International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:13, No:10, 2019

Adversarial Disentanglement Using Latent Classifier for Pose-Independent Representation

Hamed Alqahtani, Manolya Kavakli-Thorne

Abstract—The large pose discrepancy is one of the critical challenges in face recognition during video surveillance. Due to the entanglement of pose attributes with identity information, the conventional approaches for pose-independent representation lack in providing quality results in recognizing largely posed faces. In this paper, we propose a practical approach to disentangle the pose attribute from the identity information followed by synthesis of a face using a classifier network in latent space. The proposed approach employs a modified generative adversarial network framework consisting of an encoder-decoder structure embedded with a classifier in manifold space for carrying out factorization on the latent encoding. It can be further generalized to other face and non-face attributes for real-life video frames containing faces with significant attribute variations. Experimental results and comparison with state of the art in the field prove that the learned representation of the proposed approach synthesizes more compelling perceptual images through a combination of adversarial and classification losses.

Keywords—Video surveillance, disentanglement, face detection.

I. Introduction

HUMAN identity recognition is one of the most systematically researched direction in video surveillance for providing adequate security solutions. Face recognition plays a pivotal role in identifying individuals' identity. However, different face and non-face attributes pose several challenges in the face recognition process. Because, the accuracy of face recognition is affected by changes in these attributes like pose, hairstyle etc. Recently, most researchers focused on disentanglement representation. Disentanglement provides a way of separating different types of information that enable learning from distinctive features independently. With the use of disentangled representation, it is possible to separate the face identity and rest of face and non-face attributes. Several prominent researches improved face recognition task to a significant accuracy level in the last decade [1], [2], [3], [4].

Schroff et al. [5] showed human-level performance in their research. However, it has been realized that Pose-Invariant Face Recognition (PIFR) is still a challenging process. Gupta et al. [6] proved that the performance of frontal-frontal verification differs by a significant amount from frontal-profile face identification, whereas human performance differs by a small amount. The findings of their research motivated many researchers to bridge the gaps and even surpass human performance for PIFR methods.

Hamed Alqahtani is PhD student, Macquarie University, Australia (e-mail: hsqahtani@kku.edu.au).
Manolya Kavakli-Thorne is Associate Professor, Macquarie University, Australia.

The face recognition task involves learning a latent space to synthesize a frontal [7], [8], [9], [10], [11], [12] from an input profile followed by applying existing deep-learning methods to recognize individuals' identity from synthesized frontal. The task can also be done by learning of a group of models from different face attributes separately [13], [5] or by multiple single standing models for each pose[14], [15]. In this work, we propose a combined generative adversarial network framework based on PIFR and disentanglement methods to improve face recognition by learning a pose-independent representation.

Fig. 1 presents the overview of working on the proposed framework for face identification task. The proposed framework involves flipping the pose switch for any profile image that will generate a frontal with the same identity using the generative adversarial network (GAN). GAN [16] is one of the popular approaches for generating images from the learned data distribution through a min-max loss function. In this work, we propose to control the learned data distribution for face synthesis to get real-looking, high quality, and identity preserving face images. To achieve this task, we present a modified conventional Generator into an encoder-decoder architecture that converts a face image with any pose by passing it through the encoder and resulting in an encoded representation. The decoder synthesizes face from the resulting encoded representation. The learning of encoded representation, called latent space, is controlled by a classifier forcing the identity disentanglement from the pose attribute. The discriminator is also modified to perform identity and pose classification apart from distinguishing fake and real images resulting in identity preserving face images.

Original GAN consists of the learning of the image space by generator for synthesising fake images through a noise vector. In this work, we synthesize fake images using a decoder which takes a concatenation of encoded representation, pose, and noise. The encoder learns the mapping from the input image space known as pixel space to latent space called manifold. The proposed framework takes advantage of the input data distribution to find the correct manifold and noise that acts as traversal in the manifold. The pose is concatenated to control target pose at the output. It results in the generation of frontal as well as profile face images using the proposed framework.

The key contributions of this work involve 1) The proposal of a modified GAN architecture for providing a way to control learning the latent space and way to traverse it; 2) The plan of a classifier network that explicitly enforces the disentanglement in latent space; 3) Disentanglement of identity and pose for face image datasets to generalize to other face attributes;

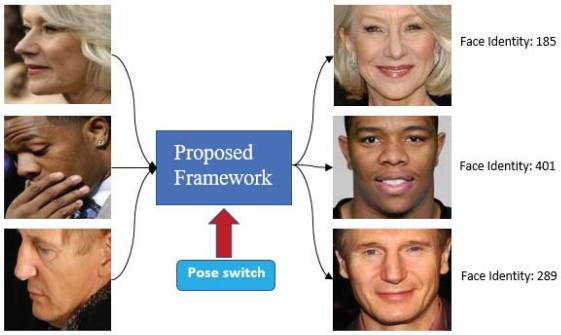

Fig. 1 For face image with any pose, the proposed framework synthesizes frontal which is photo-realistic high quality as well as identity preserving front face images which greatly improves PIFR capability

4) Evaluating the proposed framework and comparing with state-of-the-art in the field of face recognition using CFP dataset [6].

Rest of the paper is as follows. Section II presents state of the art in the field. Section III provides formal definition of the problem. It is followed by a description of the proposed framework by explaining proposed architecture, its comparison with traditional GAN architectures and research methodology in Section IV. Section V presents a setup for conducting a comprehensive set of experiments and reported the results in this work. Finally, Section VI concludes at the end of the paper.

II. Related Work

This section provides a comprehensive review of existing prominent researches for different issues of accurate face recognition.

Frontal Generation: The largely posed faces cause self-occlusion leading to increase in difficulty of generating the frontal view. The traditional computer vision methods use 2D texture warping or 3D based modelling [7], [17], [11], statistical methods[9] for frontal synthesis. Whereas, deep-learning based methods [8], [18], [10] use averaged 3D face for frontal synthesis. Kan et al. [8] employed auto-encoders to rotate the face to the front progressively while Yang et al. [18] used the recurrent connection to rotate face at fixed steps. Deep learning produces an identity-preserving synthesis and employs intermediate features for face recognition. The landmark localization method is proposed in [9] with a constrained low-rank minimization model.

Generative Adversarial Network (GAN): Generative models were first introduced by Goodfellow et al. [16]. It is based on a min-max two-player game that provides a powerful way to estimate the target distribution both two players, generator and discriminator learn simultaneously. GAN has been widely used in deep learning and computer vision field. Most researchers used GAN for generating photo-realistic image samples from the learned data distribution. Recently, several researchers modified GAN models [19], [20] for providing methods to synthesize image. The authors of [9], [21] proposed discriminator to act as a classifier, whereas Liu et al. [3] added a latent code for regularizing the output. These advancements of GANs motivate us to synthesize face image based on GAN. If x is the input image and z is the noise, then GAN optimizes the following two objective functions to create $\hat{x}$ similar to x.

$$\max_D V_D(D,G) = E_{\mathbf{x}\sim p_d(\mathbf{x})}[\log D(\mathbf{x})] + E_{\mathbf{z}\sim p_z(\mathbf{z})}[\log(1-D(G(\mathbf{z})))] \quad (1)$$

$$\max_G V_G(D,G) = E_{\mathbf{z}\sim p_z(\mathbf{z})}[\log(D(G(\mathbf{z}))] \quad (2)$$

We elaborate on different applications used in our work in the proceeding sections.

Latent Representation: Learning a suitable representation needs an appropriate objective function [22]. Initially, Huang et al. [23] proposed encoder-decoder representation. In this work, we proposed a framework on the basis of DRGAN [24] that employs an auto-encoder using disentangled representation learning. Kulkarni et al. [25] suggested DC-IGN as a way to use the architecture, but no explicit disentanglement was mentioned. These applications helped in building the prototype for our overall architecture.

Attribute Factorization: The latent space is supposed to represent input information as a whole. We control the different area to learn the pose-invariant representation. Huang et al. [26] presented a way to use a classifier to force disentanglement in the latent space.

Our work differs from prior work in some aspects. Firstly, we embed a classifier in the manifold space to separate identity and the pose information. The proposed framework involves the use of a classifier loss to train the encoder to produce required disentangled encoding and adversarial loss to generate high-quality perceptual images.

III. Problem Formulation

For a labeled dataset containing a face image x as the input image with label y has the pose information and z is the noise. The objective function comprises of two interlinked steps. The first step involves learning of a standalone identity representation disentangled from pose information. The second step synthesizes face from a learned representation with controlled target pose. In order to achieve the first objective, the classifier is trained according to the following objective function:

$$\min_C \max_{G_{enc}} L_{bce}\left(C(f(x)), y\right) = \min_C \max_{G_{enc}} L_{bce}(\tilde{y}, y) \quad (3)$$

Here, y is the real pose and $\tilde{y}$ is the classifier prediction. Classifier C is trained in a way to minimize the difference between y and $\tilde{y}$, whereas G_{enc} will try to produce such $f(x)$ which confuse the classifier and maximize the dissimilarity.

In contrast to the conventional discriminator, we employ a multi-task CNN as suggested by DRGAN [24] to perform the classification of identity and pose. Identity classification alleviates the problem of identity preservation.

When discriminator has real face image *x*, the network is trained using real identity and pose *y*; while synthetic face image from the decoder $\hat{x}$ = *G(x,ỹ,z)* is given, discriminator train itself to classify the image as fake. $\tilde{y}$ is the desired pose irrespective of the original pose *y*. The objective function *V* is defined as per following equation.

$$\max_D V_D(D,G) = E_{\mathbf{x},\mathbf{y}\sim p_d(\mathbf{x},\mathbf{y})}\left[\log D_{identity}(\mathbf{x}) + \log D_{pose}(\mathbf{x})\right] + E_{\mathbf{z}\sim p_z(\mathbf{z}),\mathbf{y}'\sim p_{y'}(\mathbf{y}')}\left[\log\left(D_{identity}(G(\mathbf{x},\mathbf{y}',\mathbf{z}))\right)\right] \quad (4)$$

Similarly, G_{enc} learns the identity representation *f(x)* and G_{dec} trained to synthesize a face image $\hat{x}$ = G_{dec}(*f(x)*, y', z).

The main objective of *G* is to fool *D* maximally in classifying synthetic images as real. Its objective function is defined as below.

$\max_G V_G(D,G) = E_{\mathbf{x},\mathbf{y}\sim p_d(\mathbf{x},\mathbf{y}),}\left[\log\left(D_{identity}(G(\mathbf{x},\mathbf{y}',\mathbf{z}))\right) + \log\left(D_{pose}(G(\mathbf{x},\mathbf{y}',\mathbf{z}))\right)\right]$

Both *G* and *D* are trained simultaneously to optimize each other. Generator strives to produce more realistic face images, and Discriminator tries to classify them as fake and provide gradient to the generator for creating more realistic images. The latent representation *f(x)* is not only useful for creating realistic face images but also provides a unique representation for a different identity.

IV. The Proposed Framework

This section provides the proposed framework, a comparison with traditional GANs, network architecture, loss functions and description of controlling disentangled attribute.

A. The Proposed Architecture

In this work, we aim to synthesize a controlled pose face image from a given face image with any pose. Fig. 2 presents the working methodology of the proposed framework in a broader sense. G_{enc} encodes the normalized input image x into a latent encoding *f(x)*. The input images lie in a pixel space of width × height × channel dimension. Encoder narrows down the pixel space into the manifold by learning the mapping of an input image distribution to latent space. The classifier ensures that *f(x)* does not contain any information regarding pose. *f(x)* is then concatenated with desired pose vector and noise vector. Noise is added to provide variation over the manifold. Different poses may have different manifolds. Thus, adding the pose vector helps to find the desired manifold and synthesize required pose image. The concatenated vector is then passed through the decoder to synthesize a face image. The target image should be photo-realistic as well as preserve the same identity of the input image. To handle this issue, we performed identity and pose classification using the discriminator in addition to the adversarial loss for high perceptual quality. The network parameters are trained by minimizing the combined loss function $L_{overall}$ explained in section III.

$$L_{overall} = \lambda_{pixel}L_{pixel} + \lambda_{adv}L_{adv} + \lambda_{cls}L_{cls} \quad (5)$$

B. Comparison with Traditional GANs Architectures

We compare the proposed framework against traditional and, most commonly used architectures as shown in Fig. 3.

- **Conditional GAN:** In case of conditional GAN [27], [28], Discriminator is also fed with class labels to synthesize images conditioned on class labels. The presence of labels to both the discriminator and the generator makes the network to learn constrained representation, which helps in finding the correct manifold. Conditional GAN classifies a real image without following the constraints to be fake. This imposes robust learning on the generator to meet the conditions. In this work, we modify discriminator to classify a real image to its true identity and pose.
- **Classification GAN:** GAN can also be used for classification rather than just generating fake images. The discriminator is trained to work as a discriminative classifier [29]. It not only distinguishes the synthetic images but also can classify the image into one of its classes. In this work, we train the discriminator to do both tasks. Firstly, it checks for the authenticity of the image, followed by the classification of the image based on its identity and its pose.
- **Adversarial Autoencoder (AAE):** The central principle of the autoencoder is to learn a mapping that compresses the input image to smaller dimensions, capturing all the necessary information to retrieve the original image back. In AAE, there is no constraint on latent data distribution, and AAE tends to find the arbitrary distribution that matches input distribution. In our framework, we proposed to control the latent space learning by a classifier network which ensures a manifold disentangled from the pose attribute. Also, there is no provision of classification in AAE while we modify discriminator for classification.

C. The Proposed Methodology

This section describes the methodology adopted in this work to conduct experiments comprehensively. It includes details of network architecture, combined loss function and controlling disentangled attributes.

1) Network Architecture: Tables I-III show the network structure for G_{enc}, G_{dec} and Classifier C respectively. Encoder and discriminator are designed based on CASIA-Net [30] with batch normalization. Since training a GAN is a highly unstable min-max game. So, we replaced sparse gradient layers of max-pool and Relu with strides convolution and ELU activation. In discriminator, a fully connected layer is added with soft-max for identity and pose classification in addition to differentiating real and fake images. The generator consists of encoder and decoder with a classifier network working on the basis of learn-able latent encoding. The latent encoding is a representation of identity information trained in a way to fool the classifier network. The classifier network is trained using latent encoding and real pose. The disentangled encoding is then concatenated with pose vector to control the target pose and with noise to provide

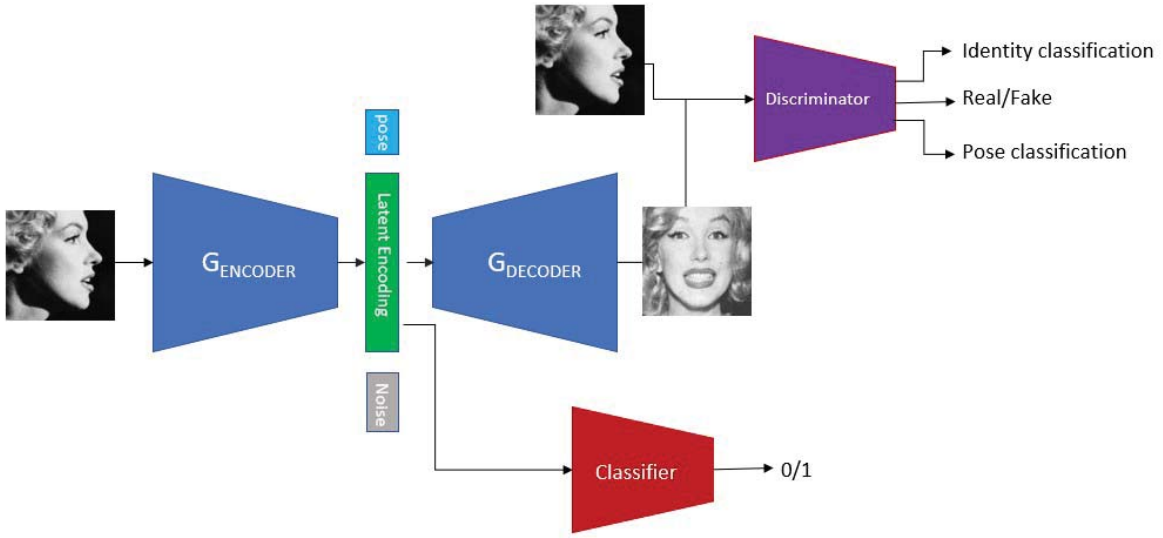

Fig. 2 Overall architecture the proposed framework. It fulfills two purposed:1) learning a standalone identity representation. 2) synthesis of frontal from the disentangled representation

TABLE I
NETWORK STRUCTURE 1

G_{enc} and D		
Layer Name	Filter Details	Output
Convolution(11)	3 x 3/1	96 x 96 x 32
Convolution(12)	3 x 3/1	96 x 96 x 64
Convolution(21)	3 x 3/2	48 x 48 x 64
Convolution(22)	3 x 3/1	48 x 48 x 64
Convolution(23)	3 x 3/1	48 x 48 x 128
Convolution(31)	3 x 3/2	24 x 24 x 128
Convolution(32)	3 x 3/1	24 x 24 x 96
Convolution(33)	3 x 3/1	24 x 24 x 192
Convolution(41)	3 x 3/2	12 x 12 x 192
Convolution(42)	3 x 3/1	12 x 12 x 128
Convolution(43)	3 x 3/1	12 x 12 x 256
Convolution(51)	3 x 3/2	6 x 6 x 256
Convolution(52)	3 x 3/1	6 x 6 x 160
Convolution(53)	3 x 3/1	6 x 6 x (320+1)
Average Pooling	6 x 6/1	1 x 1 x (320+1)
Fully Connected(D only)		(450+2+1)

important variation in the face. The decoder part is made up of combinations of fractionally-strided convolutions which transforms the concatenated encoded vector into an image $\hat{x}$, which lies in the same pixel-space of x.

2) Combined Loss Function: In order to train different parts of the architecture, we proposed to use a different combination of loss and functions carefully.

- **Pixel-level Loss:** In this work, we employ L2 loss between the ground truth face image x and synthetic image $\hat{x}$ to learn the mapping between the two.

$$L_{pixel} = L_{L2}(x, \hat{x}) = \|x - \hat{x}\|^2 \tag{6}$$

Although L2 loss cannot capture high-frequency details resulting in blurry images, it is an essential part because it accelerates the optimization and gives overall improvement.

- **Adversarial Loss:** This loss is provided by the discriminator, and it distinguishes real frontal images from the synthetic images. For input image x and synthesized image as $\hat{x}$, it is calculated as follows:

$$L_{adv} = \frac{1}{N}\sum_{n=1}^{N} -\log D(G(x)) \tag{7}$$

This helps to preserve the perceptual quality by guiding the synthesis to the frontal image manifold.

- **Classifier Loss:** The classifier is a network trained to get confused and cannot predict the correct pose from the latent encoding *f(x)*. Then the loss is defined as per following equation.

$$L_{cls} = L_{crossentropy}\left(C\left(f(x)\right), y\right) = L_{crossentropy}(\tilde{y}, y) \tag{8}$$

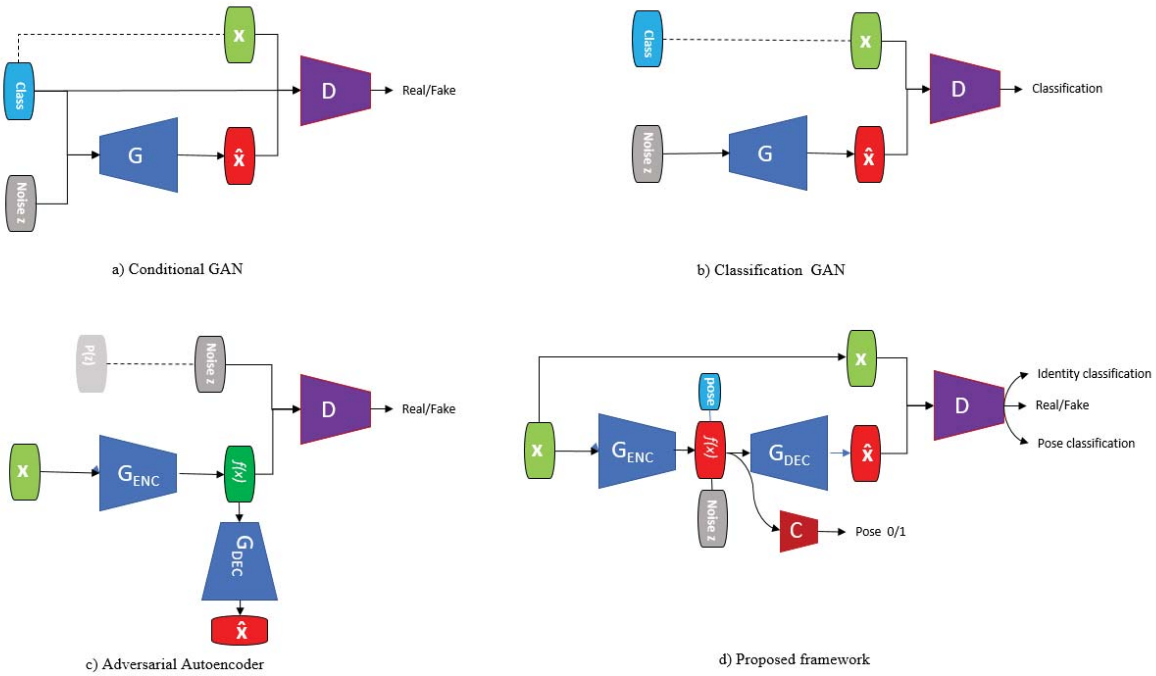

Fig. 3 Comparison of the proposed architecture with traditional GAN-based architectures

TABLE II
NETWORK STRUCTURE 2

G_{dec}		
Layer Name	Filter Details	Output
FullyConvolution		6 x 6 x 320
FullyConvolution52	3 x 3/1	6 x 6 x160
FullyConvolution51	3 x 3/1	6 x 6 x 256
FullyConvolution43	3 x 3/2	12 x 12 x 256
FullyConvolution42	3 x 3/1	12 x 12 x 128
FullyConvolution41	3 x 3/1	12 x 12 x 192
FullyConvolution33	3 x 3/2	24 x 24 x 192
FullyConvolution32	3 x 3/1	24 x 24 x 96
FullyConvolution31	3 x 3/1	24 x 24 x 128
FullyConvolution23	3 x 3/2	48 x 48 x 128
FullyConvolution22	3 x 3/1	48 x 48 x 64
FullyConvolution21	3 x 3/1	48 x 48 x 64
FullyConvolution13	3 x 3/2	96 x 96 x 64
FullyConvolution12	3 x 3/1	96 x 96 x 32
FullyConvolution11	3 x 3/1	96 x 96 x 1

TABLE III
NETWORK STRUCTURE 3

Classifier		
Layer	Input size	Output Size
Linear1	320	1000
Linear2	1000	1000
Linear2	1000	2
Softmax		

The encoder is also trained simultaneously to produce such representation which can fool the classifier network.

D. Controlling Disentangled Attribute

In order to obtain the desired pose face as an output, we concatenate a pose vector with latent encoding having identity information. After concatenating with a noise vector, it is passed to the decoder. We use pose = 1 for frontal and pose = 0 for the profile image. Thus, simply setting the pose bit to 1 in the representation space leads to generating a frontal face.

V. EXPERIMENTS AND RESULTS

This section presents an experimental setup, results of face synthesis and face recognition using the proposed framework. The proposed generative framework involves two phases. 1) The first phase presents a way to learn disentangled latent space. 2) The second phase involves the synthesis of an image-realistic pose-controlled face image while preserving the identity information of the input image. Section IV-B and IV-C present the experimental results qualitatively on face synthesis and quantitative on face recognition and section IV-D presents a visualization of the manifold to illustrate the disentanglement.

A. Experimental Setup

- **Dataset:** CFP dataset [6] is one of the most widely used face datasets containing frontal and profile images of celebrities. It consists of 500 images of the individuals, each having four different pose images and ten frontal face images. The dataset is well organized leading to a faster data pipeline. As per the evaluation metric, Face verification for both frontal-frontal (FF) pair and frontal-profile (FP) pair is carried out on ten folds with 350 similar identity pair and 350 different identity pair. The generator is provided with enough face frontal images to learn the frontal pose distribution well.

- **Implementation:** As per [30], all the faces are aligned with a 100 $\times$ 100 view and a random 96 $\times$ 96 crop is sampled for data augmentation. All images are normalized to a range of [-1,1]. The CFP dataset is organized in such a way so as to provide folder names for labels for corresponding images. The framework is implemented in PyTorch. After creating the data pipeline, we set the hyper-parameters. The batch size is taken as 64, considering enough group information and GPU capabilities. All weights are initialized from a zero-centred normal Gaussian distribution. The standard deviation is 0.02. We use Adam optimized with an initial learning rate of 0.002 and momentum set as 0.5.
  The training process of the proposed framework consists of four stages involving changes in the training periodicity of discriminator relative to generator training epochs. We start with 20 steps of Generator for one step of Determinator. After achieving the stability in the training process (after 800 epochs), we switch the periodicity to 10 that makes the discriminator learn faster. But, it may affect generator training, so we increase the periodicity to 20 in steps of 5 in each stage of training. The training seems to converge after around 4600 epochs.

B. Results

1) Face Synthesis:

- **Adversarial Loss:** Existing face rotation methods [18], [10], [31] use L2 loss for training the input-output mapping. The mapping can be frontal-frontal or profile-frontal. L2 loss produces blurry images [24] which are not perceptually good, and post-processing algorithms fail to work on the output image. This low-frequency synthesis results in a lower face recognition performance. Fig. 4 shows the adversarial generated images which contain the high-frequency details.
- **Extreme Profiles:** Most of the prior methods work well for limited pose-variance only and fail to recover mostly posed faces. Our framework learns the identity representation for extensive training data and can help to improve the frontal face from the ill-posed face. The ability to synthesize the frontal face from the image-realistic image while preserving its identity allows handling mostly posed faces. We take advantage of the known pose for training the generator, latent classifier and discriminator for classification. Not only pose, but other attributes such as ear, forehead and cheeks are preserved due to the consistency in identity. Apart from that, it also retains non-facial characteristics like spectacles, hair colour, hairstyle, etc. from the input profile image.
  In the proposed work, varying the noise or pose does not affect the identity of the face. It also enables the retention of the structure of the face, resulting in better PIFR performance. Traversing the learned manifold by varying noise generates smooth changes for the same identity , confirming that the model has learned the identity representation.
- **Face Rotation:** The proposed framework not only does the frontalization but is also trained to synthesize any output pose. Fig. 4 shows face frontalization on CFP dataset. Given extreme profiles in the top row, the synthesis is very close to the real frontal face. Extreme profiles are largely affected by pose and contain the minimum information regarding identity. Frontal synthesis is shown, and any artefacts are due to image cropping in the pre-processing. G_{dec} rotates the face at each of its layers without affecting identity information. The experimental results prove that this synthesis is superior qualitatively due to adversarial loss and improved PIFR performance due to disentangled representation and identity preservation.

TABLE IV
EXPERIMENTAL RESULTS ON CFP DATASET

Method	Frontal-Frontal	Frontal-Profile
DR-GAN	97.13 $\pm$ 0.68	90.82 $\pm$ 0.28
Current Work	98.23 $\pm$ 0.83	91.92 $\pm$ 0.59

2) Attribute Factorization: The proposed framework disentangles the information as it is trained to produce fake images to fool the discriminator. Learning a controlled underlying representation is necessary for PIFR. When the latent encoding does not have any information about pose the classifier network. It fails to classify correctly. Thus, we enforce the encoder not to put any information regarding the pose in encoding to confuse the latent classifier maximally. The framework produces identity preserving faces with high-quality visual details. We evaluate the model for face recognition performance. Unlike DR-GAN, we train the proposed framework for a single image at a time. The comparison with state-of-the-art is presented in Table IV. The experimental results prove the improved and quality results for largely posed faces using the proposed framework in comparison to the other methods.

Table IV shows the comparison of CFP dataset for input-output pair of frontal-frontal and profile-frontal. Average face verification accuracy is reported with a standard deviation over ten folds. We achieve similar results for frontal-frontal verification, and we minimize the gap further between human performance and deep learning methods.

3) Feature Visualization: To see the mapping of input to latent encoding, we employ the t-SNE framework to visualize the latent space. We project the 320-dimensional latent encoding vector onto a 2D plane and plot it. Fig. 5 shows how similar faces are grouped together based on identity and pose. Also, we observe a very narrow gap between some clusters and those represent the high similarity between two identities. We analyze the plot at multiple epochs and observe the encoder that tries to group the whole space into two clusters for profile and frontal and then cluster within that base of identity. When a face image is provided, it is encoded into a pose-invariant representation. The identity information decides the identity cluster and poses vector determine the frontal or profile face to generate a sample for providing a realistic face image with the desired pose and identity.

Fig. 4 Face frontalization on CFP dataset is shown. The top row is the input to the model. Next row is frontalized faces by the proposed framework. The bottom row represents the ground truth frontal face. The artefacts are due to pre-processing of the input. The face rotations are shown for challenging extreme profiles and in-the-wild face data

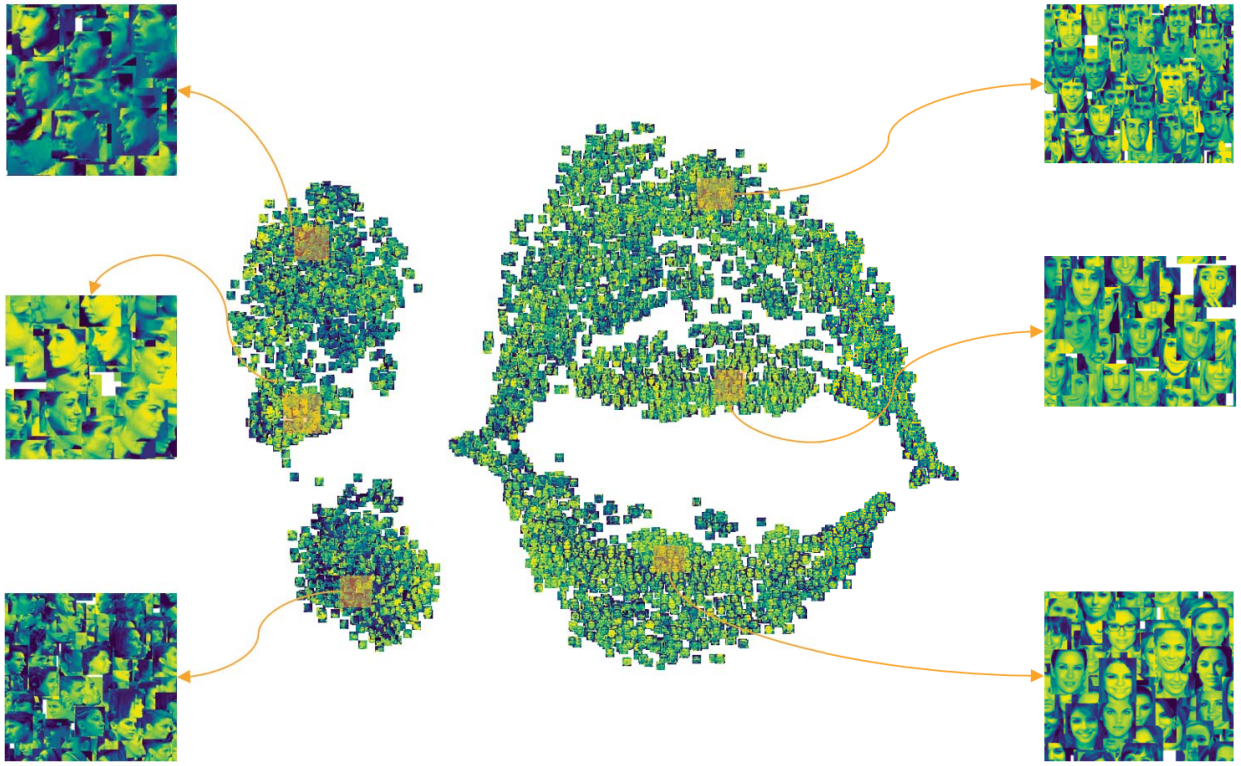

Fig. 5 Two-dimensional projection of Latent space. Different colour represents a different identity. Profile faces are grouped differently than frontal faces. Face image for one identity is shown

VI. Conclusion

This paper presents a generative framework for PIFR, Face rotation and Face synthesis. The modified GAN architecture learns to disentangle identity representation from other face attributes. This approach can be generalized to non-face attributes also. Attribute factorization using a latent classifier to help force disentanglement has been explored. It has been found that the disentangled representation preserves the identity information even for extreme profiles, and the adversarial network synthesizes perceptually clear face images. Different losses were used in the framework, and their impact was observed during training. A superior frontal synthesis capability and better face recognition performance have been found on the CFP dataset with the use of a latent classifier. The proposed framework can be referenced for further research in face attribute disentanglement. For profile face synthesis, the visual images are not perceptually appealing, but still outperforms in face recognition. In the future, we will focus on the areas to improve perceptual quality.

References

[1] X. Chai, S. Shan, X. Chen, and W. Gao, "Locally linear regression for pose-invariant face recognition," *IEEE Transactions on image processing*, vol. 16, no. 7, pp. 1716–1725, 2007.

[2] C. Ding and D. Tao, "A comprehensive survey on pose-invariant face recognition," *ACM Transactions on intelligent systems and technology (TIST)*, vol. 7, no. 3, p. 37, 2016.
[3] X. Liu and T. Chen, "Pose-robust face recognition using geometry assisted probabilistic modeling," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1. IEEE, 2005, pp. 502–509.
[4] X. Liu, J. Rittscher, and T. Chen, "Optimal pose for face recognition," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, vol. 2. IEEE, 2006, pp. 1439–1446.
[5] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 815–823.
[6] S. Sengupta, J.-C. Chen, C. Castillo, V. M. Patel, R. Chellappa, and D. W. Jacobs, "Frontal to profile face verification in the wild," in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2016, pp. 1–9.
[7] T. Hassner, S. Harel, E. Paz, and R. Enbar, "Effective face frontalization in unconstrained images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 4295–4304.
[8] M. Kan, S. Shan, H. Chang, and X. Chen, "Stacked progressive auto-encoders (spae) for face recognition across poses," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1883–1890.
[9] C. Sagonas, Y. Panagakis, S. Zafeiriou, and M. Pantic, "Robust statistical face frontalization," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 3871–3879.
[10] J. Yim, H. Jung, B. Yoo, C. Choi, D. Park, and J. Kim, "Rotating your face using multi-task deep neural network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 676–684.
[11] X. Zhu, Z. Lei, J. Yan, D. Yi, and S. Z. Li, "High-fidelity pose and expression normalization for face recognition in the wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 787–796.
[12] Z. Zhu, P. Luo, X. Wang, and X. Tang, "Multi-view perceptron: a deep model for learning face identity and view representations," in *Advances in Neural Information Processing Systems*, 2014, pp. 217–225.
[13] O. M. Parkhi, A. Vedaldi, A. Zisserman *et al.*, "Deep face recognition." in *bmvc*, vol. 1, no. 3, 2015, p. 6.
[14] C. Ding and D. Tao, "Robust face recognition via multimodal deep face representation," *IEEE Transactions on Multimedia*, vol. 17, no. 11, pp. 2049–2058, 2015.
[15] I. Masi, S. Rawls, G. Medioni, and P. Natarajan, "Pose-aware face recognition in the wild," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4838–4846.
[16] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.
[17] S. Li, X. Liu, X. Chai, H. Zhang, S. Lao, and S. Shan, "Morphable displacement field based image matching for face recognition across pose," in *European conference on computer vision*. Springer, 2012, pp. 102–115.
[18] J. Yang, S. E. Reed, M.-H. Yang, and H. Lee, "Weakly-supervised disentangling with recurrent transformations for 3d view synthesis," in *Advances in Neural Information Processing Systems*, 2015, pp. 1099–1107.
[19] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4681–4690.
[20] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015.
[21] S. Sankaranarayanan, A. Alavi, C. D. Castillo, and R. Chellappa, "Triplet probabilistic embedding for face verification and clustering," in *2016 IEEE 8th international conference on biometrics theory, applications and systems (BTAS)*. IEEE, 2016, pp. 1–8.
[22] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.
[23] F. J. Huang, Y.-L. Boureau, Y. LeCun *et al.*, "Unsupervised learning of invariant feature hierarchies with applications to object recognition," in *2007 IEEE conference on computer vision and pattern recognition*. IEEE, 2007, pp. 1–8.
[24] L. Tran, X. Yin, and X. Liu, "Disentangled representation learning gan for pose-invariant face recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1415–1424.
[25] T. D. Kulkarni, W. F. Whitney, P. Kohli, and J. Tenenbaum, "Deep convolutional inverse graphics network," in *Advances in neural information processing systems*, 2015, pp. 2539–2547.
[26] R. Huang, S. Zhang, T. Li, and R. He, "Beyond face rotation: Global and local perception gan for photorealistic and identity preserving frontal view synthesis," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2439–2448.
[27] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.
[28] H. Kwak and B.-T. Zhang, "Ways of conditioning generative adversarial networks," *arXiv preprint arXiv:1611.01455*, 2016.
[29] J. T. Springenberg, "Unsupervised and semi-supervised learning with categorical generative adversarial networks," *arXiv preprint arXiv:1511.06390*, 2015.
[30] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Learning face representation from scratch," *arXiv preprint arXiv:1411.7923*, 2014.
[31] Z. Zhu, P. Luo, X. Wang, and X. Tang, "Deep learning identity-preserving face space," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 113–120.