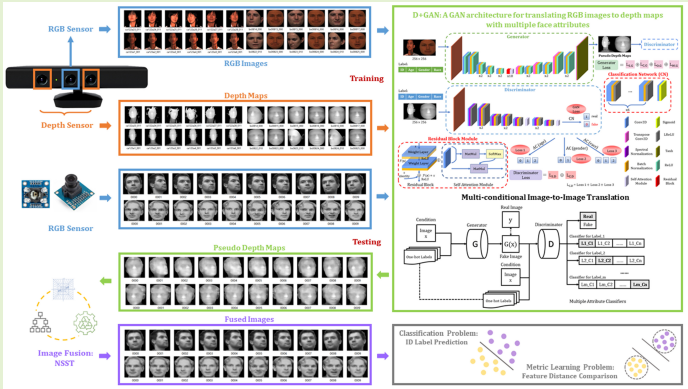
Pseudo RGB-D Face Recognition

Bo Jin, Leandro Cruz, and Nuno Gonçalves, *Member, IEEE*

***Abstract*— In the last decade, advances and popularity of low cost RGB-D sensors have enabled us to acquire depth information of objects. Consequently, researchers began to solve face recognition problems by capturing RGB-D face images using these sensors. Until now, it is not easy to acquire the depth of human faces because of limitations imposed by privacy policies, and RGB face images are still more common. Therefore, obtaining the depth map directly from the corresponding RGB image could be helpful to improve the performance of subsequent face processing tasks such as face recognition. Intelligent creatures can use a large amount of experience to obtain three-dimensional spatial information only from two-dimensional plane scenes. It is machine learning methodology which is to solve such problems that can teach computers to generate correct answers by training. To replace the depth sensors by generated pseudo depth maps, in this paper, we propose a pseudo RGB-D face recognition framework and provide data-driven ways to generate the depth maps from 2D face images. Specially, we design and implement a generative adversarial network model named "D+GAN" to perform the multi-conditional image-to-image translation with face attributes. By this means, we validate the pseudo RGB-D face recognition with experiments on various datasets. With the cooperation of image fusion technologies, especially Non-subsampled Shearlet Transform, the accuracy of face recognition has been significantly improved.**



***Index Terms*— RGB-D, face recognition, D+GAN, pseudo depth, monocular face depth estimation**

I. Introduction

DARWIN'S theory of evolution proposes natural selection which is the process of the survival of the fittest, and the elimination of the others [1]. The genetic characteristics of organisms that adapt to the environment can be preserved through natural selection, which is based on sufficient facts and has a profound effect in academic research. Nowadays, all living higher creatures have two eyes for three-dimensional positioning which is vital for foraging. In contrast, most one-eyed creatures are extinct. Human beings can still perform 3D positioning with one eye in a period of time because of a large amount of previous experience.

In recent decades, biometrics has attracted the attention of researchers because of its uniqueness, stability, versatility and difficulty to counterfeit. Because of its non-invasiveness, face recognition has become the most user-friendly biometric method, which leads to its wide applications [2] [3] [4]. However, the accuracy of RGB face recognition is commonly affected by many factors, such as lighting conditions, age, head pose variations, etc. The human vision is three-dimensional, by contrast, 2D face images that are most common lack face space stereo information. There is no doubt about the importance of facial spatial information [5]. In recent years, advances and popularity of inexpensive RGB-D sensors enable us to utilize three-dimensional information. Comparing with RGB face recognition, RGB-D face recognition which requires depth images captured by depth sensors such as Kinect [6] and PrimeSense [7] performs better in accuracy due to the effective use of spatial features [8] [9]. In modern society, although facial recognition systems are very convenient, they also give rise to many information security and privacy issues. In addition, there are no popular file formats for RGB-D data, and not as many RGB-D cameras as RGB cameras. Therefore, RGB-D face images are not easy to collect and are much less common than RGB face images.

The emergence of machine learning allows computers to imitate the human learning process to learn from historical experience to make speculations. It occurs to us that probably by utilizing machine learning algorithms we can get the models to predict the depth map from its corresponding RGB image effectively. With the development of big data and the

This work is supported by the Fundação para a Ciência e a Tecnologia (FCT) under the Project UIDB/00048/2020.

Bo Jin is with the Institute of Systems and Robotics, Department of Electrical and Computer Engineering, University of Coimbra, Coimbra 3030-290, Portugal (e-mail: jin.bo@isr.uc.pt).

Leandro Cruz is with the Institute of Systems and Robotics, Department of Electrical and Computer Engineering, University of Coimbra, Coimbra 3030-290, Portugal, and also with the Align Technology, Inc., San Jose, California 95134, United States (e-mail: lmvcruz@gmail.com).

Nuno Gonçalves is with the Institute of Systems and Robotics, Department of Electrical and Computer Engineering, University of Coimbra, Coimbra 3030-290, Portugal, and also with the Portuguese Mint and Official Printing Office (INCM), Lisbon 1000-042, Portugal (e-mail: nunogon@deec.uc.pt).

improvement of computer hardware performance, the deep learning technology that has been widely used in science and industry in recent years has more powerful reasoning performance than traditional machine learning models. So monocular depth estimation inspired us to acquire 3D information from 2D face images by deep learning. Synthesizing the above, the thought behind this paper is to generate the corresponding depth map only from the RGB face image to replace the depth map collected by the depth sensor to perform the pseudo RGB-D face recognition.

In this paper, our contributions could be summarized as follows:

1) We definitely propose and validate a pseudo RGB-D face recognition framework shown in Figure 1. Figure 1 presents a modular process. Algorithms within the module lists can be selected for preprocessing, depth-generating, image fusion and feature extraction, and therefore be combined for face recognition. The best embodiment found is provided.
2) In order to make full use of face attributes, we emphatically propose a GAN based model, D+GAN, to perform the multi-conditional image-to-image translation for transforming RGB face images to corresponding depth maps with face attribute labels.
3) Based on the obtained depth maps, we improve the face recognition performance in cooperation with image fusion technologies, especially the Non-subsampled Shearlet Transform (NSST).

The remaining of this article is organized as follows: In Chapter 2, we review the related work. In Chapter 3, we describe our proposed methods and their implementations. Our experimental results are analyzed and discussed in Chapter 4. In Chapter 5, we make a conclusion and describe a research direction for the future.

II. Related Work

Face recognition refers to the technology of identifying or verifying the identity of subjects from faces in images or videos. The history of face recognition algorithms can be traced back to the 1970s. Traditional machine learning method is to extract hand-crafted features which are designed by specialists to reduce the complexity of input data, and train a model from the input to discover the pattern to make decisions. Matthew Turk and Alex Pentland, proposed Eigenfaces method for face recognition on a smaller set of face image features approximating the set of known face images [10]. Marian Stewart Bartlett et al. proposed using the Independent Component Analysis (ICA) method for face recognition, and they showed that ICA representations were superior to Principal Components Analysis (PCA) based representations for face recognition across changes in some conditions [11]. P. Jonathon Phillips, developed a Support Vector Machine (SVM) based algorithm to generate the decision surface for face recognition [12]. In the past ten years, traditional machine learning methods have increasingly been replaced by deep learning methods based on the convolutional neural network (CNN) in face recognition. The CNN structures mainly used in face recognition are basically consistent with the ones for classification tasks in ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [13]. In order to adapt to the task of face recognition, researchers mainly focus on discovering better training loss functions. Alex Krizhevsky, Ilya Sutskever and Geoffrey E. Hinton, proposed AlexNet which is a classic CNN framework to classify a large amount of images in ILSVRC-2010 [14]. Yaniv Taigman et al. presented a DeepFace system which can reach human level performance in face recognition [15]. The backbone network of DeepFace is based on AlexNet, and the loss function used is Softmax. Christian Szegedy et al. proposed a 22-layer deep convolutional neural network GoogLeNet which is a variant of the Inception Network [16]. Florian Schroff et al. presented FaceNet which uses GoogLeNet as backbone network and the triplet loss function for training to map the face image to the Euclidean space directly [17]. Kaiming He et al. proposed ResNet which can increase the network depth to 152 layers by using residual blocks [18]. Jiankang Deng et al. presented an Additive Angular Margin Loss function aiming to enhance the discriminative power of feature embeddings learned, which could get the state of the art result for face recognition by coordinating with ResNet [19].

Similarly, in the field of RGB-D face recognition research, in recent years, researchers have used deep neural networks with CNN structures to extract face depth map features. Yuancheng Lee et al. used a 12-layer deep neural network which is firstly trained with a color face dataset, and later fine-tuned on depth face images for feature extraction to perform joint classification [9]. Donghyun Kim et al. applied a fine-tuned DCNN to extract features from 2D depth maps converted from 3D point clouds for calculating the distance for face matching [20]. Moreover, Luo Jiang, Juyong Zhang, and Bailin Deng tried to propose an attribute-aware loss function for RGB-D facial data [21].

Depth estimation to obtain a representation of the spatial structure of objects plays a crucial role in navigation, robotics, and augmented reality for inferring scene geometry from 2D images. Researchers have applied machine learning methods to estimate the depth of human faces from monocular images since the 1990s. Shang-Hong Lai, Chang-Wu Fu and Shyang Chang estimated the depth from defocus by using the raw image data in the vicinity of the edge [22]. Zhan-Li Sun, and Kin-Man Lam converted depth estimation into an independent component analysis (ICA) problem by incorporating a prior from the CANDIDE 3-D face model [23]. Zhan-Li Sun, Kin-Man Lam, and Qing-Wei Gao employed the nonlinear least-squares model to estimate the depth values of facial feature points and the pose of the 2D face image [24]. Since 2014, with the development of deep learning, researchers have successively used deep learning methods to perform monocular face depth estimation, which is similar with face recognition. Jiyun Cui et al. presented a deep neural network with a cascaded FCN and CNN architecture to estimate depth information of RGB face images [25]. Stefano Pini et al. applied a conditional Generative Adversarial Network (cGAN) for learning to translate intensity face images into their corresponding depth maps [26]. Abdullah Taha Arslan

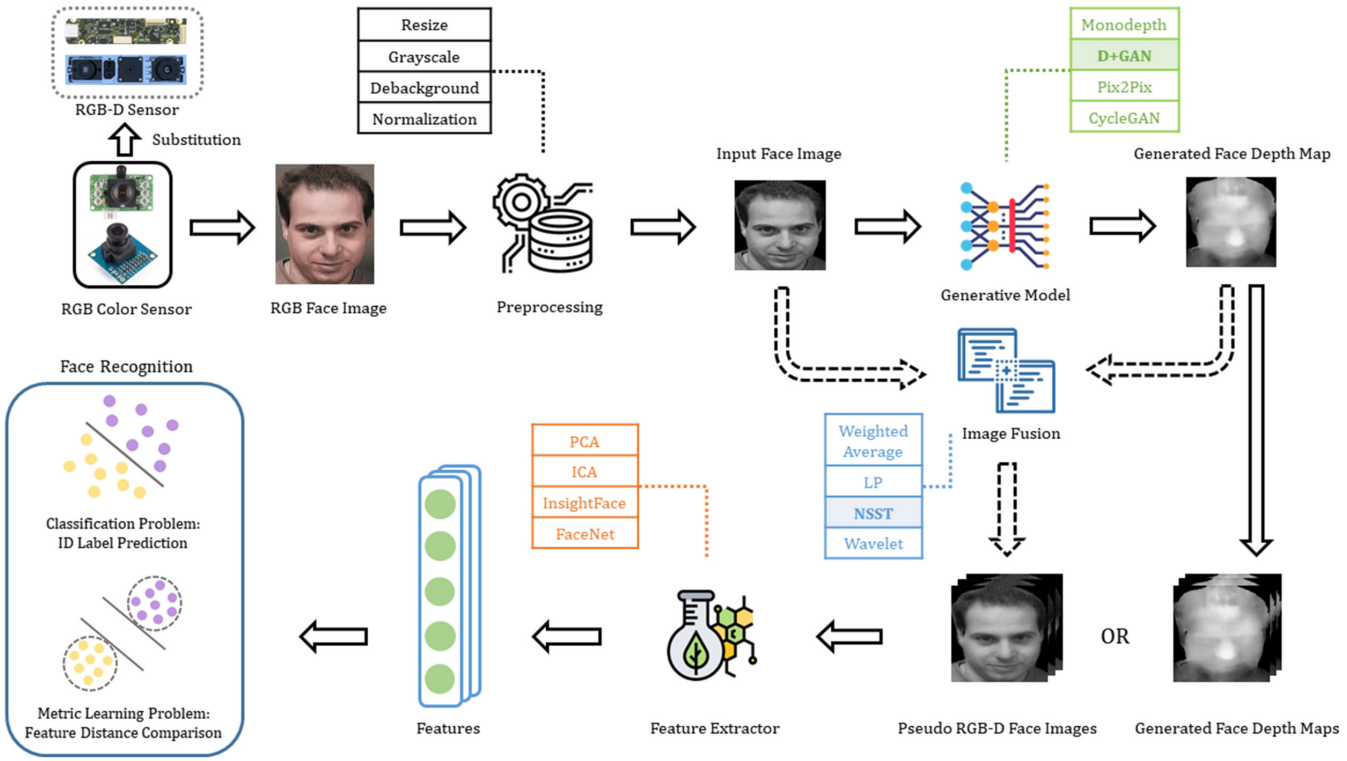


Fig. 1: Pseudo RGB-D face recognition framework

and Erol Seke applied a conditional Wasserstein GAN to perform face depth estimation [27]. Bo Jin, Leandro Cruz and Nuno Gonçalves predicted face depth maps by using pretrained models for scene depth estimation directly [28].

III. Materials and Methods

Generative Adversarial Network (GAN), proposed by Ian Goodfellow et al., is a model that learns a mapping from random noise vector to output images [29]. The original GAN consists of two parts which are a generator and a discriminator. The objective of the generator is to map input Gaussian noise into a fake image, and the discriminator is to determine whether the input image comes from the generator or not, that is, to compute the probability of the input image being false. The conditional generative adversarial network (cGAN), proposed by Mehdi Mirza and Simon Osindero, is a supervised model that can generate output images with a desired condition from random noise [30]. Pix2Pix, proposed by Phillip Isola et al., could be regarded as a special case of cGAN. It takes the 2D image as the input condition of cGAN to realize the image-to-image translation [31]. ACGAN, proposed by Augustus Odena, Christopher Olah and Jonathon Shlens, is required not only to judge whether the input image is true or not, but also to classify the category of the input image in the discriminator part [32].

For adapting our task that is generating the corresponding depth from RGB face images better, we comprehensively refer to the above network structures and cooperate with some advanced skills, and propose the D+GAN. Figure 2 indicates the main structures of cGAN, Pix2Pix, ACGAN and D+GAN. It concisely shows the difference between D+GAN and other GANs' main structures. They both control the generated images by introducing external conditions. For cGAN and ACGAN, the generator generates fake samples from random noise and conditions. For Pix2Pix, the generator generates fake images from images which could be regarded as conditions. Whereas, for D+GAN, the generator generates fake images from condition images and their corresponding labels. For cGAN and Pix2Pix, the discriminator determines whether the sample is the real sample that meet the condition. For ACGAN, the discriminator determines not only whether the sample is the real sample that meets the condition, but also the category of each sample. Whereas, for D+GAN, the discriminator determines not only whether the input sample is the real sample that corresponds the condition image, but also the multiple categories that each sample belongs to.

A. Datasets

In our experiments, there are 9290 pairs of colored images and corresponding depth maps from Bosphorus 3D Face Database [33] and CASIA 3D Face Database [34] for training the GAN models. Binghamton University 3D Facial Expression (BU-3DFE) Database [35] is only for testing.

a) Bosphorus 3D Face Database: Bosphorus 3D Face Database widely used for 3D face processing contains 105 subjects and 4666 faces in the database. One third of the subjects are professional actors or actresses. There are various expressions (up to 35), head poses (13 yaw and pitch rotations) and varieties of face occlusions for each subject. Facial data in the dataset is acquired by a 3D system based on the

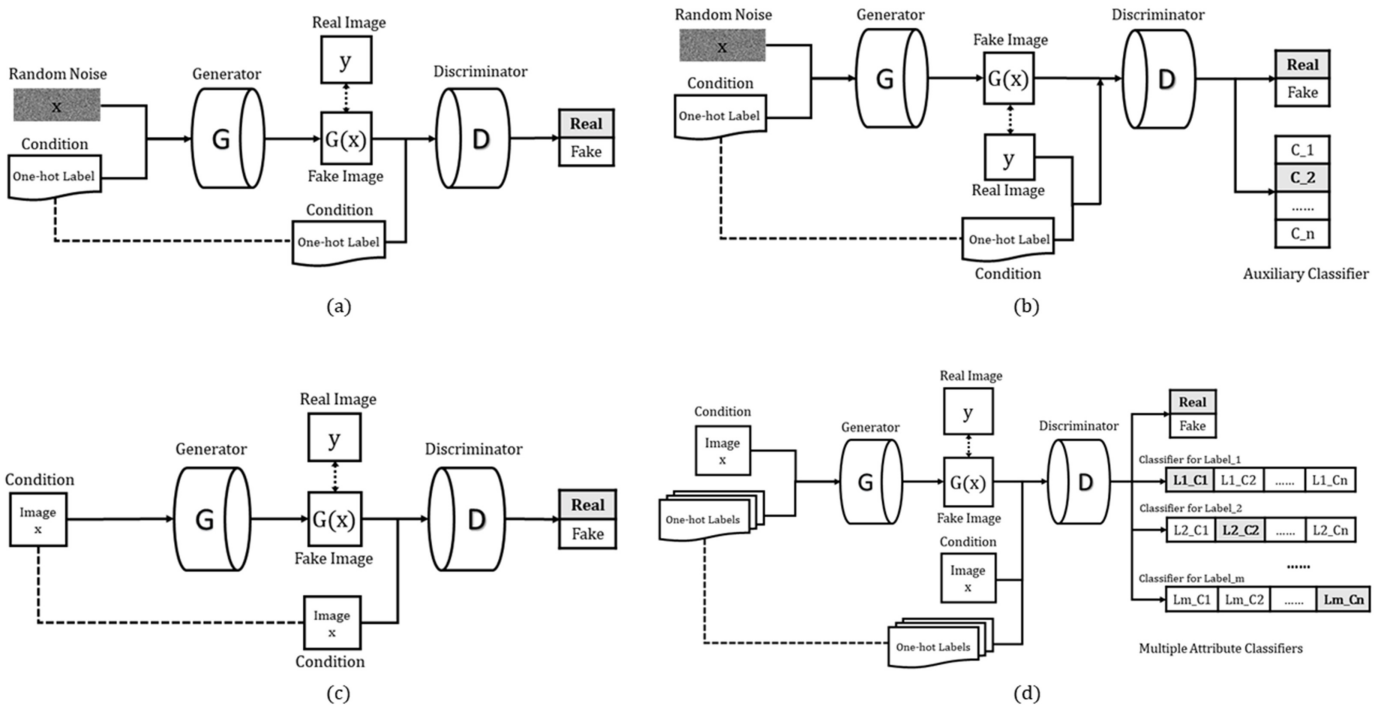


Fig. 2: Main structures of GANs. (a) cGAN (b) ACGAN (c) Pix2Pix (d) Ours: D+GAN

structured-light. The ground truth depth images and their corresponding color images are transformed from 3D point cloud files provided by the Bosphorus database.

b) CASIA 3D Face Database: CASIA 3D Face Database collected by the Chinese Academy of Sciences contains 4624 scans of 123 persons. The scans are collected by the Minolta Vivid 910 which is a non-contact 3D digitizer. Each person in the database has 37 or 38 scans which include variations of poses, expressions and illuminations. Most of the persons in the database are Mongoloid.

c) Binghamton University 3D Facial Expression (BU-3DFE) Database: There are 100 subjects in the BU-3DFE Database of which 56 are male and 44 are female. The majority of subjects were undergraduates with various races. For each subject, there are 25 3D models with seven expressions which are happiness, disgust, fear, anger, surprise, sadness and neutral with different levels of intensity.

B. Preprocessing

In practice, images always have different backgrounds which can affect the processing performance of the algorithm. Since training image pairs transformed from 3D data have black backgrounds. In this section the main purpose is to remove the image background out of the face uniformly. Firstly, the threshold is calculated by using Otsu's method [36]. Then, the image is transformed to a binary image by the threshold. Thus, 8-connected objects are labeled to locate the face based on the binary image. Next, background pixels are replaced with black pixels. Finally, an open operation which is an erosion followed by a dilation is performed to remove small objects and smooth the boundaries of larger objects of the image.

C. D+GAN

In the task of generating face depth maps from corresponding RGB images, we propose a generative adversarial network named D+GAN for making full use of the attribute information of the human face. The generator (G) is composed of residual modules [18], self-attention modules [37] and convolution neural network, and its input is a 256×256 RGB image and its facial attribute labels which include the corresponding gender, age and race categories. The output is a depth map with the same size, which realizes the mapping of image to image. The discriminator (D) is used to identify the quality of the depth map. In our design, D+GAN not only outputs the score of the depth map, but also determines gender, age and race categories. Thus the input of the discriminator is a 256×256 depth map with its labels, and the output of the discriminator contains four scalar values which represent probabilities of true or false, age, gender and race. Figure 3 shows the structure of D+GAN.

1) Generator: Specifically, the core architecture of the generator is U-shaped [38], which consists of an encoder and a decoder. The encoder is mainly used for feature extraction and feature compression of the image. It reduces the size of the input image and the number of feature parameters while increases the number of channels, which realizes the down-sampling process. The decoder with a symmetric and opposite structure to the encoder performs the encoding representation up-sampling successively and restores it to the same feature size as the encoder input.

The generator model also utilizes a skip connection in the convolutional layer between the encoder and decoder to build an information flow transmission approach, which can relieve the gradient disappearance problem effectively. The encoder

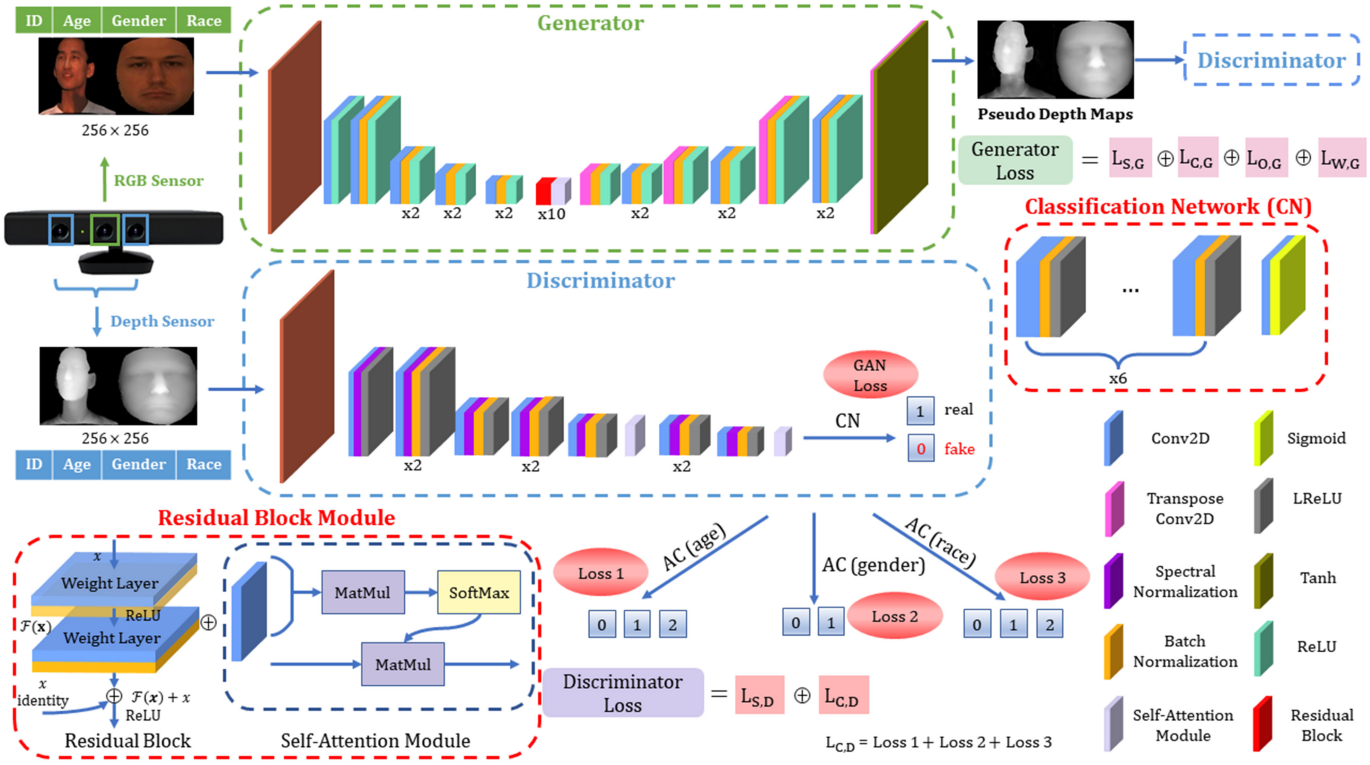


Fig. 3: D+GAN: A GAN architecture for translating RGB images to depth maps with multiple face attributes

is composed of 8 two-dimensional convolutional layers, as shown in Figure 3. The number of convolution kernels set is [64, 128, 256, 256, 512, 512, 512, 512] respectively, and the strides are set to [2, 1, 2, 1, 2, 1, 2, 1] sequentially. There are one Batch Normalization (BN) layer for normalizing input features to accelerate the convergence process and one layer with the $ReLU$ activation function for introducing the sparsity of data to suppress the overfitting after each convolutional layer except for the first one.

The decoder is mainly composed of the convolutional neural network and deconvolutional neural network. In the decoder, the convolutional neural network is designed for feature extraction, and its calculation method is the same as that of the encoder, while deconvolutional neural network is designed for increasing the size of feature maps for up-sampling. In addition, the decoder intersperses two convolutional neural layers as shown in Figure 3. The number of convolution kernels set is [512, 512, 256, 256, 256, 128, 128, 128, 64, 3] respectively, and the strides are set to [2, 1, 1, 2, 1, 1, 2, 1, 1, 2] sequentially. Layer 1, 4, 7 and 10 are the deconvolutional layers. Similarly, BN layers and $ReLU$ activation functions are added after each convolution layer except for the last one. Finally, the $tanh$ activation function is used to normalize the output depth map at [-1, 1].

a) Residual block: In order to fully extract features and increase model capacity, ten groups of residual block and self attention module combinations are used consecutively at the connection between the encoder and decoder of the generator. In our design, we use residual blocks to replace the original design of UNet. In the residual block $H(x)$, the original mapping is changed into $F(x) + x$ from $F(x)$ by using skip connections, which makes the neural network to be easier optimized. The number of convolution kernels is 256, the kernel size is 3×3, and the stride is set to 1.

b) Self-attention module: Self-attention mechanism can learn from distant blocks, so it is used in both generator and discriminator in our design. The self-attention module helps to learn multi-level and long range dependencies across image regions, which is complementary to the convolution layer. In the self-attention module, the input feature x with n channels is transformed into query ($Q = W_Q x$), key ($K = W_K x$) and value ($V = W_V x$) by convolution operations. The size of Q, K, V remains unchanged, but the number of channels becomes $n/8, n/8$ and n respectively. Next, Q, K, and V are serialized by channels so that feature map of $q_{m\times\frac{n}{8}}$, $k_{m\times\frac{n}{8}}$ and $v_{m\times n}$ are obtained respectively, where m represents the feature size. The final output of attention weight distribution is computed as:

$$attention(q, k, v) = softmax(qk^T)v \tag{1}$$

2) Discriminator: The discriminator of D+GAN consists of a backbone structure for distinguishing between true and false, and three branches for identifying face attributes of the image generated. In the backbone network, in order to provide more information exchange between channels and save computing resources, we insert a self-attention module after some higher convolutional layers as described above

before the branch node. In detail, there are ten convolutional layers where the number of convolution kernels set is [64, 64, 64, 128, 128, 128, 256, 256, 256, 512] respectively and the strides are set to [2, 1, 1, 2, 1, 1, 2, 1, 1, 2] sequentially. The size of convolution kernels is 3×3, except the first layer is 5×5. In order to make the training process more stable, we set up spectral normalization [39] in these 10 convolutional layers to make the neural network robust to input disturbances.

a) Spectral Normalization: In detail, for the weight $W_{m\times n}$ of the neural network, the spectral norm is the maximum singular value of the matrix. The maximum singular value $\sigma(W_{m\times n})$ is defined as:

$$\sigma(W_{m\times n}) = \max_{\delta} \frac{||W_{m\times n}\delta||_2}{||\delta||_2} \tag{2}$$

In practice, $\sigma(W_{m\times n})$ is approximately calculated by the power iteration, and then the weight $W_{m\times n}$ is updated to $W_{m\times n}/\sigma(W_{m\times n})$ in the forward direction during training, which is the process of spectral normalization.

The four branch networks get the output of the branch node as the input and perform different classification tasks. The first branch network is used to judge whether the depth map is true or false, which is essentially a binary classification task. Similarly, the second, third and fourth branch networks are used to classify age, gender and race respectively. In detail, the age label is divided into three categories which are 19-39 years old, 40-60 years old, and above 60 years old. The gender label is divided into two categories which are male and female. The race label is divided into three categories which are Caucasoid, Mongoloid and Negroid. These four branch networks have the same network structure except for the last layer, which are composed of seven two-dimensional convolutional layers, and their kernel size is 3×3. The number of convolution kernels in the first six layers is 512 with a stride of 1, and the number of kernels in the last layer is 2 or 3 with a stride of 2.

3) Loss function: The loss of the discriminator L_D consists of two parts. The first part $L_{S,D}$, adopted from standard GAN, is used to distinguish between training samples and generated samples, which is indicated as:

$$\begin{aligned} L_{S,D} = &\mathbb{E}_{Y\in P_{dat}(Y),X\in P_{dat}(X)}\left[logD_1\left(X,Y\right)\right] \\ &+\mathbb{E}_{X\in P_{dat}(X)}\left[log(1-D_1(G(X),X))\right] \end{aligned} \tag{3}$$

where X represents the RGB face image to be translated, Y represents the condition image corresponding to the real depth image, and P_{dat} represents the probability distribution of the corresponding dataset. D_1 represents the output of the first discriminator. For the condition real image Y and the generated image $G(X)$, the classifiers in the discriminator should be able to predict the classes it belongs to. The second part $L_{C,D}$, classification loss, is the cross entropy loss of age, gender and race classification, which is indicated as:

$$\begin{aligned} L_{C,D} = &\sum_{i=2}^{4} \mathbb{E}_{X\in P_{dat}(X)}[\log P(D_i = c|G(X))] \\ &+\mathbb{E}_{Y\in P_{dat}(Y)}[\log P(D_i = c|Y)] \end{aligned} \tag{4}$$

where D_i represents the ith discriminator, and C_i represents the corresponding label. Totally, the training loss of the discriminator, L_D, can be expressed as:

$$L_D = \lambda_1 L_{S,D} + \lambda_2 L_{C,D} \tag{5}$$

For the generator, its loss function L_G contains three parts. First, it is expected that the generated samples can deceive the discriminator, thus $L_{S,G}$ is defined as:

$$L_{S,G} = -\mathbb{E}_{X\in P_{dat}(X)}[\log D_1(G(X),X)] \tag{6}$$

In order to ensure the similarity of input and output images of the generator, L2-loss is introduced as:

$$L_{O,G} = -E_{Y\in P_{dat}(Y),X\in P_{dat}(X)}\left[\|Y - G(X)\|_2\right] \tag{7}$$

Next, the generator is expected to generate high-quality samples so that they can be correctly classified by the discriminator. Similarly, the classification loss $L_{C,G}$ is defined as:

$$L_{C,G} = \sum_{i=2}^{4} \mathbb{E}_{X\in P_{dat}(X)}[\log P(D_i = c|G(X))] \tag{8}$$

In addition, in order to avoid the over-fitting, the weight regularization term $L_{W,G}$ is introduced. It is expressed as:

$$L_{W,G} = \frac{1}{2}||W||^2 \tag{9}$$

Totally, the training loss of generator, L_G, can be expressed as:

$$L_G = \lambda_1 L_{S,G} + \lambda_2 L_{C,G} + \lambda_3 L_{O,G} + \lambda_4 L_{W,G} \tag{10}$$

IV. Experimental Results and Discussions

In this section, we not only evaluate on the face depth map generated itself, but also validate it for the face recognition task in various datasets.

A. Qualitative Results and Analysis

To perform the qualitative evaluation, we calculate some indicators on the three 3D face datasets described above to evaluate the quality of the obtained depth map. In this section, we present outputs of face depth maps generated by several state-of-the-art techniques for some examples. There are Monodepth2, DenseDepth (KITTI), DenseDepth (NYU-Depth V2), 3DMM, Pix2Pix, CycleGAN and D+GAN for comparison. In this study, Monodepth2 [40] is trained on the KITTI dataset with the mono training modality. DenseDepth (KITTI) [41] is trained successively on the ImageNet and KITTI datasets, and DenseDepth (NYU-Depth V2) is trained successively on the ImageNet and NYU-Depth V2 datasets. 3D Morphable Model (3DMM) [42] is to generate a textured 3D face with parameters including vertices, triangles and attribute based on Basel Face Model (BFM). With these parameters, we render this 3D face into the depth map via a rasterization renderer. GAN models including Pix2Pix, CycleGAN and D+GAN are all trained on the Bosphorus 3D Face Database and CASIA 3D Face Database for 20 epochs, and their training curves all converge before or around 16 epochs. Adam optimizer is used

for Pix2Pix and CycleGAN, while Adadelta optimizer is used for D+GAN.

The IDs of the example cases are bs016_LFAU_22_0 of Bosphorus 3D Face Database, 008-025 of CASIA 3D Face Database and F0010_FE03WH_F2D of BU-3DFE Database respectively. The ground truth depth image and its corresponding color image are transformed from 3D data provided.

1) Case Study: bs016_LFAU_22_0 of Bosphorus 3D Face Database: Figure 4 presents the results for the case of bs016_LFAU_22_0 of Bosphorus 3D Face Database. Figure 4a shows the RGB face image which is transformed from 3D data provided, and Figure 4b shows the ground truth face depth map which is transformed from 3D data provided. Figure 4c shows the output generated by Monodepth2. The result shows the contour of the face vaguely, and the relative depth information is not accurately expressed. Figure 4d shows the output generated by DenseDepth (KITTI). The result can only show the outline of the face, and cannot show the depth of facial details. Figure 4e shows the output generated by DenseDepth (NYU-Depth V2). The result shows the depth better, but still lacks the facial detailed depth. Figure 4f shows the output generated by 3DMM. The result shows face detailed depth information more, however the contour of eyes, nose, mouth and the face shape showed are visually very different with the ground truth. We infer that this is because 3DMM is based on an average model. Visually, Figure 4g and Figure 4h show the basically satisfactory results which are generated by Pix2Pix and CycleGAN. Figure 4i shows the best result in visual which is the output generated by D+GAN. The depth values especially in eyes, nose and mouth shown by D+GAN are more precise than Pix2Pix and CycleGAN.

The autocorrelation function is usually used as the texture measure in the image. The texture coarseness of the image is proportional to the expansion of the autocorrelation function. We assume that one image is denoted as $I(x, y)$. Autocorrelation function is defined as:

$$C(\xi, \eta, a, b) = \frac{\sum_{x=a-w}^{a+w} \sum_{y=b-w}^{b+w} I(x, y) I(x-\xi, y-\eta)}{\sum_{x=a-w}^{a+w} \sum_{y=b-w}^{b+w} [I(x, y)]^2} \tag{11}$$

where (a, b) is the pixel in the window which size is $(2w + 1) * (2w + 1)$. $\xi, \eta = \pm 0, \pm 1, \pm 2 ... \pm N$. ξ and η are shifting variables on the pixels.

In the case of bs016_LFAU_22_0 of Bosphorus 3D Face Database, autocorrelation function graphs on depth maps generated by various models are shown as Figure 5. In the autocorrelation function graph, a larger downward trend as ξ and η increasing means a larger coarseness of the corresponding image. Figure 5b shows the autocorrelation function graph of the ground truth depth map. Comparing with Figure 5a, Figure 5b has a smaller downward trend as ξ and η increasing, which means the depth map has a lower coarseness than its corresponding grayscale image. Subjectively, the spatial details of the face should be changed regularly. Comparing with Figure 5b, Figure 5c, Figure 5d and Figure 5e has a larger downward trend as ξ and η increasing, which means the depth maps generated by Monodepth2, DenseDepth (KITTI) and DenseDepth (NYU-Depth V2) have a higher coarseness than the ground truth depth map. Conversely, the shapes of Figure 5f, Figure 5g, Figure 5h and Figure 5i are similar with Figure 5b, which indicates the depth maps generated by 3DMM, Pix2Pix, CycleGAN and D+GAN have a higher quality.

In the case of bs016_LFAU_22_0 of Bosphorus 3D Face Database, local SSIM maps of the depth maps generated by various models are shown in Figure 6. The structural similarity index measure (SSIM) is to measure the similarity between two images. In the SSIM map, regions with smaller local SSIM values correspond to different regions from the reference image. Similarly, regions with larger local SSIM values correspond to uniform regions of the reference image. The reference image here is the ground truth face depth map. From Figure 6 seen, Figure 6g representing D+GAN has the most red area. Figure 6e representing Pix2Pix and Figure 6f representing CycleGAN in overall perform well except in specific areas of eyes, nose and mouth in comparison with Figure 6g. Figure 6d representing 3DMM shows a larger difference in face shape besides in eyes, nose and mouth. In addition, besides eyes, nose and mouth areas, Figure 6a representing Monodepth2, Figure 6b representing DenseDepth (KITTI) and Figure 6c representing DenseDepth (NYU-Depth V2) show a larger difference in four corners out of the face. Among these three, Figure 6c shows a less difference in the area of the human face.

2) Case Study: 008-025 of CASIA 3D Face Database: Figure 7 presents the results for the case of 008-025 of CASIA 3D Face Database. Unlike the previous example, the input image in this example is a bust. In all, the performance of each model is similar to that in the above example. Figure 7g, Figure 7h and Figure 7i representing three GAN models show a satisfactory result. Especially for Figure 7i representing D+GAN, it is difficult to see the difference from the ground truth with the naked eye. It is worth mentioning that 3DMM can only be used for the human head area (see Figure 7f).

In the case of 008-025 of CASIA 3D Face Database, autocorrelation function graphs on depth maps generated by various models are shown as Figure 8. It shows the coarseness of the generated depth map. It is worth mentioning that Figure 8 indicates the texture coarseness of the depth map of the bust should be higher than the face (see Figure 6). Comparing with Figure 8b, Figure 8c, Figure 8d and Figure 8e has a smaller downward trend as ξ and η increasing, which means the depth maps generated by Monodepth2, DenseDepth (KITTI) and DenseDepth (NYU-Depth V2) have a lower coarseness than the ground truth depth map. In contrast, Figure 8g, Figure 8h and Figure 8i representing three GAN models have similar trends with Figure 8b, which implies they retain depth information well.

In the case of 008-025 of CASIA 3D Face Database, local SSIM maps of the depth maps generated by various models are shown in Figure 9. It shows the similarity of areas in the depth maps generated. In all, the performance of each model is similar to that in the last example. It is worth mentioning that the areas of clothes and neck in the depth map generated by CycleGAN are not as satisfactory as Pix2Pix and D+GAN (see Figure 9f).

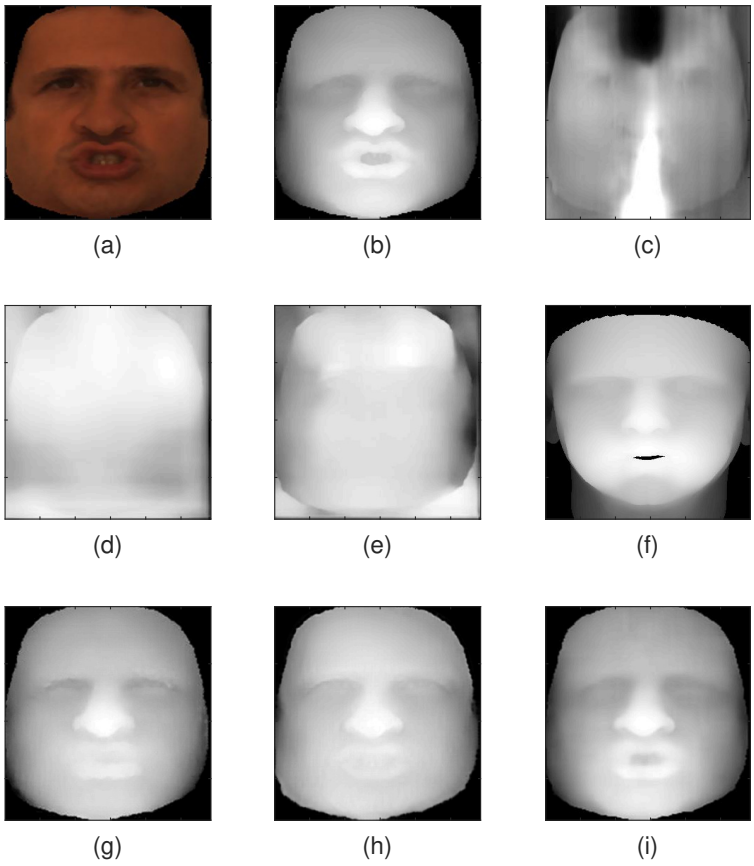


Fig. 4: Face depth maps generated by various models in the case of bs016_LFAU_22_0. (a) Input RGB image, (b) Ground truth depth map, (c) Model: Monodepth2, (d) Model: DenseDepth (KITTI), (e) Model: DenseDepth (NYU-Depth V2), (f) Model: 3DMM, (g) Model: Pix2Pix, (h) Model: CycleGAN, (i) Proposed Model: D+GAN

3) Case Study: F0010_FE03WH_F2D of BU-3DFE Database: Figure 10 presents the results for the case of F0010_FE03WH_F2D of BU-3DFE Database. It is worth mentioning that, unlike the previous examples, GAN models are not trained by BU-3DFE Database. In all, the performance of each model is similar to that in the first example. Figure 10g, Figure 10h and Figure 10i representing three GAN models show a more satisfactory result than others. In detail, Figure 10g and Figure 10h representing Pix2Pix and CycleGAN respectively show a inaccurate depth in the eyes area. However, D+GAN performs well in the eyes area (see Figure 10i). It is worth mentioning that 3DMM generates inaccurate results in the face shape again (see Figure 10f).

In the case of F0010_FE03WH_F2D of BU-3DFE Database, autocorrelation function graphs on depth maps generated by various models are shown as Figure 11. It shows the coarseness of the depth map generated. It is worth mentioning that the graph shape of Figure 11f representing 3DMM is the most similar with Figure 11b representing the ground truth in this case. Figure 11g and Figure 11h has a smaller downward trend as ξ and η increasing, which means the depth maps generated for the face by Pix2Pix and CycleGAN have a lower coarseness.

In the case of F0010_FE03WH_F2D of BU-3DFE Database, local SSIM maps of the depth maps generated by various models are shown in Figure 12. It shows the similarity of areas in the depth maps generated. In all, the performance of each model is similar to that in the previous example. It is worth mentioning that the areas of clothes and neck in the depth map generated by CycleGAN are not as satisfactory as Pix2Pix and D+GAN (see Figure 12). In comparison with Figure 12g, Figure 12i representing D+GAN performs better in the area of the eyes.

B. Quantitative Results and Analysis

In this section, quantitative analysis is carried out. The Structural Similarity Index (SSIM), Root Mean Squared Error (RMSE) and Peak Signal-to-Noise Ratio (PSNR) are selected to evaluate of the quality of the face depth map generated by several models on three datasets described before which are

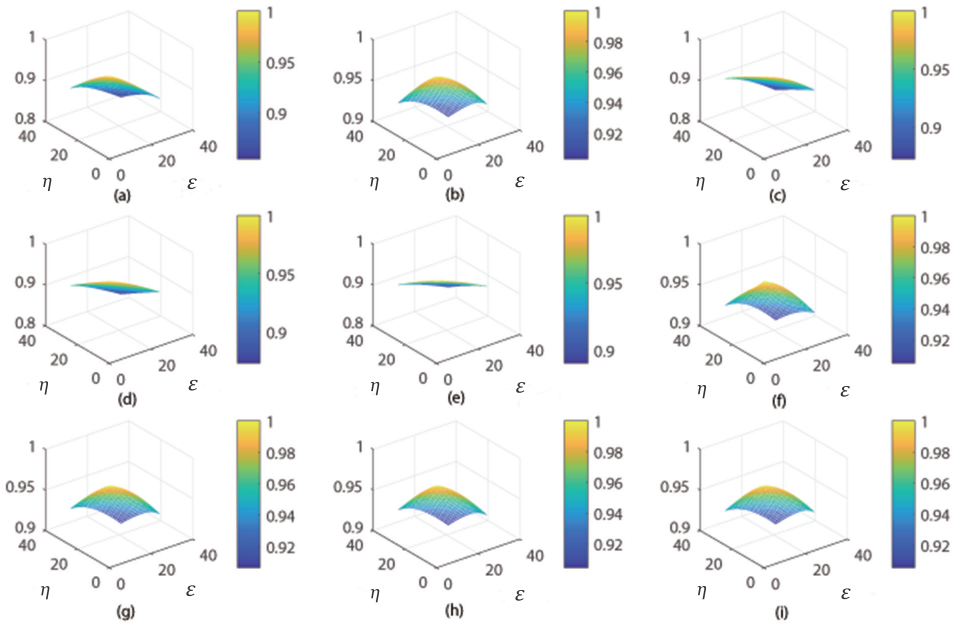


Fig. 5: Autocorrelation function graphs of various output images: (a) Original RGB image, (b) Ground truth depth map, (c) Depth map generated by Monodepth2, (d) Depth map generated by DenseDepth (KITTI), (e) Depth map generated by DenseDepth (NYU-Depth V2), (f) Depth map generated by 3DMM, (g) Depth map generated by Pix2Pix, (h) Depth map generated by CycleGAN, (i) Depth map generated by D+GAN

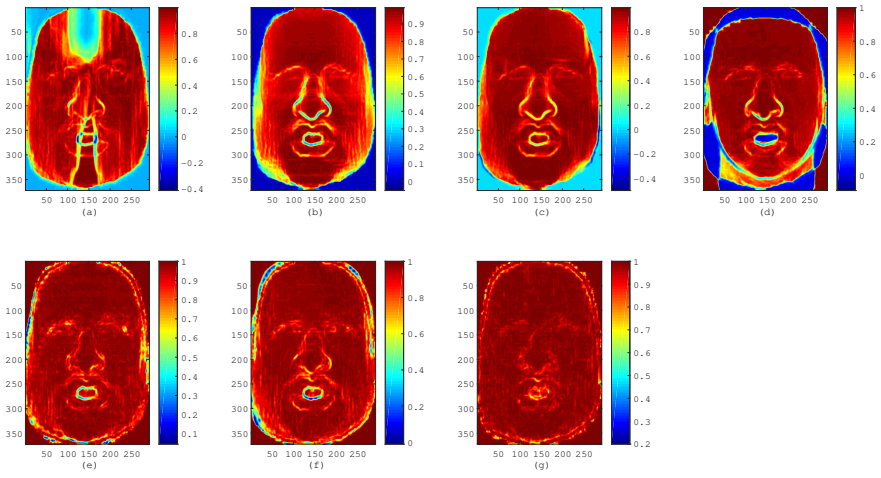


Fig. 6: Local SSIM maps of depth maps generated by various models. (a) Model: Monodepth2, (b) Model: DenseDepth (KITTI), (c) Model: DenseDepth (NYU-Depth V2), (d) Model: 3DMM, (e) Model: Pix2Pix, (f) Model: CycleGAN, (g) Proposed Model: D+GAN

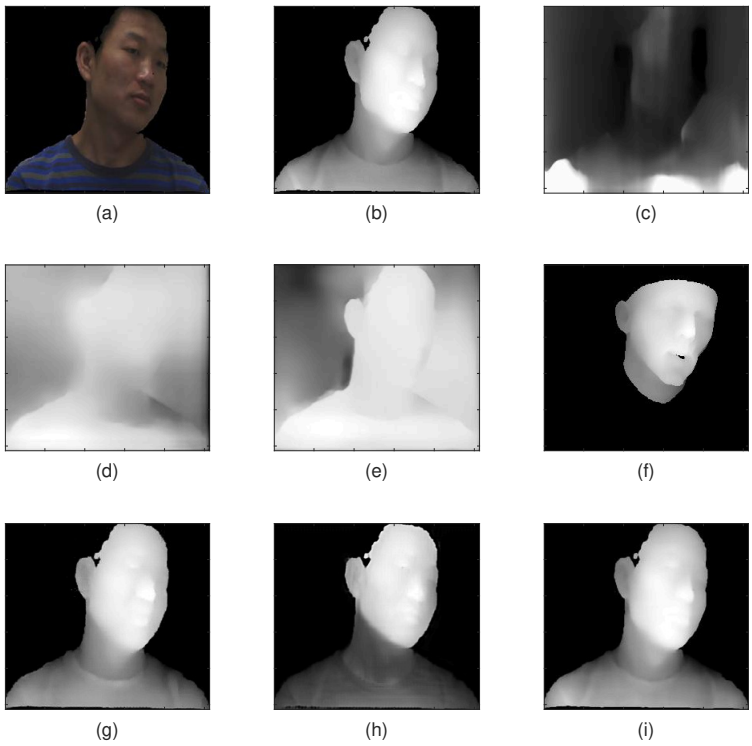


Fig. 7: Face depth maps generated by various models in the case of 008-025. (a) Input RGB image, (b) Ground truth depth map, (c) Model: Monodepth2, (d) Model: DenseDepth (KITTI), (e) Model: Densedepth (NYU-Depth V2), (f) Model: 3DMM, (g) Model: Pix2Pix, (h) Model: CycleGAN, (i) Proposed Model: D+GAN

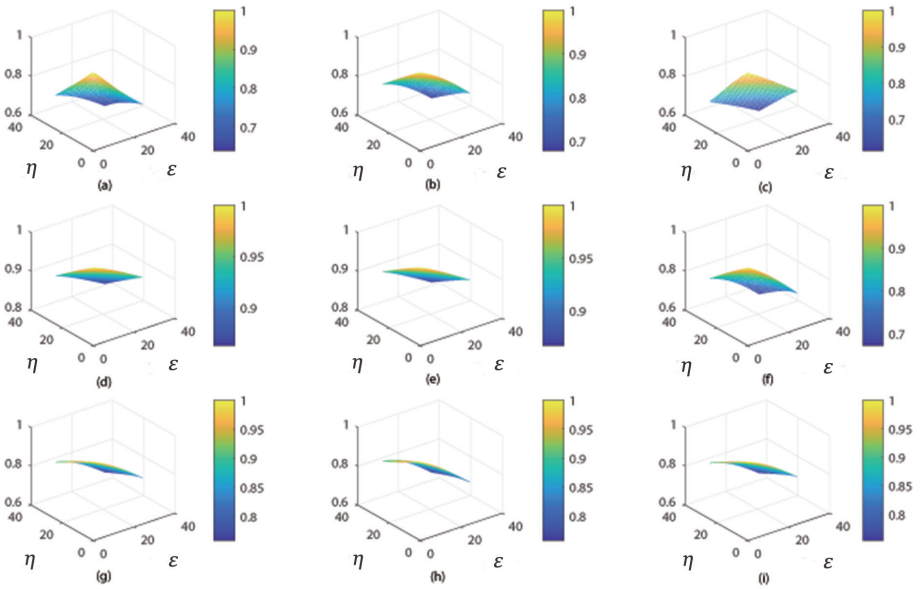


Fig. 8: Autocorrelation function graphs of various output images: (a) Original RGB image, (b) Ground truth depth map, (c) Depth map generated by Monodepth2, (d) Depth map generated by DenseDepth (KITTI), (e) Depth map generated by DenseDepth (NYU-Depth V2), (f) Depth map generated by 3DMM, (g) Depth map generated by Pix2Pix, (h) Depth map generated by CycleGAN, (i) Depth map generated by D+GAN

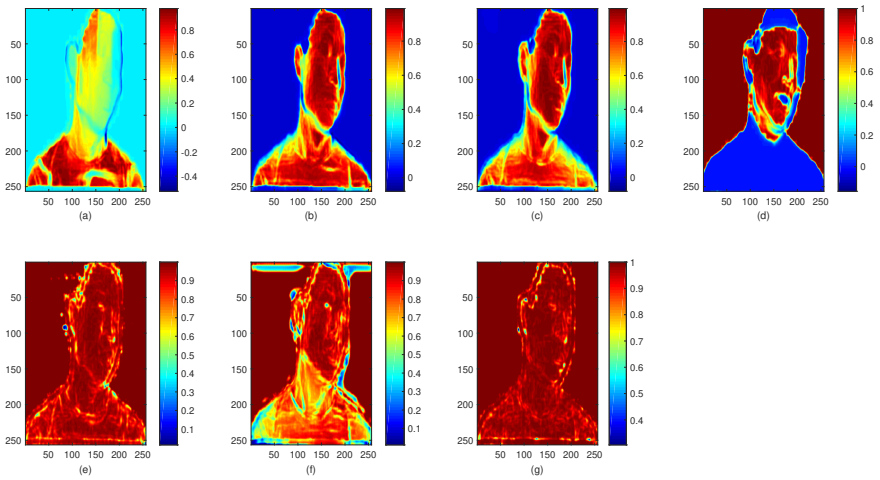


Fig. 9: Local SSIM maps of depth maps generated by various models. (a) Model: Monodepth2, (b) Model: DenseDepth (KITTI), (c) Model: DenseDepth (NYU-Depth V2), (d) Model: 3DMM, (e) Model: Pix2Pix, (f) Model: CycleGAN, (g) Proposed Model: D+GAN

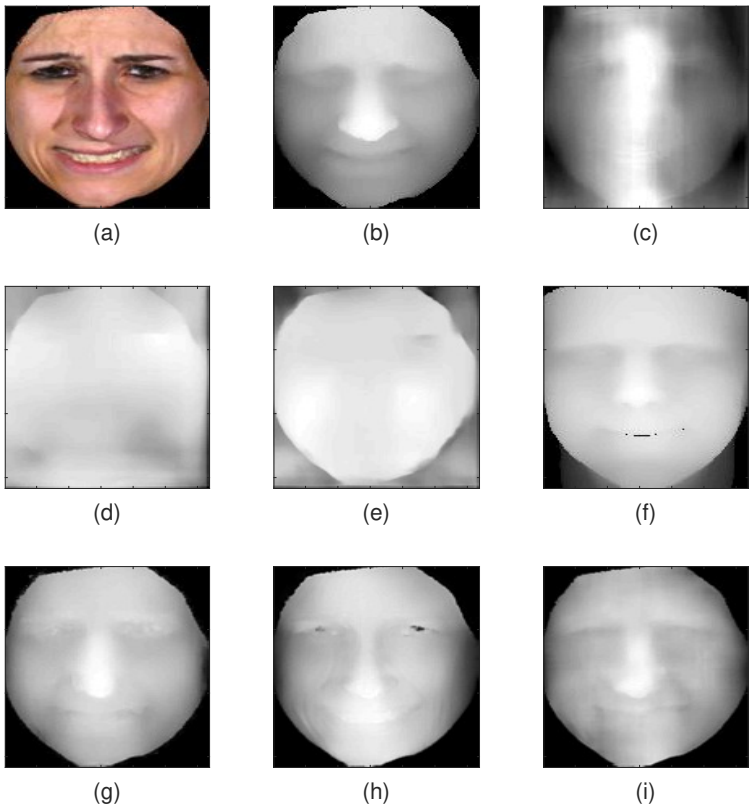


Fig. 10: Face depth maps generated by various models in the case of F0010_FE03WH_F2D. (a) Input RGB image, (b) Ground truth depth map, (c) Model: Monodepth2, (d) Model: DenseDepth (KITTI), (e) Model: DenseDepth (NYU-Depth V2), (f) Model: 3DMM, (g) Model: Pix2Pix, (h) Model: CycleGAN, (i) Proposed Model: D+GAN

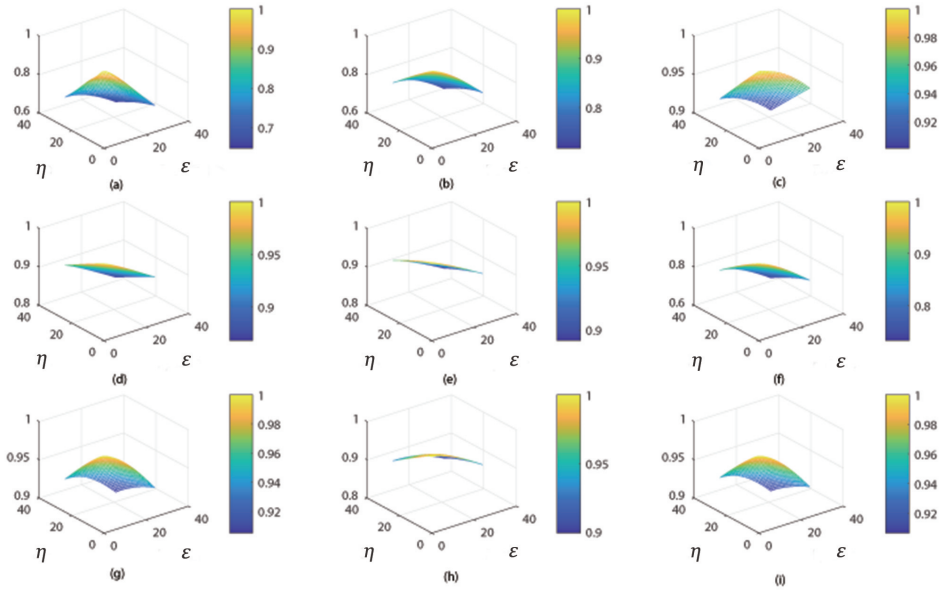


Fig. 11: Autocorrelation function graphs of various output images: (a) Original RGB image, (b) Ground truth depth map, (c) Depth map generated by Monodepth2, (d) Depth map generated by DenseDepth (KITTI), (e) Depth map generated by DenseDepth (NYU-Depth V2), (f) Depth map generated by 3DMM, (g) Depth map generated by Pix2Pix, (h) Depth map generated by CycleGAN, (i) Depth map generated by D+GAN

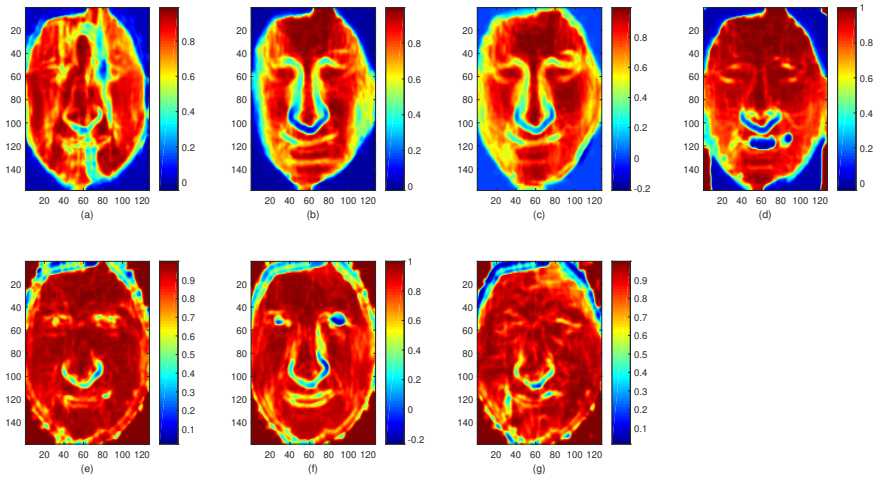


Fig. 12: Local SSIM maps of the depth maps generated by various models. (a) Model: Monodepth2, (b) Model: DenseDepth (KITTI), (c) Model: DenseDepth (NYU-Depth V2), (d) Model: 3DMM, (e) Model: Pix2Pix, (f) Model: CycleGAN, (g) Proposed Model: D+GAN

Bosphorus 3D Face Database, CASIA 3D Face Database and BU-3DFE Database.

The Structural Similarity Index (SSIM) [43] is the widely used standard for evaluating structural similarity in images that evaluates the quality of a processed image from a ground truth image. We calculate the SSIM for above six models as:

$$SSIM(a,b) = [l(a,b)]^{\alpha}[c(a,b)]^{\beta}[s(a,b)]^{\gamma} \tag{12}$$

where

$$l(a,b) = \frac{2\mu_a\mu_b + C_1}{\mu_a^2 + \mu_b^2 + C_1} \tag{13}$$

$$c(a,b) = \frac{2\sigma_a\sigma_b + C_2}{\sigma_a^2 + \sigma_b^2 + C_2} \tag{14}$$

$$s(a,b) = \frac{\sigma_{ab} + C_3}{\sigma_a\sigma_b + C_3} \tag{15}$$

In the above equations, there are two images denoted as a and b. μ_a and μ_b indicate the local mean values of corresponding images, σ_a and σ_b indicate the standard deviations and σ_{ab} indicates the cross-covariance for images.

A lower RMSE value means a more accurate result corresponding to the reference. The RMSE between images a and

b is calculated as:

$$RMSE(a,b) = \sqrt{\frac{1}{M \times N}\sum_{i=1}^{M}\sum_{i=1}^{N}(a(i,j) - b(i,j))^2} \quad (16)$$

where M and N are width and height of the image respectively.

PSNR, a logarithmic form using the decibel scale based on MSE, is widely used to quantify reconstruction quality for images. It is defined as:

$$PSNR = 10\log_{10}\frac{L^2}{MSE} = 20\log_{10}\frac{L}{RMSE} \quad (17)$$

where L is the maximum possible pixel value of the image. Here, L equals 255.

The calculated mean results in SSIM, RMSE and PSNR on datasets are presented in Table 1. Not only qualitatively, but also quantitatively, the GAN model overall outperforms other models in these three datasets. Among them, the depth map output by D+GAN can get the best SSIM, RMSE and PSNR values.

For a 256*256 image, among the above three GAN models, Pix2Pix requires 18.6G multiply-accumulate multiply-accumulate operations (MACs) approximately, CycleGAN requires about 56.8G MACs approximately [44], and D+GAN, the embodiment showed, requires about 21.6G MACs approximately. These computations are acceptable for today's GPUs. Using the GAN model to obtain high-quality spatial information of face images will take more computation, which is a trade-off.

C. Face Recognition Results and Analysis

In this section, classic machine learning and deep learning models including PCA [10], ICA [11], FaceNet [17] and InsightFace [19] are selected as face recognition methods. Five classic face recognition datasets including ORL [45], Yale [46], UMIST [47], AR [48] and FERET [49] are selected.

In order to make effective use of generated depth features in the pseudo RGB-D face recognition, image fusion algorithms are utilized. Through comparisons among Wavelet-based methods, Laplacian Pyramid and Non-subsampled Shearlet Transform (NSST) [50], NSST performs the best so as to be selected as the image fusion method for our face recognition experiments.

The shearlet system can be expressed as:

$$\Lambda_{D,S}(\Psi) = \Big\{\Psi_{j,k,l}(x) = |det(D)|^{j/2}\Psi(S^l D^j x - k) : j,l \in Z; k \in Z^2\Big\} \quad (18)$$

where j, k, and l denote the scale, shift, and direction respectively. D, the anisotropic expanding matrix, is expressed as:

$$D = \begin{bmatrix} 4 & 0 \\ 0 & 2 \end{bmatrix} \quad (19)$$

and S, the shear matrix, is expressed as:

$$S = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \quad (20)$$

The NSST performs multi-scale and multi-directional decomposition on input images by Non-subsampled Pyramids (NSPs) and shearing filters in the first place. Next, according to the made fusion strategy, the high frequency and low frequency sub-band images decomposed are transformed and combined into new sub-band images. Last, the final fused image is achieved by the inverse NSST on the new sub-band images. In our embodiment, the filter set for the Laplacian Pyramid decomposition is 'maxflat'. The vector indicating decomposition directions is set to [3, 3, 4, 4]. The vector indicating the local support of the shearing filter is set to [8, 8, 16, 16]. The fusion coefficient is set to 0.5.

Besides NSST, D+GAN is selected as the preferred embodiment for generating the pseudo face depth map in the pseudo RGB-D face recognition due to its good performance in the previous section. If the training images are sufficient, due to the great learning ability of the deep learning model, it is easy to have a 100% accuracy during testing. Therefore, in the evaluation, due to the different capabilities of ML models, we used separate experimental settings to differentiate the performance of face recognition of each model.

In experiments of testing PCA, two images of each person in the dataset are applied for testing, and the rest images of that person are for training. The number of feature face set is 30 for PCA. In this case, the mode of pseudo RGB-D face recognition improves the accuracy by 10.2%, 9.0%, 4.6%, 6.3% and 5.5% approximately on ORL, Yale, UMIST, AR and FERET dataset respectively.

In experiments of testing ICA, five images of each person in the dataset are applied for training, and rest images of that person are for testing. The number of components set is 70 for ICA. The mode of pseudo RGB-D face recognition improves the accuracy by 12.7%, 9.6%, 3.4%, 10.6% and 14.8% approximately on ORL, Yale, UMIST, AR and FERET dataset respectively.

In experiments of testing DL models including FaceNet and InsightFace, for ORL and AR datasets, 30% of the images of each person are used for training, and 70% of the images of each person are used for testing. For Yale dataset, 20% of the images of each person are used for training, and 80% of the images of each person are used for testing. For UMIST dataset, 10% of the images of each person are used for training, and 90% of the images of each person are used for testing. For FERET dataset, 40% of the images of each person are used for training, and 60% of the images of each person are used for testing. Since the number of images of each person in the ORL and YALE datasets is relatively small, and the total number of people is also relatively small. Therefore, using the pre-trained model to directly extract features, and then training a linear SVM classifier for testing could get better results. For the datasets UMIST, AR and FERET with more images, fine-tuning the pretrained network model could be used as a conventional strategy.

Table II presents the face recognition results by two modes including RGB and Pseudo RGB-D using traditional ML and advanced DL models on the five classical face recognition datasets.

Specifically, in experiments of testing the FaceNet: Incep-

TABLE I: **Quantitative Index Results**

Method	Index	Dataset		
		Bosphorus	CASIA 3D	BU-3DFE
Monodepth2	SSIM	0.660	0.205	0.585
	RMSE	60.77	99.15	54.41
	PSNR	12.46	8.205	13.42
DenseDepth (KITTI)	SSIM	0.697	0.339	0.555
	RMSE	92.70	127.7	95.91
	PSNR	8.789	6.007	8.494
DenseDepth (NYU Depth V2)	SSIM	0.728	0.334	0.570
	RMSE	74.38	123.7	86.79
	PSNR	10.70	6.283	9.361
3DMM	SSIM	0.747	0.624	0.677
	RMSE	50.20	73.27	64.82
	PSNR	14.12	10.83	11.90
Pix2Pix	SSIM	0.933	0.949	0.852
	RMSE	13.43	11.11	26.41
	PSNR	25.56	27.22	19.70
CycleGAN	SSIM	0.916	0.851	0.792
	RMSE	21.26	34.36	34.23
	PSNR	17.41	21.58	17.44
D+GAN	SSIM	0.970	0.978	0.869
	RMSE	4.122	3.803	23.99
	PSNR	35.83	36.53	20.53

tion ResNet v1 model pretrained by CASIA-WebFace, the mode of pseudo RGB-D face recognition improves the accuracy by 2.7%, 5.7%, 0.4%, 0.9% and 11.3% approximately on datasets ORL, Yale, UMIST, AR and FERET respectively. In experiments of testing the FaceNet: Inception ResNet v1 model pretrained by VGG-Face2, the mode of pseudo RGB-D face recognition improves the accuracy by 0%, 0%, 1.7%, 0.7% and 1.3% approximately on datasets ORL, Yale, UMIST, AR and FERET respectively. In experiments of testing the Insightface: IResNet34 model pretrained by MS1MV2, the mode of pseudo RGB-D face recognition improves the accuracy by 2.1%, 3.2%, 1.0%, 0.2% and 7.9% approximately on datasets ORL, Yale, UMIST, AR and FERET respectively. In experiments of testing the Insightface: IResNet100 model pretrained by MS1MV2, the mode of pseudo RGB-D face recognition improves the accuracy by 2.7%, 5.7%, 0.3%, 2.4% and 2.5% approximately on datasets ORL, Yale, UMIST, AR and FERET respectively.

Table II shows that in the face recognition experiments, the best performing results annotated in bold for each dataset of the five almost all use the mode of pseudo RGB-D face recognition. It can be concluded that pseudo RGB-D face recognition proposed is able to improve the accuracy in comparison with RGB face recognition using different classic traditional ML and DL models. Especially for traditional ML models, pseudo RGB-D face recognition mode can increase the accuracy more.

V. Conclusion

Inspired by the occurrence of RGB-D face recognition, we propose a pseudo RGB-D face recognition framework. In essence, the ML model is able to imitate the relative depth map from its corresponding RGB image by learning from big data to replace the depth sensors. We provide a D+GAN model for making increased use of face attribute information to generate the high quality face depth map. In cooperation with NSST, the pseudo RGB-D face recognition obtains an overall improvement in comparison with RGB face recognition. With the pseudo RGB-D face recognition framework, we could modularly adapt off-the-shelf algorithm models to promote the performance of RGB face recognition. In future, we will continue to discover simple and effective models to perform the monocular face depth estimation, and efficient ways to apply them to improve the biometric recognition performance.

Acknowledgment

The authors would like to thank the Portuguese Mint and Official Printing Office (INCM) and the Institute of Systems and Robotics - Coimbra. This work has been supported by the Fundação para a Ciência e a Tecnologia (FCT) under the Project UIDB/00048/2020.

References

[1] C. Darwin, *On the origin of species, 1859*. Routledge, 2004.
[2] B. Jin, "Deep learning facial diagnosis system, ZL201711255031.1," Patent, 2022, publication of CN108806792B.
[3] B. Jin, L. Cruz, and N. Gonçalves, "Deep facial diagnosis: Deep transfer learning from face recognition to facial diagnosis," *IEEE Access*, vol. 8, pp. 123 649–123 661, 2020.
[4] Y. Wang and M. Kosinski, "Deep neural networks are more accurate than humans at detecting sexual orientation from facial images." *Journal of personality and social psychology*, vol. 114, no. 2, p. 246, 2018.

TABLE II: **Experimental Results of Face Recognition**

Mode	Method	Dataset				
		ORL	Yale	UMIST	AR	FERET
RGB Face Recognition	PCA	84.9%	62.2%	69.3%	41.5%	49.1%
	ICA	79.0%	45.6%	72.9%	46.6%	55.0%
	FaceNet: Inception ResNet v1 (CASIA-WebFace)	98.6%	38.5%	90.8%	76.5%	59.8%
	FaceNet: Inception ResNet v1 (VGG-Face2)	**100.0%**	0.0%	88.0%	75.4%	56.0%
	InsightFace: IResNet34 (MS1MV2)	84.6%	92.6%	79.5%	90.1%	75.6%
	InsightFace: IResNet100 (MS1MV2)	92.9%	91.1%	77.8%	85.0%	53.4%
Pseudo RGB-D Face Recognition	PCA	93.6%	67.8%	72.6%	44.1%	51.8%
	ICA	89.0%	50.0%	76.3%	51.5%	63.1%
	FaceNet: Inception ResNet v1 (CASIA-WebFace)	**100.0%**	40.6%	**91.2%**	77.2%	66.6%
	FaceNet: Inception ResNet v1 (VGG-Face2)	**100.0%**	0.0%	89.5%	75.9%	56.7%
	InsightFace: IResNet34 (MS1MV2)	86.4%	95.6%	80.2%	**90.3%**	**81.5%**
	InsightFace: IResNet100 (MS1MV2)	95.4%	**96.3%**	78.0%	87.0%	54.7%

[5] J. Luo, F. H. Khan, I. Mori, A. de Silva, E. S. Ruezga, M. Liu, A. Pang, and J. Davis, "How much does input data type impact final face model accuracy?" in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 985–18 994.

[6] Z. Zhang, "Microsoft kinect sensor and its effect," *IEEE multimedia*, vol. 19, no. 2, pp. 4–10, 2012.

[7] M. Carfagni, R. Furferi, L. Governi, M. Servi, F. Uccheddu, and Y. Volpe, "On the performance of the intel sr300 depth camera: metrological and critical characterization," *IEEE Sensors Journal*, vol. 17, no. 14, pp. 4508–4519, 2017.

[8] G. Goswami, S. Bharadwaj, M. Vatsa, and R. Singh, "On rgb-d face recognition using kinect," in *2013 IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems (BTAS)*. IEEE, 2013, pp. 1–6.

[9] Y.-C. Lee, J. Chen, C. W. Tseng, and S.-H. Lai, "Accurate and robust face recognition from rgb-d images with a deep learning approach." in *BMVC*, vol. 1, no. 2, 2016, p. 3.

[10] M. Turk and A. Pentland, "Eigenfaces for recognition," *Journal of cognitive neuroscience*, vol. 3, no. 1, pp. 71–86, 1991.

[11] M. S. Bartlett, J. R. Movellan, and T. J. Sejnowski, "Face recognition by independent component analysis," *IEEE Transactions on neural networks*, vol. 13, no. 6, pp. 1450–1464, 2002.

[12] P. Phillips, "Support vector machines applied to face recognition," *Advances in Neural Information Processing Systems*, vol. 11, pp. 803–809, 1998.

[13] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, no. 3, pp. 211–252, 2015.

[14] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, pp. 1097–1105, 2012.

[15] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1701–1708.

[16] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[17] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 815–823.

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[19] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4690–4699.

[20] D. Kim, M. Hernandez, J. Choi, and G. Medioni, "Deep 3d face identification," in *2017 IEEE international joint conference on biometrics (IJCB)*. IEEE, 2017, pp. 133–142.

[21] L. Jiang, J. Zhang, and B. Deng, "Robust rgb-d face recognition using attribute-aware loss," *IEEE transactions on pattern analysis and machine intelligence*, vol. 42, no. 10, pp. 2552–2566, 2019.

[22] S.-H. Lai, C.-W. Fu, and S. Chang, "A generalized depth estimation algorithm with a single image," *IEEE Computer Architecture Letters*, vol. 14, no. 04, pp. 405–411, 1992.

[23] Z.-L. Sun and K.-M. Lam, "Depth estimation of face images based on the constrained ica model," *IEEE transactions on information forensics and security*, vol. 6, no. 2, pp. 360–370, 2011.

[24] Z.-L. Sun, K.-M. Lam, and Q.-W. Gao, "Depth estimation of face images using the nonlinear least-squares model," *IEEE transactions on image processing*, vol. 22, no. 1, pp. 17–30, 2012.

[25] J. Cui, H. Zhang, H. Han, S. Shan, and X. Chen, "Improving 2d face recognition via discriminative face depth estimation," in *2018 International Conference on Biometrics (ICB)*. IEEE, 2018, pp. 140–147.

[26] S. Pini, F. Grazioli, G. Borghi, R. Vezzani, and R. Cucchiara, "Learning to generate facial depth maps," in *2018 International Conference on 3D Vision (3DV)*. IEEE, 2018, pp. 634–642.

[27] A. T. Arslan and E. Seke, "Face depth estimation with conditional generative adversarial networks," *IEEE Access*, vol. 7, pp. 23 222–23 231, 2019.

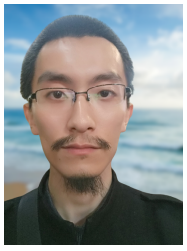
[28] B. Jin, L. Cruz, and N. Gonçalves, "Face depth prediction by the scene depth," in *2021 IEEE/ACIS 19th International Conference on Computer and Information Science (ICIS)*. IEEE, 2021, pp. 42–48.

[29] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[30] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.

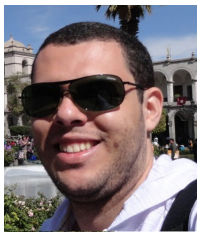
[31] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125–1134.

[32] A. Odena, C. Olah, and J. Shlens, "Conditional image synthesis with auxiliary classifier gans," in *International conference on machine learning*. PMLR, 2017, pp. 2642–2651.
[33] A. Savran, N. Alyüz, H. Dibeklioğlu, O. Çeliktutan, B. Gökberk, B. Sankur, and L. Akarun, "Bosphorus database for 3d face analysis," in *European workshop on biometrics and identity management*. Springer, 2008, pp. 47–56.
[34] CASIA, "Casia-3d face v1," Website, 2004, http://biometrics.idealtest.org/.
[35] L. Yin, X. Wei, Y. Sun, J. Wang, and M. J. Rosato, "A 3d facial expression database for facial behavior research," in *7th international conference on automatic face and gesture recognition (FGR06)*. IEEE, 2006, pp. 211–216.
[36] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE transactions on systems, man, and cybernetics*, vol. 9, no. 1, pp. 62–66, 1979.
[37] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," in *International conference on machine learning*. PMLR, 2019, pp. 7354–7363.
[38] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.
[39] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral normalization for generative adversarial networks," *arXiv preprint arXiv:1802.05957*, 2018.
[40] C. Godard, O. Mac Aodha, M. Firman, and G. J. Brostow, "Digging into self-supervised monocular depth estimation," in *Proceedings of the IEEE international conference on computer vision*, 2019, pp. 3828–3838.
[41] I. Alhashim and P. Wonka, "High quality monocular depth estimation via transfer learning," *arXiv preprint arXiv:1812.11941*, 2018.
[42] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter, "A 3d face model for pose and illumination invariant face recognition," in *2009 sixth IEEE international conference on advanced video and signal based surveillance*. Ieee, 2009, pp. 296–301.
[43] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.
[44] S. Li, M. Lin, Y. Wang, C. Fei, L. Shao, and R. Ji, "Learning efficient gans for image translation via differentiable masks and co-attention distillation," *IEEE Transactions on Multimedia*, 2022.
[45] F. S. Samaria and A. C. Harter, "Parameterisation of a stochastic model for human face identification," in *Proceedings of 1994 IEEE workshop on applications of computer vision*. IEEE, 1994, pp. 138–142.
[46] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection," in *European conference on computer vision*. Springer, 1996, pp. 43–58.
[47] D. B. Graham and N. M. Allinson, "Characterising virtual eigensignatures for general purpose face recognition," in *Face Recognition*. Springer, 1998, pp. 446–456.
[48] A. Martinez and R. Benavente, "The ar face database: Cvc technical report, 24," 1998.
[49] P. J. Phillips, H. Wechsler, J. Huang, and P. J. Rauss, "The feret database and evaluation procedure for face-recognition algorithms," *Image and vision computing*, vol. 16, no. 5, pp. 295–306, 1998.
[50] G. Easley, D. Labate, and W.-Q. Lim, "Sparse directional image representations using the discrete shearlet transform," *Applied and Computational Harmonic Analysis*, vol. 25, no. 1, pp. 25–46, 2008.



Bo Jin was born in Nanjing, China. He received the B.Sc. and M.Sc. degrees from the Department of Electrical and Computer Engineering, University of Macau, Macau SAR, China. He is currently doing the Ph.D. research with the Visual Information Security Team, Institute of Systems and Robotics, Portugal.

He published the research results related with Deep Facial Diagnosis which was awarded the national invention patent, PRC. He has a wide range of research interests, especially in computers, robotics and gene.



Leandro Cruz received his PhD degree in Math (applied to Computer Graphics) from Institute for Pure and Applied Mathematics (IMPA-Brazil). During his PhD, he visited for one year the LIRIS (Laboratoire d'InfoRmatique en Image et Systèmes d'information, France). For 2 years, he was a postdoc at IMPA, and for another two years, he was a postdoc and research manager, at ISR-University of Coimbra (Portugal).

In industry, he worked at the Portuguese Mint and National Printing Office, Siemens Process System Engineering, and currently works at Align Technology.

Nuno Gonçalves (Member, IEEE) received the Ph.D. degree in computer vision from the University of Coimbra, Portugal, in 2008. Since 2008, he has been a Tenured Assistant Professor with the Department of Electrical and Computers Engineering, Faculty of Sciences and Technologies, University of Coimbra. He is currently a Senior Researcher with the Institute of Systems and Robotics, University of Coimbra. He has been recently coordinating several projects centered on the technology transfer to the industry. In 2018, he joined the Portuguese Mint and Official Printing Office (INCM), where he coordinates innovation projects in areas, such as biometrics, facial recognition, morphing attack detection, graphical security, security coding, and robotics. He has been working in the design and introduction of new products as result of the innovation projects. He is the author of several papers and communications in high-impact journals and international conferences. His scientific career has been mainly developed in the fields of computer vision, visual information security, biometrics, computer graphics and robotics.