INFORMATICA, 2021, Vol. 32, No. 2, 425–440

DOI: https://doi.org/10.15388/21-INFOR443

PFA-GAN: Pose Face Augmentation Based on Generative Adversarial Network

Bassel ZENO[1,*], Ilya KALINOVSKIY[2], Yuri MATVEEV[1,2]

[1] *ITMO University, Kronverkskiy Prospekt 49, St. Petersburg 197101, Russia*
[2] *STC-innovations Ltd., Gelsingforsskaya Street 3, Building 11D, St. Petersburg 194044, Russia*
e-mail: basilzeno@gmail.com

Received: June 2020; accepted: January 2021

Abstract. In this work, we propose a novel framework based on Generative Adversarial Networks for pose face augmentation (PFA-GAN). It enables a controlled pose synthesis of a new face image from a source face given a driving one while preserving the identity of the source face. We introduce a method for training the framework in a fully self-supervised mode using a large-scale dataset of unconstrained face images. Besides, some augmentation strategies are presented to expand the training set. The face verification experimental results demonstrate the effectiveness of the presented augmentation strategies as all augmented datasets outperform the baseline.

Key words: generative adversarial networks, face verification, visual data augmentation.

1. Introduction

A person's face plays a key role in the identification of individual members of our highly social species due to delicate differences that make every human face unique. These variations of a face pattern inform us also about characteristics such as age, gender, and race. Over the last decade, many remarkable works based on Deep Neural Networks have demonstrated unprecedented performance on several computer vision tasks, such as facial landmark detection, face identification, face verification, face alignment, emotion classification, etc. In addition, they showed that achieving a good generalization in unconstrained conditions strongly relies on training them on large and complex datasets. Well-annotated large-scale dataset can be both expensive and time-consuming to acquire. Hiring people to manually collect images and annotate them is not efficient at all since this manual process is widely recognized as error-prone. Furthermore, the existing face image datasets suffer from the problem of insufficient data amount for each person and the unbalanced pose data distribution between the classes. In addition, there is a lack of variations comparing to the real samples in the world. To cope with insufficient facial training data, visual data augmentation provides an effective alternative. It is a technique that enables practitioners to significantly increase the diversity of data available for training models, by transforming collected real face samples. The traditional visual data augmentation methods alter the

*Corresponding author.

entire face image by transferring image pixel values to new positions or by shifting pixel colours to new values. For instance, zooming in and out, rotating or reflecting the original image, translating, applying distortion and cropping. These generic methods have some limitations. (1) They do not scale well the number of variations of facial appearances, such as make-up, lighting, and skin color. (2) Creating high-level content such as rotating head while preserving the identity is a challenging problem (Zeno *et al.*, 2019b) and it is still under study. The large discrepancy of head poses in the real world is a big challenge in face detection, identification (Farahani and Mohseni, 2019), and verification (Ribarić *et al.*, 2008), due to lighting variations and self-occlusion. Therefore, many methods were proposed to generate face images with new poses. Pose synthesis methods can be classified into a 2D geometry-based approach, a 3D geometry-based approach, and a learning-based approach. The 2D and 3D based methods appeared earlier than learning-based approaches, have obvious advantage in that they need a small amount of training data. The 2D-based methods rely on building a PCA model for a face shape to control only yaw rotations (Feng *et al.*, 2017), while the 3D based methods synthesize face images with new variations of poses using a 3D morphable face model (Crispell *et al.*, 2017; Blanz and Vetter, 1999; Zhu *et al.*, 2016; Guo *et al.*, 2017). In recent years, many learning-based methods have been proposed for face rotation, where most of them rely on a generative adversarial network (Tran *et al.*, 2017; Tian *et al.*, 2018; Cao *et al.*, 2018a; Antoniou *et al.*, 2018; Yin *et al.*, 2017; Huang *et al.*, 2017; Zeno *et al.*, 2019a). For example, the methods DRGAN (Tran *et al.*, 2017), CRGAN (Tian *et al.*, 2018) and LB-GAN (Cao *et al.*, 2018a) were proposed to rotate a face image around the yaw axis only. While DRGAN synthesizes a new pose even for extreme profiles ($\pm 90°$), CRGAN learns "complete" representations to rotate unseen faces, and LB-GAN frontalizes a face image before generating the target pose. The frontalization is a particular case of pose transformation often used to increase the accuracy of face recognition systems by rotating faces to the front view, such as in FF-GAN (Yin *et al.*, 2017) and TP-GAN (Huang *et al.*, 2017) works. Recently, Zeno *et al.* (2019a) proposed IP-GAN framework to generate a face image of any specific identity with an arbitrary target pose by explicitly disentangling identity and pose representation from a face image.

However, we argue that there are several drawbacks to the listed methods. The reposing method proposed in Crispell *et al.* (2017) produces many distortions in face structure and does not fix the background. And the 3D based approach (Blanz and Vetter, 1999) fails with large poses and it requires some additional steps to generate the hidden regions (e.g. the teeth). The augmentation methods in Zhu *et al.* (2016), Guo *et al.* (2017) reduce the realism of the generated images. On the other side, the GAN learning-based methods (Tran *et al.*, 2017; Tian *et al.*, 2018; Cao *et al.*, 2018a; Antoniou *et al.*, 2018; Yin *et al.*, 2017; Huang *et al.*, 2017) obtain impressive results, but they need additional information such as conditioning labels (e.g. indicating a head pose, 3DMM parameters). More specifically, Yin *et al.* (2017), Huang *et al.* (2017) need frontal face annotations, while (Tran *et al.*, 2017; Tian *et al.*, 2018; Cao *et al.*, 2018a; Antoniou *et al.*, 2018) need profile labels, while the IP-GAN (Zeno *et al.*, 2019a) framework does not require any pose annotations. But despite this, it failed to learn disentangled representation of pose and identity on an

unconstrained dataset of face images. Besides, the learning scheme of IP-GAN is very complex, which makes it difficult for it to converge.

To address the issues above, in this work we focus on pose face transformation for visual data augmentation using the Generative Adversarial Networks. We propose a novel GAN framework that enables a controlled synthesis of new face images from a single source face image given a driving face image while preserving the subject identity. The framework is trained in self-supervised settings using pairs of source and driving face images. To demonstrate the performance of our model, some face verification experiments are conducted using our proposed pose augmentation strategies. The framework architecture is described in Section 3, and the self-supervised training method in Section 4.

To conclude, our contributions are:

- We present the Pose Face Augmentation GAN (PFA-GAN) that can transform a pose of a source face image using another face image while preserving the identity of the source image, as well as the pose and the expression of the driving face image. The proposed framework consists of an identity encoder network, a pose encoder network, a generator, and a discriminator.
- We introduce a novel method for training the network in fully self-supervised settings using a large-scale dataset of unconstrained face images.
- We introduce some augmentation strategies that demonstrate how a baseline training set can be augmented to increase the pose variations.
- We conduct some comparative experiments on face verification. Our results clearly show that the augmented datasets based on our method outperform the baseline methods.

2. Related Work

2.1. *2D/3D Model-Based*

Feng *et al.* (2017) proposed a 2D-based method to generate profile virtual faces with out-of-plane pose variations. They built a PCA-based shape model to control only the yaw rotations since the pose varies with the same rotation direction of the original shape, i.e. left or right. Meanwhile, many approaches employed 3D face models for face pose translation (Crispell *et al.*, 2017; Blanz and Vetter, 1999; Zhu *et al.*, 2016; Guo *et al.*, 2017). Crispell *et al.* (2017) use a 3D face shape estimation method, followed by a rendering pipeline for arbitrarily reposing of faces and altering the light conditions. Although the results of the face re-lighting method are good, reposing of the face produces many distortions in its structure. In addition, the background is not fixed since it is rotated along with the direction of the face rotation. Blanz and Vetter (1999) proposed a method to estimate a 3D morphable face model by transforming the shape and the texture of a face image into a vector space representation. Then, faces with new poses and expressions can be modelled by modifying the estimated parameters to match the target 3D face model. This method is good at generating faces with small poses, but it failed with large poses due

to the serious loss of the facial texture. Furthermore, some additional steps are required at synthesizing facial expressions such as smiling to generate the hidden regions (e.g. the teeth). Zhu *et al.* (2016) introduced the 3D Dense Face Alignment (3DDFA) algorithm to solve face alignment in large poses. 3DDFA has also been used to profile faces, which means synthesizing the face appearances in profile view from medium pose samples by predicting the depth of face image. However, this augmentation method reduces the realism of the generated images. Guo *et al.* (2017) proposed a face inverse rendering method (3DFaceNet) to recover geometry and lighting from a single image. With that, they can generate new face images with different attributes. Nevertheless, their inverse rendering procedure has limitations, and it may lead to inaccurate fitting for face images (e.g. estimating the coarse face geometry and pose parameters from a face image).

2.2. *GANs-Based*

Recently, generative adversarial network model learning (Tran *et al.*, 2017; Tian *et al.*, 2018; Cao *et al.*, 2018a; Antoniou *et al.*, 2018; Yin *et al.*, 2017; Huang *et al.*, 2017; Zeno *et al.*, 2019a) demonstrated an outstanding ability to synthesize face images with new poses. Tran *et al.* (2017) introduced Disentangled Representation Learning-Generative Adversarial Network (DR-GAN), where the model takes a face image of any pose as input and outputs a synthetic face, frontal or rotated with the target pose, even for extreme profiles ($\pm 90°$). The discriminator in DR-GAN is trained also to predict the identity and the pose of the generated face. Tian *et al.* (2018) proposed the Complete Representation GAN-based method (CR-GAN) following a single-pathway design, and a two-pathway learning scheme to learn the "complete" representations. Cao *et al.* (2018a) have introduced Load Balanced Generative Adversarial Networks (LB-GAN) to rotate the yaw angle of an input face image to the target angle from a specified set of learned poses. The LB-GAN consists of two modules: a normalizer, which first frontalizes the face images, and an editor, which rotates the frontal face after that. Antoniou *et al.* (2018) introduced Data Augmentation Generative Adversarial Network (DAGAN) based on conditional GAN (cGAN). DAGAN captures the cross-class transformations since it takes any data item and generates other points of the equivalent class. A particular case of a pose transformation is the face frontalization. It is often used to increase the accuracy of face recognition systems by rotating faces to the frontal view, which is more convenient for a recognition model. Many methods have been introduced to frontalize profile faces, such as GAN-based methods (Yin *et al.*, 2017; Huang *et al.*, 2017). The FF-GAN method (Yin *et al.*, 2017) relies on any 3D knowledge for geometry shape estimation, while the TP-GAN method (Huang *et al.*, 2017) infers it through data-driven learning. TP-GAN is a Two-Pathway Generative Adversarial Network for synthesizing photorealistic frontal views from profile images by simultaneously perceiving global structures and local details. FF-GAN is a Face Frontalization Generative Adversarial Network framework, which incorporates elements from both deep 3DMM and face recognition CNNs to achieve high-quality and identity-preserving frontalization with less training data. Both TP-GAN and FF-GAN methods obtained impressive results on face frontalization, but they need explicit front-view annotations.

2.3. *IP-GAN*

Zeno *et al.* (2019a) proposed a framework for Learning Identity and Pose Disentanglement in Generative Adversarial Networks (IP-GAN). To generate a face image of any specific identity with an arbitrary target pose, IP-GAN incorporates the pose information in the synthesis process. Different from the recent work (Yin *et al.*, 2017) that uses a 3D morphable face simulator to generate pose information and the works (Tran *et al.*, 2017; Tian *et al.*, 2018; Huang *et al.*, 2017) that encode pose annotation in a one-hot vector, IP-GAN can learn such information by explicitly disentangling identity and pose representation from a face image in fully self-supervised settings. The overall architecture of the IP-GAN framework is depicted in Zeno *et al.* (2019a), and consists of five parts: 1) the identity encoder network E_I to extract the identity latent code; 2) the head pose encoder network E_P to extract the pose latent code; 3) the generative network G to produce the final output image using the combined identity latent code and the extracted pose latent code; 4) the identity classification network C to preserve the identity by measuring the posterior probability of the subject identities; 5) the discriminative network D to distinguish between real and generated images. To train these networks Zeno *et al.* (2019a) proposed a learning method in order to learn complete representations in fully self-supervised settings. The learning method consists of two learning pathways, generation, and transformation. While the generation pathway focuses on mapping the entire latent spaces of encoders to high-quality images, the transformation pathway focuses on the synthesis of new face images with the target poses. This framework has many drawbacks and when it was trained on an unconstrained dataset of face images, it failed to learn disentangled representation of pose and identity. Besides, the learning scheme is very complex that makes it difficult for the GAN to converge.

3. The Proposed Framework

Inspired by the IP-GAN model, we present in this section a novel framework (PFA-GAN) for pose face augmentation based on a generative adversarial network.

3.1. *PFA-GAN*

To simplify the proposed architecture in Zeno *et al.* (2019a), and according to the specific goal of the PFA-GAN in generating face images with new poses, we remove the Classification Network C, as there is no need to add a new task of face recognition to PFA-GAN, and preserving the subject identity in the generated face image is guaranteed by the use of the content loss function. To reduce the complexity of the learning method, we propose removing the generation pathway and focusing on the work of the transformation pathway, which consists of two sub-paths: reconstruction and transformation. The task of the head pose encoder network E_P is to learn a pose representation as it has to isolate the pose information from the other information in a face image such as age, gender, skin color, and identity. Isolating pose information in unconstrained images is a challenging

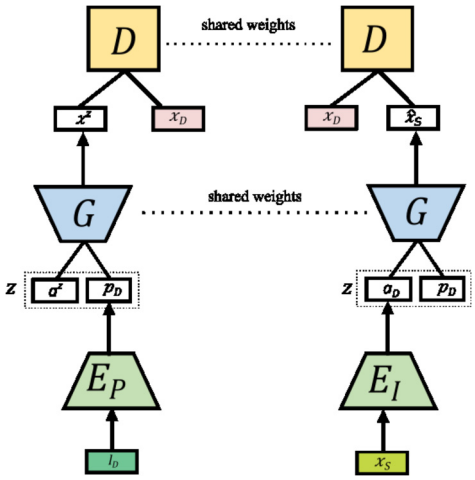

Fig. 1. The proposed framework architecture, pose encoder network, identity encoder network, generator, discriminator. Learning scheme from left to right: the reconstruction sub-path, the transformation sub-path.

task. To facilitate it, we reduce the amount of input information for the network E_P by replacing a face image with an image of its landmarks.

3.2. *Model Description*

Let $x_S \in \mathbb{X}$ be a source face image of a certain subject identity, and $x_D \in \mathbb{X}$ be a driver face image to extract the target pose features. Our goal is to generate a new face image $\acute{x}_S$ of the subject of x_S with the extracted face pose of x_D. To achieve this goal, we assume that each image $x \in \mathbb{X}$ is generated from an identity embedding vector $a \in \mathbb{A}$ and a pose embedding vector $p \in \mathbb{P}$. In other words, x_S, x_D are synthesized by the pair (a_S, p_S) and the pair (a_D, p_D), respectively. As a result, the new face image $\acute{x}_S$ is generated by the pair (a_S, p_D).

3.3. *Framework Architecture*

The proposed framework consists of the following four components, see Fig. 1:

- The pose encoder network $E_P(l_D; \Theta_P)$ receives a three-channel image of driving landmarks $l_D \in \mathbb{L}$ and maps it into a pose embedding vector p_D. Here Θ_P denotes the network parameters that are learned in a way that allows the vector p_D to only represent the pose information of the driving image. We denote with p_S, p_D the pose embedding vectors for the landmark images l_S, l_D, respectively.
- The identity encoder network $E_A(x_S; \Theta_A)$ takes a source face image x_S to extract an N-dimensional vector a_S that contains the source-specific information, such as a person's identity and skin tone information. Here Θ_A denotes the network parameters that are learned in our two sub-paths learning method.

- The generator $G(a_S, p_D; \Theta_G)$ takes the pose embedding vector p_D and the identity embedding vector a_S which is extracted from the source face image and outputs a synthesized target face image $\acute{x}_S$. During the two sub-paths learning method, the network parameters Θ_G are trained directly.
- The discriminator $D(x_D, \acute{x}_S; \Theta_{Dis})$ takes the driving face image and the generated one $\acute{x}_S$, then predicts whether the image is real or not. Here Θ_{Dis} denotes the network parameters of the discriminator.

4. The Proposed Learning Algorithm

In this section, we present our method for learning a pose face augmentation model (PFA-GAN). To achieve this goal the learning scheme is divided into two sub-paths, reconstruction and transformation, see Fig. 1. While the reconstruction sub-path aims to learn to generate a face image with the target pose, the learning goal of the transformation sub-path is to synthesize the target face image while preserving the identity of the subject. At each iteration, only one of these two sub-paths is randomly selected with a probability of 0.5.

4.1. *Reconstruction Sub-Path*

The reconstruction pathway trains the generator G, the pose encoder network E_P and the discriminator D. Here the identity encoder network E_A is not involved in the learning process since the network E_P learns the pose representations of the driving face images and the generator G tries to synthesize a face image using the driving pose embedding vector, while the identity embedding vector contains random values. Hence, given a random noise vector from noise uniform distribution $a^z \in Z$ and the pose embedding vector of the driving landmark image $p_D = E_P(l_D)$, we concatenate them in the latent space $z = [a^z, p_D]$ and feed them to the generator which aims to generate a realistic face image $x^z = G(a^z, p_D)$ under the driving pose latent vector p_D. Similar to the original GAN work (Huang *et al.*, 2007), the generative network G and the discriminative network D compete with each other in a two-player min-max game. While the discriminator D tries to distinguish real images from the output of G, the generator G tries to fool the network D. Specifically, D is trained to differentiate the fake image x^z from the real one x_D. This D minimizes:

$$L_{D-adv_{recon}} = E_{l\sim\mathcal{P}_l, a^z\sim\mathcal{P}_{a^z}}\big[D\big(G\big(a^z, E_P(l_D)\big)\big)\big] - E_{x\sim\mathcal{P}_x}\big[D(x_D)\big], \tag{1}$$

where $\mathcal{P}_l, \mathcal{P}_x$ are the real data distribution and $\mathcal{P}_{a^z}$ is the noise uniform distribution. G tries to fool D; it maximizes the following adversarial loss function:

$$L_{G-adv_{recon}} = E_{l\sim\mathcal{P}_l, a^z\sim\mathcal{P}_{a^z}}\big[D\big(G\big(a^z, E_P(l_D)\big)\big)\big]. \tag{2}$$

The pose encoder network helps the generator G to generate a high-quality image with pose of x_D, and to achieve that we reconstruct both the source and the driving face images and make use of a content-consistency loss function L_{cnt}, which measures differences in high-level content between the ground truth images x_S, x_D and the reconstructions $G(E_A(x_S), E_P(l_S))$, $G(E_A(x_D), E_P(l_D))$ using the perceptual similarity measure (Johnson *et al.*, 2016). Our content loss function uses the pre-trained VGG19 (Simonyan and Zisserman, 2015) and VGGFace (Parkhi *et al.*, 2015) networks since we extract the feature maps $\Phi^k(x)$ from several layers in these networks. Later, the loss is calculated as a weighted sum of ℓ_1-norm losses between the features of these networks:

$$L_{cnt\,S_{recon}} = \sum_{k=1}^{layers} \left\| \Phi^k\big(G\big(E_A(x_S), E_P(l_S)\big)\big) - \Phi^k(x_S) \right\|_1, \tag{3}$$

$$L_{cnt\,D_{recon}} = \sum_{k=1}^{layers} \left\| \Phi^k\big(G\big(E_A(x_D), E_P(l_D)\big)\big) - \Phi^k(x_D) \right\|_1. \tag{4}$$

We have added a regularization term that keeps the weights small, making the model simpler and avoiding overfitting:

$$L_{regular_{recon}} = \frac{1}{n} \sum_{i=1}^{n} \left\| \Theta_P^i \right\|_2, \tag{5}$$

where Θ_P denotes the parameters of the pose encoder network.

4.2. *Transformation Sub-Path*

The transformation sub-path trains the networks E_A, G, and D, but keeps the pose encoder network E_P fixed. The output of the E_A network should ensure preserving the identity of the source face image. We introduce a cross reconstruction task to make E_P and E_A disentangle the pose from the identity information. More specifically, we sample a real image pair (x_S^i, x_D^i) that shares the same identity but different appearance, poses, and facial expressions. The goal is to transform x_S^i to a new face image $\acute{x}_S^i$ where its pose matches the pose of the driving face image x_D^i. To achieve this, E_A receives x_S^i as an input and outputs a pose-invariant face representation a_S^i, while E_P takes l_D^i landmarks image of x_D^i as an input and outputs a pose representation vector p_D^i. We concatenate the embedding vectors and feed the combined vector, $z^i = (a_S^i, p_D^i) = [(E_A(x_S^i), E_P(l_D^i)]$ into the network G. The generator G should produce $\acute{x}_S^i$, the transformation of x_S^i. D is trained to distinguish the fake image $\acute{x}_S^i$ from the real one x_D^i. Thus, D minimizes:

$$L_{D-adv_{trans}} = E_{l\sim\mathcal{P}_l, x\sim\mathcal{P}_x}\big[D\big(G\big(E_A(x_S), E_P(l_D)\big)\big)\big] - E_{x\sim P_x}\big[D(x_D)\big]. \tag{6}$$

The generator tries to fool D network, it maximizes:

$$L_{G-adv_{trans}} = E_{l\sim\mathcal{P}_l, x\sim\mathcal{P}_x}\big[D\big(G\big(E_A(x_S), E_P(l_D)\big)\big)\big]. \tag{7}$$

To preserve the subject identity in the generated face image, we follow multiple feature-level warping methods instead of image-level warping. So similar to the reconstruction sub-path, the content-consistency loss is used in the transformation sub-path, since several feature maps $\Phi^k(x)$ are extracted from the pre-trained VGG19 and VGGFace networks:

$$L_{cnt_{trans}} = \sum_{k=1}^{layers} \left\| \Phi^k\left(G\left(E_A(x_S), E_P(l_D)\right)\right) - \Phi^k(x_D) \right\|_1, \tag{8}$$

$$L_{cntS_{trans}} = \sum_{k=1}^{layers} \left\| \Phi^k\left(G\left(E_A(x_S), E_P(l_S)\right)\right) - \Phi^k(x_S) \right\|_1. \tag{9}$$

To avoid overfitting problem, we add the following regularization loss function to keep the weights small in the identity encoder network:

$$L_{regular_{trans}} = \frac{1}{n} \sum_{i=1}^{n} \left\| \Theta_A^i \right\|_2. \tag{10}$$

where Θ_A denotes the parameters of the identity encoder network.

4.3. *The Overall Loss Function*

The final loss function is a weighted sum of all losses defined in Eqs. (1)–(10):

$$\begin{aligned} Loss_{recon} &= \lambda_{adv}(L_{D-adv_{recon}} + L_{G-adv_{recon}}) + \lambda_{cnt}(L_{cntS_{recon}} + L_{cntD_{recon}}) \\ &\quad + \lambda_{reg} L_{regular_{recon}}, \end{aligned} \tag{11}$$

$$\begin{aligned} Loss_{trans} &= \lambda_{adv}(L_{D-adv_{trans}} + L_{G-adv_{trans}}) + \lambda_{cnt}(L_{cnt_{trans}} + L_{cntS_{trans}}) \\ &\quad + \lambda_{reg} L_{regular_{trans}}, \end{aligned} \tag{12}$$

where λ_{adv}, λ_{cnt}, and λ_{reg} are weights that control the importance of loss terms. The overall loss will be:

$$Loss_{overall} = r Loss_{recon} + (1 - r) Loss_{trans} \tag{13}$$

since $r \in \{0, 1\}$ is a random binary value that is updated before each learning iteration.

5. Experiments

5.1. *Dataset*

The PFA-GAN is trained on a subset of the MS-Celeb-1M (Guo *et al.*, 2016) dataset, which contains about 5M images of 80K celebrities with unbalanced viewpoint distributions and with a very large appearance variation (e.g. due to gender, race, age, or even

makeups). We use 36K face images belonging to 528 different identities, while no pose or identity annotations are employed in the training process. For each face image, we first detect the facial region using the multi-task cascaded CNN detector (MTCNN) (Zhang *et al.*, 2016) and then align and resize the detected face to 128×128 pixels.

5.2. *Implementation Details*

We use the same implementations of the generator G and the discriminator D in IP-GAN that were introduced by Tian *et al.* (2018). For the pose encoder network E_P and the identity encoder network E_A, we use ResNet50 (He *et al.*, 2016) network architecture, where the skip connections in the network allow to learn the desired representation (e.g. the identity or the pose) since the performance of the upper layers will be at least as good as the lower layers. The following parameters were used: $\lambda_{cnt} = 1$, $\lambda_{adv} = 0.005$, and $\lambda_{reg} = 0.001$. The values of the random noise a^z are in the range $[-1, +1]$. To implement the model, a set of Pytorch deep learning tools was used. The batch size was set in 16 and one Nvidia graphic card (GTX 1080 Ti) was used. The Adam optimizer (Kingma and Ba, 2015) was used and configured with the learning rate of 0.0005, and the momentum of $[0, 0.9]$.

5.3. *Interpolation of Pose Latent Space*

In this section, we demonstrate that a pose of the generated face images can be gradually changed with the latent vector. We call this phenomenon face pose morphing. We have tested our model on the selected subset from the MS-Celeb-1M (Guo *et al.*, 2016) dataset. We first choose a pair of images x_S and x_D, and then extract the pose latent vectors p_S and p_D using the pose encoder network E_P. Then, we obtain a series of pose embedding vectors $\tilde{p}_i$ by linear interpolation, i.e.:

$$\tilde{p}_i = \alpha_i p_S + (1 - \alpha_i) p_D, \tag{14}$$

where $\alpha_i \in [0, 1]$, $i = 1, \ldots, k$; k is the number of interpolated images. Finally, we concatenate each interpolated pose vector $\tilde{p}_i$ with the extracted identity embedding vector a_S and feed the combined vector into the generator G to synthesize an interpolated face image $\tilde{x}_i = G(a_S, \tilde{p}_i)$. Figure 2 presents the results of the face pose morphing using $k = 10$, since every row shows how a face pose is gradually morphing into the next one. The last column denotes the landmark image of the driving face.

5.4. *Visual Data Augmentation Strategies*

The traditional visual data augmentation methods alter the entire face image by transferring image pixel values to new positions or by shifting pixel colours to new values. These generic methods ignore high-level content such as moving the head or adding a smile, so in this section, we show the effectiveness of using our model as an alternative face specific augmentation method. The ability of the PFA-GAN model to perform a controlled

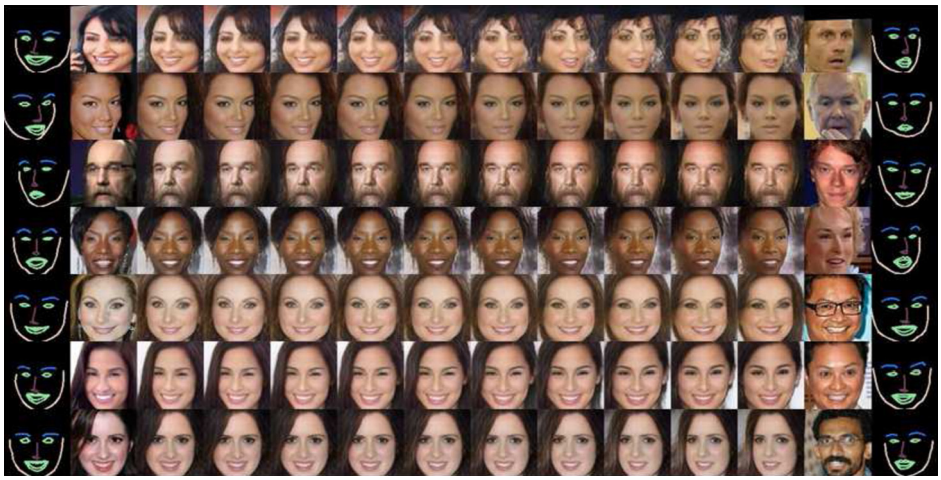

Fig. 2. Interpolation of the pose latent space.

synthesis of a face image allows enlarging the volume of data for training or testing by generating new face images with new poses. So, using our proposed model, for each image in the dataset, an unlimited number of images can be generated for the same identity subject with a great variety of face poses. Assuming the original dataset is R, pose face augmentation can be represented by the following transformation:

$$f : R \longrightarrow T, \tag{15}$$

where T is the augmented dataset of R. Then the dataset is expanded as a combination of the original dataset and the augmented one:

$$\acute{R} = R \cup T. \tag{16}$$

We introduce three visual data augmentation strategies, each one extends the original training dataset with a new augmented dataset whose images have a degree of difference in the pose face from the original ones.

- First augmentation strategy (Aug-S1). For each image in our dataset, we choose a random driving face image, and by following the interpolation technique described in Section 5.3, we choose the interpolated pose vector $\tilde{p}_3$ which has a slight difference from the original one as a driving pose. Then we feed $\tilde{p}_3$ along with a_S to synthesize an augmented face image $\tilde{x}_S = G(a_S, \tilde{p}_3)$ in the augmented dataset T_1. Therefore, the dataset of this augmentation strategy will be:

$$\acute{R}_1 = R \cup T_1. \tag{17}$$

- Second augmentation strategy (Aug-S2). Similar to the first augmentation strategy, we select the interpolated pose vector $\tilde{p}_6$, which differs more than the $\tilde{p}_3$ to synthesize

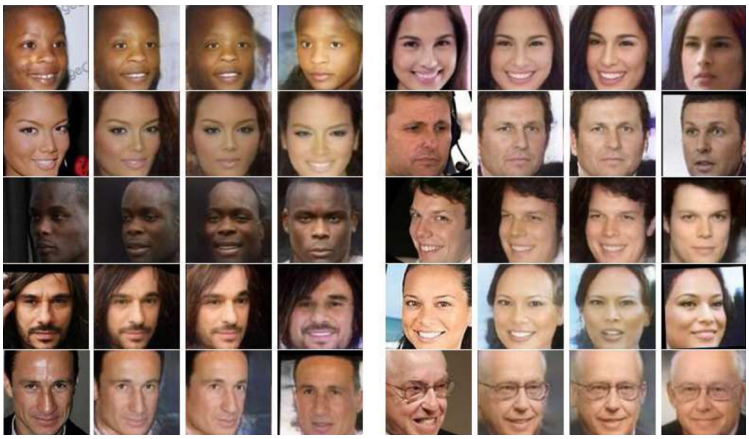

Fig. 3. Face image examples from the original and augmented datasets. From left to right: the original dataset R, the augmented datasets T_1, T_2, T_3 in the second, third and fourth columns, respectively.

an augmented face image $\tilde{x}_S = G(a_S, \tilde{p}_6)$. Consequently, the dataset of the second augmentation strategy will be:

$$\acute{R}_2 = R \cup T_2. \tag{18}$$

- Third augmentation strategy (Aug-S3). The generated images may have a large degree of variation from the original with regard to head pose and facial expressions. That's why for each source face image in our dataset, we randomly select a driving image to extract the pose embedding vector p_r and feed the combined vector $[a_S, p_r]$ to the generator. The dataset of this augmentation strategy will be:

$$\acute{R}_3 = R \cup T_3. \tag{19}$$

Figure 3 shows examples of face images from the augmented datasets. We can note the pose variation between them.

5.5. *Face Verification Task*

In this subsection, we evaluate whether the augmented datasets will improve the performance of the face verification task or not. In general, face verification needs the following steps: training a convolution neural network classifier on a dataset, then using it as a feature extraction network to extract the embedding vectors for a pair of face images from testing datasets. Next, the extracted two vectors are sent to the distance function to calculate the similarity between them, and according to the threshold, the function judges whether it is the face of the same person or not. Two classifiers are used, C_1 and C_2, the backbones Restnet50, Restnet101 are chosen for the C_1, C_2, respectively. Both classifiers use ArcFace (Deng *et al.*, 2019) and Focal loss function. We use different datasets for face

Table 1
Characteristics of the training and testing datasets.

Dataset	Number of people	Total images
R	529	36000
$\acute{R}_1$	529	72000
$\acute{R}_2$	529	72000
$\acute{R}_3$	529	72000
LFW	5749	13233
CFP-FP	500	2000
CFP-FF	500	5000
AgeDB	570	16488
CALFW	4025	12174
CPLFW	3884	11652
VGGFace2-FP	500	11000

Table 2
Verification accuracy after training the classifier C_1.

Classifier	Training dataset	LFW	CFP-FP	CFP-FF	AgeDB	CAL-FW	CPLFW-FP	VGG-Face2
C_1	R	89.77	78.39	88.73	69.23	72.32	70.15	80.74
C_1	$\acute{R}_1$	91.00	80.04	89.57	70.17	72.25	70.72	81.08
C_1	$\acute{R}_2$	90.88	80.34	**89.81**	70.65	71.80	**70.78**	**81.36**
C_1	$\acute{R}_3$	**91.53**	**81.23**	89.70	**71.20**	**72.43**	70.28	81.16

Table 3
Verification accuracy after training the classifier C_2.

Classifier	Training dataset	LFW	CFP-FP	CFP-FF	AgeDB	CAL-FW	CPLFW-FP	VGG-Face2
C_2	R	89.38	77.97	88.39	68.30	70.50	69.90	80.06
C_2	$\acute{R}_1$	90.90	80.13	89.67	70.38	72.03	70.90	80.44
C_2	$\acute{R}_2$	91.23	81.17	89.73	69.23	72.18	71.42	81.56
C_2	$\acute{R}_3$	**91.68**	**81.70**	**89.77**	**70.93**	**72.67**	**71.58**	**81.64**

verification, such as LFW (Huang *et al.*, 2007), CFP-FP (Sengupta *et al.*, 2016), AgeDB (Moschoglou *et al.*, 2017), CFP-FF (Sengupta *et al.*, 2016) and VGGFace2-FP (Cao *et al.*, 2018b). Apart from the most widely used LFW dataset, we also report the performance of our augmentation model on the recent large-pose and large-age datasets (e.g. CPLFW. Zheng and Deng, 2018 and CALFW, Zheng *et al.*, 2017). Table 1 shows the statistics of both training and testing datasets in a verification scenario.

We feed the augmented datasets to the classifiers (C_1, C_2) for training, then we use the learned model to extract embedding vectors for each image in the testing datasets. Therefore, the verification accuracies are calculated as shown in Table 2 and Table 3, and it is clear, that the verification accuracy is higher than it on the dataset without augmentation. For instance, the verification accuracy on AgeDB increased from using the CNN model (RestNet50) that trained on the augmented dataset (Aug-D3). As a comparison be-

tween the proposed augmentation strategies, the difference between augmented datasets (T_1, T_2, T_3) is that the pose face in each of them is transformed from the original dataset baseline with a different degree of change, from small, as in the case of the T_1, to randomness, as in the case of the T_3. Consequently, by comparing the tables' results, it can be noted that a small change in pose transformation improves the results of face verification, but when the change is random as in the augmentation strategy (Aug-D3), the results of verification using the more deeper feature extractor Resnet101 are the best. Since on the publicly available dataset CFP-FP, the increase in verification accuracy has been achieved up to 4.5%.

6. Conclusion

In this paper, we proposed a self-supervised framework PFA-GAN based on Generative Adversarial Networks to control the pose of a given face image using another face image, where the identity of the source image is preserved in the generated one. This framework makes no assumptions about the pose of the source images since the proposed training method allows us to train the overall networks in fully self-supervised settings using a large-scale unconstrained face images dataset. Finally, we use the trained model as a tool for visual data augmentation. Our PFA-GAN framework demonstrates the ability to synthesize photorealistic and identity-preserving faces with arbitrary poses, which improve face recognition tasks. The face verification experimental results demonstrate the effectiveness of the proposed framework for pose face augmentation as all augmented datasets outperform the baseline. Furthermore, to the best of our knowledge, we are the first to train such a model using a large-scale unconstrained dataset of face images. One exciting avenue for future work is to improve the network architecture by utilizing operations such as adaptive instance normalization (AdaIN) and to train our framework on other datasets larger than ours.

Funding

This work was financially supported by the Government of the Russian Federation (Grant 08-08).

References

Antoniou, A., Storkey, A., Edwards, H. (2018). *Data Augmentation Generative Adversarial Networks. Iclr.*

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, SIGGRAPH, 1999.*

Cao, J., Hu, Y., Yu, B., He, R., Sun, Z. (2018a). Load balanced GANs for multi-view face image synthesis. abs/1802.07447.

Cao, Q., Shen, L., Xie, W., Parkhi, O.M., Zisserman, A. (2018b). VGGFace2: a dataset for recognising faces across pose and age. In: *Proceedings – 13th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2018.*

Crispell, D., Biris, O., Crosswhite, N., Byrne, J., Mundy, J.L. (2017). Dataset augmentation for pose and lighting invariant face recognition. arXiv:1704.04326 [cs.CV].

Deng, J., Guo, J., Xue, N., Zafeiriou, S. (2019). ArcFace: additive angular margin loss for deep face recognition. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*.

Farahani, A., Mohseni, H. (2019). Multi-pose face recognition using pairwise supervised dictionary learning. *Informatica*, 30, 647–670.

Feng, Z.H., Kittler, J., Christmas, W., Huber, P., Wu, X.J. (2017). Dynamic attention-controlled cascaded shape regression exploiting training data augmentation and fuzzy-set sample weighting. In: *Proceedings – 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*.

Guo, Y., Zhang, L., Hu, Y., He, X., Gao, J. (2016). MS-celeb-1M: a dataset and benchmark for large-scale face recognition. In: *Lecture Notes in Computer Science* (including subseries *Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics*).

Guo, Y., Zhang, J., Cai, J., Jiang, B., Zheng, J. (2017). 3DFaceNet: real-time dense face reconstruction via synthesizing photo-realistic face images. arXiv:1708.00980.

He, K., Zhang, X., Ren, S., Sun, J. (2016). Deep residual learning for image recognition. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*.

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. 07-49, University of Massachusetts, Amherst.

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*.

Johnson, J., Alahi, A., Fei-Fei, L. (2016). Perceptual losses for real-time style transfer and super-resolution. In: *Lecture Notes in Computer Science* (including subseries *Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics*).

Kingma, D.P., Ba, J.L. (2015). Adam: a method for stochastic optimization. In: *3rd International Conference on Learning Representations, ICLR 2015 – Conference Track Proceedings*.

Moschoglou, S., Papaioannou, A., Sagonas, C., Deng, J., Kotsia, I., Zafeiriou, S. (2017). AgeDB: the first manually collected, in-the-wild age database. In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*.

Parkhi, O.M., Vedaldi, A., Zisserman, A. (2015). Deep face recognition. In: *British Machine Vision Conference*, Vol. 1, pp. 41.1–41.12.

Ribarić, S., Fratrić, I., Kiš, K. (2008). A novel biometric personal verification system based on the combination of palmprints and faces. *Informatica*, 19(1), 81–100.

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: *2016 IEEE Winter Conference on Applications of Computer Vision, WACV 2016*.

Simonyan, K., Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In: *3rd International Conference on Learning Representations, ICLR 2015 – Conference Track Proceedings*.

Tian, Y., Peng, X., Zhao, L., Zhang, S., Metaxas, D.N. (2018). CR-GAN: learning complete representations for multi-view generation. In: *IJCAI International Joint Conference on Artificial Intelligence*.

Tran, L., Yin, X., Liu, X. (2017). Disentangled representation learning GAN for pose-invariant face recognition. In: *Proceedings – 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*.

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*.

Zeno, B., Kalinovskiy, I., Matveev, Y. (2019a). IP-GAN: learning identity and pose disentanglement in generative adversarial networks. In: *Lecture Notes in Computer Science* (including subseries *Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics*).

Zeno, B.H., Kalinovskiy, I.A., Matveev, Y.N. (2019b). Identity preserving face synthesis using generative adversarial networks. In: *ACM International Conference Proceeding Series*.

Zhang, K., Zhang, Z., Li, Z., Qiao, Y. (2016). Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10), 1499–1503.

Zheng, T., Deng, W. (2018). Cross-pose LFW: a database for studying cross-pose face recognition in unconstrained environments.

Zheng, T., Deng, W., Hu, J. (2017). Cross-age LFW: a database for studying cross-age face recognition in unconstrained environments. ArXiv: abs/1708.08197.

Zhu, X., Lei, Z., Liu, X., Shi, H., Li, S.Z. (2016). Face alignment across large poses: a 3D solution. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*.

B. Zeno is currently a PhD student at the St. Petersburg National Research University of Information Technologies, Mechanics and Optics, Russia (ITMO University). He received his MSc degree in computer science, in 2016, from Belgorod State Technological University, Belgorod, Russia. He received the BS degree in artificial intelligence, in 2009, from Damascus University, Damascus, Syria. In 2013 he started working as an assistant professor at the Department of Artificial Intelligence, Damascus University. Since 2018 he has been working in the International Research Laboratory "Multimodal Biometric and Speech Systems", ITMO University. His research interests include pattern recognition, machine learning, and computer vision.

I. Kalinovskiy received his PhD degree in computer science in 2016 from Computer Engineering Department of Tomsk Polytechnic University, Russia. From 2017 he has been a researcher at the R&D Department of STC-innovations Ltd. He worked on problems of face detection and recognition, liveness detection, 3D-reconstruction. Currently, he is the lead of speech synthesis research group. He is the author of more than 25 articles. His research interests include genetic and evolutionary algorithms, object detection and tracking, face recognition, speech synthesis techniques using deep neural networks.

Y. Matveev received his PhD (1985) and DSc (1995) degrees in computer science from the St. Petersburg National Research University of Information Technologies, Mechanics and Optics, Russia (ITMO University). He is currently a Chief Research Officer at STC-innovations Ltd. (STC Group, Russia), a full professor and the head of the International Research Laboratory for Multimodal Biometric and Speech Systems at ITMO University. He is the author of two books, more than 150 articles and 20 national patents. His research interests include speaker recognition, face recognition, multimodal biometrics. Prof. Matveev is a member of IEEE and ISCA.