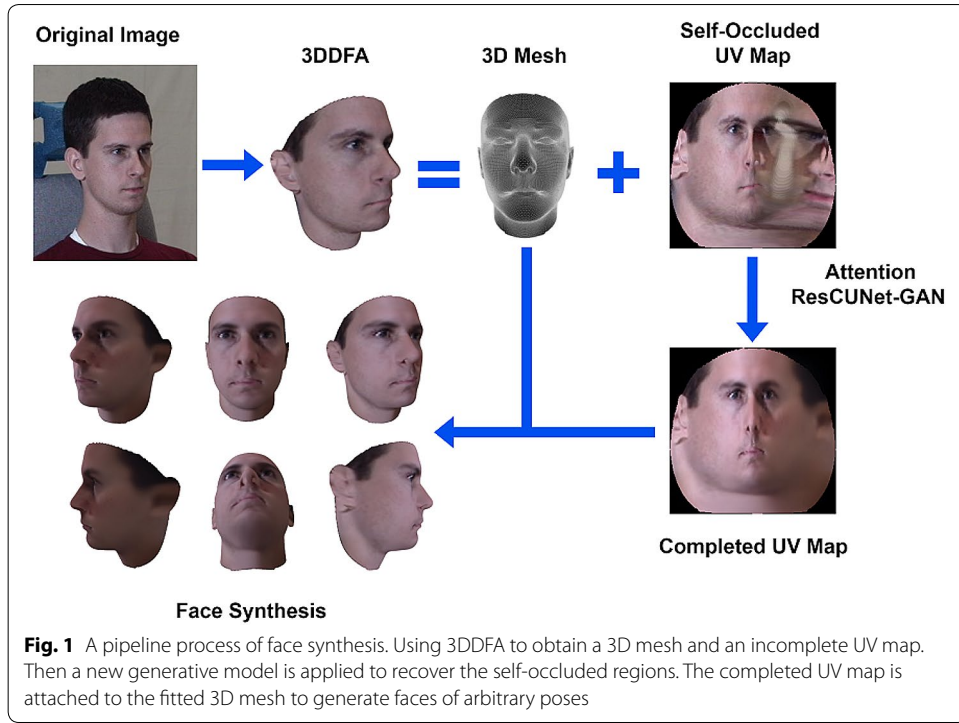
Our proposed method

Following [35], we use 3DDFA [48] to fitting 2D images to retrieve UV maps and 3D meshes. With a non-frontal face, the UV map generated by 3DDFA is always incomplete due to self-occlusion. Hence, we propose a new generative model architecture called Attention ResCUNet-GAN to improve the performance of the original UV-GAN [35] in filling up the missing contents of the UV map, which in turns helps to synthesize facial images of arbitrary poses. The overall pipeline process to synthesize more faces of various poses is depicted in Fig. 1.

3DDFA fitting

3D morphable model

Blanz and Vetter [49] introduce the 3D morphable model (3DMM) to recover the 3D face from a 2D image. Assuming that a 3D face scan with N vertexes can be represented as a $3N \times 1$ vector $\mathbf{S} = [x_1, y_1, z_1, \ldots, x_N, y_N, z_N]^T \in \mathbb{R}^{3N}$, where $[x_i, y_i, z_i]^T$ are the object-centered Cartesian coordinates of the i-th vertex. Given a dataset of such 3D face scans, one would like to represent them as a smaller set of variables. The authors in [49] propose to use a two-stage principle component analysis (PCA) to estimate the shape identity parameters along with expression parameters of the 3D faces. Suppose that, after the first stage, we keep first n_s principal components and $\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_{n_s}$ are the corresponding orthonormal basis, then a 3D face $\mathbf{S}$ can be represented as follows:

Fig. 1 A pipeline process of face synthesis. Using 3DDFA to obtain a 3D mesh and an incomplete UV map. Then a new generative model is applied to recover the self-occluded regions. The completed UV map is attached to the fitted 3D mesh to generate faces of arbitrary poses

$$\mathbf{S} = \bar{\mathbf{S}} + \sum_{i=1}^{n_s} \mathbf{s}_i \alpha_i \tag{1}$$

where $\bar{\mathbf{S}} \in \mathbb{R}^{3N}$ are the mean shape vector across the dataset of 3D face scans and $\boldsymbol{\alpha} = [\alpha_1, \dots, \alpha_{n_s}]$ are the shape parameters.

In the second stage, a new PCA is trained on the offsets between expression scans and neutral scans. After this stage, the final shape a representation is follows:

$$\mathbf{S} = \bar{\mathbf{S}} + \sum_{i=1}^{n_s} \mathbf{s}_i \alpha_i + \sum_{i=1}^{n_e} \mathbf{e}_i \beta_i, \tag{2}$$

where $\mathbf{e}_i, i = 1, \dots, n_e$ are the orthonormal basis of first n_e principal components, and $\boldsymbol{\beta} = [\beta_1, \dots, \beta_{n_e}]$ are the expression parameters.
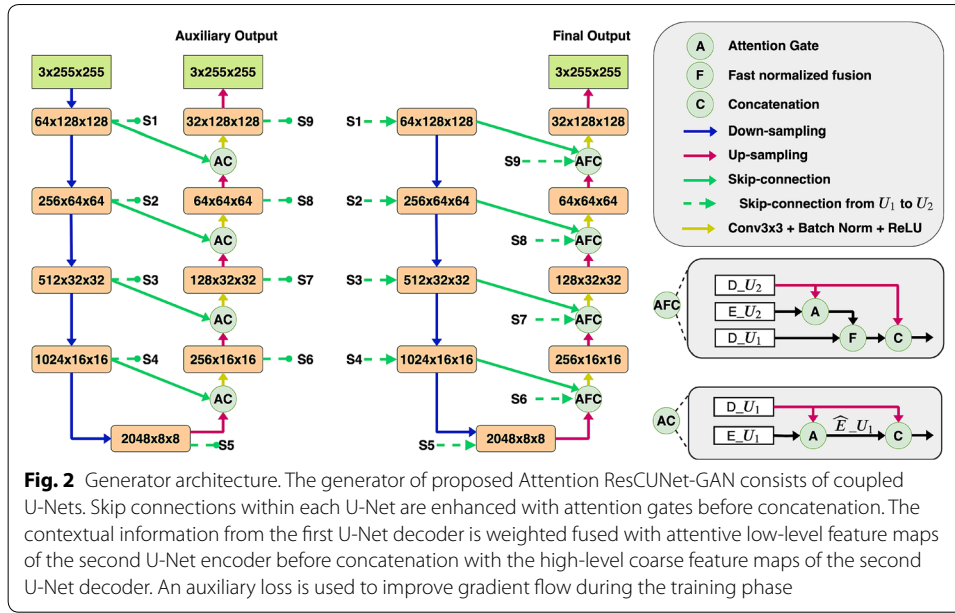
After the 3D face is constructed, a rigid transformation is applied on the shape from the barycentric coordinate to camera based world coordinate. Each 3D vertex $\mathbf{v} = [x, y, z]^T$ is rotated and translated as follows:

$$\mathbf{v}_{cam} = \mathbf{R}\mathbf{v} + \mathbf{t}, \tag{3}$$

where $\mathbf{R} \in \mathbb{R}^{3 \times 3}$ and $\mathbf{t} = [t_x, t_y, t_z]^T$ are the 3D rotation and translation components, respectively.

Finally, each 3D point can be projected into its 2D location in the image plane with scale orthographic projection:

$$\mathbf{v}_p = f * \mathbf{PR} * \mathbf{v}_{cam} + \mathbf{t}_{2d}, \tag{4}$$

Na et al. Hum. Cent. Comput. Inf. Sci.   (2020) 10:45

Page 5 of 17

Fig. 2 Generator architecture. The generator of proposed Attention ResCUNet-GAN consists of coupled U-Nets. Skip connections within each U-Net are enhanced with attention gates before concatenation. The contextual information from the first U-Net decoder is weighted fused with attentive low-level feature maps of the second U-Net encoder before concatenation with the high-level coarse feature maps of the second U-Net decoder. An auxiliary loss is used to improve gradient flow during the training phase

where f is the scale factor, $\mathbf{Pr} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$ is the orthographic projection matrix and $\mathbf{t}_{2d}$ is the principal point that is set to the image center.

Suppose that the set of all the model parameters are denoted by $\mathbf{p} = [f, \mathbf{R}, \mathbf{t}_{2d}, \boldsymbol{\alpha}, \boldsymbol{\beta}]$.

3DDFA method

Method 3DDFA associates Cascaded Regression and a Convolutional Neural Network (CNN). Cascaded CNN can be formulated as:

$$\mathbf{p}^{k+1} = \mathbf{p}^k + Net^k(Feat(I, \mathbf{p}^k)), \tag{5}$$

where $\mathbf{p}^k$ is the model parameters at the k-th iteration, which is updated by applying a CNN-based regressor Net^k on the shape indexed feature $Feat$ that depends on the input image $\mathbf{I}$ and the current parameters $\mathbf{p}^k$.

The purpose of the CNN regressors is to predict the parameter update $\Delta\mathbf{p}$ to shift the initial parameter $\mathbf{p}^0$ as close as possible to the ground truth $\mathbf{p}^g$. In term of objective function, [48] proposes to use the Optimized Weighted Parameter Distance Cost (OWPDC):
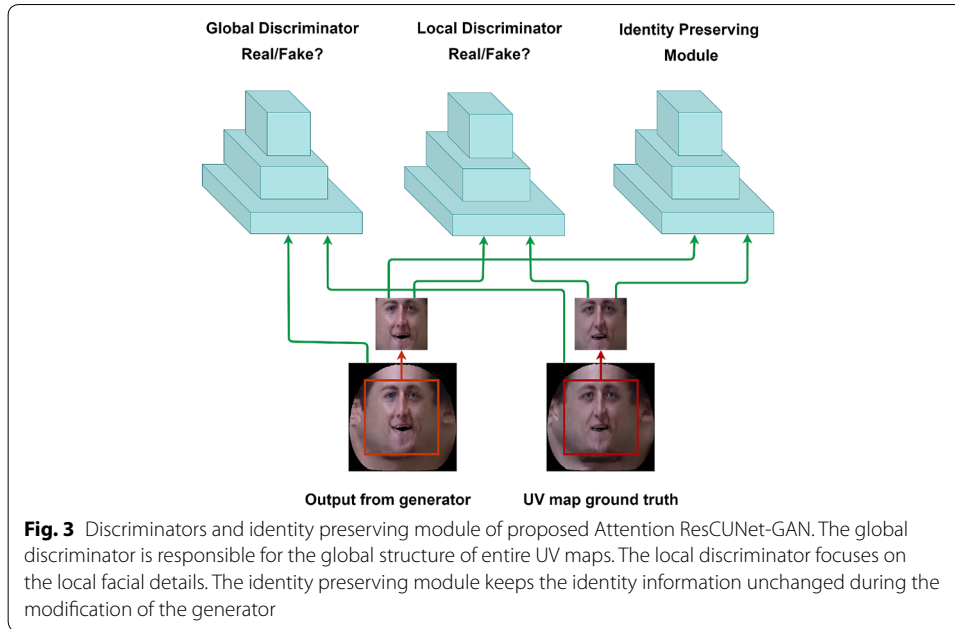
$$\begin{aligned} E_{owpdc} = (\Delta\mathbf{p} + \mathbf{p}^0 - \mathbf{p}^g)^T \mathrm{diag}(\mathbf{w}^*) \\ (\Delta\mathbf{p} + \mathbf{p}^0 - \mathbf{p}^g), \end{aligned} \tag{6}$$

where $\mathbf{w}^*$ is the optimized parameter importance vector.

Proposed network for UV map completion

The proposed Attention ResCUNet-GAN consists of a generator (Fig. 2), two discriminators, and an identity preserving module (Fig. 3). The global discriminator deals with the global structure of entire complete UV maps, while the local discriminator focuses on the local details of the face region.
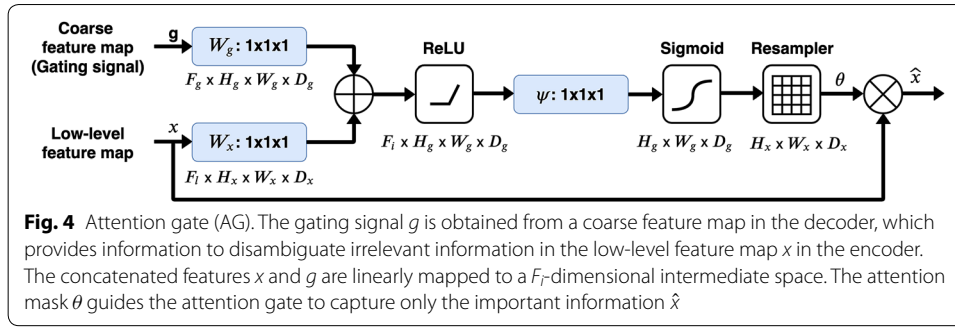
Na *et al. Hum. Cent. Comput. Inf. Sci.*   (2020) 10:45

Page 6 of 17

Fig. 3 Discriminators and identity preserving module of proposed Attention ResCUNet-GAN. The global discriminator is responsible for the global structure of entire UV maps. The local discriminator focuses on the local facial details. The identity preserving module keeps the identity information unchanged during the modification of the generator

Generator network

An incomplete UV map is fed into Attention ResCUNet-GAN Generator, which acts as an auto-encoder to reconstruct missing regions. We use the following reconstruction loss as in [35]:

$$L_{rec} = \frac{1}{W * H} \sum_{i=1}^{W} \sum_{j=1}^{H} |G(I_{i,j}^P) - I_{i,j}^F|, \tag{7}$$

where I^P is the input incomplete UV map, $G(I^P)$ is the output from the generator, and I^F is the ground truth texture.

The generator (Fig. 2) consists of coupled U-Nets [47]. A drawback of the UV-GAN's generator is the plain convolutional backbone, which is shown to be rapidly degraded as the network depth increases [40]. Therefore, here we leverage the residual architecture in [40] to build a deeper backbone that is capable of extracting better high-level features without suffering from the degradation problem. Particularly, in terms of the backbone network for encoders, we use ResNet-50 [40] consisting of multiple bottleneck residual blocks, each of which is a stack of three successive layers with 1×1, 3×3, 1×1 convolutions. Batch normalization is used right after each convolution and before activation layers. We use skip connections within each U-Net to transfer low-level information from the encoder to high-level contextual features in the decoder. Attention gates [46] are used to suppress irrelevant low-level information from encoders. Figure 4 illustrates how a coarse feature map can guide another low-level feature map to ignore irrelevant information.

To combine features across two U-Nets, one can apply a direct depth-wise concatenation of the coarse feature maps D_U_1, D_U_2 extracted from the decoders of both U-Nets and the attentive information $\hat{E}\_U_2$ extracted from an attention gate of the encoder of the second U-Net. In such a combination, the latest feature map D_U_2, which is thought

Fig. 4 Attention gate (AG). The gating signal *g* is obtained from a coarse feature map in the decoder, which provides information to disambiguate irrelevant information in the low-level feature map *x* in the encoder. The concatenated features *x* and *g* are linearly mapped to a F_l-dimensional intermediate space. The attention mask θ guides the attention gate to capture only the important information $\hat{x}$

to obtain more contextual information, would play the most crucial role regarding the contribution to the final output. However, such a direct concatenation always requires more memory. Thus, before concatenating with D_U_2, here we apply fast normalized fusion [50] to combine D_U_1 and $\hat{E}\_U_2$ as follows:

$$\hat{D}\_U_1 = \frac{w_1 \times D\_U_1 + w_2 \times \hat{E}\_U_2}{w_1 + w_2 + \epsilon} \tag{8}$$

where w_1, w_2 are learnable scalar weights that can be trained via normal back propagation algorithm and $\epsilon = 0.0001$ is a small value to avoid numerical instability. Parameters are ensured to be positive by applying Relu activation after them.

Global and local discriminators   Global discriminator enforces maintaining the surrounding context of the facial image. Meanwhile, the local discriminator focuses on the central face region to enforce better recovering local details such as eye, nose, mouth and so on. We keep the same architectures for the discriminators as described in [35]. The following typical adversarial loss is used:

$$
\begin{aligned}
L_{adv} = \min_G \max_D E_{x\ p_d(x),y\ p_d(y)}[log(D(x,y))] \\
+ E_{x\ p_d(x),z\ p_d(z)}[log(1 - D(G(x,z),y))],
\end{aligned}
\tag{9}
$$

where $p_d(x), p_d(y), p_d(z)$ denote the distributions of incomplete UV maps x, complete UV maps y and the Gaussian noise z, respectively.

Identity preserving module   The synthetic faces must not only be photorealistic but also preserve identity information, which plays a crucial role in generation-based face recognition. To this end, the following identity loss [35] is used:

$$L_{id} = \parallel F(I^F) - F(G(I^p)) \parallel_2^2, \tag{10}$$

where $F(.)$ denotes the embedding features extracted by the last layer before softmax in a pretrained CNN. Here in terms of embedding feature extractor, we use FaceNet pretrained on VGGFace2 dataset, which contains 3.31M face images of 9131 identities. This feature extractor is frozen during training. The identity preserving module in Eq. (10) enforces the embedding features of faces in the UV map ground truth I^F and the generated UV map $G(I^P)$ to be close to each other. The dimension of the embedding features is 512.

Final loss function   Overall, the total loss function is a weighted sum of the abovementioned losses:

$$L_{total} = L_{rec} + \lambda_1 L_{adv}^{local} + \lambda_2 L_{adv}^{global} + \lambda_3 L_{id}, \tag{11}$$

where $\lambda_1, \lambda_2, \lambda_3$ are the weights that control the importance factors of different losses.

Moreover, a similar auxiliary loss is also applied to the intermediate output of the generator right after the end of the first U-Net decoder. The auxiliary loss strengthens the gradient flow to the layers of the first U-Net so that the parameters in the first U-Net can be trained more efficiently. Therefore, the final loss can be expressed as follows:

$$L_{final} = L_{total} + \eta L_{total}^{aux}, \tag{12}$$

where η is a parameter regulating the contribution of the auxiliary loss.

Experiments and evaluation

Datasets and settings

We train our Attention ResCUNet-GAN on the Multi-PIE dataset [51]. All subjects in this dataset were taken in 15 viewpoints, 19 illumination conditions, and many facial expressions. Totally, there are more than 750,000 images of 337 people.

For every subject with each illumination condition and facial expression, we feed 15 facial images captured from 15 viewpoints to the 3DDFA model to retrieve separate incomplete UV maps. We then select the incomplete UV maps with yaw angles of $0°$, $-30°, +30°$ and merge them using Poisson blending [52] (Fig. 5) to create the corresponding ground-truth UV map. In that way, we can ideally create 15 pairs of images for training the generator. Each of these pairs consists of an incomplete and a ground-truth UV map. However, in some cases, when the quality of an input facial image is not good enough, the 3DDFA model can not successfully detect the face landmarks; thus, the corresponding 3D mesh and incomplete UV map can not be created. Therefore, such cases are ignored in the training phase. All generated UV maps are rescaled to 256×256 to fit the input size of our ResCUNet-GAN.

In addition to the proposed Attention ResCUNet-GAN, we also try a normal ResCUNet-GAN that has a similar architecture but without any attention gates and fast normalized fusion. In this ResCUNet-GAN, the concatenation is applied to all skip connections. Our networks are implemented in Pytorch. It takes three days for training each network on a server with two GPU RTX 2080Ti. We train each network for 100 epochs with a batch size of 16 and a learning rate of 10^{-4}. We empirically set the importance factors as follows: $\eta = 0.3, \lambda_1 = \lambda_2 = 0.5, \lambda_3 = 0.01$.

In order to evaluate the effectiveness of the proposed method, we conduct experiments on pose-invariant face recognition on different benchmarks. Casia Web Face is a facial dataset that consists of 453,453 images over 10,575 identities. LFW (Labeled Faces in the Wild) is a well-known dataset for face verification in-the-wild. LFW contains more than 13,000 images of 1680 identities, and each identity has two or more images of various poses. CPLFW (Cross-Pose LFW) is an extended version of LFW, which is more difficult due to different illuminations, occlusions, and expressions. CFP dataset consists of 500 subjects, each of which has ten frontal and four profile images. There are two

Na *et al. Hum. Cent. Comput. Inf. Sci.*   (2020) 10:45

Page 9 of 17

Fig. 5 The creation of ground-truth complete UV maps. Three facial images with yaw angles of 0°, −30°, +30° are fed to the 3DDFA model to create three incomplete UV maps which are then merged by Poisson blending to generate the ground-truth complete UV map
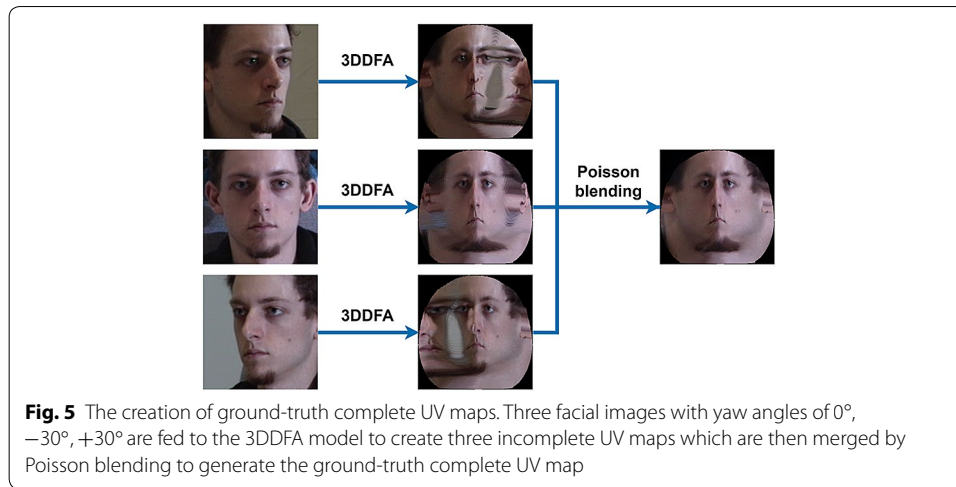
Table 1 Performance comparison of different methods on the Multi-PIE dataset

Model	SSIM	PSNR
UV-GAN [35]	0.61	13.67
Our ResCUNet-GAN (w/o attention gates and fast normalized fusion)	0.66	15.67
Our Attention ResCUNet-GAN	0.685	15.974

evaluation protocols regarding the CFP dataset: frontal–frontal (FF) and frontal-profile (FP) face verification. Each of them has ten folders with 350 same-person pairs and 350 different-person pairs.

Image reconstruction

We use two metrics to evaluate the quality of output from the Attention ResCUNet-GAN. The first metric is the structural similarity (SSIM), which is designed for measuring the similarity between images. The second one is the peak signal-to-noise ratio (PSNR), which is commonly used to measure the quality of reconstruction. Table 1 shows that our method achieves better results than the original UV-GAN according to both metrics SSIM and PSNR.

Figures 6 and 7 show the results of UV map completion on the test data taken from the Multi-PIE, where the UV map ground truths are available. For frontal input faces, the results of different methods look similar to each other. However, for profile input faces, the results are quite different. UV-GAN produces the worst UV maps. Normal ResCU-Net-GAN yields better results, and Attention ResCUNet-GAN gives the most realistic ones with a smooth texture. Note that the intermediate output obtained from the first U-Net of the Attention ResCUNet-GAN still yields better results than UV-GAN's. The results from some in-the-wild input images are shown in Fig. 8. One can see that Attention ResCUNet-GAN yields significantly better results than other ones, especially compared to the original UV-GAN.

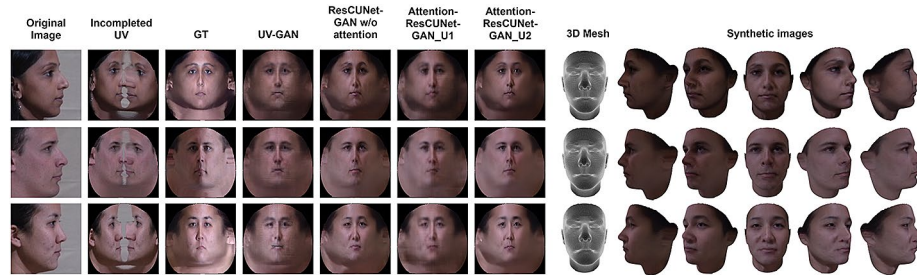Na *et al. Hum. Cent. Comput. Inf. Sci.*   (2020) 10:45

Page 10 of 17

Fig. 6 Results with frontal input images. Incomplete UV maps are generated using 3DDFA. Next columns are ground truth UV maps, results of UV-GAN, results of normal ResCUNet-GAN, intermediate results of Attention ResCUNet-GAN (after the first U-Net) and final results of Attention ResCUNet-GAN (after the second U-Net), respectively. The most right block shows some synthetic images generated based on the final results of Attention ResCUNet-GAN

Fig. 7 Results with profile input images. Incomplete UV maps are generated using 3DDFA. Next columns are ground truth UV maps, results of UV-GAN, results of normal ResCUNet-GAN, intermediate results of Attention ResCUNet-GAN (after the first U-Net) and final results of Attention ResCUNet-GAN (after the second U-Net), respectively. The most right block shows some synthetic images generated based on the final results of Attention ResCUNet-GAN
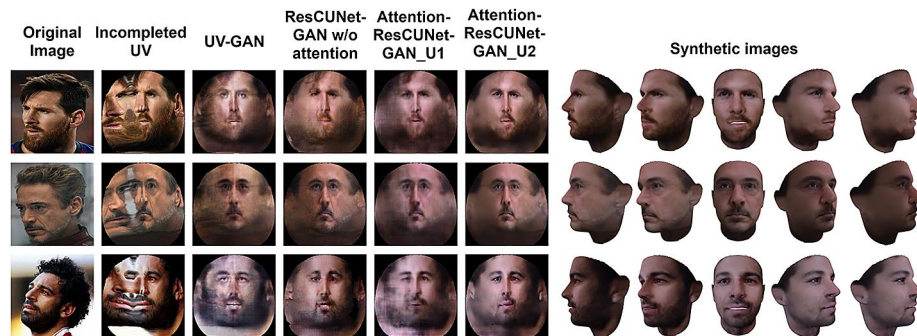
Fig. 8 Results with in-the-wild input images. Incomplete UV maps are generated using 3DDFA. The ground truth UV maps are unavailable. The next columns are the results of UV-GAN, results of normal ResCUNet-GAN, intermediate results of Attention ResCUNet-GAN (after the first U-Net), and final results of Attention ResCUNet-GAN (after the second U-Net), respectively. The right block shows some synthetic images generated based on the final results of Attention ResCUNet-GAN

In Figs. 9, 10 and 11, we show side-by-side synthetic images generated from the UV map reconstructed by UV-GAN and the proposed Attention ResCUNet-GAN, respectively. One can see that our model yields qualitatively better results than the original UV-GAN, especially for profile and in-the-wild input images.

Na *et al. Hum. Cent. Comput. Inf. Sci.*   (2020) 10:45

Page 11 of 17

Fig. 9 Synthetic images for frontal input images. The left block corresponds to the result of UV-GAN. The right block corresponds to the final result of Attention ResCUNet-GAN (after the second U-Net)
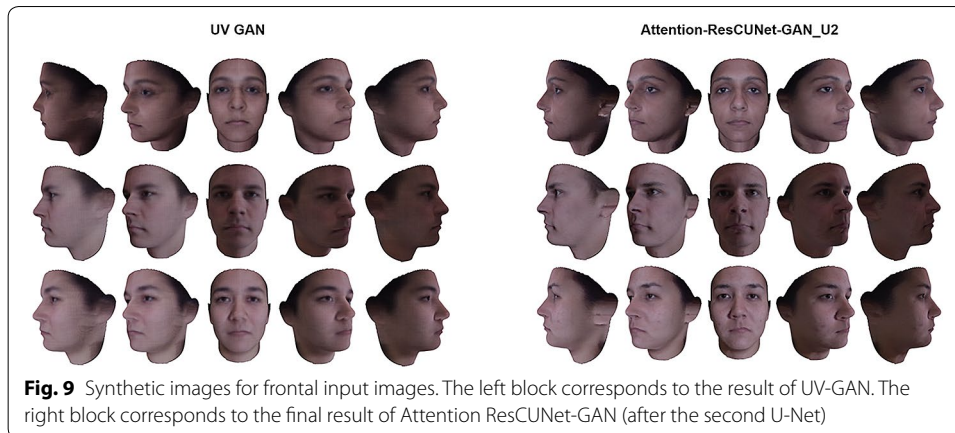
The facial images in the Multi-PIE dataset are not diverse enough to reflect the real data distribution. Thus, in-the-wild faces occluded by strange things or with too much makeup can lead to some failures of the model, as illustrated in Fig. 12.

Attention map visualization

The attention coefficients of the proposed Attention ResCUNet-GAN are visualized in Fig. 13. These attention coefficients are obtained in the attention gate of the AFC node that takes $S9$ as input (see Fig. 2). One can see that the attention maps try to ignore the visible face regions, focusing only on the missing regions of incomplete UV maps.

Pose invariance face recognition

We compare our methods with UV-GAN on the Multi-PIE dataset in the face verification task. We take facial images from different pose ranged from 0° to 75° and frontalize them using UV-GAN and our methods. We then use a face detector [53] to crop the central faces from the generated complete UV maps and push the cropped faces through ArcFace [54] to verify if the synthetic frontal face and the ground truth one belong to the same subject or not. The verification results are shown on Table 2. One can see that the verification accuracy falls down along with the increase of pose. Nevertheless, our proposed ResCUNet-GANs (and even ResUNet-GAN with one U-Net block) always produces better frontal faces in term of preserving identity. Attention ResCUNet-GAN outperforms other methods by orders of magnitude on all profile poses. Surprisingly, for frontal faces, Attention ResCUNet-GAN yields little degraded results than normal ResCUNet-GAN. The reason may be that the useful information, which is necessary for the recognition task, in frontal images is almost comprehensive. Hence, a complicated transformer with attention gates and fast normalized fusion might unintendedly diminish some useful information and leads to the degradation in the verification accuracy.

In the next experiment, we train a face recognition model on the CASIA dataset and evaluate its performance in the face verification task on other different datasets. Firstly, we train a face deep feature extractor with ResNet-101 backbone and arcface [54] loss on the CASIA dataset augmented by using Attention ResCUNet-GAN. For each identity in CASIA, we generate different profile faces from the frontal one, ranging from −80° to

Fig. 10 Synthetic images for profile input images. The left block corresponds to the result of UV-GAN. The right block corresponds to the final result of Attention ResCUNet-GAN (after the second U-Net)

Fig. 11 Synthetic images for in-the-wild input images. The left block corresponds to the result of UV-GAN. The right block corresponds to the final result of Attention ResCUNet-GAN (after the second U-Net)

80° with the step of 20°. For each identity, we synthesize approximately 300 frontal and profile images. We train the network with a batch size of 128 for 30 epochs. The learned model is then used for the verification task on the LFW, CPLFW and CFP datasets. Note that for the CFP dataset, we consider two verification types: frontal–frontal means to verify two frontal faces, and frontal-profile means to verify a frontal face and a profile one. We use k-fold cross-validation to evaluate the face verification task. Particularly, each dataset will be divided into 10 groups ($k = 10$). Each group is considered as a test set in turn, while the remaining groups are used to tune the best verification threshold. In total, we have ten runs for each face verification dataset. The mean accuracy and the standard deviation over ten runs are reported. Tables 3 and 4 show that data augmentation using the proposed Attention ResCUNet-GAN improves the performance of the recognition model. Note that the LFW dataset does not pay much attention to cross-pose face verification, and most faces in this dataset are nearly frontal. Therefore, a heavy facial pose augmentation using generative networks for training the recognition model is probably not really necessary. In fact, the verification performance on the LFW dataset over ten runs slightly fluctuates when we apply the proposed generative model for data augmentation. The standard deviation of accuracy increases from 0.032 to 0.391 (see
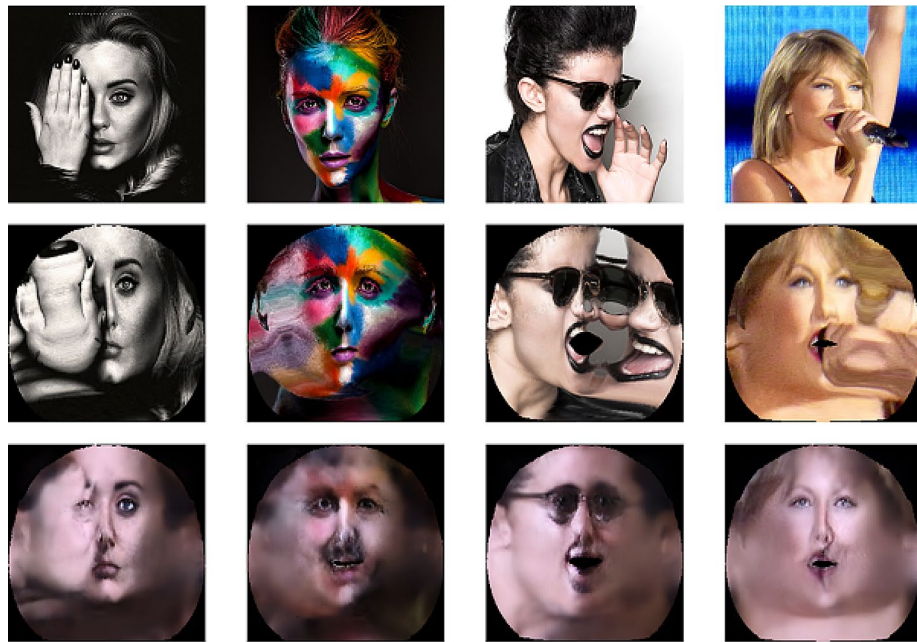
Fig. 12 Some failed cases when the input facial images are "abnormal" with respect to the training data. The top row shows the input images, the second row contains incomplete UV map and the third row displays the completed UV maps generated by our Attention ResCUNet-GAN

Table 2 Verification results on different poses on the Multi-PIE dataset

Model	0°	±15°	±30°	±45°	±60°	±75°
UV-GAN [3]	0.688	0.655	0.631	0.605	0.553	0.512
Our ResCUNet-GAN (w/o attention gates and fast normalized fusion)	0.983	0.840	0.801	0.761	0.732	0.705
Our Attention ResCUNet-GAN	0.955	0.874	0.837	0.794	0.768	0.742

Table 3 Verification accuracy (%) comparison on the LFW and CPLFW datasets

Model	LFW	CPLFW
HUMAN-individual	97.27	81.21
CASIA-sm-augUVGAN [35]	99.22	–
CASIA-arcface	99.43 ± 0.032	90.30 ± 1.797
CASIA-arcface-augAttentionResCUNetGAN	99.47 ± 0.391	90.45 ± 1.812

Table 3). However, in overall, using Attention ResCUNet-GAN still helps to improve the average cross-validation accuracy. In contrast to the LFW dataset, the CPLFW dataset has lots of positive face pairs with different poses to enlarge intra-class variance. In this case, our model results in more stable improvements, where the standard deviation of accuracy is almost the same as if the data augmentation is not used.

Table 4 Verification accuracy (%) comparison on the CFP dataset

Model	Frontal–frontal	Frontal-profile
CASIA-center [35]	98.34 ± 0.44	87.77 ± 2.39
CASIA-Sphere [35]	98.64 ± 0.24	84.39 ± 2.59
CASIA-sm [35]	98.59 ± 0.21	87.74 ± 1.07
CASIA-sm-aug1 [35]	98.25 ± 0.42	90.14 ± 1.53
CASIA-sm-augUV-GAN [35]	98.83 ± 0.27	93.09 ± 1.72
Profile2Frontal	–	93.55 ± 1.67
Frontal2Profile	–	93.72 ± 1.59
Template2Template	–	94.05 ± 1.73
CASIA-arcface-augAttentionResCUNetGAN	99.47 ± 0.29	97.0 ± 0.82

Fig. 13 Attention map visualization. The first column contains UV maps generated by 3DDFA network, the second column contains generated UV maps overlaid by attention masks, and the last column illustrates attention coefficients only
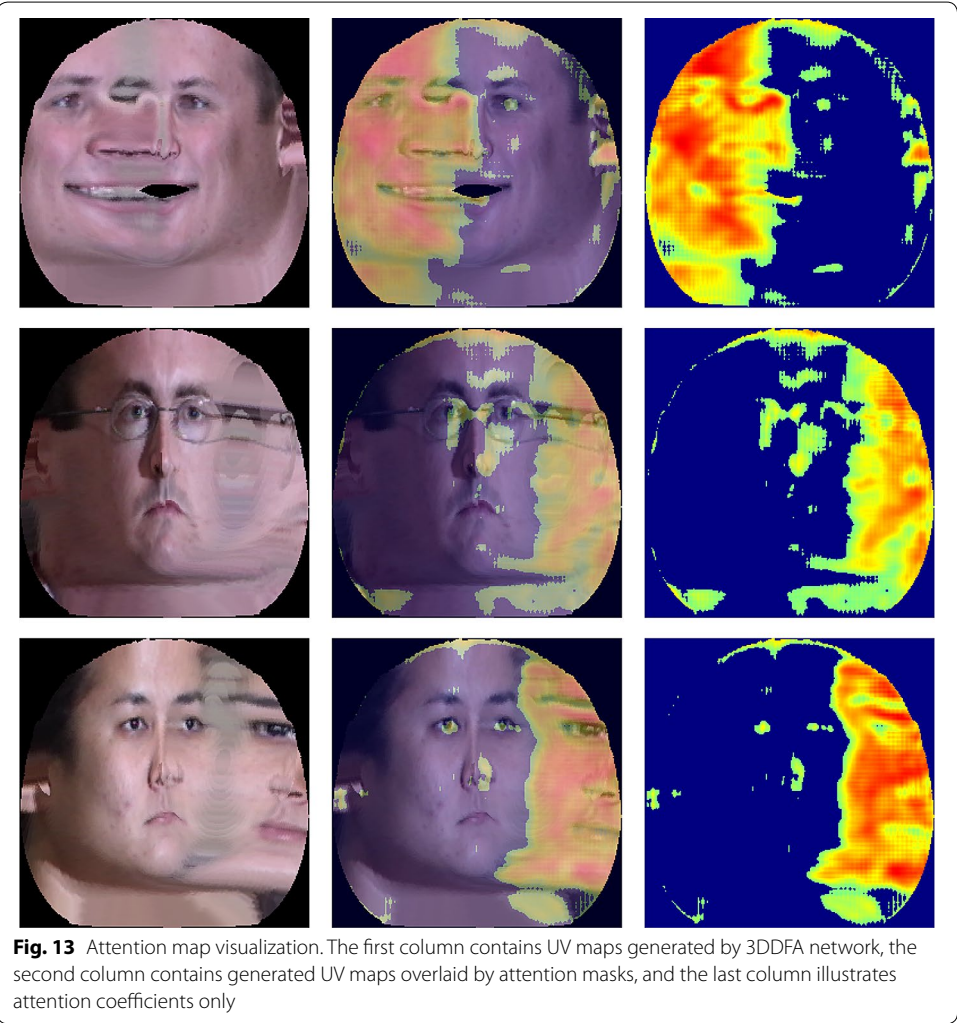
Fig. 14 Some samples of positive pairs from the CFP dataset

The CFP dataset focuses on the pose variation in terms of extreme pose where many details of faces are occluded (see Fig. 14). One can see from Table 4, our Attention ResCUNetGAN considerably improves the performance of the face recognition model, especially for the frontal-profile subtask.

Conclusions and future work

In this paper, we introduce a novel generative model called Attention ResCUNet-GAN to generate complete facial UV maps, which allows us to synthesize various faces of arbitrary poses and improve pose-invariant face recognition performance. We leverage the residual connections in ResNet, intra-block and extra-block feature fusion in coupled UNets to enhance the generator. The skip connections within each U-Net are amplified with attention gates, while the contextual feature maps from two U-Nets are fused with trainable scalar weights. We jointly train global and local adversarial losses with identity preserving loss. The experiments show that the proposed Attention ResCUNet-GAN outperforms the original UV-GAN by order of magnitude in terms of both reconstruction metrics and the performance on the pose-invariant face verification task.

In future work, we would like to exploit some recent efficient backbones such as EfficientNet [55] to improve the performance of the proposed approach. More complex short-cut connections [45, 56] can also be utilized to improve gradient flow and stimulate feature reuse within the network.

or implied, of the CCDC Pacific and CCDC ARL and the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation hereon.

Author details
[1] Chosun University, 309 Pilmun-daero, Gwangju 61452, South Korea. [2] Hanoi University of Science and Technology, 1 Dai Co Viet, Hanoi, Vietnam. [3] Vietnam Academy of Science and Technology, 18 Hoang Quoc Viet, Hanoi, Vietnam.

References
1. Masi I, Wu Y, Hassner T, Natarajan P (2018) Deep face recognition: a survey. In: 2018 31st SIBGRAPI conference on graphics, patterns and images (SIBGRAPI), IEEE, pp 471–478
2. Zhou S, Xiao S (2018) 3d face recognition: a survey. Hum-Cent Comput Inf Sci 8(1):35
3. Jafri R, Arabnia HR (2009) A survey of face recognition techniques. J Inf Process Syst 5(2):41–68
4. Tran L, Yin X, Liu X (2017) Disentangled representation learning gan for pose-invariant face recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1415–1424
5. Parkhi OM, Vedaldi A, Zisserman A (2015) Deep face recognition
6. Schroff F, Kalenichenko D, Philbin J (2015) Facenet: a unified embedding for face recognition and clustering. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 815–823
7. Sun Y, Chen Y, Wang X, Tang X (2014) Deep learning face representation by joint identification-verification. In: Advances in neural information processing systems, pp 1988–1996
8. Taigman Y, Yang M, Ranzato M, Wolf L (2014) Deepface: closing the gap to human-level performance in face verification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1701–1708
9. Yang J, Reed SE, Yang M-H, Lee H (2015) Weakly-supervised disentangling with recurrent transformations for 3d view synthesis. In: Advances in neural information processing systems, pp 1099–1107
10. Sayan M, Mohamed A-M, Shihab SA (2020) Multimodal biometrics recognition from facial video with missing modalities using deep learning. J Inf Process Syst 16(1):6–29
11. Sang DV, Van Dat N, et al (2017) Facial expression recognition using deep convolutional neural networks. In: 2017 9th international conference on knowledge and systems engineering (KSE), IEEE, pp 130–135
12. Hai-Duong N, Sun-Hee K, Guee-Sang L, Hyung-Jeong N, Yang abd In-Seop, Soo-Hyung K (2019) Facial expression recognition using a temporal ensemble of multi-level convolutional neural networks. IEEE Trans Affect Comput
13. Blanco-Gonzalo R, Poh N, Wong R, Sanchez-Reillo R (2015) Time evolution of face recognition in accessible scenarios. Hum-Cent Comput Inf Sci 5(1):24
14. Masi I, Tran AT, Hassner T, Leksut JT, Medioni G (2016) Do we really need to collect millions of faces for effective face recognition? In: European conference on computer vision, Springer, pp 579–596
15. Sagonas C, Panagakis Y, Zafeiriou S, Pantic M (2015) Robust statistical face frontalization. In: Proceedings of the IEEE international conference on computer vision, pp 3871–3879
16. Kan M, Shan S, Chang H Chen X (2014) Stacked progressive auto-encoders (spae) for face recognition across poses. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1883–1890
17. Hassner T, Harel S, Paz E, Enbar R (2015) Effective face frontalization in unconstrained images. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4295–4304
18. Peng X, Yu X, Sohn K, Metaxas DN, Chandraker M (2017) Reconstruction-based disentanglement for pose-invariant face recognition. In: Proceedings of the IEEE international conference on computer vision, pp 1623–1632
19. Chongxuan L, Xu T, Zhu J, Zhang B (2017) Triple generative adversarial nets. In: Advances in neural information processing systems, pp 4088–4098
20. Yeh R, Chen C, Lim TY, Hasegawa-Johnson M, Do MN (2016) Semantic image inpainting with perceptual and contextual losses, 2(3). arXiv preprint arXiv:1607.07539
21. Yu J, Lin Z, Yang J, Shen X, Lu X, Huang TS (2018) Generative image inpainting with contextual attention. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 5505–5514
22. Luan F, Paris S, Shechtman E, Bala K (2017) Deep photo style transfer. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4990–4998
23. Zhu J-Y, Park T, Isola P, Efros AA (2017) Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision, pp 2223–2232
24. Karras T, Aila T, Laine S, Lehtinen J (2017) Progressive growing of gans for improved quality, stability, and variation. arXiv preprint arXiv:1710.10196
25. Karras T, Laine S, Aila T (2019) A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4401–4410
26. Ledig C, Theis L, Huszár F, Caballero J, Cunningham A, Acosta A, Aitken A, Tejani A, Totz J, Wang Z, et al (2017) Photo-realistic single image super-resolution using a generative adversarial network. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4681–4690
27. Tran L, Yin X, Liu X (2018) Representation learning by rotating your faces. IEEE Trans Pattern Anal Mach Intell 41(12):3007–3021

28. Wang Q, Fan H, Sun G, Ren W, Tang Y (2020) Recurrent generative adversarial network for face completion. IEEE Trans Multimed

29. 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: Proceedings of the IEEE international conference on computer vision, pp 2439–2448

30. Yin X, Yu X, Sohn K, Liu X, Chandraker M (2017) Towards large-pose face frontalization in the wild. In: Proceedings of the IEEE international conference on computer vision, pp 3990–3999

31. 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: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2207–2216

32. Duan Q, Zhang L (2020) Look more into occlusion: realistic face frontalization and recognition with boostgan. IEEE Trans Neural Netw Learn Syst

33. Zhao J, Xing J, Xiong L, Yan S, Feng J (2020) Recognizing profile faces by imagining frontal view. Int J Comput Vis 128(2):460–478

34. Zhang F, Zhang T, Mao Q, Xu C (2020) Geometry guided pose-invariant facial expression recognition. IEEE Trans Image Process 29:4445–4460

35. Deng J, Cheng S, Xue N, Zhou Y, Zafeiriou S (2018) Uv-gan: adversarial facial uv map completion for pose-invariant face recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7093–7102

36. Booth J, Antonakos E, Ploumpis S, Trigeorgis G, Panagakis Y, Zafeiriou S (2017) 3d face morphable models "in-the-wild". In: 2017 IEEE conference on computer vision and pattern recognition (CVPR), IEEE, pp 5464–5473

37. Isola P, Zhu J-Y, Zhou T, Efros AA (2017) Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1125–1134

38. Isola P, Zhu J-Y, Zhou T, Efros AA (2017) Image-to-image translation with conditional adversarial networks. CVPR, Salt Lake City

39. 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, Springer, pp 234–241

40. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778

41. Xue N, Deng J, Cheng S, Panagakis Y, Zafeiriou S (2019) Side information for face completion: a robust PCA approach. IEEE Trans Pattern Anal Mach Intell 41(10):2349–2364

42. He R, Cao J, Song L, Sun Z, Tan T (2019) Adversarial cross-spectral face completion for nir-vis face recognition. IEEE Trans Pattern Anal Mach Intell 42(5):1025–1037

43. Shah S, Ghosh P, Davis LS, Goldstein T (2018) Stacked u-nets: a no-frills approach to natural image segmentation. arXiv preprint arXiv:1804.10343

44. Newell A, Yang K, Deng J (2016) Stacked hourglass networks for human pose estimation. In: European conference on computer vision, Springer, pp 483–499

45. Ibtehaz N, Rahman MS (2020) Multiresunet: rethinking the u-net architecture for multimodal biomedical image segmentation. Neural Netw 121:74–87

46. Schlemper J, Oktay O, Schaap M, Heinrich M, Kainz B, Glocker B, Rueckert D (2019) Attention gated networks: learning to leverage salient regions in medical images. Med Image Anal 53:197–207

47. Tang Z, Peng X, Geng S, Zhu Y, Metaxas DN (2019) Cu-net: coupled u-nets. In: 29th British machine vision conference, BMVC 2018

48. Zhu X, Liu X, Lei Z, Li SZ (2017) Face alignment in full pose range: a 3d total solution. IEEE Trans Pattern Anal Mach Intell 41(1):78–92

49. Blanz V, Vetter T (1999) A morphable model for the synthesis of 3d faces. In: Proceedings of the 26th annual conference on computer graphics and interactive techniques, pp 187–194

50. Tan M, Pang R, Le QV (2019) Efficientdet: scalable and efficient object detection. arXiv preprint arXiv:1911.09070

51. Gross R, Matthews I, Cohn J, Kanade T, Baker S (2010) Multi-pie. Image Vis Comput 28(5):807–813

52. Pérez P, Gangnet M, Blake A (2003) Poisson image editing. In: ACM SIGGRAPH 2003 papers, pp 313–318

53. Zhang K, Zhang Z, Li Z, Qiao Y (2016) Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Process Lett 23(10):1499–1503. https://doi.org/10.1109/LSP.2016.2603342

54. Deng J, Guo J, Xue N, Zafeiriou S (2019) Arcface: additive angular margin loss for deep face recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4690–4699

55. Tan M, Le QV (2019) Efficientnet: rethinking model scaling for convolutional neural networks. arXiv preprint arXiv:1905.11946

56. Zhou Z, Siddiquee MMR, Tajbakhsh N, Liang J (2018) Unet++: a nested u-net architecture for medical image segmentation. In: Deep learning in medical image analysis and multimodal learning for clinical decision support, pp 3–11

Publisher's Note