IE-WGAN: A model with Identity Extracting for Face Frontal Synthesis

Yanrui Zhu¹, Yonghong Chen² and Yuxin Ren³*

¹ Group of Inspur, Jinan, Shandong Province, 250000, China
²Public Passenger Transport Management Service Centre, Jinan, Shandong Province, 250000, China
³School of Control Science and Engineering, Shandong University, Jinan, Shandong Province, 250000, China
*Corresponding author’s e-mail: renyuxin@sdu.edu.cn

Abstract: Face pose affects the effect of the frontal face synthesis, we propose a model used for frontal face synthesis based on WGAN-GP. This model includes identity extracting module, which is used to supervise the training of the face generation module. On the one hand, the model improves the quality of synthetic face images, on the other hand, it can accelerate the convergence speed of network training. We conduct verification experiments on CelebA data sets, and the results show that this model improves the graphic quality of frontal synthesis.

1. Introduction
In recent years, with the continuous improvement of computer hardware performance and the development of convolutional neural network, deep learning has greatly promoted the development of face recognition. Compared with the traditional feature description, the advanced face features extracted by deep network can represent the face identity information more accurately, and achieve better face recognition effect.

Pose change has always been an important direction in the field of face recognition. Because pose deflection can cause the topological structure change of face image, it is very challenging to synthesize the front view from a single pose deflection image.

3D morphable model [1] (3DMM) models the shape of human face and its appearance as PCA space. Yim etc. [2] proposed a multi-task CNN to predict the identity information of the face image with posture deflection. In this method, the face is rotated with cascade network structure, and loss function at the pixel level is used to regularize the reconstructed image. Zhu etc. [3] developed a novel structure and learning goal to decompose the identity information and pose representation when estimating the front view. DR-GAN [4] used the face classification model to regularize the identity, and used the pose code as the input of the encoder to generate the image with a specific pose. FF-GAN [5] proposed a generative adversarial networks for frontal face based on deep 3D deformation model (3DMM). They extracted 3DMM coefficients of face through deep convolution module, and combined them with input image to input GAN module to generate frontal face image. TP-GAN [6] proposed a two-way generative adversarial networks for frontal face synthesis. Hu etc. [7] proposed CAPG-GAN to generate face images with arbitrary pose, which encoded the positions of key points on the face to represent the head pose information, and then carried out model training. At present, the structure of
frontal face synthesis model is usually too complex, and needs too many parameters. The training and verification of model consume a very long time. The training also requires paired input, which leads to more problems of training data constraints. At the same time, the synthetic face images are often not natural enough, and the effect of network recognition is not satisfactory.

2. IE-WGAN
IE-WGAN(Identity-Extracting WGAN), The model integrates the identity extracting module, which is a pre-trained face recognition classifier. It can extract effective face identity features and supervise the training of face generation module. In addition, this model makes full use of the prior knowledge of face symmetry, and designs a symmetry feature extraction module, which not only improves the quality of synthetic face images, but also speeds up the convergence of network training.

![Fig 1. Illustration of IE-WGAN](image)

2.1. Generator networks
The Generator networks consist of two main parts:

The down sampling encoder: each convolution layer is followed by a residual block, and the output feature map of the fully connected layer is operated by maxout to extract the effective features of the image.

The up sampling decoder: it mainly includes three parts: the first part is a simple deconvolution structure, which is used to up sample the feature FC2; the second part is composed of stacked deconvolution layers, which are used to reconstruct the image, and each deconvolution layer is connected with two residuals; the third part is mainly composed of some convolution layers used to recover the face image.

The generator carries out the multi-scale feature fusion by skip-layer connection. The encoder of the generator networks takes the face images of any pose as the input, and extracts the input features through multiple down sampling convolution layers. Its specific structure is shown in Table 1.

| Layer   | Input   | Filter/stride | Output Size |
|---------|---------|---------------|-------------|
| conv0   | $I^p$   | 7’7/1         | 64’64’64    |
| conv1   | conv0   | 5’5/2         | 32’32’64    |
| conv2   | conv1   | 3’3/2         | 16’16’128   |
| conv3   | conv2   | 3’3/2         | 8’8’256     |
| conv4   | conv3   | 3’3/2         | 4’4’512     |
| fc1     | conv4   | -             | 512         |
| maxout  | fc1     | -             | 256         |
| fc2     | maxout  | -             | 4’4’64      |
| de0_1   | fc2     | 4’4/4         | 16’16’32    |
| de0_2   | de0_1   | 2’2/2         | 32’32’16    |
| de0_3   | de0_2   | 2’2/2         | 64’64’8     |
2.2. Discriminator networks

The input of the discriminator is the face images synthesized by the generator or the real face images, and its output is a one-dimensional vector. 0 represents the synthesized face images and 1 represents the real face images. Because our goal is to synthesize the frontal face images, we use the real frontal face images as the target data set of the discriminator.

The discriminator networks are composed of five convolution layers and a linear layer. Every convolution layer contains a convolution operation, a pooling layer and a ReLU activation function layer. In the specific process, the discriminator networks should avoid batch normalization operation, because batch normalization operation is to create association between samples in the same batch, and statistics will be calculated among irrelevant images in the same training batch, which will weaken some specific details of a single image. These operations will bring negative effects to image generation tasks, whose input and output are both pixel level images, such as image style conversion, face synthesis and so on. In order to solve this problem, layer normalization is used instead of batch normalization in this model to realize normalization and accelerate network convergence.

During the network training process, the discriminator will optimize the following objective functions:

\[
L_D = -E_{I^r \in \mathcal{R}} \log(D(I^r)) - E_{I^s \in \mathcal{K}} \log(1 - D(G(I^s, z)))
\]  

\(L_D\) is objective function of discriminator, \(\mathcal{R}\) is real face images set and \(\mathcal{K}\) is synthetic face images set.\n
In the training process, the goal of the generator is to generate images to interfere with the judgment of the discriminant network, and the goal of the discriminant network is to distinguish the synthesized images from the real images as far as possible.

The training is a game process of generator and discriminator. The loss function in the whole training process is as follows:

\[
L_{gan} = -E_{I^r \in \mathcal{R}} \log(D(G(I^r, z)))
\]

In order to overcome the problem of model collapse or not easy to converge, WGAN-GP model is used to set an additional loss function to limit the gradient of the discriminator. The formula is as follows:

\[
L_{gp} = \|
abla_I D(I^s) - 1 \|^2
\]

Combining the loss function with the loss function of WGAN discriminator, a new discriminator objective function can be obtained:

\[
L = E_{I^r \in \mathcal{R}} \log(D(G(I^r, z))) - E_{I^s \in \mathcal{K}} \log(1 - D(G(I^s, z))) + \lambda E_{I^r \in \mathcal{R}} \left\| \nabla_I D(I^r) - 1 \right\|^2
\]

\(\mathcal{Z}\) is random mixing of real images and synthesized images. \(\lambda\) is weight of gradient cost.

Experimental results show that adding gradient penalty can significantly improve the training speed and solve the problem of slow convergence of original WGAN.
2.3. Classifier networks

In the process of synthesizing the frontal face images from the side deflection images, it is very important how to retain the identity information of the input images. In order to achieve this goal, we introduce an identity extracting module to extract the identity features of the input images. We use VggFace network as the face recognition pre-training model, and The cross entropy loss function is used as the identity loss function of the input images. The formula is as follows.

\[ L_{id} = \sum_{j} \left( -p_{j} \log s_{j} \right) \]  

\( N \) is the total number of sample categories, \( j \) is the real label of the image to be classified using one-hot coding. \( s_{j} \) represents the \( j \)-th value of output vector \( S \) of softmax layer, \( p_{j} \) is the probability that this sample belongs to the \( j \)-th category.

3. Experiments

IE-WGAN provides an effective model to synthesize natural appearance frontal view image from any head posture.

In order to evaluate the performance of this model, we compare IE-WGAN with the most advanced methods on CelebA database, including DR-GAN, HPEN [8] and LFW-3D [9].

CelebA data set is a large-scale facial attribute database, which contains more than 200000 celebrity images, and each image has 40 attribute annotations. The images in the database cover large pose changes and background clutter. CelebA has a wide range of types, a large number and rich annotations, including 10177 identities, 202599 facial images and 5 landmark positions, and 40 binary attribute annotations for each image.

As Fig.2 shows, the synthesized face images are realistic synthesis, while the images effectively maintain the characteristics of the input images, such as hairstyle and expression.

As Fig.3 shows, the performance of IE-WGAN model is much better than LFW-3D and HPEN, but poorer than HPEN model in images edge blur.
4. Conclusion
In this paper, we present a novel model based on WGAN-GP. Compared with the traditional face synthesis methods based on GAN framework, the IE-WGAN model integrates the identity extracting module, which can extract effective face identity features and supervise the training of face generation module. In addition, this model makes full use of the prior knowledge of face symmetry, and designs a symmetry feature extraction module, which not only improves the visual quality of synthetic face, but also speeds up the convergence of network training. Experiments on CelebA data sets show that the model improves the quality of frontal face synthesis.

References
[1] Blanz V, Vetter T, Rockwood A. A Morphable Model for the Synthesis of 3D Faces[J]. ACM Siggraph, 2002:187-194.
[2] Yim J, Jung H, Yoo B I, et al. Rotating your face using multi-task deep neural network[C]// 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2015.
[3] Sun Y , Wang X , Tang X . Deep Learning Face Representation from Predicting 10, 000 Classes.[M]// The Semantic Web. Latest Advances and New Domains. Springer International Publishing, 2016.
[4] Luan Tran, Xi Yin, Xiaoming Liu. Disentangled Representation Learning GAN for Pose-Invariant Face Recognition[C]// IEEE Computer Vision and Pattern Recognition (CVPR 2017). IEEE, 2017:1283-1292.
[5] Yin X , Yu X , Sohn K , et al. Towards Large-Pose Face Frontalization in the Wild[A]. //IEEE International Conference on Computer Vision[C]. 2017:1-10.
[6] Huang R , Zhang S , Li T , et al. Beyond Face Rotation: Global and Local Perception GAN for Photorealistic and Identity Preserving Frontal View Synthesis[C]// 2017 IEEE International Conference on Computer Vision (ICCV). IEEE, 2017.
[7] Yibo Hu, Xiang Wu, Bing Yu et al. Pose-Guided Photorealistic Face Rotation. In CVPR, 2018.
[8] 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 CVPR, 2015.
[9] T. Hassner, S. Harel, E. Paz, and R. Enbar. Effective face frontalization in unconstrained images. In CVPR 2015.