# 2D facial landmark localization method for multi-view face synthesis image using a two-pathway generative adversarial network approach

**Mahmood HB Alhlffee** [Corresp., 1] , **Yea-Shuan Huang** [Corresp., 2] , **Yi-An Chen** [3]

[1] College of Computer Science and Electrical Engineering, Chung-Hua University, Hsinchu, Hsinchu, Taiwan

[2] College of Computer Science and information Engineering, Chung Hua University, Hsinchu, Hsinchu, Taiwan

[3] College of Computer Science, Chung-Hua University, Hsinchu, Hsinchu, Taiwan

Corresponding Authors: Mahmood HB Alhlffee, Yea-Shuan Huang
Email address: mahmood.bidir1985@gmail.com, yeashuan@chu.edu.tw

One of the key challenges in facial recognition is multi-view face synthesis from a single face image. The existing generative adversarial network (GAN) deep learning methods have been proven to be effective in performing facial recognition with a set of pre-processing, post-processing and feature representation techniques to bring a frontal view into the same position in-order to achieve highaccuracy face identification. However, these methods still perform relatively weak in generating high quality frontal-face image samples under extreme face pose scenarios. The novel framework architecture of the two-pathway generative adversarial network (TP-GAN), has made commendable progress in the face synthesis model, making it possible to perceive global structure and local details in an unsupervised manner. More importantly, the TP - GAN solves the problems of photorealistic frontal view synthesis by relying on texture details of the landmark detection and synthesis function, which limits its ability to achieve the desired performance in generating high-quality frontal face image samples under extreme pose. We propose, in this paper, a landmark feature-based method (LFM) for robust pose-invariant facial recognition, which aims to improve image resolution quality of the generated frontal faces under a variety of facial poses. We therefore augment the existing TP-GAN generative global pathway with a well-constructed 2D face landmark localization to cooperate with the local pathway structure in a landmark sharing manner to incorporate empirical face pose into the learning process, and improve the encoder-decoder global pathway structure for better representation of facial image features by establishing robust feature extractors that select meaningful features that ease the operational workflow toward achieving a balanced learning strategy, thus significantly improving the photorealistic face image resolution. We verify the effectiveness of our proposed method on both Multi-PIE and FEI datasets. The quantitative and qualitative experimental results show that our proposed

method not only generates high quality perceptual images under extreme poses but also significantly improves upon the TP-GAN results.

1 **Manuscript Title 2D facial landmark localization**
2 **method for multi-view face synthesis image using a**
3 **two-pathway generative adversarial network approach**

5 Mahmood HB Alhlffee [1], Yea-Shuan Huang [2], Yi-An Chen [3]

7 [1] College of Computer Science and Electrical Engineering, Chung-Hua University,

8 Hsinchu, Hsinchu, Taiwan

9 [2] College of Computer Science and information Engineering, Chung Hua University,

10 Hsinchu, Hsinchu, Taiwan

11 [3] College of Computer Science, Chung-Hua University, Hsinchu, Hsinchu, Taiwan

13 Corresponding Author:

14 Mahmood HB Alhlffee [1],Yea-Shuan Huang [2]

15 [1]College of Computer Science and information Engineering, Chung Hua University,

16 Hsinchu, Hsinchu, Taiwan

17 [2]College of Computer Science and information Engineering, Chung Hua University,

18 Hsinchu, Hsinchu, Taiwan

19 Email address [1]: mahmood.bidir1985@gmail.com

20 Email address [2]: yeashuan@chu.edu.tw

22
23
24

## Abstract

One of the key challenges in facial recognition is multi-view face synthesis from a single face image. The existing generative adversarial network (GAN) deep learning methods have been proven to be effective in performing facial recognition with a set of pre-processing, post-processing and feature representation techniques to bring a frontal view into the same position in-order to achieve high-accuracy face identification. However, these methods still perform relatively weak in generating high quality frontal-face image samples under extreme face pose scenarios. The novel framework architecture of the two-pathway generative adversarial network (TP-GAN), has made commendable progress in the face synthesis model, making it possible to perceive global structure and local details in an unsupervised manner. More importantly, the TP - GAN solves the problems of photorealistic frontal view synthesis by relying on texture details of the landmark detection and synthesis function, which limits its ability to achieve the desired performance in generating high-quality frontal face image samples under extreme pose. We propose, in this paper, a landmark feature-based method (LFM) for robust pose-invariant facial recognition, which aims to improve image resolution quality of the generated frontal faces under a variety of facial poses. We therefore augment the existing TP-GAN generative global pathway with a well-constructed 2D face landmark localization to cooperate with the local pathway structure in a landmark sharing manner to incorporate empirical face pose into the learning process, and improve the encoder-decoder global pathway structure for better representation of facial image features by establishing robust feature extractors that select meaningful features that ease the operational workflow toward achieving a balanced learning strategy, thus significantly improving the photorealistic face image resolution. We verify the effectiveness of our proposed method on both Multi-PIE and FEI datasets. The quantitative and qualitative experimental results show that our proposed method not only generates high quality perceptual images under extreme poses but also significantly improves upon the TP-GAN results.

## Introduction

Face recognition is one of the most commonly used biometric systems for identifying individuals and objects on digital media platforms. Due to changes in posture, illumination, and occlusion, face recognition faces multiple challenges. The challenge of posture changes comes into play when the entire face cannot be seen in an image. Normally, this situation may happen when a person is not facing the camera during surveillance and photo tagging. In order to overcome these difficulties, several promising face recognition algorithms based on deep learning have been developed, including generative adversarial networks (GANs). These methods have been shown to work more efficiently and accurately than humans at detection and recognition tasks. In such methods, pre-processing, post-processing, and multitask learning or feature representation techniques are combined to provide high accuracy results on a wide range of benchmark data sets (Junho et al., 2015; Chao et al., 2015; Xi et al., 2017; Jian et al., 2018). The main hurdle to these methods is multi-view face synthesis from a single face image (Bassel, Ilya & Yuri, 2021; Chenxu et al., 2021; Yi et al., 2021; Hang et al., 2020; Luan, Xi & Xiaoming, 2018; Rui et al.,

66   2017). Furthermore, a recent study (Soumyadip et al., 2016) emphasized that compared with
67   frontal face images with yaw variation less than 10 degrees, the accuracy of recognizing face
68   images with yaw variation more than 60 degrees is reduced by 10%. The results indicate that
69   pose variation continues to be a challenge for many real-world facial recognition applications.
70   The existing approaches to these challenges can be divided into two main groups. In a first
71   approach, frontalization of the input image is used to synthesize frontal-view faces (Meina et al.,
72   2014; Tal et al., 2015; Christos et al., 2015; Junho et al., 2015), meaning that traditional facial
73   recognition methods are applicable. Meanwhile, the second approach focuses on learning
74   discriminative representations directly from non-frontal faces through either a one-joint model or
75   multiple pose-specific models (Omkar, Andrea & Andrew, 2015; Florian, Dmitry & James,
76   2015). It is necessary to explore the above approaches in more detail before proceeding. For the
77   first approach, the conventional approaches often make use of robust local descriptors, (such as
78   John, 1985; Lowe, 1999; Ahonen, Hadid & Pietikäinen, 2006; Dalal & Triggs, 2015), to account
79   for local distortions and then adapt the metric learning method to achieve pose invariance.
80   Moreover, local descriptors are often used (Kilian & Lawrence, 2009; Tsai-Wen et al., 2013)
81   approaches to eliminate distortions locally, followed by a metric learning method to prove pose
82   invariance. However, due to the tradeoff between invariance and discriminability, this type of
83   approach is relatively weak in handling images with extreme poses. A second approach, often
84   known as face rotation, uses one-joint models or multiple pose-specific models to learn
85   discriminative representations directly from non-frontal faces. These methods have shown good
86   results for near-frontal face images, but they typically perform poorly for profile face images
87   because of severe texture loss and artifacts. Due to this poor performance, researchers have been
88   working to find more effective methods to reconstruct positive facial images from data (Yaniv et
89   al., 2014; Amin & Xiaoming, 2017; Xi et al., 2017). For instance, (Junho et al., 2015), adopted a
90   multi-task model to improves identity preservation over a single task model from paired training
91   data. Later on (Luan, Xi & Xiaoming, 2017; Rui et al., 2017), their main contribution was a
92   novel two-pathway GAN architecture tasks for photorealistic and identity preserving frontal
93   view synthesis starting from a single face image. Recent work by (Bassel, Ilya & Yuri, 2021;
94   Chenxu et al., 2021; Yi et al.,2021) has demonstrated advances in the field of face recognition.
95   During pose face transformation, however, some of the synthetic faces appeared incomplete and
96   lacked fine detail. So far, the TP-GAN (Rui et al., 2017) has made significant progress in the
97   face synthesis model, which can perceive global structure and local details simultaneously in an
98   unsupervised manner. More importantly, TP-GAN solves the photorealistic frontal view
99   synthesis problems by collecting more details on local features for a global encoder-decoder
100   network along with synthesis functions to learn multi-view face synthesis from a single face
101   image. However, we argue that TP-GAN has two major limitations. First, it is critically
102   dependent on texture details of the landmark detection. To be more specific, this method focuses
103   on the inference of the global structure and the transformation of the local texture details, as their
104   corresponding feature maps, to produce the final synthesis. The image visual quality results
105   indicate that these techniques alone have the following deficiencies: A color bias can be

106   observed between the synthetic frontal face obtained by TP-GAN method and the input
107   corresponding to non-frontal input. In some cases, the synthetic faces are even incomplete and
108   fall short in terms of fine detail. Therefore, the quality of the synthesized images still cannot
109   meet the requirements for performing specific facial analysis tasks, such as facial recognition and
110   face verification. Second, it uses a global structure, four local network architectures and
111   synthesis functions for face frontalization, where training and inference are unstable under large
112   data distribution, which makes it ineffective for synthesising arbitrary poses. The goal of this
113   paper is to address these challenges through a landmark feature-based method (LFM) for robust
114   pose-invariant facial recognition to improve image resolution under extreme facial poses.
115
116   In this paper, we make the following contributions:
117   The LFM is a newly introduced method for the existing generative global pathway structure that
118   utilizes a 2D face landmark localization to cooperate with the local pathway structure in a
119   landmark sharing manner to incorporate empirical face pose into the learning process. LFM of
120   target facial details provides guidance to arbitrary pose synthesis, whereas the four-local patch
121   network architecture remains unchanged to capture the input facial local perception information.
122   The LFM provides an easy way for transforming and fitting two-dimensional face models in
123   order to achieve target pose variation and learn face synthesis information from generated
124   images.
125   In order to better represent facial image features, we use a denoising autoencoder (DAE) to
126   modify the structure of the generator's global-path encoder and decoder. The goal of this
127   modification is to train the encoder decoder with multiple noise levels so that it can learn about
128   the missing texture face details. Adding noise to the image pixels causes them to diffuse away
129   from the manifold. As we apply DAE to the diffused image pixels, it attempts to pull the data
130   points back onto the manifold. This implies that DAE implicitly learns the statistical structure of
131   the data by learning a vector field from locations with no data points back to the data manifold.
132   As a result, encoder-decoders must infer missing pieces and retrieve the denoised version in
133   order to achieve balanced learning behavior.
134   We optimize the training process using an accurate parameter configuration for a complex
135   distribution of facial image data. By re-configuring the parameters (such as the learning rate,
136   batch size, number of epochs, and etc.), the GAN performance can be better optimized during the
137   training process. Occasionally, unstable "un-optimized" training for the synthetic image problem
138   results in unreliable images for extreme facial positions.
139
140   **Related Work**

141   In this section, we focus on the most recent studies which are related to the multi-view face
142   synthesis problem using deep learning approaches. The deep learning approaches including face
143   normalization, generative adversarial network and facial landmark detection, are reviewed.
144
145   **Face Normalization**

146   Face normalization, or multi-view face synthesis from a single face image, is a unique challenge
147   for computer vision systems due to its ill-posed problem. The existing solutions to address this
148   challenge can be classified into three categories: 2D/3D local texture warping methods (Tal et
149   al., 2015; Xiangyu et al., 2015), statistical methods (Christos et al., 2015; Li et al., 2014), and
150   deep learning methods (Xi et al., 2017; Luan, Xi & Xiaoming, 2017). (Tal et al., 2015),
151   employed a single 3D reference surface for all query faces in order to produce face
152   frontalization. (Xiangyu et al., 2015), employed a pose and expression normalization method to
153   recover the canonical-view. (Christos et al., 2015), proposed a joint frontal view synthesis and
154   landmark localization method. (Li et al., 2014), the authors concentrated on local binary pattern-
155   like feature extraction. (Xi et al., 2017), proposed a novel deep 3DMM-conditioned face
156   frontalization GAN in order to achieve identity-preserving frontalization and high-quality images
157   by using a single input image with a $90^0$ face pose. (Luan, Xi & Xiaoming, 2017), proposed a
158   single-pathway framework called the disentangled representation learning-generative adversarial
159   network (DR-GAN) to learn identity features that are invariant to viewpoints, etc.
160
161   **Generative Adversarial Networks (GANs)**

162   The GAN is one of the most interesting research frameworks that is used for deep generative
163   models proposed by (Ian et al., 2014). The theory behind the GAN framework can be seen as a
164   two-player non-cooperative game to improve the learning model. A GAN model has two main
165   components, generator (G) and discriminator (D). G generates a set of images that is as plausible
166   as possible in order to confuse the D, while the D works to distinguish the real generated images
167   from the fake. The convergence is achieved by alternately training them. The main difference
168   between GANs and traditional generative models is that GANs generate whole images rather
169   than pixel by pixel. In a GAN framework, the generator consists of two dense layers and a
170   dropout layer. A normal distribution is used to sample the noise vectors and feed them into the
171   generator networks. The discriminator can be any supervised learning model. GANs have been
172   proven effective for a wide range of applications, such as image synthesis (Rui et al., 2017; Yu et
173   al., 2018; Yu et al., 2020), image super-resolution (Christian et al., 2017), image-to-image
174   translation (Jun-Yan et al., 2017), etc. Several effective GAN models have been proposed to
175   cope with the most complex unconstrained face image situations, such as changes in pose,
176   lighting and expression. For instance, (Alec, Luke & Soumith, 2016), proposed a deep
177   convolutional GAN to integrate a convolutional network into the GAN model to achieve more
178   realistic face image generation. (Mehdi & Simon, 2014), proposed a conditional version of the
179   generative adversarial net framework in both generator and discriminator. (Augustus,
180   Christopher & Jonathon, 2017), presented an improved version of the Cycle-GAN model called
181   "pixel2pixel" to handle the image-to-image translation problems by using labels to train the
182   generator and discriminator. (David, Thomas & Luke, 2017), proposed a boundary equilibrium
183   generative adversarial network (BE-GAN) method, which focuses on the image generation task
184   to produce high-quality image resolution, etc.
185

**Facial Landmark Detection**

The face landmark detection algorithm is one of most successful and fundamental components in a variety of face applications, such as object detection and facial recognition. The methods used for facial landmark detection can be divided into three major groups; holistic methods, constrained local methods, and regression-based methods. In the past decade, deep learning models have proven to be a highly effective way to improve landmark detection. Several existing methods are considered to be good baseline solutions to the 2D face alignment problem for faces with controlled pose variation (Xuehan & Fernando, 2013; Georgios, 2015; Xiangyu et al., 2016; Adrian & Georgios, 2017). (Xuehan & Fernando, 2013), proposed supervised descent method, which learns the general descent directions in a supervised manner. (Georgios, 2015), in their method, a sequence of Jacobian matrices and hessian matrices is determined by using regression. (Xiangyu et al., 2016), proposed a 3D model with cascaded convolutional neural network to solve the self-occlusion problem. (Adrian & Georgios, 2017), proposed a guided-by-2D landmarks convolutional neural network that converts 2D annotations into 3D annotations, etc.

We can summarize some important points from our related work. Despite the fact that the existed methods produced good results on the specific face image datasets for which they were designed and provided robust alignment across poses, they are difficult to replicate if they are applied alone to different datasets. This is especially true for tasks like facial normalization or other face synthesis tasks, where deep structure learning methods still fail to generate high-quality image samples under extreme pose scenarios, which results in significantly inferior final results.

**Proposed Method**

In this section, we shall first briefly describe the existing TP-GAN architecture and then describe our proposed LFM method in detail.

**TP-GAN Architecture**

Based on the structure shown in Fig 1, the TP-GAN framework architecture consists of two stages. The first stage is a generator of two-pathways CNN $G_{\theta_G}$ that is parameterized by $\theta_G$. Each pathway has encoder-decoder $\{G_{\theta_E}, G_{\theta_D}\}$ structure and combination of loss functions, a local pathway $\{G_{\theta_E^l}, G_{\theta_D^l}\}$ of four landmark patch networks $G_{\theta_i^l}, i \in \{0, 1, 2, 3\}$ to capture the local texture around four facial landmarks, and one global network $\{G_{\theta_E^g}, G_{\theta_D^g}\}$ to process the global face structure. Furthermore, the bottleneck layer $(G_g)$, which is the output of $G_{\theta_E^g}$, is typically used for classification tasks with the cross-entropy loss $L_{cross-entropy}$. A global pathway helps to integrate facial features with their long-range dependencies and, therefore, to create faces that preserve identities, especially in cases of faces with large pose angles. In this way, we can learn a

222    richer feature representation and generate inferences that incorporate both contextual
223    dependencies and local consistency. The loss functions, including pixel-wise loss, symmetry
224    loss, adversarial loss, and identity preserving loss, are used to guide an identity preserving
225    inference of frontal view synthesis. The discriminator $D_{\theta_D}$ is used to distinguish real facial
226    images $I^F$ or 'ground-truth (GT) frontal view' from synthesized frontal face images $G_{\theta_G}(I^P)$ or
227    'synthesized-frontal (SF) view'. A second stage involves a light-CNN model that is used to
228    compute face dataset's identity-preserving properties. For a more detailed description, see (Rui et
229    al., 2017).
230

231    **LFM for Generative Global Pathway Structure**

232    To the best of our knowledge, this is the first study to integrate an LFM with the existing TP-
233    GAN global pathway structure for training and evaluation purposes. In this work, we exploit a
234    landmark detection mechanism (Adrian & Georgios, 2017) that proposed for 2D-to-3D facial
235    landmark localization to help our model obtain a high quality frontal-face image resolution. Face
236    landmarks are the most compressed representation of a face that maintains information such as
237    pose and facial structure. There are many situations where landmarks can provide advanced face-
238    related analyses without using whole face images. The landmark method used in this study was
239    explored at (Xi Y et al., 2017; Shuicheng & Jiashi. 2018; Xing D, Vishwanath. Sindagi & Vishal.
240    2018). These methods can achieve high accuracy of face alignment by cascaded regression
241    methods. Methods like these work well when particular poses are chosen without taking other
242    factors into consideration, such as facial characteristics. We found that facial characteristics can
243    play an imperative role in improving the results of the current state of the art. By adding
244    landmarks to augment the synthesized faces, recognition accuracy will be improved since these
245    landmarks rely on generative models to enhance the information contained within them. The
246    process for generating facial images is shown in Fig 1. We will discuss our Fig 2 architecture in
247    the subsequent paragraph. We perform a face detection to locate the face in the Multi-PIE and
248    FEI datasets. The face detection can be achieved by using a Multi-Task Cascade CNN through
249    the MTCNN library (Kaipeng Z et al.,2016). After that, cropping and processing of the profile
250    image. A local pathway of four landmark patch networks $G_{\theta_i^l}$, $i \in \{0, 1, 2, 3\}$ to capture the local

251    texture around four facial landmarks. Each patch learns a set of filters for rotating the center-
252    cropped patch (after rotation, the facial landmarks remain in the center) to its corresponding
253    frontal view. Then, we used a multiple feature map to combine the four facial tensors into one.
254    Each tensor feature is placed at a "template landmark location" and a max-out fusing strategy is
255    used to ensure that stitches on overlapping areas are minimized. Then, a 2D zero padding
256    technique is used to fill out the rows and columns around the template landmark location with
257    zeros. Nevertheless, local landmark detection alone cannot provide accurate texture detail for a
258    face that has a different shape, or multiple views, because all generated local synthesis faces
259    have the same fixed patch (or template) centralized location, regardless of their shape

260    characteristics. The challenge becomes even greater with these shapes when the face is under
261    extreme poses. A texture can be defined as a function of the spatial variation of brightness
262    intensity of pixels in an image. Each texture level represents a variable, with variations such as
263    smoothness, coarseness, regularity, and etc., of each surface oriented in different directions. Our
264    work focuses on two important phenomena: rotation and noise "noise is a term used to describe
265    image information that varies randomly in brightness or color". As a result, if the methods used
266    to eliminate these common phenomena are unreliable, the results will be less accurate; therefore,
267    in practice, the methods used to create the images should be as robust and stable as possible. In
268    addition, the images may differ in position, viewpoint, and light intensity, all of which can
269    influence the final results, challenging texture detail capturing. In order to overcome these
270    challenges, we must adapt a method that can capture and restore the missing texture information.
271    The key to solving this problem is a landmark localization method based on regression. Our
272    work utilized Face Alignment Network (FAN). A FAN framework is based on the HourGlass
273    (HG) architecture, which integrates four Hourglass models to model human pose through
274    hierarchical, parallel, and multi-scale integration to improve texture maps by reconstructing self-
275    occluded parts of faces. The landmark detection algorithm captures and restores the texture
276    details of the synthesised face image by repositioning the appearance spot of the mismatched or
277    drifted patch "template landmark location". Figure 3 illustrates some examples. Our method
278    allows us to treat faces that have a variety of shape characteristics. In this way, the spatial
279    variations, smoothness, and coarseness that arise due to mismatched or drifted pixels between
280    local and global synthesized faces are eliminated. Typical landmark templates are approximately
281    the same size as a local patch network, but each region has its own structure, texture, and filter.
282    Next, we combined the local synthesis image with the global synthesis image (or two textures)
283    for data augmentation. Every patch of our FAN has its own augmented channels, and each patch
284    has its own RGB along with a depth map (D) input for each 2D local synthesis image. In this
285    way, the texture details help us to build a more robust model around the face patch region and
286    enhance generalization. Even though landmark feature extraction may result in some
287    incongruous or over-smoothing due to noise, it still remains an important method for
288    incorporating pose information during learning.
289        The landmark detection algorithm for our synthesis face image was built using 68 points. We
290    then reconstructed those synthesis image into four uniform patches (or templates), $Leye = G_{\theta_0^g}$
291    ,$i \in 0$, $Reye = G_{\theta_1^g}, i \in 1$, $Nose = G_{\theta_2^g}, i \in 2$ and $Mouth = G_{\theta_3^g}, i \in 3$, and each patch is comprised
292    of convolutional components. Each patch region has its own filter, which contains different
293    texture details, regain size and structure information. Individual filters provide more details about
294    specific areas in an image, such as pixels or small areas with a high contrast or that are different
295    in color or intensity from the surrounding pixels or areas. Then, one layer of convolutional neural
296    networks is used to coordinate and extract intermediate features. For our method to work more
297    effectively, we remove layers 5, 6, and 7 from ConvNet $\{G_{\theta_D^g}\}$ and replace them after the

298 concatenation stage. Those layers' act as a visualization feature map for a specific input of a
299 fontal-face image in order to increase the amount of visual information kept by subsampling
300 layers' structure. Then, we merged the features tensors of the local and global pathways into one
301 tensor to produce the final synthesis face. Table 1 shows the workflow of all operational steps.
302 The landmark method provides useful information for large-pose regions, e.g., $90^0$, which helps
303 our model to produce more realistic images.
304

**Global Pathway Encoder-Decoder Structure**

306 In this section we describe our encoder-decoder formulation. Inspired by the work of (Jimei Y et
307 al., 2016), for the DAE, our aim is to train the encoder-decoder with multiple noise levels in
308 order to learn more about the missing texture face details of the input face image and preserve
309 the identity of the frontal-view image $I^F$ from the profile image $I^P$. The encoder-decoder
310 mechanism has to discover and capture information between the dimensions of the input in order
311 to infer missing pieces and recover the denoised version. In a subsequent paragraph, we will
312 discuss encoder-decoders in more technical details. The idea starts with assuming that the input
313 data points (image pixels) lies on a manifold in $\mathbb{R}^N$, where $\mathbb{R}^N$ is the a set of databases. Adding
314 noise to the data image pixel results in diffusing away from the manifold. When we apply DAE
315 to the diffused data image pixels, it tries to pull the data point back onto the manifold. Therefore,
316 DAE learns a vector field pointing from locations with no data point back to the data manifold,
317 implying that it implicitly learns the statistical structure of the data. However, a sparse coding
318 model has been shown to be a good model for image denoising. We assume that group sparse
319 coding, which generalizes standard sparse coding, is effective for image denoising as well, and
320 we will view from a DAE perspective. The encoding function of sparse coding occurs in the
321 inference process, where the network infers the latent variable $s$ from noisy input $\hat{x}$. Each
322 individual symbol is defined here. Let $f_e$ be the RGB components of the input face image, $\Phi$ is
323 the method that splits the input image into its RBG components, $\wedge$ is a set of weights and bias
324 for the DAE, and $a$ is the activation function. In our case, the iterative shrinkage-thresholding
325 algorithm (ISTA) is used to perform inference and is formulated as follows:
326

$$s = f_e(\tilde{x}; \{\Phi, \wedge, a\}) = ISTA\,(\tilde{x}; \Phi, \wedge, a) \quad (1)$$

328

329 The decoding function is the network's reconstruction of the input from the latent variable.
330

$$\hat{x} = f_d\,(s; \Phi) = \Phi_s \quad (2)$$

332

333 where $f_d$ is a denoising function.

334 The DAE method can be used to learn $\wedge$ through the following. For each input data point as
335 shown in Fig 4, we construct a noisy input by adding Gaussian white noise (GWN) to the
336 original input profile images, as given in: $\tilde{x} = x + v$, where $v \sim N\,(0, \sigma^2\,I)$. Here, $I$ is an $N \times N$

337    identity matrix, where $N$ is the size of input data (batch output of the generator), and $\sigma^2$ is the
338    noise variance which is the same in all directions. We use group sparse coding to denoise $\tilde{x}$, as
339    described in Equation (1) and Equation (2). The DAE is formulated as follows:

340

341
$$E_{DAE} = \| x - \hat{x} \|_2^2 \quad (3)$$

342

343    We define DAE's reconstruction error as a square error between $x$ and $\hat{x}$ .
344    Generally, the cost function can be another form of differentiable error measure, shown as
345    follows:

346

347
$$\Delta \wedge \,\, \propto - \,\, \partial E_{DAE} / \partial \wedge \quad (4)$$

348

349    where $\partial E_{DAE} / \partial \wedge \rightarrow$ differentiate DAE's reconstruction error with respect to weights, and $\Delta \wedge$
350    is the change in weights, $\wedge$ can then be learned by doing gradient descent on the DAE's cost
351    function.
352       Essentially, the $G_{\theta_E^g}$ encodes input data $x \in \mathbb{R}^N$ into a hidden representation: $h = f_e(x;\theta) \in$
353    $\mathbb{R}^M$. $f_e$ is the encoder, $\mathbb{R}^M$ is the set of database, and $\theta$ is learning parameters of the DAE
354    function. Then, it decodes $G_{\theta_D^g}$ the hidden representation into a reconstruction of the input data: $\hat{x}$
355    $= f_d(h;\theta)$. The objective for learning the parameter $\theta$ of an $\{G_{\theta_E^g}, G_{\theta_D^g}\}$ is to minimize the
356    reconstruction error between $\hat{x}$ and $x$. Usually, there is some constraint on $\{G_{\theta_E^g}, G_{\theta_D^g}\}$ to prevent
357    it from learning an identity transformation. For example, the dimension of $h$ is much smaller than
358    the input data's dimension *i.e.* $M \ll N$, then $\{G_{\theta_E^g}, G_{\theta_D^g}\}$ will function similarly to principle
359    component analysis (PCA). The hidden representation $\|h\|_1$ is small, then $G_{\theta_E^g}$ functions
360    similarly to sparse coding. The DAE tries to remove noise from input data. Let $\tilde{x} = x + v$ be a
361    noisy input by adding $v$ to an original input $x$. DAE takes $\tilde{x}$ as an input, then outputs a denoising
362    signal $\hat{x} = f_d(f^e(\tilde{x}; \theta);\theta)$. The objective function of a such technique is to minimize the error
363    between $x$ and $\hat{x}$ by adjusting $\theta$, i.e., it tries to reconstruct the actual content well while not
364    reconstructing noise. DAE can also be viewed as a generative model.

365

366    **Adversarial Networks**

367    Following (Ian et al., 2014) work, adversaries network consists of two components $(G)$ and $(D)$.
368    "The loss function reflects the difference in distribution between the generated and original
369    data". We will first review some technical aspects of the training process, and then the
370    adversaries' network. In frontal view synthesis, the aim is to generate a photorealistic and

371   identity-preserving frontal view image ($I^F$) from a face image under a different pose, i.e. a

372   profile image ($I^P$). During the training phase of such networks, pairs of corresponding $\{I^F, I^P\}$

373   from multiple identities $y$ are required. Input $I^P$ and output $I^F$ are both based on a pixel space of

374   size $W \times H \times C$ with a color channel $C$. We aim to learn a synthesis function that can output a

375   frontal view when given a profile image. This section will be omitted since it was already

376   explained in TP-GAN architecture. Optimizing the network parameters ($G_{\theta_G}$) starts with

377   minimizing the specifically designed synthesis loss ($L_{syn}$) and the aforementioned $L_{cross-entropy}$

378   . For a training set with $N$ training pairs of $\{I^F_n, I^P_n\}$, the optimization problem is expressed as

379   follows:

380

381
$$\hat{\theta}_G = \frac{1}{N} \underset{\theta_G}{\mathrm{argmin}} \sum_{n=1}^{N} \{L_{syn}(G_{\theta_G}(I^P_n), I^F_n) + \alpha L_{cross-entropy}(G_{\theta^g_E}(I^p_n), y_n)\} \quad (5)$$

382

383   where $\alpha$ is a weighting parameter and $L_{syn}$ is a weighted sum of individual losses that together

384   constrain the image to reside within the desired manifold. Each individual loss function will be

385   explained in the comprehensive loss functions section.

386      In order to generate the best images, we need a very good generator and discriminator. The

387   reason for this is that if our generator is not good enough, we won't be able to fool the

388   discriminator, resulting in no convergence. A bad discriminator will also classify images that

389   make no sense as real, which means our model never trains, and we never produce the desired

390   output. The image can be generated by sampling values from a Gaussian distribution and feeding

391   them into the generator network. Based on a game-theoretical approach, our objective function is

392   a minimax function.

393

394
$$\min_{\theta_G} \max_{\theta_D} [\mathbb{E}_{I^F \sim P(I^F)} \log D_{\theta_D}(I^F) + \mathbb{E}_{I^P \sim P(I^P)} \log (1 - D_{\theta_D}(G_{\theta_G}(I^P)))] \quad (6)$$

395

396   where $I^F$ presents as real frontal face images

397     $D_{\theta_D}$ presents as the discriminator

398     $G_{\theta_G}$ presents as the generator

399   $G_{\theta_G}(I^P)$ presents as synthesized frontal face images.

400   Using the discriminator to maximize the objective function allows us to perform gradient descent

401   on it. The generator tries to minimize its objective function, so we can use gradient descent to

402   compute it. In order to train the network, gradient ascent and descent must be alternated.

403

404
$$\max_{\theta_D} [\mathbb{E}_{I^F \sim P(I^F)} \log D_{\theta_D}(I^F) + \mathbb{E}_{I^P \sim P(I^P)} \log (1 - D_{\theta_D}(G_{\theta_G}(I^P)))] \quad (7)$$

405

406 Gradient ascent on $D$.

407

408
$$\min_{\theta_G} \mathbb{E}_{I^P \sim P(I^P)} \; log\,(1 - D_{\theta_D}(G_{\theta_G}(I^P))) (8)$$

409

410 Gradient descent on $G$.

411 Minimax problem allows discriminate to maximize adversarial networks, so that we can perform

412 gradient ascent on these networks; whereas generator tries to minimize adversarial networks, so

413 that we can perform gradient descent on these networks. In practice, Equation (6) might not

414 provide enough gradients for $G$ to learn well. During the early stages of learning, when $G$ is poor,

415 $D$ can reject samples with a high degree of confidence, since they are clearly different from the

416 training data. In this case, $log\,(1 - D_{\theta_D}(G_{\theta_G}(I^P)))$ saturates. As an alternative to training $G$ to

417 minimize $log\,(1 - D_{\theta_D}(G_{\theta_G}(I^P)))$, we can train $G$ to maximize $log\,D_{\theta_D}(I^F)$. As a result of this

418 objective function, the discriminator and generator have much stronger gradients at the start of

419 the learning process.

420   When a single discriminating network is trained, the refiner network will tend to

421 overemphasize certain features in order to fool the current discriminator network, resulting in

422 drift and producing artifacts. We can conclude that any local patch sampled from the refined

423 image should have similar statistics to the real image patches. We can therefore define a

424 discriminator network that classifies each local image patch separately rather than defining a

425 global discriminator network. The division limits both the receptive field and the capacity of the

426 discriminator network, as well as providing a large number of samples per image for learning the

427 discriminator network. The discriminator in our implementation is a fully convolutional network

428 that outputs a probabilistic $N \times N$ ("$N = 2$" represented as a synthesized face image) map instead

429 of one scalar value to distinguish between a ground truth frontal view (GT) and a synthesized

430 frontal view (SF). Our discriminator loss is defined through the discrepancy between the model

431 distribution and the data distribution, using an adversarial loss ($L_{adv}$). By assigning each

432 probability value to a particular region, the $D_{\theta_D}$ can now concentrate on a single semantic region

433 rather than the whole face.

434

435 **Comprehensive Loss Functions**

436 In addition to the existing TP-GAN synthesis loss functions, which are a weighted sum of four

437 individual loss functions ($L_{px}$,"$L_{sym}$,$L_{ip}$,$L_{tv}$ and $L_{part}$"), a classification loss $L_{classify}$ is further

438 added to our method in order to get better results. Each individual loss function is presented

439 below, and the used symbols are defined here. Let $I$ be an output image.

440 Let $W$ and $H$ be the width and height of $I$, $(x,y)$ be a pixel coordinate of a 2D image, $I^{pred}$ be the

441 predicted (i.e., synthesized) frontal-face image of $I$, $I^{gt}$ be the representative frontal-face image

442     of the ground-truth category of $I$. $I_{local}^{gt}$ be the composition image of the four local facial patches

443     of $I^{gt}$, and $I_{local}^{gt}$ be the composition image of the four local facial patches of $I$.

444

445     **Pixel-Wise Loss ($L_{px}$)**

446

447
$$L_{px} = \frac{1}{W \times H} \sum_{x=1}^{W} \sum_{y=1}^{H} |I^{gt}(x,y) - I^{pred}(x,y)| \quad (9)$$

448

449     Although pixel wise loss may bring some over-smooth effects to the refined results, it is still an

450     essential part for both accelerated optimization and superior performance.

451

452     **Symmetry Loss ($L_{sym}$)**

453

454
$$L_{sym} = \frac{1}{W/2 \times H} \sum_{x=1}^{W/2} \sum_{y=1}^{H} |I^{pred}(x,y) - I^{pred}(W-(x-1),y)| \quad (10)$$

455

456     $L_{sym}$ is used to calculate the symmetry of the synthesized face image because a face image is

457     generally considered to be a symmetrical pattern.

458

459     **Identity Preserving Loss ($L_{ip}$)**

460

461
$$L_{ip} = \sum_{i=1}^{2} \frac{1}{W_l \times H_l} \sum_{x=1}^{W_l} \sum_{y=1}^{H_l} \left|F_l^{gt}(x,y) - F_l^{pred}(x,y)\right| \quad (11)$$

462

463     where $F_l^{gt}$ and $F_l^{pred}$ respectively denote the feature map of the last two layers of the light-CNN

464     net (Xiang et al., 2015) of $I^{gt}$ and $I^{pred}$, $W_l$ and $H_l$ are the width and height of the feature map of

465     the last $l-th$ layer ($l = 1, 2$) of the light-CNN net. It is expected that a good synthesized frontal-

466     face image will have similar characteristics to its corresponding real frontal-face image. We

467     employ a fully connected layer of the pre-trained light-CNN net for the feature extraction of the

468     pre-trained recognition network. The pre-trained model will leverage the loss to enforce identity-

469     preserving frontal view synthesis.

470

471     **Adversarial Loss ($L_{adv}$)**

472

473
$$L_{adv} = \frac{1}{N}\sum_{n=1}^{N} - \log D_{\theta_D}(G_{\theta_G}(I_n^P)) \quad (12)$$

474

475 $L_{adv}$ is used to make the real frontal-face image $I^F$ and a synthesized frontal face images $G_{\theta_G}(I_n^P)$

476 indistinguishable, so that the synthesized frontal face image achieves a visually pleasing effect.

477

478 **Total Variation Loss ($L_{tv}$)**

479 Generally, the face images synthesized by two pathways generative adversarial networks have

480 unfavorable visual artifacts, which deteriorates the visualization and recognition performance.

481 Imposing $L_{tv}$ on the final synthesized face images can help to alleviate this issue. The $L_{tv}$ loss is

482 calculated as follows:

483

484
$$L_{tv} = \sum_{x=1}^{W}\sum_{y=1}^{H} \left|I_{x,y}^{pred} - I_{x-1,y}^{pred}\right| + \left|I_{x,y}^{pred} - I_{x,y-1}^{pred}\right| \quad (13)$$

485

486 $L_{tv}$ will generate a smooth synthesized face image.

487

488 **Classification Loss ($L_{classify}$)**

489

490
$$L_{classify} = -\sum_{i} y_i^{true} \log_2\left(y_i^{pred}\right) \quad (14)$$

491

492 where $i$ is the class index, $y^{true}$ denotes the tensor of the one-hot true target of $I$, and $y^{pred}$ is the

493 predicted probability tensor. $L_{classify}$ is a cross-entropy loss which is used to ensure the

494 synthesized frontal-face image can be classified correctly.

495

496 **Local Pixel-Wise Loss ($L_{part}$)**

497

498
$$L_{part} = \frac{1}{W \times H}\sum_{x=1}^{W}\sum_{y=1}^{H} |I_{local}^{gt}(x,y) - I_{local}^{pred}(x,y)| \quad (15)$$

499

500 This loss is to calculate the total average pixel difference between $I_{local}^{gt}$ and $I_{local}^{pred}$, and it is not

501 addressed in (Rui et al., 2017) work.

502

503    **Generator Loss Functions ($L_{generator}$)**

504    The generator loss function of the proposed method is a weighted sum of all the losses defined
505    above:

506
507

$$L_{generator} = L_{px} + \lambda_1 L_{sym} + \lambda_2 L_{ip} + \lambda_3 L_{tv} + \lambda_4 L_{classyify} + \lambda_5 L_{part} \quad (16)$$

508
509    where $\lambda_i$ ($i = 1\sim5$) are weights that coordinate the different losses, and they are set to be

510    $\lambda_1 = 0.1$, $\lambda_2 = 0.001$, $\lambda_3 = 0.0001$, $\lambda_4 = 0.1$, and $\lambda_5 = 0.3$ in our experiments.

511

512    **Optimizing the Training Process**

513    In order to optimize the training process, we propose some modifications to the TP-GAN
514    parameters as shown in Table 2 in order to improve the performance in learning the frontal-face
515    data distribution. In our experiment, we consider few parameters: learning rate, batch size,
516    number of epochs, and loss functions. We chose those parameters based on our experience,
517    knowledge and observations. The adopted learning rate improve the module loss accuracy for
518    both $G$ and $D$. The batch size is the number of examples from the training dataset used in the
519    estimation of the error gradient. This parameter determines how the learning algorithm will
520    behave. We have found that using a larger batch size has adversely affected our method
521    performance. As a result, during initial training the discriminator may be overwhelmed by too
522    many examples. This will lead to poor training performance. The number of training epochs is a
523    key advantage of machine learning. As the number of epochs increases, the performance will be
524    improved and the outcomes will be astounding. However, the disadvantage is that it takes a long
525    time to train a large number of epochs. These parameters are essential to improve LFMTP-
526    GAN's representation learning, gaining high-precision performance and reducing visual artifacts
527    when synthesising frontal-face images.
528

529    **Experiments**

530    We conducted extensive experiments to verify the effectiveness of our method by comparing it
531    with the TP-GAN. The evaluation protocol includes frontal face image resolution and accuracy
532    preserving face identity.
533

534    **Experimental Settings**

535    Both LFMTP-GAN and TP-GAN models are tested and trained on the Multi-PIE and FEI
536    datasets. Multi-PIE (Ralph et al., 2010) is a large face dataset with $75000 +$ images for 337
537    identities across a variety of different poses, illuminations and expression conditions captured in
538    a constrained environment. Multi-PIE has 15 poses ranging from $\pm\,90^0$, and 20 illumination
539    levels for each subject. All 20 illuminations were taken within a few seconds: two without any

540  flash illumination, followed by an 18 image with each flash firing independently. Figure 5 shows
541  an example of our model results. "The intensity of light is determined by the brightness of the
542  flash and the background. For example, a bright or dark flash, shadow reflection, or a white or
543  blue background will affect intensity. This depends on the recording equipment and the
544  positioning." To minimize the number of saturated pixels in flash illuminated images, all
545  cameras have been set to have a pixel value of 128 for the brightest pixel in an image without
546  flash illumination. In the same way, the diffusers in front of each flash were added. The color
547  balance was also manually adjusted so that the images looked similar. FEI is a Brazilian
548  unlabeled face dataset with 2800 + images for 200 identities across a variety of different poses
549  captured in a constrained environment. The face images were taken between "June 2005 and
550  March 2006" at the artificial intelligence laboratory at são bernardo do campo, são paulo, brazil.
551  The FEI images were taken against a white homogenous background in an upright frontal
552  position with a $\pm 90^0$ range of profile poses; and different illuminations, distinct appearances
553  and hairstyles were included for each subject. Our method shares the same implementation
554  concept as TP-GAN but totally has different parameters settings. The training lasts for 10-to-18
555  days in each system for each dataset. The training model and source code will be released after
556  this article is accepted.
557

558  **Visual Quality**

559  In this subsection, we compare LFMTP-GAN with TP-GAN. Figure 6 and Fig 7 shows the
560  comparison images, where the first column is the profile-face images under different face poses,
561  the second column is the synthesized frontal-face images by TP-GAN, the third column is the
562  synthesized frontal-face images by LFMTP-GAN, and the last column is one randomly selected
563  frontal-face image of the category corresponding to the profile-face image. The yaw angle of the
564  input face image are chosen circularly from ($15^0,30^0,45^0,60^0,75^0$ and $90^0$). Obviously, the
565  resolution of the LFMTP-GAN images looks better than TP-GAN. This reveals that the use of a
566  proper 2D landmark localization algorithm could significantly improve the image quality of the
567  synthesized images, provide rich textural details, and contain fewer blur effects. Therefore, the
568  visualization results shown in Fig 6 and Fig 7 demonstrate the effectiveness of our method across
569  a variety of poses and datasets. Similarly, Fig 8 shows the results of these datasets in close-up
570  image with $90^0$ facial pose.
571      Previous frontal view synthesis methods are usually based on a posture range of $\pm 60^0$. It is
572  generally believed that if the posture is greater than $60^0$, it is difficult to reconstruct the image of
573  the front view. Nonetheless, we will show that with enough training data and a properly designed
574  loss function, this is achievable. In Fig 9 and Fig 10, we show that LFMTP-GAN can recover
575  identity-preserving frontal faces from any pose, as well as comparing with state-of-the-art face
576  frontalization methods, it performs better. In addition, our geometry estimation method does not
577  require 3D geometry knowledge because it is driven by data alone.
578

**Identity Preserving**

580    To quantitatively demonstrate the identity preserving ability of the proposed method, we
581    evaluate the classification accuracy of synthesized frontal-face images on both Multi-PIE and
582    FEI databases, and show their classification accuracy (%) in Table 3 across views and
583    illuminations. The experiments were conducted by first employing light-CNN to extract deep
584    features and then using the cosine-distance metric to compute the similarity of these features.
585    The light-CNN model was trained on MS-Celeb-1M (Microsoft Celeb, 2016) which is a large-
586    scale face dataset, and fine-tuned on the images from Multi-PIE and FEI. Therefore, the light-
587    CNN results on the profile images $I^P$ serves as our baseline. Our method produces better results
588    than TP-GAN as the pose angle is increased. Our approach has shown improvements in frontal-
589    frontal face recognition. Moreover, Table 4 illustrates that our method is superior to many
590    existing state-of-the-art approaches.
591        Many deep learning methods have been proposed for frontal view synthesis, but none of them
592    have been proved to be sufficient for recognition tasks. Chao et al., 2015; Yibo et al., 2018 relied
593    on direct methods such as CNN for face recognition, which will definitely reduce rather than
594    improve performance. It is therefore important to verify whether our synthesis results can
595    improve recognition performance (whether "recognition via generation" works) or not. The next
596    section presents the loss curves for the two models.
597

**Model Loss Curve Performance**

599    This section provides a comparison with TP-GAN. We analyze the effects of our model on three
600    tradeoff parameters named generator loss, pixel-wise loss, and identity preserving accuracy. 80%
601    of the face image subjects from the Multi PIE and FEI datasets were used for training and
602    evolution purposes. 90% of the image subjects were used for training, while 10% were used for
603    testing. The recognition accuracy and corresponding loss curves are shown in Fig 10. We can
604    clearly see from the curves that, the proposed method improves the TP-GAN model and provides
605    a much better performance on both datasets. In particular, the number of epochs exceeds 4500,
606    the loss performance decreased sharply in LFMTP-GAN model, while the loss performance
607    decreased slightly in the TP-GAN model. Our optimization learning curves was calculated
608    according to the metric by which the parameters of the model were optimized, i.e., loss. More
609    importantly, our method still produces visually convincing results (as shown in Fig 6, Fig 7 and
610    Fig 9) even under extreme face poses, its recognition performance is about 1.2% higher than that
611    of TP-GAN.
612

**Results and Discussions**

614    The goal of this method is to match the appearance of each query face by marking the partially
615    face surface of the generated image, such as the eyes, nose and mouth. In theory, this would have
616    allowed the TP-GAN method to better preserve facial appearance in the updated, synthesized
617    views. The statement holds true when the face is considered to have similar shape characteristics.

618    Considering that all human faces have unique shape characteristics, this may actually be
619    counterproductive and harmful, rather than improving face recognition. We believe that it is
620    necessary to integrate the four important local areas (eyes, nose and mouth) into their right
621    positions on the whole face image, which is already proved to be correct from our experiment
622    results. A majority of face recognition systems require a complete image to be recognized,
623    however, recovering the entire image is difficult when parts of the face are missing. This makes
624    it difficult to achieve good performance. We demonstrate how our approach can help enhance
625    face recognition by focusing on these areas of the face and outperform other methods in the same
626    context. Landmark detection is widely used in a variety of applications including object
627    detection, texture classification, image retrieval and etc. TP-GAN already has landmark detection
628    implemented for detecting the four face patches in its early stages as shown in Fig 1. Such a
629    method might be valuable in obtaining further textual details that can help in recovering those
630    facial areas and face shapes that have different characteristics. In our proposal, we offer a
631    relatively simple yet effective method for restoring the texture details of a synthesised face image
632    by repositioning the appearance areas of four landmark patches rather than the entire face. We
633    show a different method for facial recognition (Table 4) for faces with extreme poses. In four
634    major poses we achieve rank-1 recognition rates (75%, 60%, 45%, and 30%). Furthermore, when
635    it comes to global $\{G_{\theta_E^g}, G_{\theta_D^g}\}$, we found that optimizing for the corrupted images resulted in a

636    better convergence rate than optimizing for the clean images in order to achieve balanced
637    learning behavior. Therefore, we neither extensively modify the $\{G_{\theta_E^g}, G_{\theta_D^g}\}$, nor include an

638    external neural network in our modification, because that integration will increase the network
639    complexity and cause training limitations. We obtained better results by introducing noise to the
640    image before feeding it to the $\{G_{\theta_E^g}, G_{\theta_D^g}\}$ during optimization.

641

642    **Conclusion**

643    In this paper, we propose an LFM model for synthesizing a frontal-face image from a single
644    image to further enhance the frontal-face images quality of the TP-GAN model. To accomplish
645    our goal smoothly, we expand the existing generative global pathway with a well-constructed 2D
646    face landmark localization to cooperate with the local pathway structure in a landmark sharing
647    manner to incorporate empirical face pose into the learning process, and improve the encoder-
648    decoder global pathway structure for better facial image features representation. Compared with
649    TP-GAN, our method can generate frontal images with rich texture details and preserve the
650    identity information. Face landmark localization allows us to restore the missing information of
651    the real face image from the synthetic frontal images, and provide a rich texture detail. The
652    quantitative and qualitative experimental results of the Multi-PIE and FEI datasets show that our
653    proposed method can not only generate high-quality perceptual facial images in extreme poses,
654    but also significantly improves the TP-GAN results. Although LFMTP-GAN method achieves a
655    high-quality image resolution output, there is still room for improvement by choosing different

656   optimization algorithms, such as loss functions, or introducing some different techniques for
657   facial analysis and recognition. Our future research is to apply different error functions or
658   different face analysis and recognition techniques, combined with two pathway structures, to
659   achieve a super-resolution generative model and high-precision performance.
660

661   **References**

662   Junho Y, Heechul J, ByungIn Y, Changkyu C, Dusik P, Junmo K. 2015. Rotating Your Face
663   Using Multi-Task Deep Neural Network. In: 2015 IEEE Conference on Computer Vision and
664   Pattern Recognition (CVPR). Boston, MA, USA: 676 – 684 DOI:
665   10.1109/CVPR.2015.7298667.
666   Chao X, Xiaowei Z, Danhang T, Karlekar J, Shuicheng Y, Tae-Kyun K. 2015. Conditional
667   Convolutional Neural Network for Modality-aware Face Recognition. In: 2015 IEEE
668   International Conference on Computer Vision (ICCV). Santiago, Chile: 3667 – 3675 DOI:
669   10.1109/ICCV.2015.418.
670   Xi Y, Xiang Y, Kihyuk S, Xiaoming L, Manmohan C. 2017. Towards Large-Pose Face
671   Frontalization in the Wild. In: 2017 IEEE International Conference on Computer Vision
672   (ICCV). Venice, Italy: 3990 – 3999 DOI: 10.1109/ICCV.2017.430.
673   Jian Z, Yu C, Yan X, Lin X, Jianshu L, Fang Z, Karlekar J, Sugiri P, Shengmei S, Junliang X,
674   Shuicheng Y, Jiashi F. 2018. Towards Pose Invariant Face Recognition in the Wild. In: 2018
675   IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake, UT, USA:
676   2207 – 2216 DOI: 10.1109/CVPR.2018.00235.
677   Bassel Z, Ilya K, Yuri M. 2021. PFA-GAN: Pose Face Augmentation Based on Generative
678   Adversarial Network. Vilnius University, vol. 32, no. 2: 425 – 440 DOI:
679   https://doi.org/10.15388/21-INFOR443
680   Chenxu Z, Saifeng N, Zhipeng F, Hongbo L, Ming Z, Madhukar B, Xiaohu G. 2021. 3D
681   Talking Face with Personalized Pose Dynamics. In: IEEE Transactions on Visualization and
682   Computer Graphics. 1 – 25 DOI: 10.1109/TVCG.2021.3117484.
683   Yi Z, Keren F, Cong H, Peng C. 2021. Identity-and-pose-guided Generative Adversarial
684   Network for Face Rotation. 33 – 47 DOI: 10.1016/J.NEUCOM.2021.04.007.
685   Hang Z, Jihao L, Ziwei L, Yu L, Xiaogang. 2020. Rotate-and-Render: Unsupervised
686   Photorealistic Face Rotation from Single-View Images. In: 2020 IEEE/CVF Conference on
687   Computer Vision and Pattern Recognition (CVPR). Seattle, WA, USA: 5911 – 5920 DOI:
688   10.1109/CVPR42600.2020.00595.
689   Yibo H, Xiang W, Bing Y, Ran H, Zhenan S. 2018. Pose-Guided Photorealistic Face
690   Rotation. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt
691   Lake City, UT, USA: 18 – 23 DOI: 10.1109/CVPR.2018.00876.
692   Luan T, Xi Y, Xiaoming L. 2018. Representation Learning by Rotating Your Faces. IEEE
693   Transactions on Pattern Analysis and Machine Intelligence, vol. 41, ls. 12: 3007 – 3021 DOI:
694   10.1109/TPAMI.2018.2868350.

695    Rui H, Shu Z, Tianyu L, Ran H. 2017. Beyond Face Rotation: Global and Local Perception
696    GAN for Photorealistic and Identity Preserving Frontal View Synthesis. In: 2017 IEEE
697    International Conference on Computer Vision (ICCV). Venice, Italy: 2439 – 2448 DOI:
698    10.1109/ICCV.2017.267.
699    Soumyadip S, Jun-Cheng C, Carlos C, Vishal M. P, Rama C, David W J. 2016. Frontal to
700    Profile Face Verification in The Wild. In: 2016 IEEE Winter Conference on Applications of
701    Computer Vision (WACV). Lake Placid, NY, USA: 1 – 9 DOI:
702    10.1109/WACV.2016.7477558.
703    Peipei L, Xiang W, Yibo H, Ran H, Zhenan S. 2019. M2fpa: A multi-yaw Multi-Pitch High-
704    Quality Database and Benchmark for Facial Pose Analysis: 1 – 9 Available at
705    https://arxiv.org/abs/1904.00168
706    Meina K, Shiguang S, Hong C, Xilin C. 2014. Stacked Progressive Auto-Encoders (SPAE)
707    for Face Recognition Across Poses. In: 2014 IEEE Conference on Computer Vision and
708    Pattern Recognition. Columbus, OH, USA: 1883 – 1890 DOI: 10.1109/CVPR.2014.243.
709    Tal H, Shai H, Eran P, Roee E. 2015. Effective Face Frontalization in Unconstrained Images.
710    In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston,
711    USA: 1 – 10 Available at https://arxiv.org/abs/1411.7964
712    Christos S, Yannis P, Stefanos Z, Maja P. 2015. Robust Statistical Face Frontalization. In:
713    2015 IEEE International Conference on Computer Vision (ICCV). Santiago, Chile: 3871 –
714    3879 DOI: 10.1109/ICCV.2015.441.
715    Omkar M P, Andrea V, Andrew Z. 2015. Deep Face Recognition. In: Proceeding of the
716    British Machine Vision Conference (BMVC). England, UK: 1 – 12 DOI: 10.5244/C.29.41.
717    BMVA Press.
718    Florian S, Dmitry K, James P. 2015. FaceNet: A Unified Embedding for Face Recognition
719    and Clustering. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition
720    (CVPR). Boston, MA, USA: 1 – 10 DOI: 10.1109/CVPR.2015.7298682.
721    John G D. 1985. Uncertainty Relation for Resolution in Space, Spatial Frequency, and
722    Orientation Optimized by Two-dimensional Visual Cortical Filters. Journal of Optical
723    American, vol. 2, ls. 7: 1160 – 1169 DOI: 10.1364/josaa.2.001160.
724    David – G Lowe. 1999. Object Recognition from Local Scale-invariant Features. In:
725    Proceeding of the 7th IEEE International Conference on Computer Vision. Kerkyra, Greece:
726    1150 – 1157 DOI: 10.1109/ICCV.1999.790410.
727    Ahonen T, Hadid A, Pietikäinen M. 2006. Face Description with Local Binary Patterns:
728    Application to Face Recognition. IEEE Transaction on Pattern Analysis and Machine
729    Intelligence, vol. 28, no. 12: 2037 – 2041 DOI: 10.1109/TPAMI.2006.244.
730    Dalal N, Triggs B. 2005. Histograms of Oriented Gradients for Human Detection. In: 2005
731    IEEE Computer Society Conference on Computer Vision and Pattern Recognition
732    (CVPR'05). San Diego, CA, USA: 886 – 893 DOI: 10.1109/CVPR.2005.177.

733  Kilian Q W, Lawrence K S. 2009. Distance Metric Learning for Large Margin Nearest
734  Neighbor Classification. Journal of Machine Learning Research (2009) 207-244, vol. 10: 1 –
735  38.

736  Tsai-Wen C, Trevor J W, Yi S, Stefan R P, Sabine L R, Amy B, Eric R S, Rex A K, Michael
737  B O, Vivek J, Loren L Karel S, Douglas S K. 2013. Ultrasensitive Fluorescent Proteins for
738  Imaging Neuronal Activity. Journal Natural Research, vol. 499, Is. 7458: 295 – 300 DOI:
739  10.1038/nature12354.

740  Yaniv T, Ming Y, Marc'Aurelio R, Lior W. 2014. DeepFace: Closing the Gap to Human-
741  Level Performance in Face Verification. In: 2014 IEEE Conference on Computer Vision and
742  Pattern Recognition (CVPR). Columbus, OH, USA: 1 – 8 DOI: 10.1109/CVPR.2014.220.

743  Amin J, Xiaoming L. 2017. Pose-invariant Face Alignment via CNN-based Dense 3D Model
744  Fitting. International Journal of Computer Vision, vol. 125, no. 2: 187 – 203 DOI:
745  10.1109/CVPR.2016.454.

746  Kaipeng Z, Zhanpeng Z, Zhifeng L, Yu Q. 2016. Joint Face Detection and Alignment using
747  Multi-Task Cascaded Convolutional Networks. In: Proceeding of the 4th International
748  Conference on Information Science and Control Engineering (ICISCE). Changsha, China: 1–
749  5 DOI: 10.1109/ICISCE.2017.95

750  Jimei Y, Scott R, Ming-Hsuan Y, Honglak L. 2016. Weakly-supervised Disentangling with
751  Recurrent Transformations for 3D View Synthesis. In: Proceeding of the 28th International
752  Conference on Neural Information Processing Systems, vol. 1: 1099 – 1107.

753  Luan T, Xi Y, Xiaoming L. 2017. Disentangled Representation Learning GAN for Pose –
754  invariant Face Recognition. In: 2017 IEEE Conference on Computer Vision and Pattern
755  Recognition (CVPR). Honolulu, HI, USA: 1415 – 1424 DOI: 10.1109/CVPR.2017.141.

756  Xiang W, Ran H, Zhenan S, Tieniu T. 2015. A light CNN for Deep Face Representation with
757  Noisy Labels. IEEE Transactions on Information Forensics and Security, vol. 13, ls. 11: 2884
758  – 2896 DOI: 10.1109/TIFS.2018.2833032.

759  Yu T, Xi P, Long Z, Shaoting Z, Dimitris N M. 2018. CR-GAN: Learning Complete
760  Representations for Multi-View Generation. In: Proceeding of the 27th International Joint
761  Conference on Artificial Intelligence (IJCAI-18): 942 – 948 Available at
762  https://arxiv.org/abs/1806.11191v1

763  Yu Y, Songyao J, Joseph P R, Yun F. 2020. Dual-Attention GAN for Large-pose Face
764  Frontalization. In: 2020 15th IEEE International Conference on Automatic Face and Gesture
765  Recognition (FG2020): 24 – 31 DOI: 10.1109/FG47880.2020.00004.

766  Ian J G, Jean P A, Mehdi M, Bing X, David W F, Sherjil O, Aaron C, Yoshua B. 2014.
767  Generative Adversarial Networks. In: International Conference on Neural Information
768  Processing Systems (NIPS): 2672 – 2680 Available at https://arxiv.org/abs/1406.2661

769  Christian L, Lucas T, Ferenc H, Jose C, Andrew C, Alejandro A, Andrew A, Alykhan T,
770  Johannes T, Zehan W, Wenzhe S. 2017. Photo-Realistic Single Image Super-Resolution
771  Using a Generative Adversarial Network. In: 2017 IEEE Conference on Computer Vision and
772  Pattern Recognition (CVPR). Honolulu, HI, USA: 1 – 19 DOI: 10.1109/CVPR.2017.19.

773 Alec R, Luke M, Soumith C. 2016. Unsupervised Representation Learning with Deep
774 Convolutional Generative Adversarial Networks: 1 – 16 Available at
775 https://arxiv.org/pdf/1511.06434.pdf

776 Mehdi M, Simon O. 2014. Conditional Generative Adversarial Nets: 1411 – 1784 Available
777 at https://arxiv.org/abs/1411.1784

778 Augustus O, Christopher O, Jonathon S. 2017. Conditional Image Synthesis with Auxiliary
779 Classifier GANs. In: Proceeding of the 34th International Conference on Machine Learning,
780 (PMLR 70:2642-2651), vol. 70. Sydney, Australia: 2642 – 2651 Available at
781 https://arxiv.org/abs/1610.09585

782 David B, Thomas S, Luke M. 2017. BEGAN: Boundary Equilibrium Generative Adversarial
783 Networks: 1 – 10 Available at https://arxiv.org/abs/1703.10717

784 Anthreas A, Amos S, Harrison E. 2018. Data Augmentation Generative Adversarial
785 Networks: 1 – 14 Available at https://arxiv.org/abs/1711.04340

786 Pascal V, Hugo L, Yoshua B, Pierre-Antoine M. 2008. Extracting and Composing Robust
787 Features with Denoising Autoencoders. In: Proceeding of the 25th International Conference on
788 Machine Learning, Helsinki, Finland: 1 – 8 DOI: 10.1145/1390156.1390294.

789 Jun-Yan Z, Taesung P, Phillip I, Alexei A E. 2017. Unpaired Image-to-Image Translation
790 Using Cycle – Consistent Adversarial Networks. In: 2017 IEEE International Conference on
791 Computer Vision (ICCV). Venice, Italy: 1 – 18 DOI: 10.1109/ICCV.2017.244.

792 Xiangyu Z, Zhen L, Junjie Y, Dong Y, Stan Z L. 2015. High-fidelity Pose and Expression
793 Normalization for Face Recognition in The Wild. In: 2015 IEEE Conference on Computer
794 Vision and Pattern Recognition (CVPR). Boston, MA, USA: 787 – 796 DOI:
795 10.1109/CVPR.2015.7298679.

796 Li L, Paul F, Guoying Z, Matti P. 2014. Extended Local Binary Pattern Fusion for Face
797 Recognition. In: 2014 IEEE International Conference on Image Processing (ICIP). Paris,
798 France: 718 – 722 DOI: 10.1109/ICIP.2014.7025144.

799 Xuehan X, Fernando De la T. 2013. Supervised Descent Method and Its Applications to Face
800 Alignment. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
801 Portland, OR, USA: 1 – 8 DOI: 10.1109/CVPR.2013.75.

802 Georgios T. 2015. Project-out Cascaded Regression with an Application to Face Alignment.
803 In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston,
804 MA, USA: 3659 – 3667 DOI: 10.1109/CVPR.2015.7298989.

805 Sanjeev A, Andrej R, Yi Z. 2018. Do GANs Learn the Distribution? Some Theory and
806 Empirics. In: Proceeding of 6th International Conference on Learning Representations (ICLR
807 2018): 1 – 11 Available at https://openreview.net/forum?id=BJehNfW0-

808 Cong H, Zhen-Hua F, Xiao-Jun W, Josef K. 2020. Dual Encoder-Decoder based Generative
809 Adversarial Networks for Disentangled Facial Representation Learning. IEEE Transactions,
810 vol. 8: 130159 – 130171 DOI: 10.1109/ACCESS.2020.3009512

811 Kevin R, Aurelien L, Sebastian N, Thomas H. 2017. Stabilizing Training of Generative
812 Adversarial Networks Through Regularization. In: Proceeding of the 31st Conference on

813     Neural Information Processing Systems (NIPS 2017). Long Beach, CA, USA: 1 – 11

814     Available at https://arxiv.org/abs/1705.09367

815     Adrian B, Georgios T. 2017. How Far are we from Solving The 2D & 3D Face Alignment

816     Problem? (and a Dataset of 230,000 3D Facial Landmarks). In: 2017 IEEE International

817     Conference on Computer Vision (ICCV). Venice, Italy: 1021 – 1030 DOI:

818     10.1109/ICCV.2017.116.

819     Xing D, Vishwanath A. Sindagi, Vishal M-P. 2018. GP-GAN: Gender Preserving GAN for

820     Synthesizing Faces from Landmarks: 1 – 7 Available at https://arxiv.org/pdf/1710.00962.pdf

821     Justin J, Alexandre A, Li F-F. 2016. Perceptual Losses for Real-time Style Transfer and

822     Super-Resolution. In: Proceeding of 14th European Conference on Computer Vision.

823     Amsterdam, Netherlands: 1 – 17 Available at https://arxiv.org/abs/1603.08155

824     Ralph G M, Jeffrey C, Takeo K, Simon B. 2010. Multi-PIE. Image and Vision Computing. In:

825     Proceeding International Conference Auto Face Gesture Recognition, vol. 28, no. 5: 1 – 23

826     DOI: 10.1016/j.imavis.2009.08.002.

827     FEI. 2005-2006. Artificial Intelligence Laboratory of FEI, São Bernardo do Campo, São

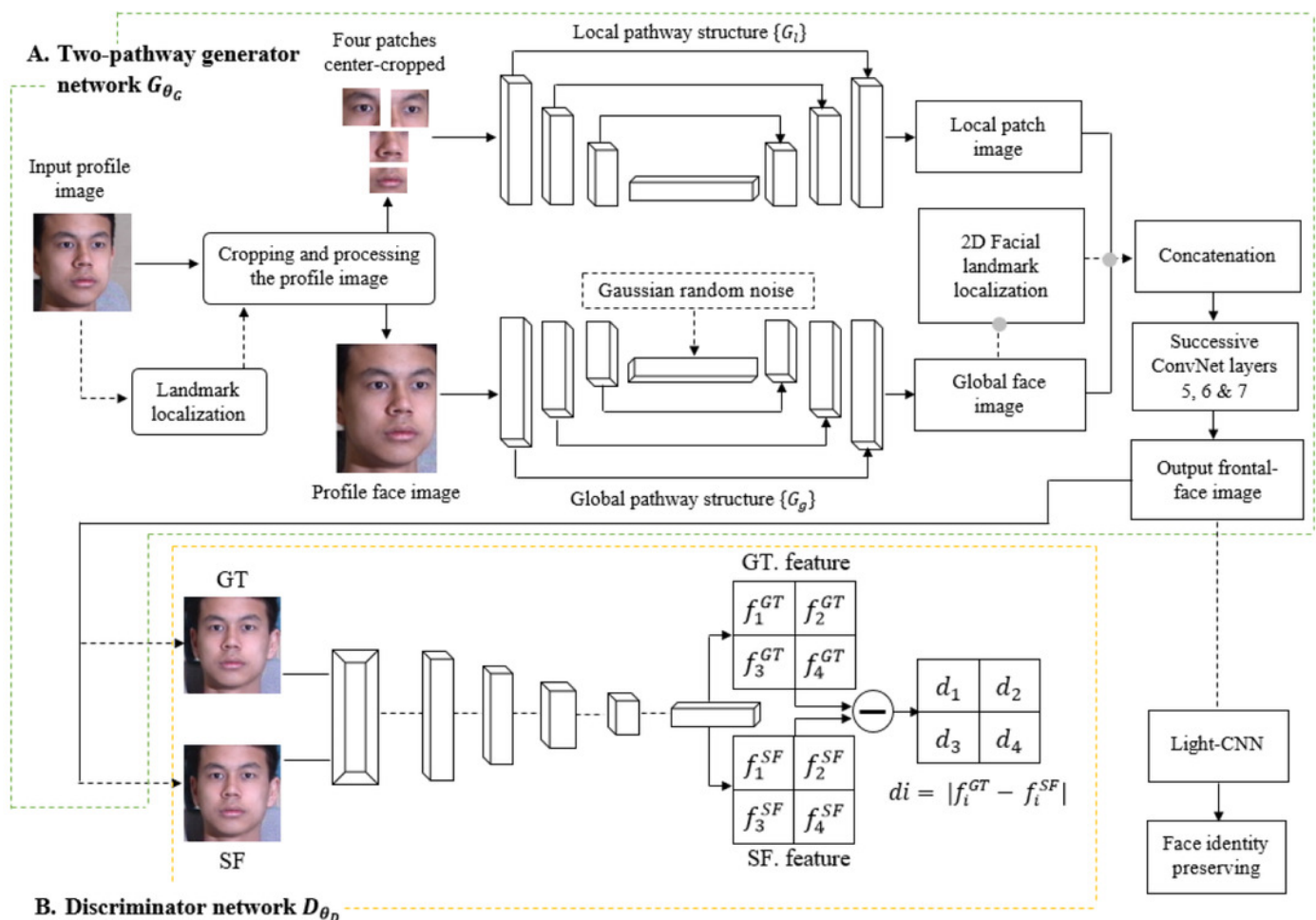
828     Paulo, Brazil: Available at https://fei.edu.br/~cet/facedatabase.html

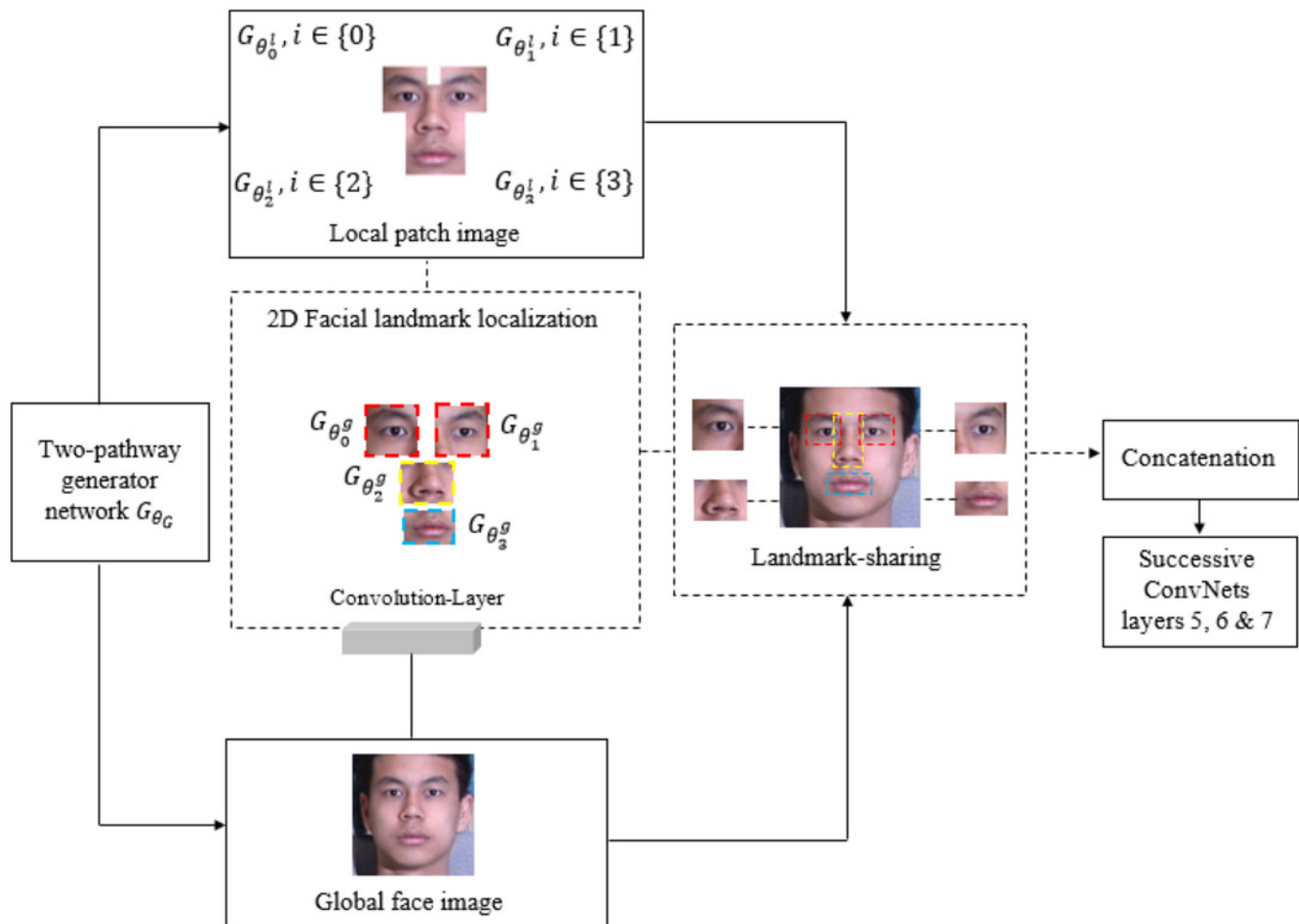
829     Microsoft Celeb (MS-Celeb-1M). 2016: Available at https://www.microsoft.com/en-

830     us/research/project/ms-celeb-1m-challenge-recognizing-one-million-celebrities-real-world

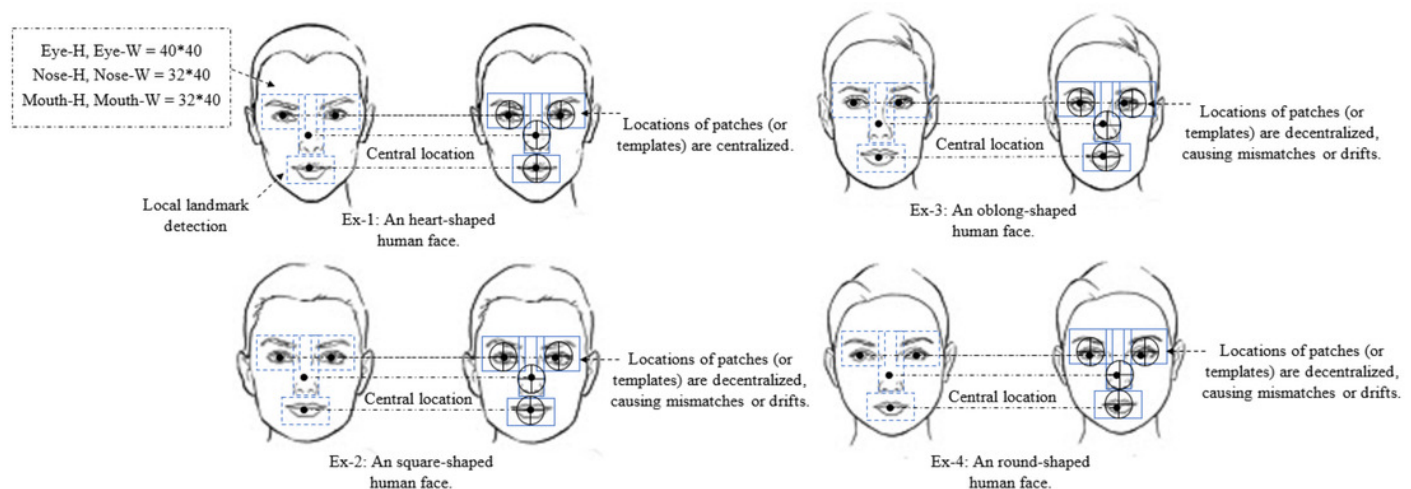
# Figure 1

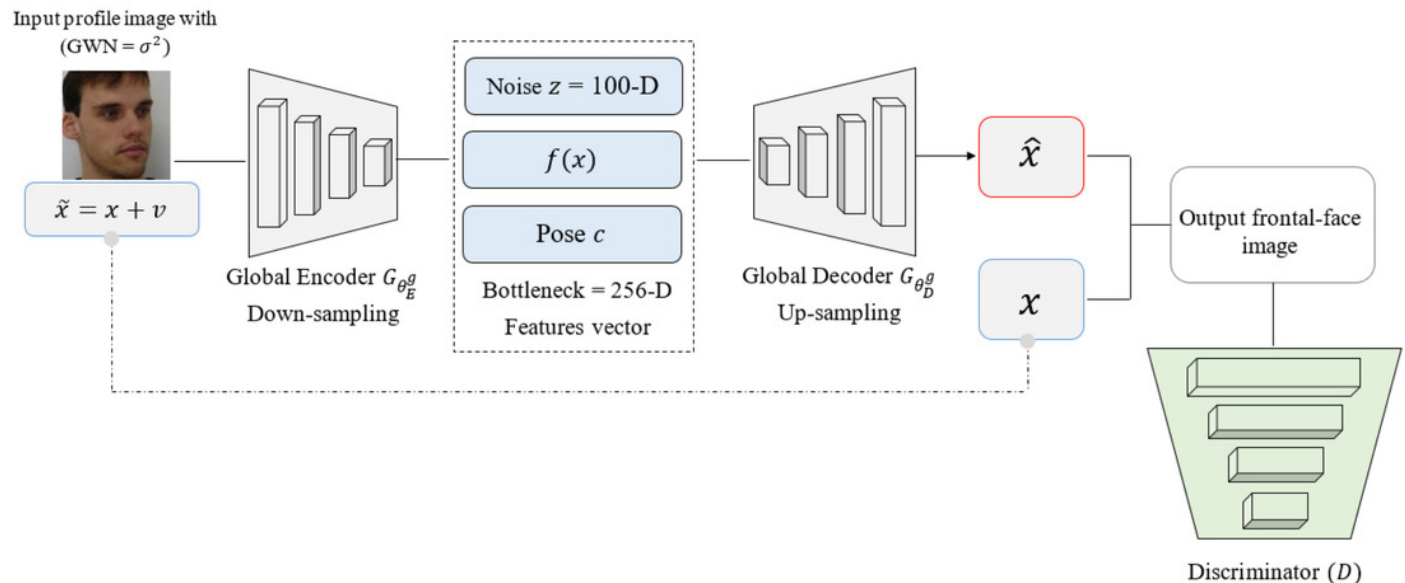
The structure of TP-GAN.

The final output was obtained by integrating the global pathway with a 2D facial landmark localization to collaborate with the local pathway in a landmark sharing fashion.

# Figure 2

The structure of the 2D facial landmark localization method.

# Figure 3

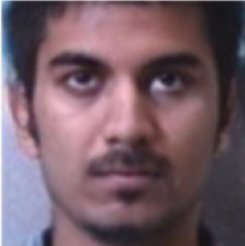
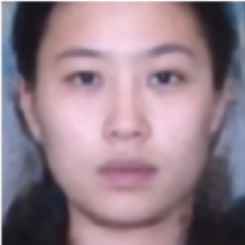
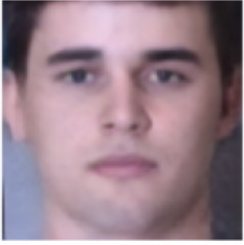
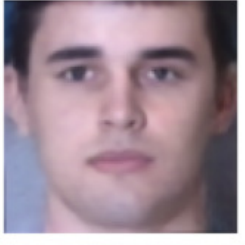
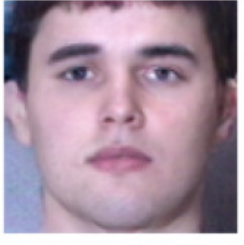
The detection of different facial shapes characteristics.

Example1., The appearance of the face is very accurate because all of the generated local synthesis patches (or templates) are centralized. In examples 2, 3, and 4 the appearance of the face is inaccurate since all of the generated local synthesis patches (or templates) are decentralized, resulting in mismatches or drifts. Faces in examples 2, 3, and 4 can be treated using our method regardless of their shape.

# Figure 4

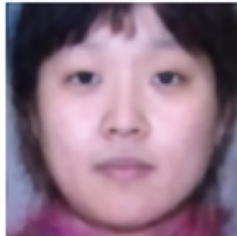
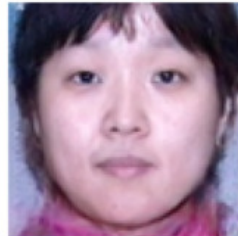
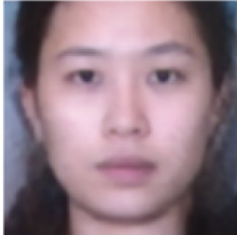
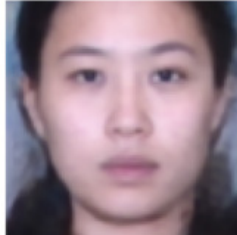
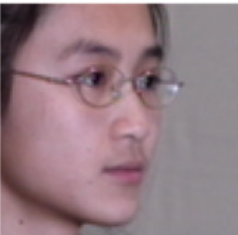
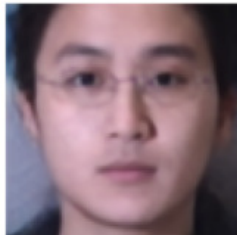
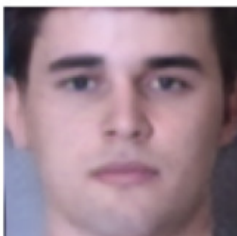
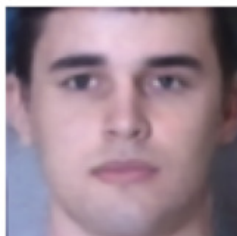
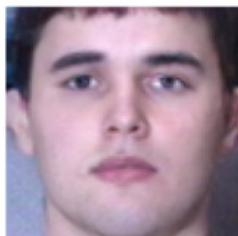
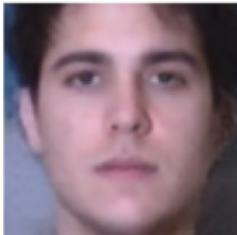
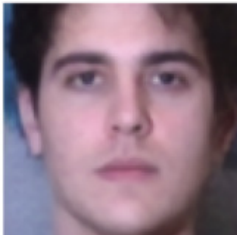
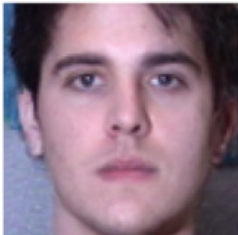
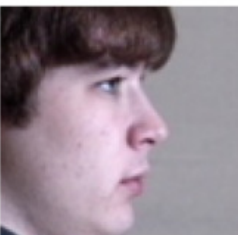
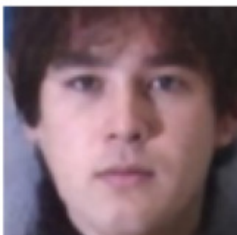
The global pathway structure. The structure is based on encoder down-sampling and decoder up-sampling. The bottleneck with 256-D features vector remains the same.

# Figure 5

Examples of different levels of illumination.

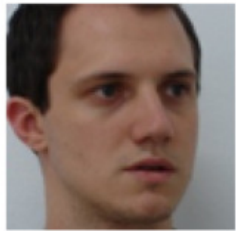
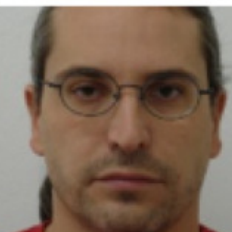
The level of illumination, including brightness, exposure, contrast, and shadows and etc.

Some quality effects can also be observed, such as sharpness, smoothness, blurriness, etc.

| Pose range | Profile image | TP-GAN | LFMTP-GAN | Ground-Truth |
|---|---|---|---|---|
| $15^0$ | | | | |
| $45^0$ | | | | |
| $60^0$ | | | | |
| $75^0$ | | | | |
| $90^0$ | | | | |

# Figure 6

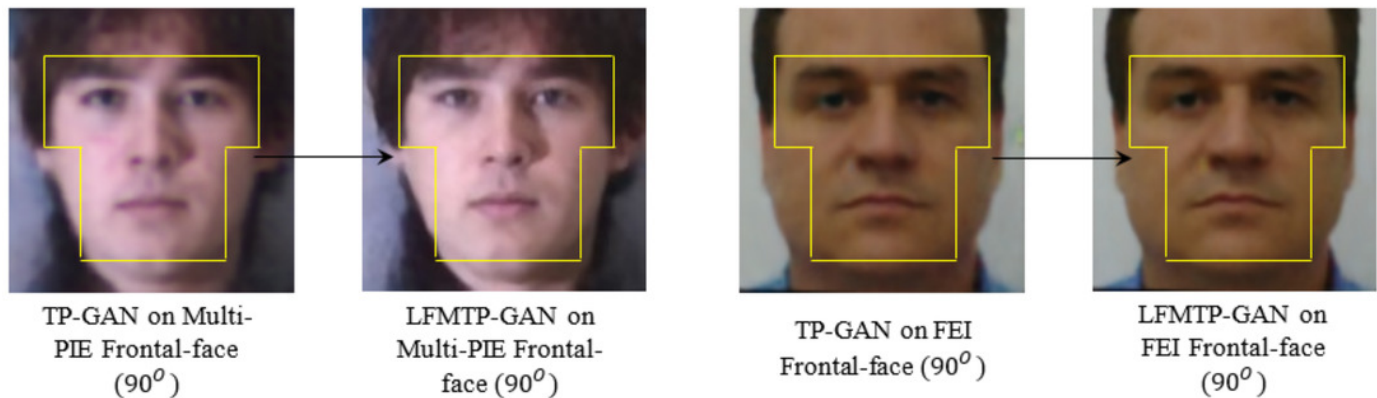
Comparison of TP-GAN and LFMTP-GAN by synthesizing frontal-face images under different poses and illuminations on Multi-PIE face database. It can be seen that LFMTPGAN generates better visual quality images.

| Pose range | Profile image | TP-GAN | LFMTP-GAN | Ground-Truth |
|---|---|---|---|---|
| $15^0$ | | | | |
| $30^0$ | | | | |
| $45^0$ | | | | |
| $60^0$ | | | | |
| $75^0$ | | | | |
| $90^0$ | | | | |

# Figure 7

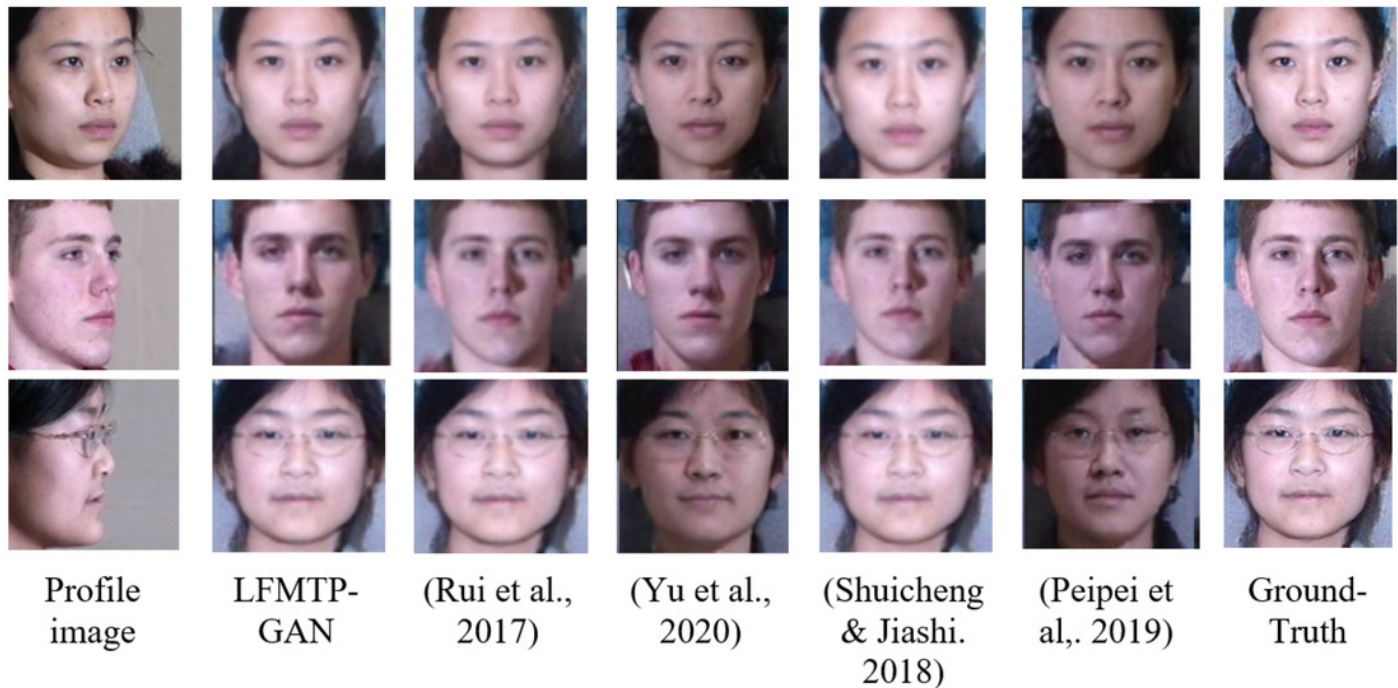
Comparison of TP-GAN and LFMTP-GAN by synthesizing frontal-face images under different poses and illuminations on FEI face database. It can be seen that LFMTP-GAN generates better visual quality images with less blurry visualization effect.

| Pose range | Profile image | TP-GAN | LFMTP-GAN | Ground-Truth |
|---|---|---|---|---|
| $15^0$ | | | | |
| $30^0$ | | | | |
| $45^0$ | | | | |
| $60^0$ | | | | |
| $75^0$ | | | | |
| $90^0$ | | | | |

# Figure 8

A comparison of the resolution of facial images taken under $90^{0}$ face pose, and illuminations condition. The facial visualization area is inside the yellow map.

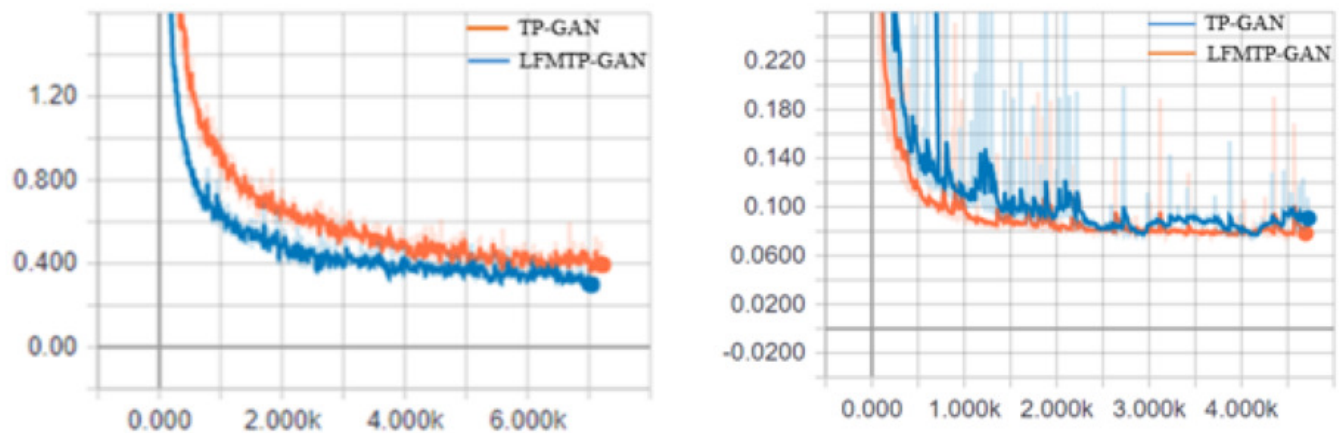
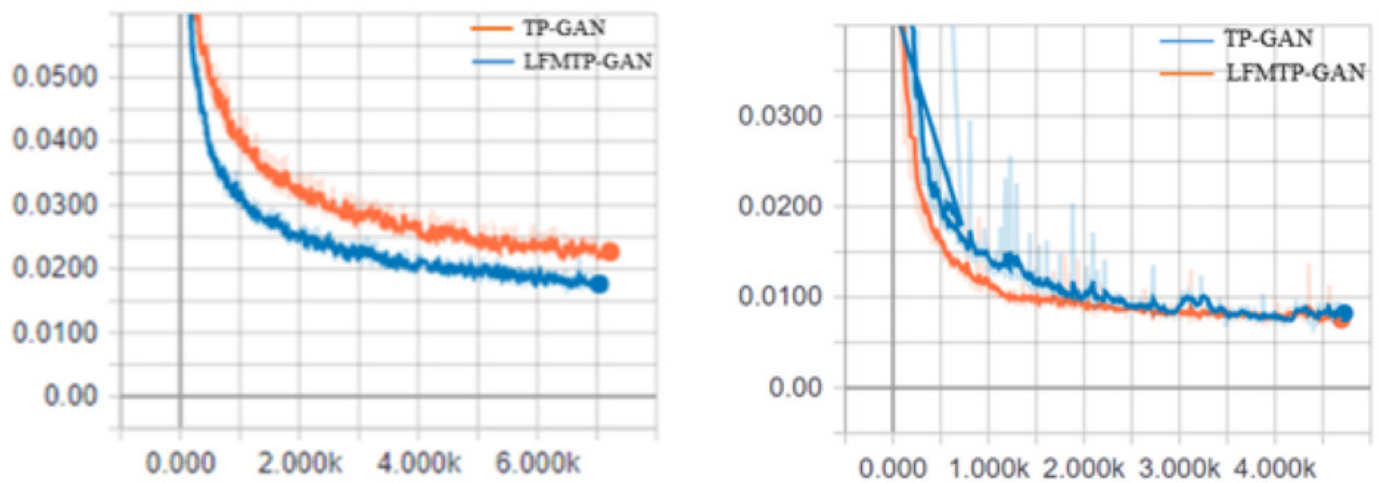
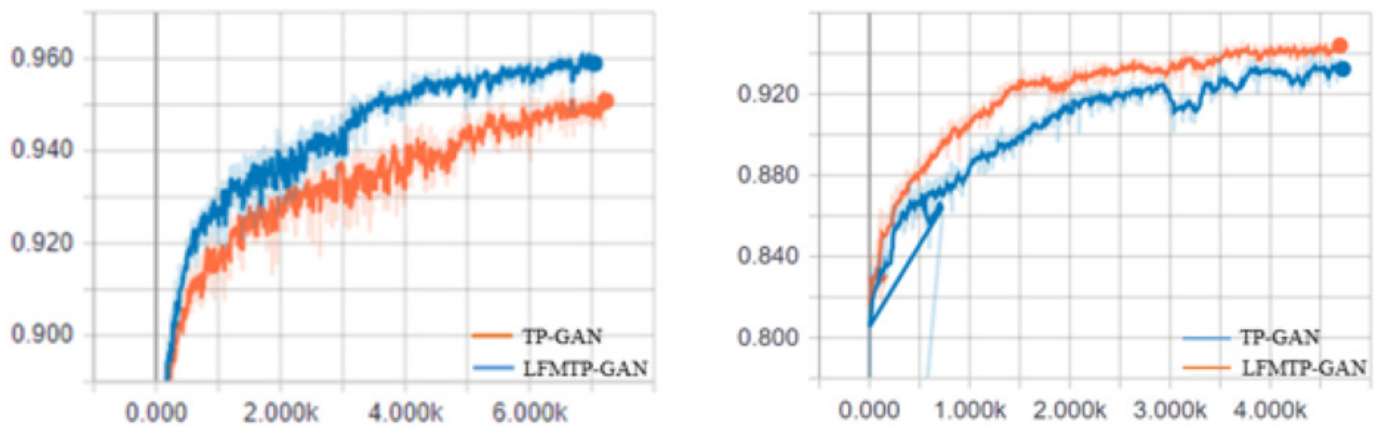
TP-GAN on Multi-PIE Frontal-face $(90^{0})$    LFMTP-GAN on Multi-PIE Frontal-face $(90^{0})$    TP-GAN on FEI Frontal-face $(90^{0})$    LFMTP-GAN on FEI Frontal-face $(90^{0})$

# Figure 9

A comparison of LFMTP-GAN algorithms with different techniques for face frontalization on Multi-PIE dataset.

| Profile image | LFMTP-GAN | (Rui et al., 2017) | (Yu et al., 2020) | (Shuicheng & Jiashi. 2018) | (Peipei et al,. 2019) | Ground-Truth |
|---|---|---|---|---|---|---|

# Figure 10

The TP-GAN and LFMTP-GAN loss curve plots.

Where A is the generator loss curve, and B is the pixel-wise loss curve. The horizontal axis indicates the number of epochs, which is the number of times that entire training data has been trained. The vertical axis indicates how well the model performed after each epoch; the lower the loss, the better a model. C is the identity preserving accuracy curve, which is a quality metric that measures how accurate it is to preserve a user's identity; the higher the accuracy, the better a model.

## A   Generator loss

## B   Pixel-wise loss

## C   Identity preserving accuracy

An overview of the workflow of the 2D facial landmarks localization including all its operational steps. These steps are the key to our method that results in a successful implementation.

| Steps | Process |
|---|---|
| 2D landmark detection<br><br>↓<br><br>Convolutional neural network<br><br>↓<br><br>Data augmentation<br><br>↓<br><br>5, 6 and 7 layers | 1- Face Detection: the goal of this step is to identify faces that are generated by local and global pathways.<br>2- Facial landmarks such as the eye centers, tip of the nose, and mouth are located.<br>3- The feature extractor encodes identity information into a high-dimension descriptor. |
| | 4- The purpose of this layer is to coordinate and extract intermediate features. |
| | 5- This technique is used to enhance the synthesis image textures detail by adding slightly modified copies of already existing data or by creating new synthetic data based on existing data. |
| | 6- 5, 6 and 7 act as a visual feature map for specific inputs of fontal-face images in order to retain more visual information by subsampling layers' structure. |

2

The structure and parameters setting in the LFMTP-GAN. Those values contribute to the learning process and help our model to achieve its goals.

| Parameters name | Corresponding value |
|---|---|
| Batch Size $(G_{\theta_G})$ | 4 |
| Batch Size $(D_{\theta_D})$ | 10 |
| Epoch Steps | 4500 + |
| Learning Rate $(lr)$ | 0.01 |
| Adversarial Loss $(L_{adv})$ | 0.005 |

Cross-validation of face recognition rates on Multi-PIE and FEI datasets using weights that coordinate different loss value parameters and Table 1 settings. The run time is measured in minutes and indicates how long it will take for the model to run with

The run time is measured in minutes and indicates how long it will take for the model to run with more than 70 epochs.

| Multi-PIE Dataset | | | | |
|---|---|---|---|---|
| Method | Train | Valid | Cross-valid ($L_s$) | Time (m) |
| TP-GAN | 95.2 % | 91.7 % | 0.165 | 373 |
| LFMTP-GAN | 96.1 % | 92.7 % | 0.164 | 290 |
| FEI Dataset | | | | |
| Method | Train | Valid | Cross-valid ($L_s$) | Time (m) |
| TP-GAN | 94.0 % | 90.0 % | 0.179 | 303 |
| LFMTP-GAN | 95.0 % | 91.3 % | 0.164 | 315 |

2

The recognition rate (%) across views and illuminations based on the Multi-PIE dataset under Table 1 settings. This table compares different methods of facial recognition. Our method is capable of outperforming and achieving a better result.

This table compares different methods of facial recognition. Our method is capable of outperforming and achieving a better result.

| Method \ Angle | LFMTP-GAN | (Rui et al., 2017) | (Luan, Xi & Xiaoming. 2017) | (Yibo et al., 2018) | (Chao et al., 2015) | (Xiang et al., 2015) | (Junho et al., 2015) |
|---|---|---|---|---|---|---|---|
| $\pm\,90^0$ | 65.23 % | 64.03 % | --- | **66.05 %** | 47.26 % | 5.51 % | --- |
| $\pm\,75^0$ | **85.30 %** | 84.10 % | --- | 83.05 % | 60.66 % | 24.18 % | --- |
| $\pm\,60^0$ | **94.13 %** | 92.93 % | 83.2 % | 90.63 % | 74.38 % | 62.09 % | --- |
| $\pm\,45^0$ | **98.80 %** | 98.58 % | 86.2 % | 97.33 % | 89.02 % | 92.13 % | 71.65 % |
| $\pm\,30^0$ | **99.88 %** | 99.85 % | 90.1 % | 99.65 % | 94.05 % | 97.38 % | 81.05 % |
| $\pm\,15^0$ | 99.80 % | 99.78 % | 94.0 % | **99.82 %** | 96.97 % | 98.68 % | 89.45 % |
| Average ACC | **90.523 %** | 89.878 % | 88.375 % | 89.421 % | 77.056 % | 63.328 % | 80.716 % |