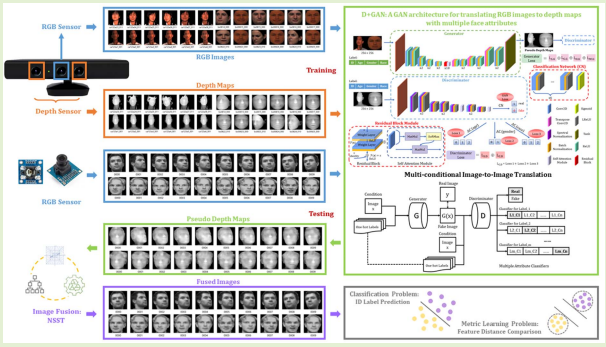
Pseudo RGB-D Face Recognition

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 3D spatial information only from 2D 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 article, 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 multiconditional 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 (NSST), the accuracy of face recognition has been significantly improved.**



***Index Terms*— Depth plus generative adversarial network (D+GAN), face recognition, monocular face depth estimation, pseudo-depth, RGB-D.**

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 3D 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 noninvasiveness, 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 able to perceive the 3D world. 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 Red, Green, Blue-Depth (RGB-D) sensors enable us to utilize 3D information. Compared with RGB face recognition, RGB-D face recognition that 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.

Manuscript received 12 July 2022; accepted 23 July 2022. Date of publication 16 August 2022; date of current version 14 November 2022. This work was supported by the Fundação para a Ciência e a Tecnologia (FCT) under Project UIDB/00048/2020. The associate editor coordinating the review of this article and approving it for publication was Prof. Yu-Dong Zhang. *(Corresponding author: Bo Jin.)*

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

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

Digital Object Identifier 10.1109/JSEN.2022.3197235

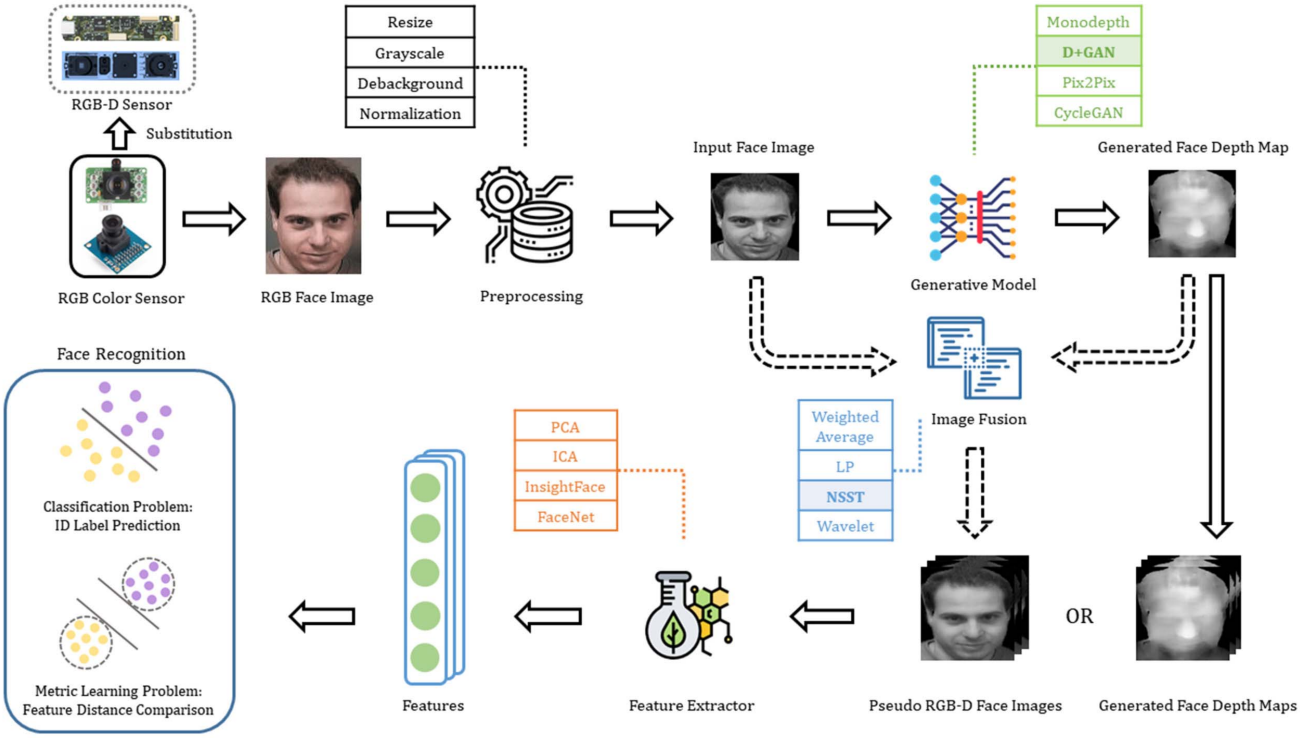


Fig. 1. Pseudo RGB-D face recognition framework.

The emergence of machine learning (ML) allows computers to imitate the human learning process to learn from historical experience to make speculations. It occurs to us that probably by utilizing ML 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 improvement of computer hardware performance, the deep learning (DL) technology that has been widely used in science and industry in recent years has more powerful reasoning performance than traditional ML algorithms. So monocular depth estimation inspired us to acquire 3D information from 2D face images by DL. Synthesizing the aforementioned facts, the thought behind this article 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 article, our contributions could be summarized as follows:

1) We definitely propose and validate a pseudo RGB-D face recognition framework shown in Fig. 1. Fig. 1 presents a modular process. Algorithms within the module lists can be selected for preprocessing, depth-generating, image fusion, and feature extraction, and therefore can 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 generative adversarial network (GAN)-based model, D+GAN, to perform the multiconditional 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 Section II, we review the related works. In Section III, we describe our proposed methods and their implementations. Our experimental results are analyzed and discussed in Section IV. In Section V, 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. The traditional ML method is to extract hand-crafted features which are designed by specialists to reduce the complexity of input data, and to train a model from the input to discover the pattern to make decisions. Turk and Pentland [10] proposed the Eigenfaces method for face recognition on a smaller set of face image features approximating the set of known face images. Bartlett *et al.* [11] proposed using the independent component analysis (ICA) method for face recognition, and they showed that ICA representations were superior to principal component analysis (PCA)-based representations for face recognition across changes in some conditions. Phillips [12] developed a support vector machine (SVM)-based algorithm to generate the decision surface for face recognition. In the past ten years, traditional ML methods have increasingly been replaced

by DL 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. Krizhevsky *et al.* [14] proposed AlexNet which is a classic CNN framework to classify a large amount of images in ILSVRC-2010. Taigman *et al.* [15] presented a DeepFace system, which can reach human-level performance in face recognition. The backbone network of DeepFace is based on AlexNet, and the loss function used is Softmax. Szegedy *et al.* [16] proposed a 22-layer deep CNN (DCNN), GoogLeNet, which is a variant of the inception network. Schroff *et al.* [17] presented FaceNet, which uses GoogLeNet as the backbone network and the triplet loss function for training to map the face image to the Euclidean space directly. He *et al.* [18] proposed ResNet, which can increase the network depth to 152 layers by using residual blocks. Deng *et al.* [19] 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.

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. Lee *et al.* [9] used a 12-layer deep neural network which is first trained with a color face dataset and later fine-tuned on depth face images for feature extraction to perform joint classification. Kim *et al.* [20] applied a fine-tuned DCNN to extract features from 2D depth maps converted from 3D point clouds for calculating the distance for face matching. Moreover, Jiang *et al.* [21] tried to propose an attribute-aware loss function for RGB-D facial data.

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 ML methods to estimate the depth of human faces from monocular images since the 1990s. Lai *et al.* [22] estimated the depth from defocus by using the raw image data in the vicinity of the edge. Sun and Lam [23] converted depth estimation into an ICA problem by incorporating a prior from the CANDIDE 3D face model. Sun *et al.* [24] employed the nonlinear least-squares model to estimate the depth values of facial feature points and the pose of the 2D face image. Since 2014, with the development of DL, researchers have successively used DL methods to perform monocular face depth estimation, which is similar to face recognition. Cui *et al.* [25] presented a deep neural network with a cascaded fully convolutional network (FCN) and CNN architecture to estimate depth information of RGB face images. Pini *et al.* [26] applied a conditional GAN (cGAN) for learning to translate intensity face images into their corresponding depth maps. Arslan and Seke *et al.* [27] applied a conditional Wasserstein GAN to perform face depth estimation. Jin *et al.* [28] predicted face depth maps by using pretrained models for scene depth estimation directly.

III. Materials and Methods

GAN, proposed by Goodfellow *et al.* [29], is a model that learns a mapping from random noise vectors to output images. The original GAN consists of two parts, which are a generator and a discriminator. The objective of the generator is to map the input Gaussian noise into a fake image, and the objective of 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 cGAN, proposed by Mirza and Osindero [30], is a supervised model that can generate output images with a desired condition from random noise. Pix2Pix, proposed by Isola *et al.* [31], 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. Auxiliary classifier GAN (ACGAN), proposed by Odena *et al.* [32], 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.

For adapting our task that is generating the corresponding depth from RGB face images better, we comprehensively refer to the aforementioned network structures and cooperate with some advanced skills, and propose the D+GAN. Fig. 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 meets 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 to 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.

1) 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 are acquired by a 3D system based on the structured light. The ground-truth depth images and their corresponding color images are transformed from 3D point cloud files provided by the Bosphorus database.

2) CASIA 3D Face Database: The CASIA 3D Face Database collected by the Chinese Academy of Sciences contains

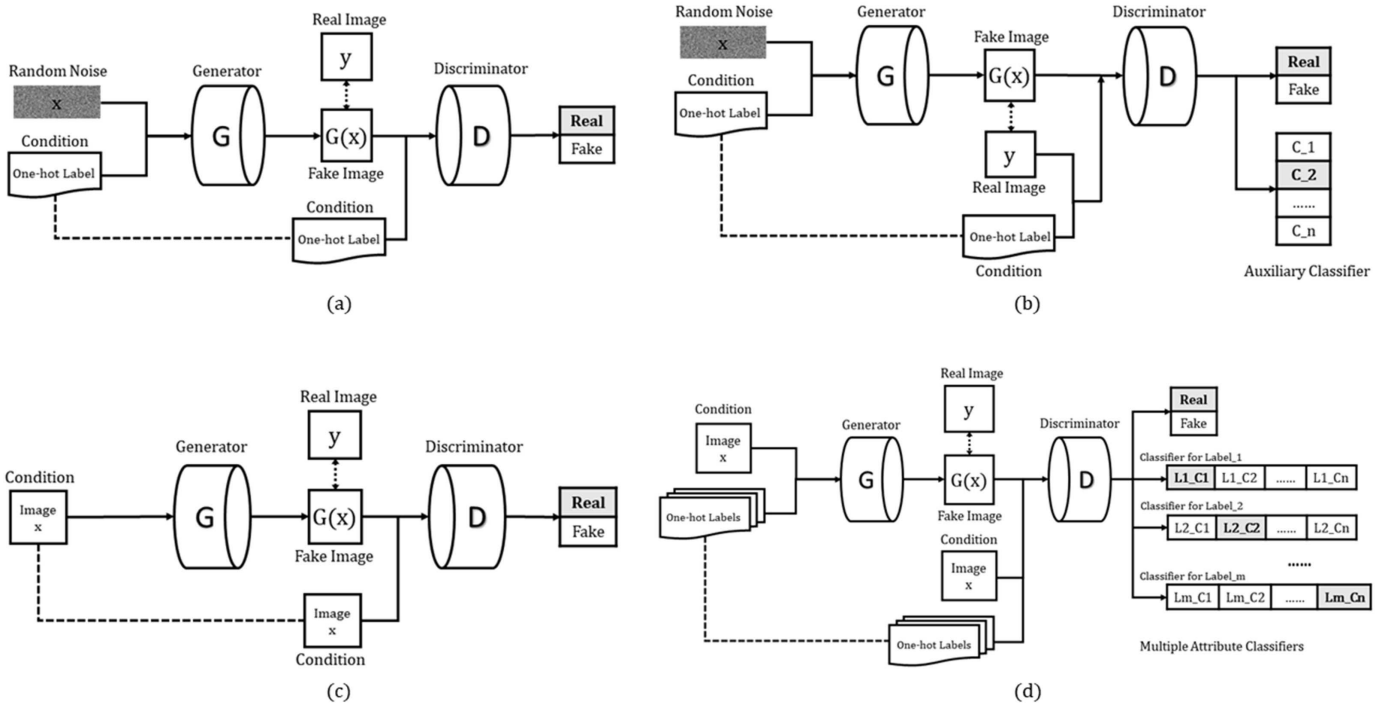


Fig. 2. Main structures of GANs. (a) cGAN. (b) ACGAN. (c) Pix2Pix. (d) Ours: D+GAN.

4624 scans of 123 persons. The scans are collected by the Minolta Vivid 910, which is a noncontact 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.

3) 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. First, 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 GAN 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 convolutional 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. Fig. 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 the decoder to build an information flow transmission approach, which can relieve the gradient disappearance problem effectively. The encoder is composed of eight 2D convolutional layers, as shown in Fig. 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.

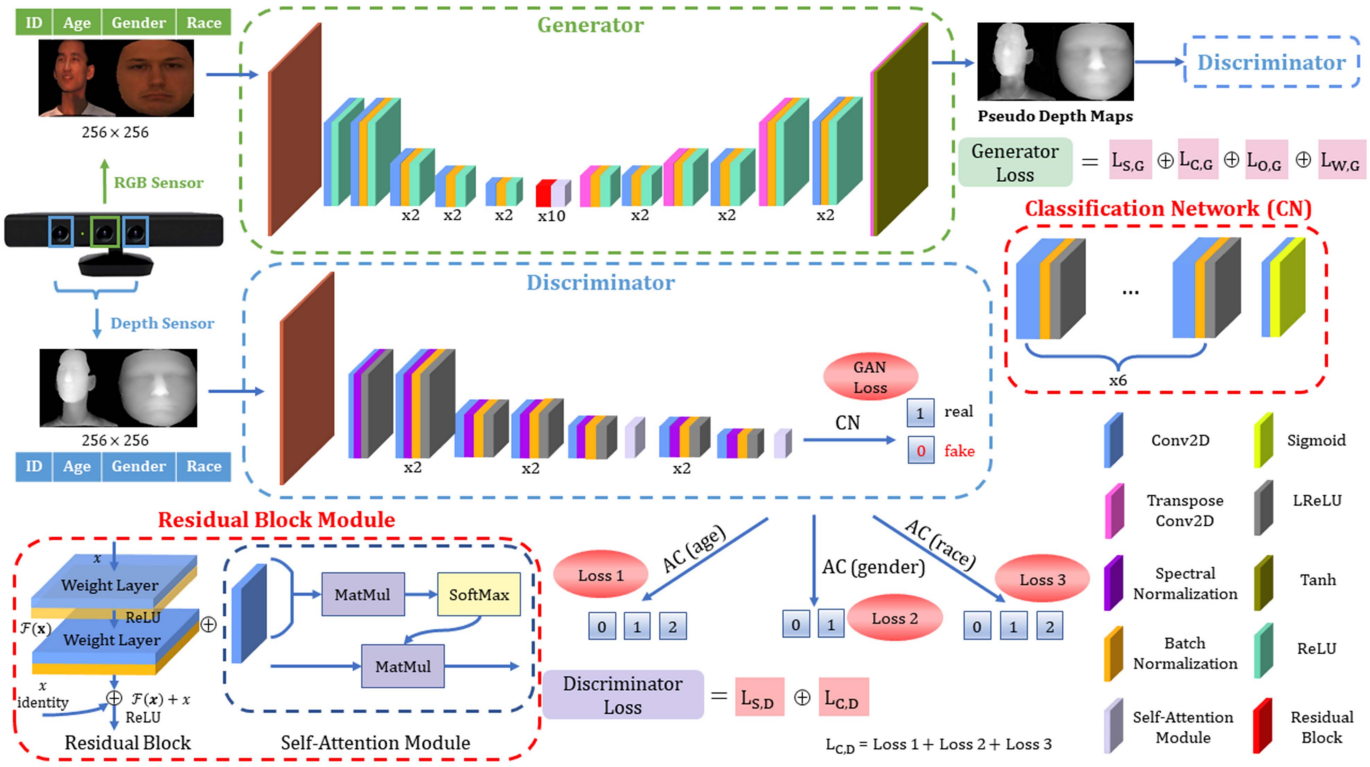


Fig. 3. D+GAN: a GAN architecture for translating RGB images to depth maps with multiple face attributes.

There is one batch normalization (BN) layer for normalizing input features to accelerate the convergence process and one layer with the rectified linear unit (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 CNN and the deconvolutional neural network. In the decoder, the CNN is designed for feature extraction, and its calculation method is the same as that of the encoder, while the 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 Fig. 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. Layers 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 the 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 easily 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: The 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 multilevel and long-range dependencies across image regions, which is complementary to the convolutional 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, and 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 $q_{m\times(n/8)}$, $k_{m\times(n/8)}$ and $v_{m\times n}$ are obtained, respectively, where m represents the feature size. The final output of attention weight distribution is computed as follows:

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

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 earlier before the branch node. In detail, there are ten convolutional layers where the number of convolution kernel 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 that the first layer is 5×5. In order to make the training process more stable, we set up spectral

normalization [39] in these ten 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 follows:

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

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 2D 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 follows:

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

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 follows:

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

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 follows:

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

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 follows:

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

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

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

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 follows:

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

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

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

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

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

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 earlier 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), 3D Morphable Model (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. The 3DMM [42] is to generate a textured 3D face with parameters, including vertices, triangles, and attribute based on the 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. The Adam optimizer is used for Pix2Pix and CycleGAN, while the Adadelta optimizer is used for D+GAN.

The identifiers (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.

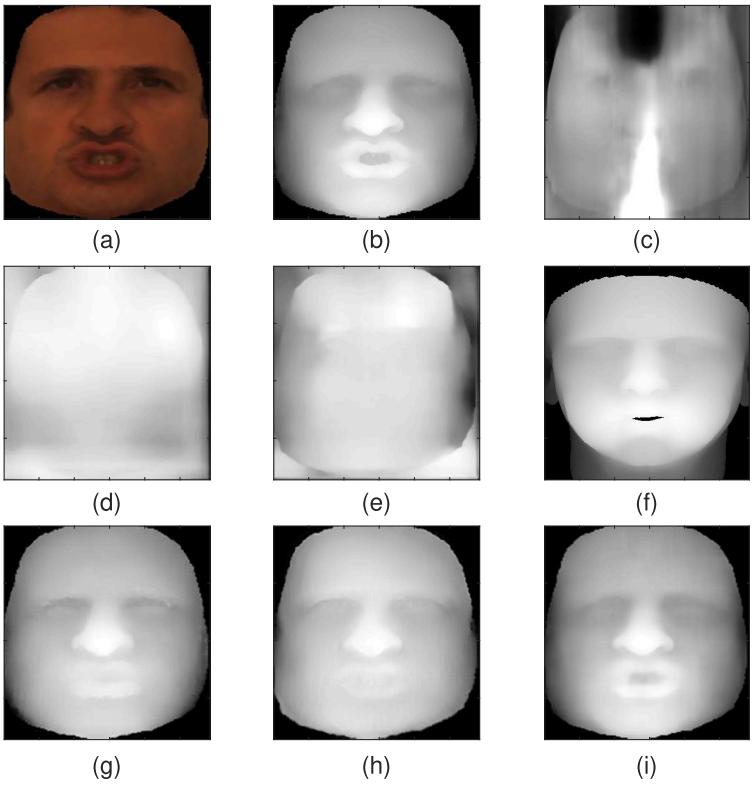


Fig. 4. Face depth maps generated by various models in the case of bs016_LFAU_22_0. (a) Input RGB image. (b) Ground-truth depth map. (c) Model: Monodepth2. (d) Model: DenseDepth (KITTI). (e) Model: DenseDepth (NYU-Depth V2). (f) Model: 3DMM. (g) Model: Pix2Pix. (h) Model: CycleGAN. (i) Proposed Model: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

1) Case Study: bs016_LFAU_22_0 of Bosphorus 3D Face Database: Fig. 4 presents the results for the case of bs016_LFAU_22_0 of Bosphorus 3D Face Database. Fig. 4(a) shows the RGB face image, which is transformed from 3D data provided, and Fig. 4(b) shows the ground-truth face depth map, which is transformed from 3D data provided. Fig. 4(c) shows the output generated by Monodepth2. The result shows the contour of the face vaguely, and the relative depth information is not accurately expressed. Fig. 4(d) 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. Fig. 4(e) shows the output generated by DenseDepth (NYU-Depth V2). The result shows the depth better but still lacks the facial detailed depth. Fig. 4(f) 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 is visually very different from the ground-truth. We infer that this is because 3DMM is based on an average model. Visually, Fig. 4(g) and Fig. 4(h) show the basically satisfactory results which are generated by Pix2Pix and CycleGAN. Fig. 4(i) 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 $I(x, y)$. The autocorrelation function is defined as follows:

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

where (a, b) is the pixel in the window with the size of $(2w+1) \times (2w+1)$. $\xi, \eta = \pm 0, \pm 1, \pm 2, \ldots, \pm N$. 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 in Fig. 5. In the autocorrelation function graph, a larger downward trend as ξ and η increasing means a larger coarseness of the corresponding image. Fig. 5(b) shows the autocorrelation function graph of the ground-truth depth map. Compared with Fig. 5(a), Fig. 5(b) has a smaller downward trend as ξ and η increasing, which means that the depth map has a lower coarseness than its corresponding grayscale image. Subjectively, the spatial details of the face should be changed regularly. Compared with Fig. 5(b), Fig. 5(c), (d), and (e) have a larger downward trend as ξ and η increasing, which means that 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 Fig. 5(f)–(i) are similar to the shapes of Fig. 5(b), which indicates that the depth maps generated by 3DMM, Pix2Pix, CycleGAN, and D+GAN have a higher quality.

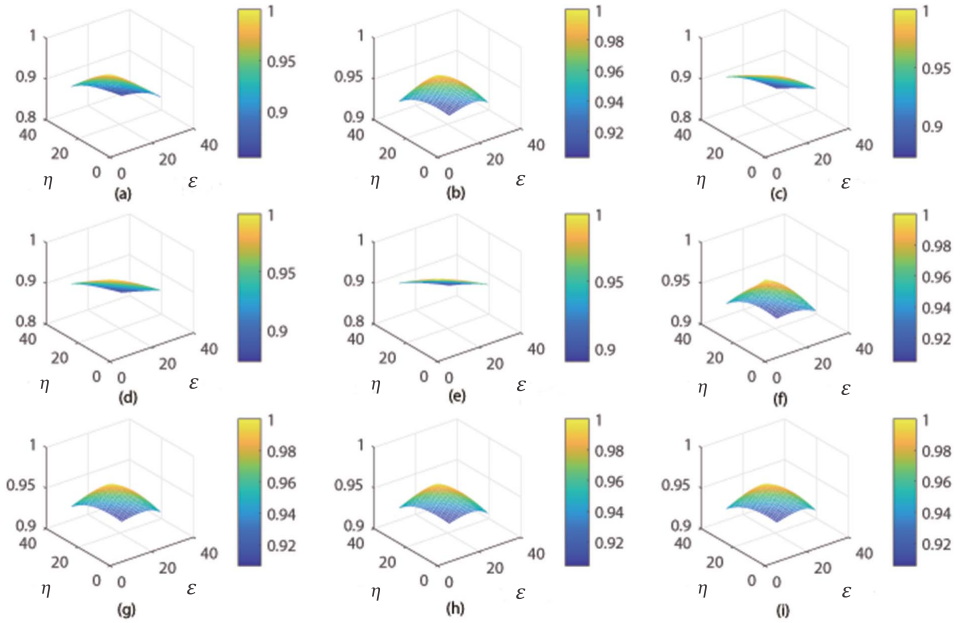


Fig. 5. Autocorrelation function graphs of various output images: (a) Original RGB image, (b) Ground-truth depth map, (c) Depth map generated by Monodepth2, (d) Depth map generated by DenseDepth (KITTI), (e) Depth map generated by DenseDepth (NYU-Depth V2), (f) Depth map generated by 3DMM, (g) Depth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

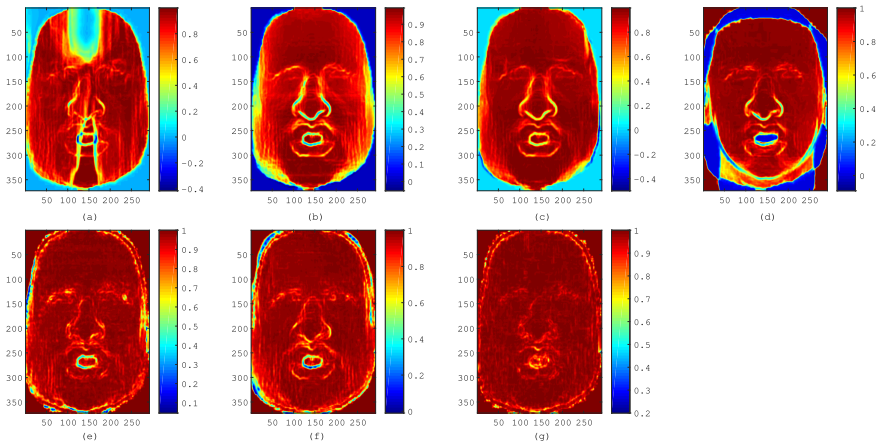


Fig. 6. Local SSIM maps of depth maps generated by various models. (a) Model: Monodepth2. (b) Model: DenseDepth (KITTI). (c) Model: DenseDepth (NYU-Depth V2). (d) Model: 3DMM. (e) Model: Pix2Pix. (f) Model: CycleGAN. (g) Proposed model: D+GAN.

In the case of bs016_LFAU_22_0 of Bosphorus 3D Face Database, local structural similarity index measure (SSIM) maps of the depth maps generated by various models are shown in Fig. 6. The 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 Fig. 6 seen, Fig. 6(g) representing D+GAN has the most red area. Fig. 6(e) representing Pix2Pix and Fig. 6(f) representing CycleGAN in overall perform well except in specific areas of eyes, nose, and mouth in comparison with Fig. 6(g). Fig. 6(d) representing 3DMM shows a larger difference in face shape besides in eyes, nose, and mouth. In addition, besides eyes, nose, and mouth areas, Fig. 6(a) representing Monodepth2, Fig. 6(b) representing DenseDepth (KITTI), and Fig. 6(c) representing DenseDepth (NYU-Depth V2) show a larger difference in four corners out of the face. Among these three, Fig. 6(c) shows a less difference in the area of the human face.

2) Case Study: 008-025 of CASIA 3D Face Database: Fig. 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 aforementioned example. Fig. 7(g)–Fig. 7(i) representing three GAN models show a satisfactory result. Especially for Fig. 7(i) 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 Fig. 7(f)].

In the case of 008-025 of CASIA 3D Face Database, autocorrelation function graphs of depth maps generated by various models are shown in Fig. 8. It shows the coarseness

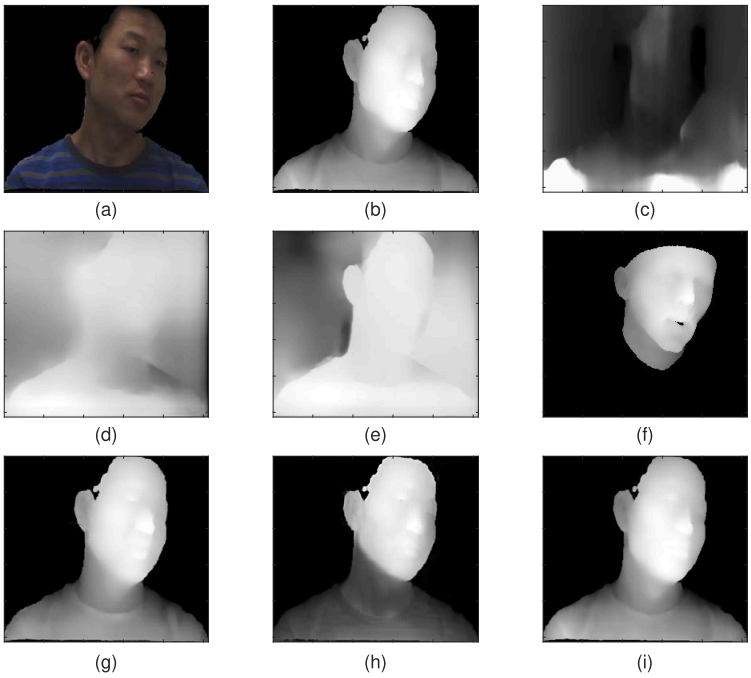


Fig. 7. Face depth maps generated by various models in the case of 008-025. (a) Input RGB image. (b) Ground-truth depth map. (c) Model: Monodepth2. (d) Model: DenseDepth (KITTI). (e) Model: DenseDepth (NYU-Depth V2). (f) Model: 3DMM. (g) Model: Pix2Pix. (h) Model: CycleGAN. (i) Proposed model: D+GAN.

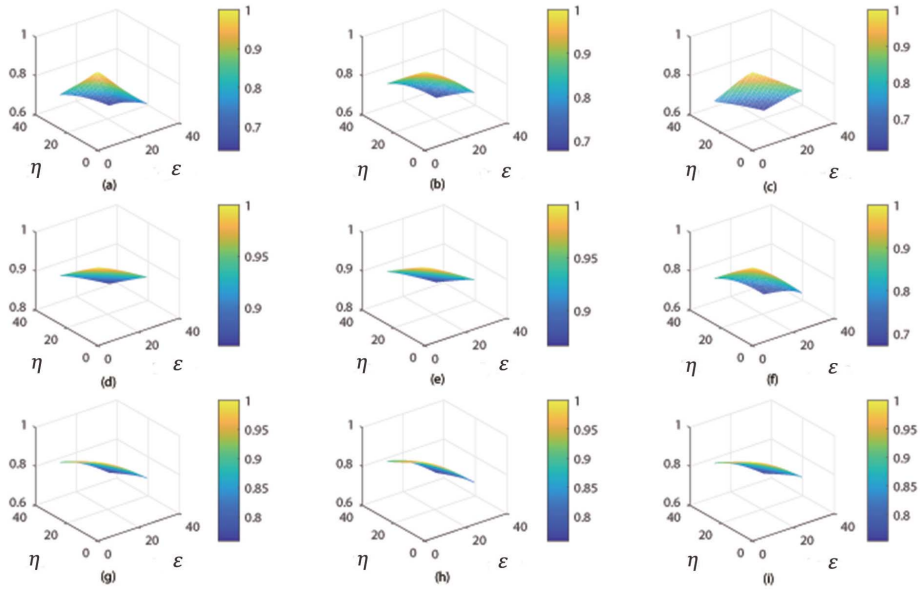


Fig. 8. Autocorrelation function graphs of various output images: (a) Original RGB image, (b) Ground-truth depth map, (c) Depth map generated by Monodepth2, (d) Depth map generated by DenseDepth (KITTI), (e) Depth map generated by DenseDepth (NYU-Depth V2), (f) Depth map generated by 3DMM, (g) Depth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

of the generated depth map. It is worth mentioning that Fig. 8 indicates that the texture coarseness of the depth map of the bust should be higher than the face (see Fig. 5). Compared with Fig. 8(b), Fig. 8(d) and Fig. 8(e) have a smaller downward trend as ξ and η increasing, which means that the depth maps generated by DenseDepth (KITTI) and DenseDepth (NYU-Depth V2) have a lower coarseness than the ground-truth depth map. In contrast, Fig. 8(g)–Fig. 8(i) representing three GAN models have similar trends to Fig. 8(b), which implies that they retain depth information well.

In the case of 008-025 of CASIA 3D Face Database, the local SSIM maps of the depth maps generated by various models are shown in Fig. 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

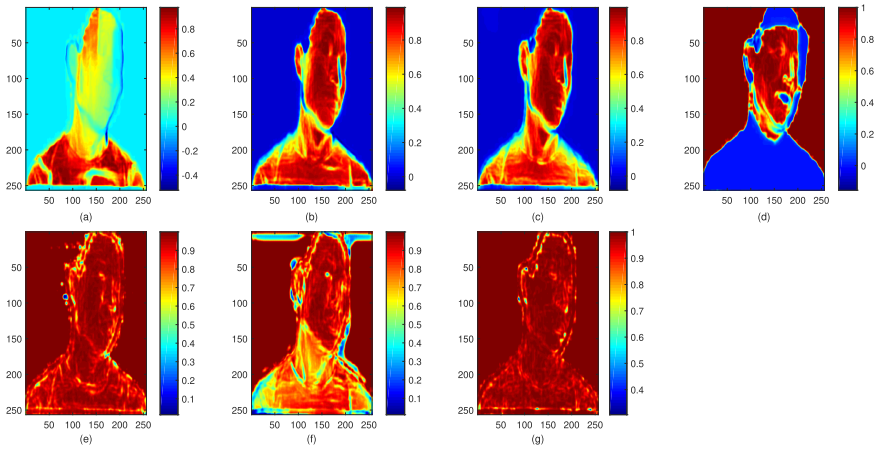


Fig. 9. Local SSIM maps of depth maps generated by various models. (a) Model: Monodepth2. (b) Model: DenseDepth (KITTI). (c) Model: DenseDepth (NYU-Depth V2). (d) Model: 3DMM. (e) Model: Pix2Pix. (f) Model: CycleGAN. (g) Proposed model: D+GAN.

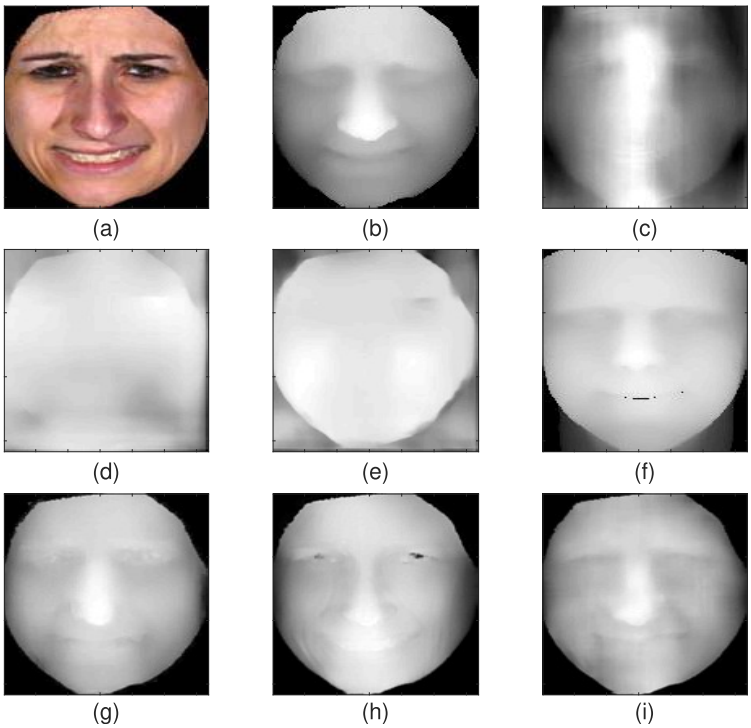


Fig. 10. Face depth maps generated by various models in the case of F0010_FE03WH_F2D. (a) Input RGB image. (b) Ground-truth depth map. (c) Model: Monodepth2. (d) Model: DenseDepth (KITTI). (e) Model: DenseDepth (NYU-Depth V2). (f) Model: 3DMM. (g) Model: Pix2Pix. (h) Model: CycleGAN. (i) Proposed model: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

generated by CycleGAN are not as satisfactory as Pix2Pix and D+GAN [see Fig. 9(f)].

3) Case Study: F0010_FE03WH_F2D of BU-3DFE Database: Fig. 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 the BU-3DFE Database. In all, the performance of each model is similar to that in the first example. Fig. 10(g)–Fig. 10(i) representing three GAN models show a more satisfactory result than others. In detail, Fig. 10(g) and (h) representing Pix2Pix and CycleGAN, respectively, show an inaccurate depth in the eyes area. However, D+GAN performs well in the eyes area [see Fig. 10(i)]. It is worth mentioning that 3DMM generates inaccurate results in the face shape again [see Fig. 10(f)].

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

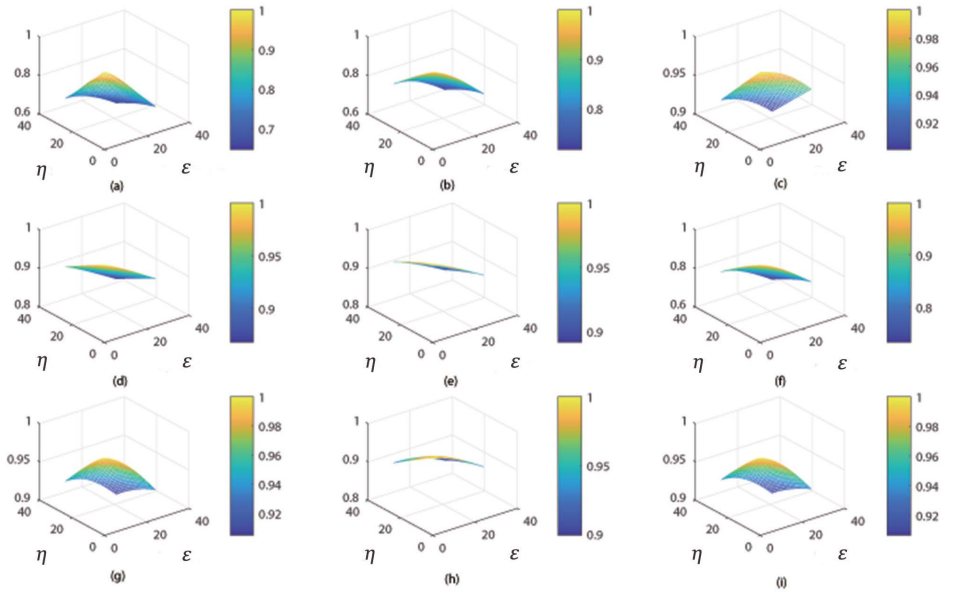


Fig. 11. Autocorrelation function graphs of various output images: (a) Original RGB image, (b) Ground-truth depth map, (c) Depth map generated by Monodepth2, (d) Depth map generated by DenseDepth (KITTI), (e) Depth map generated by DenseDepth (NYU-Depth V2), (f) Depth map generated by 3DMM, (g) Depth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

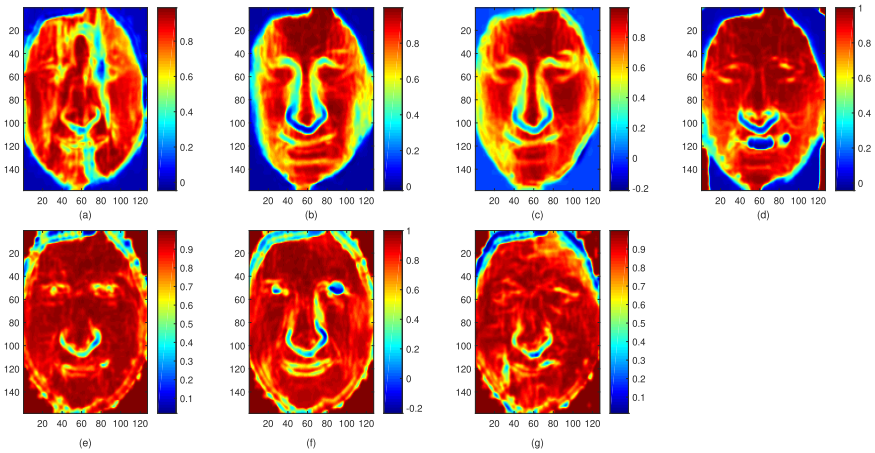


Fig. 12. Local SSIM maps of the depth maps generated by various models. (a) Model: Monodepth2. (b) Model: DenseDepth (KITTI). (c) Model: DenseDepth (NYU-Depth V2). (d) Model: 3DMM. (e) Model: Pix2Pix. (f) Model: CycleGAN. (g) Proposed model: D+GAN.

models are shown in Fig. 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 Fig. 12(f)). In comparison with Fig. 12(e), Fig. 12(g) representing D+GAN performs better in the area of the eyes.

B. Quantitative Results and Analysis

In this section, a quantitative analysis is carried out. The SSIM, root-mean-squared error (RMSE), and peak signal-to-noise ratio (PSNR) are selected to evaluate the quality of the face depth map generated by several models on three datasets described before which are the Bosphorus 3D Face Database, CASIA 3D Face Database and BU-3DFE Database.

The 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 aforementioned six models as follows:

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

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 aforementioned 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.

Weights $\alpha > 0$, $\beta > 0$ and $\gamma > 0$. C_1, C_2 and C_3 are all constants to avoid the denominator being 0.

A lower RMSE value means a more accurate result corresponding to the reference. The RMSE between images a and b is calculated as follows:

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

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

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

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

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 I. 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 aforementioned three GAN models, Pix2Pix requires 18.6G 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 ML and DL 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 the wavelet transform-based methods, the Laplacian pyramid, and 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 follows:

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

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

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

and S, the shear matrix, is expressed as follows:

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

The NSST performs multiscale and multidirectional 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 subband images decomposed are transformed and combined into new subband images. Finally, the final fused image is achieved by the inverse NSST on the new subband 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 Section IV-B. If the training images are sufficient, due to the great learning ability of the DL 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 datasets, 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 component 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 datasets, 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 the 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, the total number of people is also relatively small. Therefore, using the pretrained 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

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

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%

advanced DL models on the five classical face recognition datasets.

Specifically, in experiments of testing the FaceNet: Inception 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 the pseudo RGB-D face recognition proposed is able to improve the accuracy in comparison with RGB face recognition by using different classic traditional ML and DL models. Especially for traditional ML models, the 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.

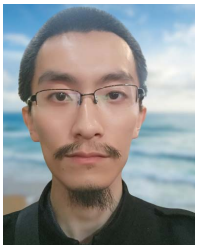
Acknowledgment

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*. Evanston, IL, USA: Routledge, 2004.
[2] B. Jin, "Deep learning facial diagnosis system, ZL201711255031.1," CN Patent 10 880 679 2B, 2022.
[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. 123649–123661, 2020.
[4] Y. Wang and M. Kosinski, "Deep neural networks are more accurate than humans at detecting sexual orientation from facial images," *J. Personality Social Psychol.*, vol. 114, no. 2, p. 246, 2018.
[5] J. Luo *et al.*, "How much does input data type impact final face model accuracy?" in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 18985–18994.
[6] Z. Zhang, "Microsoft Kinect sensor and its effect," *IEEE Multimedia*, vol. 19, no. 2, pp. 4–10, Feb. 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 J.*, vol. 17, no. 14, pp. 4508–4519, Jul. 2017.
[8] G. Goswami, S. Bharadwaj, M. Vatsa, and R. Singh, "On RGB-D face recognition using Kinect," in *Proc. IEEE 6th Int. Conf. Biometrics, Theory, Appl. Syst. (BTAS)*, Sep. 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 *Proc. BMVC*, 2016, vol. 1, no. 2, pp. 1–14.
[10] M. Turk and A. Pentland, "Eigenfaces for recognition," *J. Cognit. Neurosci.*, 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 Trans. Neural Netw.*, vol. 13, no. 6, pp. 1450–1464, Nov. 2002.
[12] P. Phillips, "Support vector machines applied to face recognition," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 11, 1998, pp. 803–809.
[13] O. Russakovsky *et al.*, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.
[14] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 25, Dec. 2012, pp. 1097–1105.
[15] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "DeepFace: Closing the gap to human-level performance in face verification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1701–1708.
[16] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.
[17] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 815–823.
[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[19] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4690–4699.
[20] D. Kim, M. Hernandez, J. Choi, and G. Medioni, "Deep 3D face identification," in *Proc. IEEE Int. Joint Conf. Biometrics (IJCB)*, Oct. 2017, pp. 133–142.
[21] L. Jiang, J. Zhang, and B. Deng, "Robust RGB-D face recognition using attribute-aware loss," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 10, pp. 2552–2566, Oct. 2020.
[22] S.-S. Lai, C.-C. Fu, and S. Chang, "A generalized depth estimation algorithm with a single image," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 14, no. 4, pp. 405–411, Apr. 1992.
[23] Z.-L. Sun and K.-M. Lam, "Depth estimation of face images based on the constrained ICA model," *IEEE Trans. Inf. Forensics Security*, vol. 6, no. 2, pp. 360–370, Jun. 2011.
[24] Z. L. Sun, K. M. Lam, and Q. W. Gao, "Depth estimation of face images using the nonlinear least-squares model," *IEEE Trans. Image Process.*, vol. 22, no. 1, pp. 17–30, Jan. 2013.
[25] J. Cui, H. Zhang, H. Han, S. Shan, and X. Chen, "Improving 2D face recognition via discriminative face depth estimation," in *Proc. Int. Conf. Biometrics (ICB)*, Feb. 2018, pp. 140–147.
[26] S. Pini, F. Grazioli, G. Borghi, R. Vezzani, and R. Cucchiara, "Learning to generate facial depth maps," in *Proc. Int. Conf. 3D Vis. (3DV)*, Sep. 2018, pp. 634–642.
[27] A. T. Arslan and E. Seke, "Face depth estimation with conditional generative adversarial networks," *IEEE Access*, vol. 7, pp. 23222–23231, 2019.
[28] B. Jin, L. Cruz, and N. Gonçalves, "Face depth prediction by the scene depth," in *Proc. IEEE/ACIS 19th Int. Conf. Comput. Inf. Sci. (ICIS)*, Jun. 2021, pp. 42–48.
[29] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 27, 2014, pp. 1–9.
[30] M. Mirza and S. Osindero, "Conditional generative adversarial nets," 2014, *arXiv:1411.1784*.
[31] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1125–1134.
[32] A. Odena, C. Olah, and J. Shlens, "Conditional image synthesis with auxiliary classifier GANs," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 2642–2651.

[33] A. Savran *et al.*, "Bosphorus database for 3D face analysis," in *Proc. Eur. Workshop Biometrics Identity Manag.* Berlin, Germany: Springer, 2008, pp. 47–56.
[34] The Institute of Automation of the Chinese Academy of Sciences (CASIA). *CASIA-3D Face V1*. [Online]. Available: 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 *Proc. 7th Int. Conf. Autom. Face Gesture Recognit. (FGR)*, 2006, pp. 211–216.
[36] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-9, no. 1, pp. 62–66, Jan. 1979.
[37] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 7354–7363.
[38] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2015, pp. 234–241.
[39] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral normalization for generative adversarial networks," 2018, *arXiv:1802.05957*.
[40] C. Godard, O. M. Aodha, M. Firman, and G. Brostow, "Digging into self-supervised monocular depth estimation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3828–3838.
[41] I. Alhashim and P. Wonka, "High quality monocular depth estimation via transfer learning," 2018, *arXiv:1812.11941*.
[42] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter, "A 3D face model for pose and illumination invariant face recognition," in *Proc. 6th IEEE Int. Conf. Adv. Video Signal Based Surveill.*, Sep. 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 Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 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 Trans. Multimedia*, early access, Mar. 7, 2022, doi: 10.1109/TMM.2022.3156699.
[45] F. S. Samaria and A. C. Harter, "Parameterisation of a stochastic model for human face identification," in *Proc. IEEE Workshop Appl. Comput. Vis.*, Dec. 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 *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 1996, pp. 43–58.
[47] D. B. Graham and N. M. Allinson, "Characterising virtual eigensignatures for general purpose face recognition," in *Face Recognition*. Berlin, Germany: Springer, 1998, pp. 446–456.
[48] A. Martinez and R. Benavente, "The AR face database," Auton. Univ. Barcelona, Barcelona, Spain, CVC Tech. Rep., 24, 1998. [Online]. Available: https://portalrecerca.uab.cat/en/publications/the-ar-face-database-cvc-technical-report-24
[49] P. J. Phillips, H. Wechsler, J. Huang, and P. J. Rauss, "The FERET database and evaluation procedure for face-recognition algorithms," *Image Vis. Comput.*, 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," *Appl. Comput. Harmon. Anal.*, vol. 25, no. 1, pp. 25–46, Jul. 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 was doing Ph.D. research with the Visual Information Security Team, Institute of Systems and Robotics (ISR), Portugal.

He published research results related to Deep Facial Diagnosis, which was awarded the national invention patent, the People's Republic of China (PRC). He has a wide range of research interests, especially in computers, robotics, and genes.

Leandro Cruz received the Ph.D. degree in mathematics (applied to computer graphics) from the Institute for Pure and Applied Mathematics (IMPA), Rio de Janeiro, Brazil.

During his Ph.D. degree, he visited the Laboratoire d'InfoRmatique en Image et Systèmes d'information (LIRIS), Écully, France, for one year. For two years, he was a Postdoctoral Researcher at IMPA, and for another two years, he was a Postdoctoral Researcher and the Research Manager at The Institute of Systems and Robotics (ISR)-University of Coimbra, Coimbra, Portugal. In industry, he worked at Portuguese Mint and National Printing Office, Lisbon, Portugal, Siemens Process Systems Engineering, London, U.K., and currently works at Align Technology, San Jose, CA, USA.

Nuno Gonçalves (Member, IEEE) received the Ph.D. degree in computer vision from the University of Coimbra, Coimbra, Portugal, in 2008.

Since 2008, he has been a Tenured Assistant Professor with the Department of Electrical and Computer Engineering, Faculty of Sciences and Technologies, University of Coimbra, where he is currently a Senior Researcher with the Institute of Systems and Robotics (ISR). He has been recently coordinating several projects centered on the technology transfer to the industry. In 2018, he joined Portuguese Mint and Official Printing Office (INCM), Lisbon, Portugal, 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 a 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.