RGB-D Face Recognition with Identity-Style Disentanglement and Depth Augmentation

Meng-Tzu Chiu[1], Hsun-Ying Cheng[1], Chien-Yi Wang[2], Shang-Hong Lai[1,2],

Abstract—Deep learning approaches achieve highly accurate face recognition by training the models with huge face image datasets. Unlike 2D face image datasets, there is a lack of large 3D face datasets available to the public. Existing public 3D face datasets were usually collected with few subjects, leading to the over-fitting problem. This paper proposes two CNN models to improve the RGB-D face recognition task. The first is a segmentation-aware depth estimation network, called DepthNet, which estimates depth maps from RGB face images by exploiting semantic segmentation for more accurate face region localization. The other is a novel segmentation-guided RGB-D face recognition model that contains an RGB recognition branch, a depth map recognition branch, and an auxiliary segmentation mask branch. In our multi-modality face recognition model, a feature disentanglement scheme is employed to factorize the feature representation into identity-related and style-related components. DepthNet is applied to augment a large 2D face image dataset to a large RGB-D face dataset, which is used for training our RGB-D face recognition model. Our experimental results show that DepthNet can produce more reliable depth maps from face images with the segmentation mask. Our multi-modality face recognition model fully exploits the depth map and outperforms state-of-the-art methods on several public 3D face datasets with challenging variations.

Index Terms—Depth Estimation, 3D Face Recognition, RGB-D Face Recognition, Multi-modality Face Recognition, Face Representation Learning, Disentangled Representation Learning.

1 Introduction

FACE recognition can be applied to many different tasks in the real world, such as video surveillance, biometric identification, security verification, etc. 2D face recognition has been a rapidly developing application for many years, and the deep learning based method has achieved very high accuracy in most public datasets. However, it has been shown that face recognition is still very challenging under large pose variations [1]. Some proposed face-frontalization methods [13] [35] can normalize profile face images to frontal pose images to overcome variations in head pose. More and more methods have recently focused on RGB-D face recognition since it achieves enhanced recognition performance than 2D face recognition methods. Unlike the 2D face recognition approach that uses only RGB images as input, RGB-D face recognition includes depth as additional information, thus leading to more robust performance against challenging variations such as large pose, expression, illumination, and occlusion.

However, developing a 3D model requires multiple cameras or extra depth sensors, which makes it an expensive option compared to acquiring 2D images. Therefore, the lack of large 3D or RGB-D face datasets available to the public is the main reason why the development of RGB-D face recognition is slower than 2D face recognition. The numbers of subjects in most 3D or RGB-D face datasets are much smaller than those in 2D face datasets. Numerous 2D datasets contain more than thousands of identities and millions of images [46] [5] [18]. In contrast, existing public 3D face datasets usually contain only hundreds of subjects or at most thousands of images [47] [34] [11]. It is easy to fall into the over-fitting problem when we only use a limited number of subjects in a 3D dataset to train a face recognition model.

To address the problem of lacking large 3D face datasets for model training, many works [53] [24] [4] applied different data augmentation methods to train their face recognition models. [53] presented a unique protocol that merges the publicly available 3D face datasets to generate a large 3D face datasets for large-scale face recognition testing. [24] changed the values of expression parameters of the 3DMM model and randomly generated rigid transformations matrices to the input 3D point cloud to synthesize expression and pose variations. [4] applied pose augmentation on 3D scan and performed resolution and transformational augmentation on range images to enlarge the training set size. [46] generated new identities by morphing two 3D face models of different identities. These methods construct the augmented face data with virtual identity, and it is tough to generate realistic identity-preserving intra-person variations for the synthesized 3D face data for virtual identities.

Our previous work [7] overcomes the limitations mentioned earlier by converting a 2D face dataset to an RGB-D face dataset to address inadequate numbers of subjects in RGB-D face datasets for model training. Our system has two major parts, i.e., the depth estimation module (DepthNet) and the multi-modality face recognition module. As an auxiliary task, we include a face semantic segmentation branch into the depth estimation network model. The module can correctly recognize the local facial features to estimate realistic face depth images. The proposed multi-modality face recognition model takes RGB face images, segmentation

- [1] *Department of Computer Science, National Tsing Hua University, Taiwan*
- [2] *Microsoft Artificial Intelligence R&D Center, Taiwan*

Manuscript received XX XX, XXXX; revised XX XX, XXXX.

masks, and generated depth maps as input, and this model can achieve high-accuracy RGB-D face recognition. Thus, we can convert a large 2D face dataset to the corresponding RGB-D face dataset with the same number of subjects and intra-variations for training the RGB-D face recognition model.

The RGB-D face recognition network in [7] is trained and tested on the augmented depth images. However, a domain gap exists between the augmented depth map and the ground truth depth map, and it could cause decreasing in RGB-D face recognition accuracy. This paper presents an extended version of our preliminary work [7] and proposes a novel identity-style disentanglement framework to minimize such domain shift. The framework of our method is shown in Fig. 1. The main contributions of this work can be summarized as follows:

1) We propose a novel depth estimation CNN model called DepthNet, which includes semantic segmentation to estimate a more accurate depth map than the existing face depth estimation networks.
2) By applying the proposed DepthNet to a large RGB face image dataset, we obtain the corresponding RGB-D dataset with a large number of subjects and large intra-variations, which can be used for training accurate RGB-D face recognition models.
3) We propose a novel multi-modality face recognition model which contains an RGB recognition branch, a depth map recognition branch, and an auxiliary segmentation mask branch with a spatial attention module to overcome challenging variations in expression, pose, and occlusion.
4) We introduce three well-designed objective functions to learn a discriminative representation for the RGB-D face recognition task. The cross-modal focal loss ensures the model training focuses on the samples that both recognition branches can not classify correctly. The semantic alignment loss engages the extracted identity feature to represent the same semantic meaning as the subject identity among different modalities. The feature disentanglement loss aims to disentangle the feature representation into identity-related information and style-related components to ease the domain gap between the augmented depth map and the ground truth map.
5) Experiments on several public 3D face datasets demonstrate that the proposed multi-modality face recognition model outperforms the state-of-the-art methods for RGB-D face recognition.

2 Related Works

2.1 3D Dataset

3D face data could provide additional 3D geometric information than a 2D face image. Nonetheless, the development of 3D face recognition is somehow limited by the availability of large 3D face datasets to the research community. Most of the largest 3D face datasets do not contain enough subjects for training highly accurate 3D face recognition models. For example, the ND-2006 dataset [11] consists of only 888 subjects with totally 13,450 scans, and FRGCv2 [34] consists of only 466 subjects with totally 4007 scans. The reason is the time-consuming process of collecting 3D data and the lack of 3D face data in public. In contrast, most of the large-scale 2D face datasets usually contain a lot more than thousands of identities and millions of images [46] [5] [18].

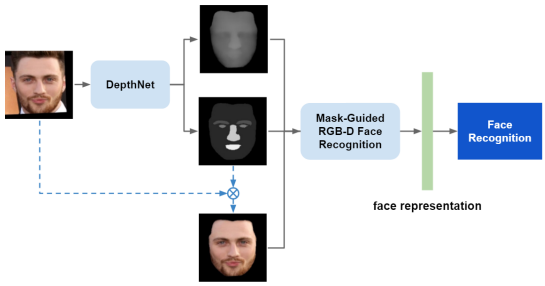

Fig. 1. The pipeline of the proposed RGB-D face recognition system

2.2 3D Data Augmentation

Many 3D face recognition works focused on developing different 3D data augmentation methods to tackle 3D face data scarcity. Kim et al. [24] use the BFM [33] model to synthesize several different facial expressions from a single 3D face scan. The BFM model fits each point cloud from the FRGCv2 dataset [34] to produce 25 expressions for each face model by modifying the expression parameters. Also, they generated eight random patches for each 2D depth map to simulate occlusions. Gilani et al. [53] simultaneously interpolate the facial identity and facial expression spaces over 3D scans to generate millions of 3D facial models with different virtual identities. Zhang et al. [51] applied GPMM to generate a large 3D face training dataset. Moreover, it further constrains the face sampling area to compensate for the inconsistent distribution between generated data and real faces.

The 3D face data augmentation methods mentioned above either sample from a low-dimensional identity and expression parametric space for a synthesized 3D face model, such as GPMM, or interpolates new 3D face models from actual 3D face scans. However, it is still unclear how effectively the data synthesis of new virtual identities will benefit the training of face recognition models. This paper proposes a new CNN-based face depth estimation model, DepthNet, to convert an existing 2D face dataset to an RGB-D face dataset from 2D face images. Our DepthNet is developed for this specific image-to-image translation problem by estimating associated depth maps from an RGB face image. Besides, the DepthNet does not need to perform computation on the 3D point clouds method, which is time-consuming and computationally expensive.

2.3 RGB-D Face Recognition

Besides the lack of large 3D face datasets available to the public, as mentioned above, it is also imperative to have an effective way to pass the 3D data to the neural network, which is still under research. Therefore, many works [24] [25] [38] [44] pre-train the CNN model on 2D face datasets first, and then fine-tune their model on the relatively small 3D dataset. Gilani et al. [53] proposed the first deep CNN

model specifically designed for RGB-D face recognition that takes depth, azimuth, and elevation angles of the normal vector as a 3-channel input. Jiang et al. [23] normalized the depth values to the same range as the RGB values and proposed an attribute-aware loss function for CNN-based face recognition to improve the accuracy of recognition results. Li et al. [28] presented a fusion CNN, which took six types of 2D facial attribute maps (i.e., geometry map, three normal maps, curvature map, and texture map) as input for RGB-D facial expression recognition. Instead of using depth data as input, Zhang et al. [51] proposed a data-free 3D face recognition method that only used synthesized unreal data from 3D Morphable Model to train a deep point cloud network.

Recently, several researchers have focused on multi-modal image fusion for RGB-D face recognition. Chowdhury et al. [8] used an autoencoder architecture to extract a facial representation from RGB and depth modalities. Moreover, their recent work [15] further includes the Stacked Mapping Model and the Joint Hierarchical Feature Learning to learn a shared multi-modal representation of RGB and depth data. Although [8] and [15] included an autoencoder network to reconstruct the depth images, the generated depth images are treated as auxiliary information and could not be applied to depth augmentation. In addition, they require an extra classifier to perform the RGB-D recognition task. Note that these methods were trained on RGB-D face datasets that contain small numbers of subjects compared to the popular 2D face datasets which usually contain large numbers of subjects. In contrast, our DepthNet is used to convert an existing large 2D face dataset to a large RGB-D face dataset for training an accurate RGB-D face recognition model, since DepthNet could provide more accurate depth estimation from a face image than the existing depth estimation works by exploiting semantic segmentation information. Furthermore, the proposed multi-modality RGB-D face recognition network has much better generalization ability by including the identity-style disentanglement, which can be directly tested on all other 3D face datasets without fine-tuning.

2.4 Disentangled representation learning

The main idea about feature disentanglement is factorizing data into interpretable components of variations. With disentangled feature learning, the model could separate only the task-discriminative information in the latent space. Many previous works applied disentangled learning techniques in face images. [49] introduced a network that partitions latent representation of facial images into liveness-related information and liveness-irrelated information for face anti-spoofing task. The disentangled representation learning module (DR-Net) proposed in [41] extracted a domain-related PAD feature and a subject-related feature to improve the robustness of cross-domain face presentation attack detection. [21] disentangled learned latent face representations into identity and age for age components to minimize the effect of age variations for face recognition. [50] presented an expression embedding framework that separates facial identity attribute from the learned representation to tackle the expression embedding problem.

Fig. 2. Samples of the face images and its corresponding semantic segmentation mask images.

3 Proposed Method

We aim to build a robust RGB-D face recognition model by generating a large RGB-D face dataset from a large 2D face image dataset. We propose a new CNN model, called DepthNet, for generating the associated depth map and segmentation mask from an input face image to achieve this goal. Thus, the DepthNet can be applied to generate a large RGB-D face dataset from an existing 2D face dataset for improving the training of RGB-D face recognition models. Our system consists of two modules: (1) the DepthNet and (2) the multi-modality RGB-D face recognition model. In Fig. 1, it is clear to understand the whole process of our method. For each 2D image, we apply FAN face alignment [3] as the first step. Second, the augmented depth image and semantic segmentation mask image are generated by the DepthNet. Third, we set the background pixels of the RGB image as zero according to the semantic segmentation mask image. Finally, the face representation is computed by a multi-modality RGB-D face recognition model for RGB-D face recognition. Our RGB-D face recognition model can also take the acquired depth map as the input by simply replacing the augmented depth image in Fig. 1 with the actual depth image.

3.1 DepthNet

The proposed DepthNet model aims to provide more accurate depth map estimation from a face image than the existing face depth estimation networks by exploiting the semantic segmentation information. The semantic segmentation mask partitions a face image into seven different parts: background, skin, brows, eyes, glasses, nose, and mouth, as shown in Fig. 2 to assist depth estimation task. The proposed DepthNet includes a generator and three discriminators as illustrated in Fig. 3. With the semantic segmentation mask images, the DepthNet will generate depth images that are similar to ground truth depth images. The generator consists of three networks: the face encoder, the face decoder, and the auxiliary decoder. This generator is an UNet encoder-decoder architecture with a skip-connection module as proposed in [36]. The decoder directly uses the features from the encoder with the skip-connection module. For a given source face image X_{input}, which passes through the face encoder and the auxiliary decoder to encode image x_{input} information. The face decoder generates the reconstructed image $\hat{y}_{rgb}$ from the hidden latent representation.

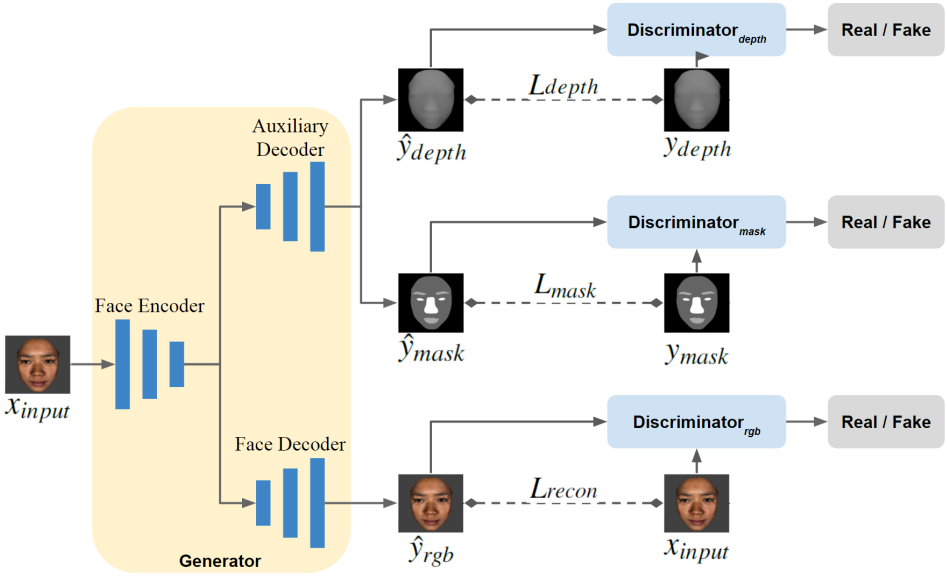

Fig. 3. **Architecture of the proposed DepthNet model.** Each input image ($\hat{X}_{input}$) is first fed into the encoder to encode the face feature vector. Then, the two branches of the decoder generate the semantic segmentation mask ($\hat{y}_{mask}$), the depth map ($\hat{y}_{depth}$), and the reconstructed image ($\hat{y}_{rgb}$) from the embedded features.

We adopt the L1 loss to minimize the difference between the reconstructed image $\hat{y}_{rgb}$ and the source face image X_{input} as follows:

$$L_{recon} = E_{x\sim P_x}[\|x_{input} - \hat{y}_{rgb}\|_1] \tag{1}$$

Meanwhile, the auxiliary decoder generates the corresponding depth map $\hat{y}_{depth}$ and semantic segmentation mask $\hat{y}_{mask}$ of the input face image X_{input}. We design a shared weight architecture to simultaneously output the segmentation mask and depth. The L1 loss is applied to minimize the distance between the generated depth image $\hat{y}_{depth}$ and ground truth depth y_{depth} as follows:

$$L_{depth} = E_{x\sim P_x}[\|y_{depth} - \hat{y}_{depth}\|_1] \tag{2}$$

As for the reconstructed semantic segmentation mask, we adopt the binary cross-entropy loss to enforce the output from the auxiliary decoder to be similar to the ground-truth semantic segmentation. It is given by

$$\begin{aligned} L_{mask} =& E_{x\sim P_x} - (y_{mask} \cdot log(\hat{y}_{mask}) \\ &+ (1 - y_{mask}) \cdot log(1 - \hat{y}_{mask})) \end{aligned} \tag{3}$$

where $\hat{y}_{mask}$ denotes the generated segmentation mask for the input face image, and y_{mask} is the ground truth segmentation mask. We also leverage the adversarial loss to train the RGB discriminator D_{rgb}, depth discriminator D_d, and mask discriminator D_m, which constrains the decoders to generate more accurate results from the learned hidden latent representation. The loss function is given by

$$\begin{aligned} L_{adv}^{Gen} =& E_{y\sim P_y}[(D_{rgb}(\hat{y}_{rgb}) - 1)^2] \\ &+ E_{y\sim P_y}[(D_d(\hat{y}_{depth}) - 1)^2] \\ &+ E_{y\sim P_y}[(D_m(\hat{y}_{mask}) - 1)^2] \end{aligned} \tag{4}$$

$$\begin{aligned} L_{adv}^{Dis} =& E_{y\sim P_y}[(D_{rgb}(x_{input}) - 1)^2] \\ &+ E_{x\sim P_x}[(D_{rgb}(\hat{y}_{rgb}))^2] \\ &+ E_{y\sim P_y}[(D_d(y_{depth}) - 1)^2] \\ &+ E_{x\sim P_x}[(D_d(\hat{y}_{depth}))^2] \\ &+ E_{y\sim P_y}[(D_m(y_{mask}) - 1)^2] \\ &+ E_{x\sim P_x}[(D_m(\hat{y}_{mask}))^2] \end{aligned} \tag{5}$$

The overall loss function for training the DepthNet is given as follows:

$$L_{Total} = L_G + L_{adv}^{Dis} \tag{6}$$

$$L_G = \lambda_1 L_{depth} + \lambda_2 L_{mask} + \lambda_3 L_{adv}^{Gen} + L_{recon} \tag{7}$$

where λ_1, λ_2 and λ_3 are the weights used to balance the three loss terms.

3.2 Multi-Modal RGB-D Face Recognition

The proposed multi-modality RGB-D face recognition network aims to learn a better representation for the face recognition task. Since we can obtain a paired RGB image and depth image from our DepthNet, we can observe that the RGB image and depth image should share the same semantic meaning for representing the same identity, and the major difference between RGB image and depth image is the style (RGB or depth) of them. Therefore, we should be able to obtain the same identity information from both RGB and depth images. Thus, we employ disentangled feature learning to separate latent representation into identity-related and style-related components. Fig. 4 depicts the overall architecture of our multi-modality RGB-D face recognition network, which is composed of the segmentation-guided RGB-D face recognition network and the representation disentanglement module.

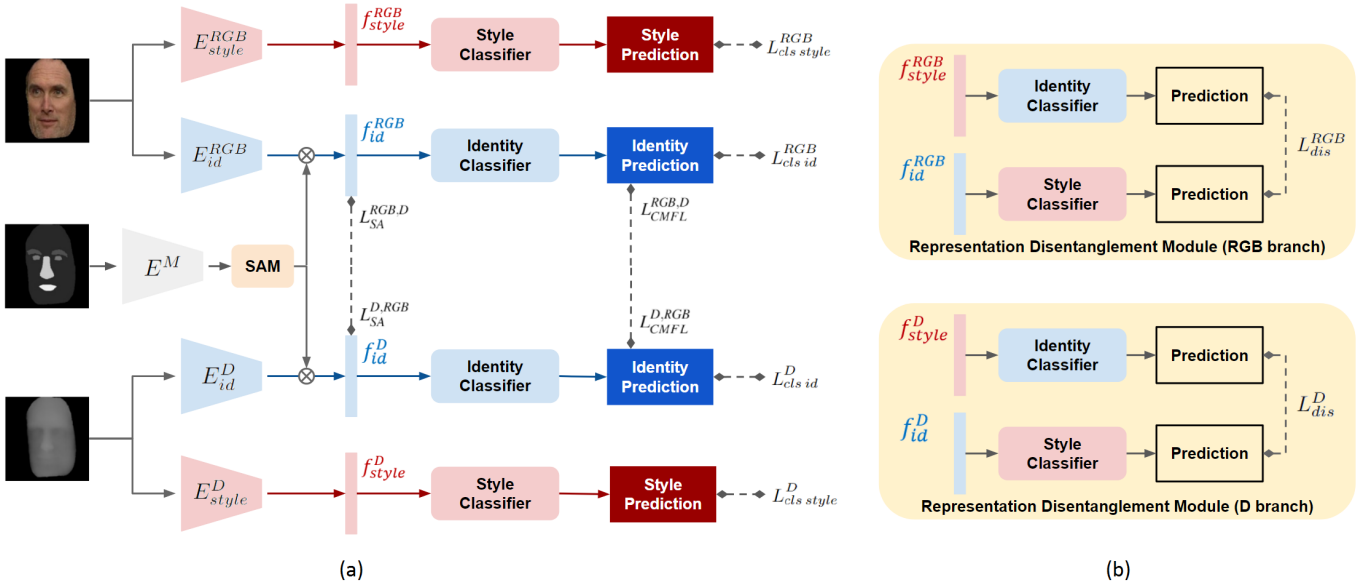

Fig. 4. **The proposed multi-modality RGB-D face recognition network architecture.** (a) The segmentation-guided RGB-D face recognition network architecture. (b) The representation disentanglement module.

3.2.1 Segmentation-Guided RGB-D Face Recognition

Our segmentation-guided RGB-D face recognition network architecture contains an RGB recognition branch, a depth map recognition branch, and an auxiliary segmentation mask branch with spatial attention module proposed in [42], as shown in Fig. 4(a). Inside each recognition branch, two encoders extract a pair of disentangled features: an identity-related component and a style-related component. As in our previous work [7], we choose the Squeeze-and-Excitation Networks SENet [22] as the backbone encoder network. At the training stage, the RGB recognition branch extracts identity representation f_{id}^{RGB} and the style representation f_{style}^{RGB} by the identity encoder E_{id}^{RGB} and the style encoder E_{style}^{RGB} accordingly. Similarly, the depth map recognition branch extracts corresponding identity feature f_{id}^{D} and the style feature f_{style}^{D} by the identity encoder E_{id}^{D} and the style encoder E_{style}^{D}. The auxiliary segmentation mask branch extracts different levels of feature map from the segmentation mask by E^{M}. It then applies spatial attention module (SAM) on those feature maps to aid RGB and D branches while training. This SAM is share-weighted across the RGB recognition branch and D recognition branch. It can provide auxiliary information from the segmentation branch to help recognition branches focus on the informative parts on segmentation feature maps. Finally, the style classifier predicts whether the input is an RGB image or a depth image; and the identity classifier with ArcFace [10] additive angular margin loss predicts a vector of probabilities with one value for each possible identity.

The proposed segmentation-guided RGB-D face recognition network is a two-stream-multi-head architecture, and we apply the cross-entropy loss as identity classification loss $L_{cls\,id}$ and style classification loss $L_{cls\,style}$ on individual branches. We also adopt the cross-modal focal loss $L_{CMFL}^{m,n}$ in [14] to learn robust identity representations jointly, which is defined as

$$L_{CMFL}^{m,n} = -\alpha(1 - w(m_t, n_t))^{\gamma} \log(m_t), \tag{8}$$

where α and γ are tunable hyper-parameters, and the function w is defined by

$$w(m_t, n_t) = n_t \frac{2m_t n_t}{m_t + n_t}, \tag{9}$$

where m_t and n_t denote the identity classification probabilities after fully connected layer in current branch m and the other branch n, respectively. The CMFL contributed by branch n will decrease when branch n can predict with high confidence. Please note that we only apply this cross-modal focal loss on the extracted identity features, because we pay more attention on learning an identity-efficient representation to improve RGB-D face recognition performance.

Although the inputs to the segmentation-guided RGB-D face recognition network could be a different combination of modalities, their inputs should represent the same semantic meaning as the subject identity. Inspired by [2], we add another semantic alignment loss $L_{SA}^{m,n}$ to share semantics for the extracted identity feature vectors f_m and f_n, given by

$$L_{SA}^{m,n} = \rho^{m,n}(1 - \text{cosine\_similarity}(f_m, f_n)) \tag{10}$$

where $\rho^{m,n}$ is the focal regularization parameter to make sure the network will only transfer information from the more accurate network to the weaker network. For current modality m and the other modality n, we use the difference of identity classification losses between m and n to measure the performance of the network, and it is denoted as $L_{cls\,id}^{m} - L_{cls\,id}^{n}$. If the difference is positive, then it means modality m is weaker than modality n. The model will enforce f_m to be

similar to f_n. The focal regularization parameter is defined as follows

$$\rho^{m,n} = \begin{cases} e^{\beta(L^m_{cls\,id} - L^n_{cls\,id})} - 1, & \text{if } L^m_{cls\,id} > L^n_{cls} \\ 0, & \text{if } L^m_{cls\,id} \leq L^n_{cls\,id} \end{cases} \tag{11}$$

where β is a positive focusing parameter. As the same idea to the cross-modal focal loss, we only apply this semantic alignment loss on the extracted identity features, to learn an identity-efficient representation for RGB-D face recognition task.

3.2.2 Representation Disentanglement Module

Motivated by [17], we employ a representation disentanglement loss to minimize the correlation between the disentangled features. As illustrated in Fig. 4(b), now we have two kinds of features: identity-related feature and style-related feature in each modality; and two corresponding classifiers: identity classifier and style classifier. We pass the style-related feature into the identity classifier and pass the identity-related feature into the style classifier to mitigate the discrepancies between the two disentangled features. Taking the RGB branch in Fig. 4(b) as an example, we aim to train the identity classifier that outputs a correct identity prediction with the extracted identity feature as input, but it cannot correctly classify the identity with the style feature f^{RGB}_{style}. Because the style feature should not include any identity-related information, the identity classifier gives the prediction with a maximum chaos distribution that is uniform since it has the highest entropy. Similarly, when passing the identity feature f^{RGB}_{id} into the style classifier will give a prediction that its probability distribution is very similar to a uniform distribution. For the same reason, we input the style feature f^{D}_{style} into the identity classifier and input the identity feature f^{D}_{id} into the style classifier to obtain the disentanglement loss in the D branch.

Thus, the identity feature only contains the subject-related information while the style feature contains only the style-related information. The feature disentanglement loss aims to reduce the interference among the extracted features, which is defined as

$$\begin{aligned} L^{RGB}_{dis} = &\frac{1}{N}\sum_{n=1}^{N} \log \frac{e^{C_{id}(f^{RGB}_{style})_n}}{\sum_{j=1}^{N} e^{C_{id}(f^{RGB}_{style})_j}} \\ &+ \frac{1}{M}\sum_{m=1}^{M} \log \frac{e^{C_{style}(f^{RGB}_{id})_n}}{\sum_{j=1}^{M} e^{C_{style}(f^{RGB}_{id})_j}} \end{aligned} \tag{12}$$

$$\begin{aligned} L^{D}_{dis} = &\frac{1}{N}\sum_{n=1}^{N} \log \frac{e^{C_{id}(f^{D}_{style})_n}}{\sum_{j=1}^{N} e^{C_{id}(f^{D}_{style})_j}} \\ &+ \frac{1}{M}\sum_{m=1}^{M} \log \frac{e^{C_{style}(f^{D}_{id})_n}}{\sum_{j=1}^{M} e^{C_{style}(f^{D}_{id})_j}} \end{aligned} \tag{13}$$

where N is the total number of classes for the face recognition and M is the total number of classes for style classification. C_{id} denotes the share-weight identity classifier, and C_{style} denotes the share-weight style classifier. With the proposed disentanglement, we can train a pair of representations that are independent of each other. The overall loss functions in branch RGB and D are given as

$$\begin{aligned} L^{RGB}_{total} = &\lambda_0 L^{RGB}_{cls} + \lambda_1 L^{RGB,D}_{CMFL} \\ &+ \lambda_2 L^{RGB,D}_{SA} + \lambda_3 L^{RGB}_{dis} \end{aligned} \tag{14}$$

$$\begin{aligned} L^{D}_{total} = &\lambda_0 L^{D}_{cls} + \lambda_1 L^{D,RGB}_{CMFL} \\ &+ \lambda_2 L^{D,RGB}_{SA} + \lambda_3 L^{D}_{dis} \end{aligned} \tag{15}$$

The total loss of RGB branch L^{RGB}_{total} optimizes the parameters of the RGB recognition branch and auxiliary segmentation branch. Similarly, the total loss of D branch L^{D}_{total} optimizes the parameters of the D recognition branch and auxiliary segmentation branch.

4 Experimental Results

4.1 Experimental Datasets

Here we introduce the datasets that were used in the training and evaluation. The depth estimation model, DepthNet, is trained with BU-3DFE 3D database [29] dataset. We evaluate the depth estimation performance on BU-3DFE [29], Bosphorus [37], and FRGCv2 [34] datasets. The 2D VGGFace2 [5] face dataset is augmented to the corresponding RGB-D face dataset by applying the DepthNet model and is then used to train the segmentation-guided RGB-D face recognition model. While evaluating the RGB-D face recognition model, we experiment on several public 3D face datasets, including BU-3DFE [29], Texas FR3D [19], Bosphorus [37], FRGCv2 [34], and Lock3DFace [48] datasets. In order to mitigate the depth domain shift caused by different depth acquisition devices, we further fine-tune the DepthNet with other public 3D face databases when conducting the face recognition experiments. We split each 3D face dataset into training and testing sets. The training set is used for fine-tuning the DepthNet, and the testing set is for evaluation. The strategy of training-testing partition will be explained in detail in the following subsection.

4.1.1 3D Face Datasets

BU-3DFE 3D database [29] includes 100 subjects with 2,500 scans. There is no pose variation in this database, and it contains only facial expression variations. Each identity performs seven expressions with four intensity levels for each expression except for the neutral one. The BU-3DFE dataset is used for the depth estimation model DepthNet and the RGB-D face recognition model. For the DepthNet, we leave ten subjects out for depth estimation evaluation, and the remaining as the training set as in [29]. Therefore, there will be 250 face images from ten individuals in the testing set. While reporting the rank-1 identification accuracy of the RGB-D face recognition model, we follow the evaluation protocol as [16]. We select the first neutral scan of each identity as the gallery images so that we have 100 images in the gallery and 2400 images in the probe.

Texas FR3D database [19] contains 1,149 scans of 118 subjects. All the scans in Texas FR3D are frontal with different expressions. As suggested in the paper, we split the dataset into a training set with 360 face images of twelve subjects and a testing set including the remaining

face images. We further selected the first images for all the 118 subjects as the gallery from this testing set and put the remaining 789 images as the probe.

Bosphorus database [37] consists of 4,666 scans of 105 subjects in various poses, expressions and occlusions. There are two settings for evaluation. Both settings select the first neutral image of each subject to form a gallery set with 105 images. As for the probe images, *Setting-1* adopts the other 2,797 images that convey only expression variations, and *Setting-2* takes all the remaining 4,561 images for identification. Following the baseline result in [37], we only include the 105 images in the gallery set as the training set for fine-tuning the DepthNet and the others as the testing set.

FRGCv2 database [34] contains images from 466 subjects collected in 4,007 scans with two facial expression variances (e.g., neutral and smile). We select the first neutral images from all the 466 subjects as the gallery and take the remaining 3,541 images as the probe. Like the Bosphorus database, we set the 466 images in the gallery as a training set.

BUAA Lock3DFace database [48] contains 5711 RGB-D face videos of 509 subjects with variations in facial expression, pose, occlusion and time-lapse. [48] provides a standard evaluation protocol with the given four variations. The first neutral image of each subject in Session-1 (S-1) is selected as the gallery, and then divide the remaining into four test sets. Probe_Set_1 for images with expression changes in S-1; Probe_Set_2 for images with pose variations in S-1; Probe_Set_3 for images with occlusions in S-1; and Probe_Set_4 for all images in Session-2 (S-2). In their extended work [32], the authors provided a partition method in which all the neutral face scans in Session-1 (S-1) are selected to be the training set.

4.1.2 Training Data Preparation

Since we aim to make DepthNet learn how to convert an RGB face image to the corresponding depth image, we adopt 90 identities from BU-3DFE 3D face database [29] as the training data. Also, the pose augmentation is implemented by rotating the original frontal face point cloud along with the yaw and pitch axis. Finally, each person's original 25 3D models are augmented to 2,275 models and then projected to 2D depth images. Next, We modify the BiSeNet [6] model to generate the semantic segmentation mask for an input face image and take the results as the pseudo ground truth segmentation mask. The segmentation mask consists of seven channels representing different labels: background, skin, brows, eyes, glasses, nose, and mouth. Then, we use these RGB-D images and the corresponding pseudo-ground-truth segmentation masks to train our DepthNet.

For RGB-D face recognition, we select VGGFace2 [5] as our RGB face training dataset. It contains 9,131 subjects and a total of 3.31 million RGB images. As mentioned in Section 4.1, we fine-tune the proposed DepthNet with each 3D face database while evaluating the RGB-D face recognition results, which indicates that there will be a total of five fine-tuned DepthNet models that produce the corresponding depth images and segmentation masks for all 3D face databases (BU-3DFE, Texas FR3D, Bosphorus, FRGCv2, and Lock3DFace). Then we train the segmentation-guided RGB-D face recognition network with the fine-tuned DepthNet, respectively. The augmented depth images will be gray images with a channel equal to one and a seven-channel image representing the segmentation mask. We can generate an even larger RGB-D dataset for model training by using a larger RGB dataset. However, due to the consideration of memory and training time, we use the VGGFace2 dataset for conversion into the RGB-D dataset to train our face recognition model.

4.2 Implementation Details

For training DepthNet, we adopt Adam as the optimizer with setting $\beta_1 = 0.5$, $\beta_2 = 0.999$ and learning rate γ is set to 0.0002. For the hyper-parameter of the loss function in (7), we set $\lambda_1 = 100$, $\lambda_2 = 100$ and $\lambda_3 = 1$. We train the DepthNet model on a GTX1080Ti GPU card with batch size equal to 16 and image size 256x256.

When training the RGB-D face recognition model, we use SGD as the optimizer with momentum = 0.9, weight decay = 0.0005 and learning rate = 0.1 divided by 10 at 6, 10, 17 epochs. The SAM is applied on feature map with size of 56x56x64 and 14x14x256. We set $\alpha = 1$ and $\gamma = 3$ in (8) and set $\beta = 2$ in (11). In our experiments, we set N in (12) and (13) equal to 9,131, the total subjects of the VGGFace2 dataset. And M in (12) and (13) is set to 2 since we only have two kinds of style (RGB and depth). For the hyper-parameter of the loss function in (14) and (15), we set $\lambda_0 = 0.5$, $\lambda_1 = 0.5$, $\lambda_2 = 0.05$ and $\lambda_3 = 0.5$. We train RGB-D face recognition on 2 Tesla V100 GPUs with batch size 256 and image size 112x112. At the testing stage, instead of simply concatenating f_{id}^{RGB} and f_{id}^{D}, we compute two cosine similarity scores and perform score-level fusion by averaging two cosine similarity scores to give the final prediction. Our experimental result shows that combining all modalities provides the most accurate result for face recognition.

4.3 DepthNet Evaluation

This section demonstrates some results of our proposed depth estimation method. Our proposed DepthNet aims to produce additional augmented depth images and segmentation mask images for 2D datasets. As a result, in Fig. 5, we depict some examples of applying our fine-tuned DepthNet (with BU3DFE) to the VGGFace2 2D face database, which is the training set for our RGB-D face recognition model. The

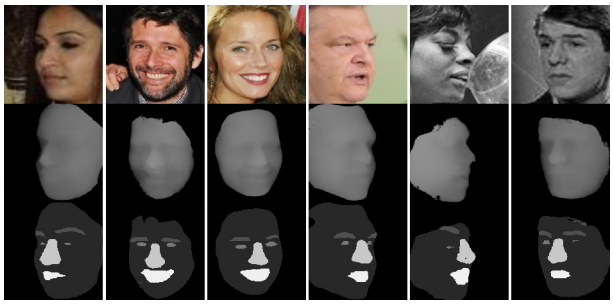

Fig. 5. The generated depth and segmentation images of VGGFace2 of DepthNet that fine-tuned with BU3DFE database. Rows from top to bottom: RGB images, augmented depth images, and segmentation images.

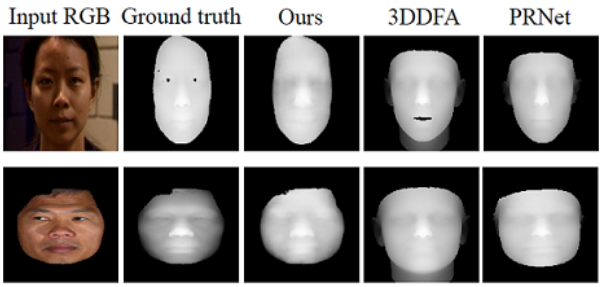

Fig. 6. Depth estimation results by using different methods on some sample images in FRGCv2 dataset (top row) and Texas dataset (bottom row).

TABLE 1
Quantitative comparison of depth estimation errors for different methods: MSE between ground truth depth and estimated depth images.

Method	BU3DFE	FRGCv2	Bosphorus
3DDFA [52]	125.78	597.17	540.48
PRNet [12]	74.86	**216.29**	615.99
Ours	**16.65**	435.88	**421.46**

results show that our method can produce well-preserved face contour and face shape features of different expressions.

In Table 1, we compute Mean Square Error (MSE) between the generated depth images and ground truth depth images with comparison with two 3D face depth estimation methods, 3DDFA [43] and PRNet [45]. Although we employ fine-tuning on the DepthNet while conducting identification experiments, we do not do the same procedure when evaluating the depth estimation results. Because the 3DDFA and PRNet do not fine-tune their models on the testing database, we do not report the fine-tuning result for having a fair comparison. Therefore, we only calculate the MSE in the intersection of ground truth depth and all predicted depth images from our DepthNet, 3DDFA, and PRNet. All the predicted depth images are normalized to the same pixel scale varying in [0, 255].

For the BU3DFE database, it is evident that we have the best performance on the testing set of BU3DFE partially because we train DepthNet with the training set of BU3DFE, which leads to negligible bias between training and testing data. For the FRGCv2 dataset, although the estimation results by our model are not the best among the three methods, our DepthNet model can generate a more accurate depth image around the face contour than the other two methods, as shown in Fig. 6. This is because our model includes semantic segmentation together into the depth estimation model. Our DepthNet is trained with BU3DFE, acquired with a structured-light-based 3D sensor. The Bosphorus 3D images were also acquired using a structured-light-based device. However, the FRGCv2 3D images were captured by a laser-based sensor. As a result, the improvement is not as significant as the others. Our DepthNet achieves the best performance on the Bosphorus dataset, which contains large pose variations, and our DepthNet was trained with such variations. In Fig. 7, we further illustrate how our DepthNet provides superior depth estimation for face images with large poses.

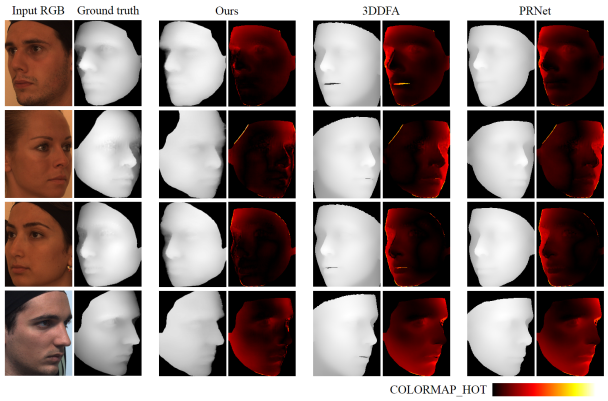

Fig. 7. Depth estimation results by using different methods on the Bosphorus dataset. We utilize hot colormap to illustrate the MSE. The darker the color, the smaller the error.

TABLE 2
The rank-1 identification accuracy on public 3D face databases.

Method	BU3DFE	Texas	Bosphorus-1	Bosphorus-2
Li *et al.* [27]	-	-	98.8	96.6
Lei *et al.* [26]	93.25	-	98.9	-
Mian *et al.* [31]	95.9	98.0	-	96.4
Lin *et al.* [30]	96.2	-	99.71	-
Kim *et al.* [24]	95.0	-	99.2	-
FR3DNet [53]	98.64	100	-	96.18
Ours w/ D*	**100**	**100**	**100**	**98.18**
Ours w/ Dgt	100	100	99.97	97.59

The other two methods have large deviations near the face profile regions. With an additional semantic segmentation branch, our DepthNet can recognize the facial regions from the image to generate an accurate and plausible depth map consistent with the RGB face image.

4.4 Segmentation-Guided RGB-D Face Recognition Evaluation

Our segmentation-guided RGB-D face recognition model has excellent generalization ability on other 3D datasets, even though it is trained with a depth-augmented 2D dataset. We only fine-tune the DepthNet to ensure the augmented depth has a similar distribution to the ground truth depth. Our network is trained on VGGFace2 [5] and directly tested on all other 3D face datasets without any fine-tuning. Our segmentation-guided RGB-D face recognition model takes RGB face image, augmented depth image generated by DepthNet (D*), and augmented segmentation mask generated by DepthNet (M*) as input. We demonstrate the rank-1 identification results on some public 3D face databases in Table 2. We can pass either the augmented depth image D* generating from DepthNet or the ground truth depth image Dgt that came with the database itself to our network by simply replacing the input depth image. The result that comes from augmented depth is denoted as "Ours w/ D*," while the result that comes from ground truth depth is denoted as "Ours w/ Dgt." For all the datasets, our method consistently provides state-of-the-art RGB-D face recognition accuracy. Especially for Bosphorus-2, which has large pose variations, the proposed method outperforms the other methods by around 2% accuracy.

Table 3 further shows that our model can also be applied to different modalities. We compare our results with other RGB-D face recognition methods and report the rank-1 identification accuracy on FRGCv2 and Bosphorus-1. Our segmentation-guided recognition model trains the RGB and D branches jointly; the training data that includes the augmented depth or segmentation mask images of VGGFace2 are denoted as VGGFace2*. Our DepthNet can effectively transform a 2D face image into the corresponding RGB-D image to resolve the problem that the existing public 3D face database usually has inadequate subjects or intra-person variations. Jiang *et al.* [23] proposed an attribute-aware loss function and a newly collected RGB-D face database with 60K subjects to improve the accuracy of RGB-D face recognition results. We can observe that our proposed method, trained with the RGB-D dataset with augmented depth, segmentation masks, and 9K subjects, is superior to the model trained with ground truth depth images and many more subjects. With the proposed method, we can obtain an RGB-D database with sufficient subjects from an existing 2D face database and do not need to collect a new 3D face database.

The rank-1 identification accuracy for the Lock3DFace dataset is shown in Table 4; we report both results obtained from the augmented depth and ground truth depth. Especially in the subset with pose variations, our result achieves a 95.46% accuracy which is significantly better (around +25%) than others. For occlusion variations such as covering the face with hand or glasses, we reach an accuracy of 95.22% obtaining about +10% performance gain. In the subset over time scenario, our method also provides an accuracy of 92.23%, which exceeds others by +11%. In general, our segmentation-guided RGB-D recognition model achieves a much higher (+8%) average accuracy of 95.73% compared to other state-of-the-art methods. This indicates that our segmentation-guided FR model fully exploits the augmented depth and segmentation mask information and is more robust against pose variation than other RGB-D face recognition methods. It is worth noting that some other methods include part of the Lock3DFace in their training set. However, our RGB-D face recognition model was directly tested on Lock3DFace without any fine-tuning. We can say that the depth images generated by the fine-tuned DepthNet help the recognition model obtain better results.

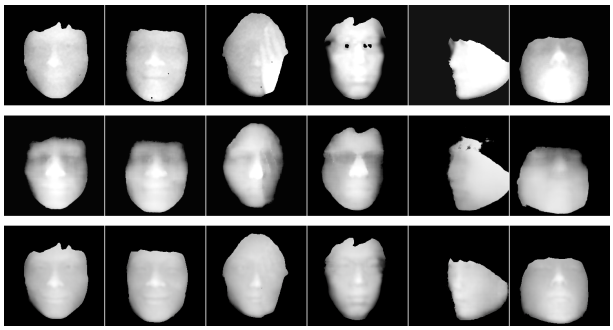

Fig. 8. Depth estimation results of DepthNeton Lock3DFace dataset. First row: ground truth depth images; Second row: generated depth images with out fine-tuning; and Last row: generated depth images with fine-tuning.

Our segmentation-guided RGB-D face recognition network is trained with the augmented 2D face database VGGFace2*; a domain gap exists between the generated depth map and the ground truth map. As a result, We fine-tune the DepthNet with other public 3D face databases when conducting the face recognition experiments to minimize such domain shift caused by different depth acquisition devices among training data and testing data. We report the rank-1 identification accuracy with ground truth depth input for the Lock3DFace dataset in Table 5 and its visualization in Fig. 8 to analyze how the fine-tuned DepthNet has affected the recognition results. The method "Ours_ft" means to fine-tune the DepthNet with the Lock3DFace database; thus, the fine-tuned DepthNet could predict depth images that have a similar depth distribution with the ground truth images. Finally, we use the RGB face image and the images generated from fine-tuned DpthNet for RGB-D face recognition training. It is evident that fine-tuning DepthNet with the target dataset could assist the segmentation-guided RGB-D face recognition model reduce the domain shift between the generated depth map and the ground truth depth map. For the pose variations subset, the result with fine-tuning achieves a 93.89% accuracy which is better (around +7%) than the one without fine-tuning. In the subset with occlusion, we reach an accuracy of 93.13% obtaining about +21% performance gain. In the subset with time scenario, the fine-tuning provides an accuracy of 91.86%, which improves

TABLE 3
Rank-1 identification accuracy applied to different modalities. VGGFace2* denotes the augmented data of VGGFace2 produced by DepthNet. D* and M* denotes the augmented depth map and segmentation mask generated by DepthNet, and Dgt denotes the ground truth depth map.

Method	Training data	Subjects	Testing Modality	FRGCv2	Bosphorus-1
VGG-Face [5]	Private [53]	100	RGB	87.92	96.39
Jiang *et al.* [23]	TRAINING-SET-I [23]	60,000	RGB	95.69	96.08
Ours	VGGFace2* [5]	9,131	RGB + M*	**99.66**	**99.86**
Li *et al.* [27]	Part of Bosphorus [37]	105	Depth	96.30	95.40
FR3DNet [53]	Private [53]	100	Depth	97.06	96.18
Jiang *et al.* [23]	TRAINING-SET-I [23]	60,000	Depth	97.45	**99.37**
Ours w/ D*	VGGFace2* [5]	9,131	D* + M*	**97.51**	99.21
Ours w/ Dgt	VGGFace2* [5]	9,131	Dgt + M*	92.32	96.00
Li *et al.* [27]	Part of FRGCv2 [34]	466	RGB + Depth	95.20	99.40
Jiang *et al.* [23]	TRAINING-SET-I [23]	60,000	RGB + Depth	98.52	99.52
Ours w/ D*	VGGFace2* [5]	9,131	RGB + D* + M*	**99.80**	**100**
Ours w/ Dgt	VGGFace2* [5]	9,131	RGB + Dgt + M*	99.69	99.97

TABLE 4
The rank-1 identification accuracy on Lock3DFace databases. D* and M* denotes the augmented depth map and segmentation mask generated by DepthNet, and Dgt denotes the ground truth depth map.

Method	Input	Accuracy				
		Expression	Pose	Occlusion	Time	Average
He *et al.* [20]	RGB	96.3	58.4	74.7	75.5	76.2
Hu *et al.* [22]	RGB	98.2	60.7	77.9	78.3	78.7
Cui *et al.* [9]	RGB + D	97.3	54.6	69.6	66.1	71.9
Mu *et al.* [32]	RGB + 3D Model	98.2	70.4	78.1	65.3	84.2
Uppal *et al.* [39]	RGB + D	99.4	70.6	85.8	81.1	87.3
Ours	RGB + D* + M*	**100**	**95.46**	**95.22**	**92.23**	**95.73**
Ours	RGB + Dgt + M*	99.69	93.89	93.13	91.86	94.74

TABLE 5
The rank-1 identification accuracy on Lock3DFace databases. Ours stands for the recognition network is trained without fine-tuning the DepthNet, and Ours_ft stnads for the recognition network is trained with fine-tuning DepthNet. D* and M* denotes the augmented depth map and segmentation mask generated by DepthNet, and Dgt denotes the ground truth depth map.

Method	Input	Accuracy				
		Expression	Pose	Occlusion	Time	Average
Ours	RGB + Dgt + M*	99.61	86.98	71.51	79.22	84.88
Ours_ft	RGB + Dgt + M*	99.69	93.89	93.13	91.86	94.74

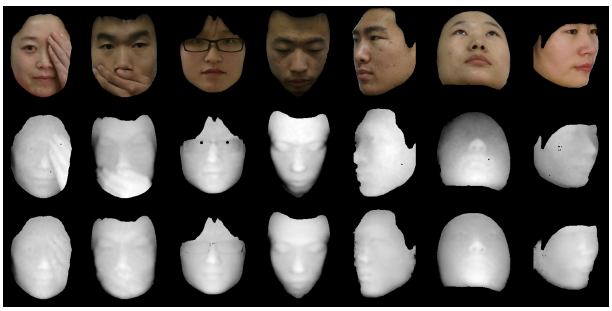

Fig. 9. Samples from the Lock3DFace dataset that are correctly recognized with augmented depth but miss-classified with the ground truth depth. First row: RGB images; Second row: ground truth depth images; and Last row: generated depth images from DepthNet.

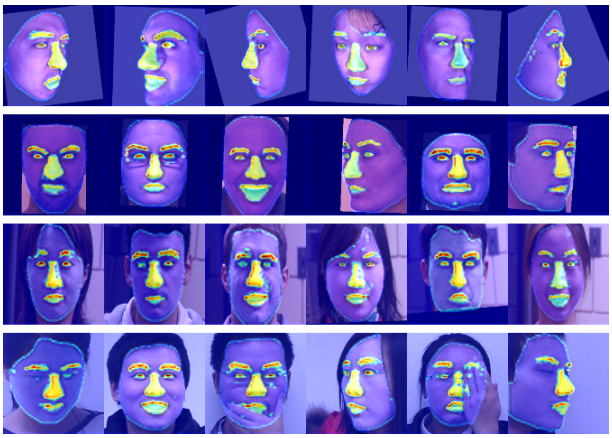

Fig. 10. Examples of spatial attention maps for the four RGB-D face datasets. First row: BU-3DFE dataset; Second row: Bosphorus datase; Third row: FRGCv2 dataset; and Last row: Lock3DFace dataset.

the model accuracy by +12%. In general, the fine-tuning provides much higher (almost +10%) average accuracy of 94.74% compared to the one without fine-tuning.

The proposed RGB-D face recognition network is trained on the augmented depth images. When passing the ground truth depth image into the network, there will be some performance decrease caused by a domain gap between the augmented depth map and the ground truth depth map. From Table 4 and Table 5, we can observe that how accurate the depth estimation plays an important role in the identification results. The fine-tuned DepthNet could predict depth images similar to the ground-truth depth map and achieve better identification results than those without fine-tuned. Although with fine-tuned DepthNet, a domain gap still exists, causing a slight accuracy decrease (up to -2%) if we input the ground truth depth images into the recognition network. Here we show samples from the Lock3DFace dataset to illustrate such observation in Fig. 9. Those samples are correctly recognized with augmented depth images but miss-classified with the ground truth depth images. Although our DepthNet can estimate a well-preserved face contour and face shape features compared to other methods, we still can notice large errors in depth estimation near the border and the occlusion regions (zoom in for details). Our DepthNet is a pixel-to-pixel image transformation network. It tends to predict a smoother result around the border regions and the regions occluded by hands and glasses, which will cause a lower identification accuracy when passing the ground truth depth image into the segmentation-guided RGB-D face recognition model.

Fig. 10 demonstrates visualization results of the segmentation-guided spatial attention module on some public 3D datasets. The result shows some samples with expression, pose, and occlusion variations. The segmentation mask branch provides auxiliary information to the spatial attention module. Therefore, we can observe that the attention has selectively focused on the informative parts such as eyes, nose, eyebrows, and lips for RGB-D face recognition.

TABLE 6
The comparison of MSEs of depth estimation with and without including the semantic segmentation task.

Method	BU3DFE	FRGCv2	Bosphorus
DepthNet *w/o* mask	84.62	880.99	848.88
DepthNet *w* mask	**42.66**	**605.48**	**839.86**

5 Ablation Study

This section validates that each component in the proposed method is essential and plays a vital role in achieving high identification accuracy.

5.1 Effect of the Segmentation Mask on DepthNet

We first analyze the effects of the segmentation mask branch in the proposed DepthNet. Table 6 demonstrates significant improvement of the depth estimation by including the semantic segmentation into the model. Unlike section 4.3, we directly calculate the MSE between the estimated depth image and the ground truth image. We can observe that the generated depth images have similar values without adding the segmentation mask branch. It is challenging to locate the facial features from the depth images. However, with the addition of the semantic segmentation branch, it can focus on the face features and provide precise depth estimation. We can easily perceive both profile and frontal images' expression with the segmentation mask.

5.2 Effect of the Training Modalities on RGB-D Face Recognition Model

We report the rank-1 face identification results of three baseline models for face recognition as a reference for the ablation study. The RGB recognition baseline network with SENet backbone and ArcFace takes only RGB images as training data, which only contains the RGB recognition part of our segmentation-guided recognition network and uses cross-entropy loss as the objective function, denoted as Baseline RGB. Similar to the RGB baseline network, the D recognition baseline model is the recognition network with SENet backbone and ArcFace that takes only generated depth images as training data, which is the depth map recognition part of our segmentation-guided recognition network with cross-entropy loss function, denoted as Baseline D. The other is the RGB-D recognition baseline network, which takes RGB images and depth images as input. It has the same architecture as the proposed segmentation-guided recognition network without auxiliary segmentation mask branch and only apply cross-entropy loss as the objective function, denoted as Baseline RGB-D. And then, we add the auxiliary segmentation mask branch to those three baseline models, denoted as Baseline + mask branch, to examine the effect of the auxiliary segmentation mask branch with SAM on RGB-D face recognition results.

The experimental comparison of our models with different combinations of input image modalities are presented in Table 7. For the upper part of the table, the first row gives the results of our model with augmented depth images, the second row gives the results with RGB images as input, and the third row gives the results with RGB and the augmented depth images as input. For the lower part of the table, the first row is the result of our model with augmented depth and the augmented segmentation mask images as input, the second row gives the results of our model with RGB and the augmented segmentation mask as input, and the third row shows the results with RGB, the augmented depth, and the augmented segmentation mask images as input to our model. In general, adding the auxiliary segmentation mask branch with SAM helps achieve a better accuracy, which indicates the auxiliary segmentation mask branch could supply additional face information, which is essential to our segmentation-guided RGB-D face recognition method. We also notice that both RGB-D baseline models have worse performance than the RGB baseline model. It is caused by the negative transfer mentioned in Section 3.2, the network could transfer information from the weaker network to a more accurate network and degrade the final prediction results. Therefore, we will have a thorough ablation study about the three designed loss functions in the following section.

5.3 Effect of Training Loss Functions on Segmentation-guided RGB-D Face Recognition Model

We introduce one basic loss function and three additional loss functions for our model training in Section 3.2: the cross-entropy loss L_{cls}, the semantic alignment loss L_{SA}, the cross-modal focal loss L_{CMFL}, and the disentanglement loss L_{dis}. We set the RGB-D baseline network with the auxiliary segmentation mask branch and only apply cross-entropy loss as the objective function in Table 8 as the standard model, denoted as Baseline. Moreover, add the additional losses in our segmentation-guided face recognition model one at a time. In Table 8, we can discover that the semantic alignment loss L_{SA} reaches a 94.25% accuracy which is +8% than baseline with generated depth images and a +17% accuracy gain than baseline with ground truth depth images. The focal regularization parameter of the semantic alignment loss avoids the negative transfer and fully exploits the augmented depth and segmentation mask information. Including the cross-modal focal loss will result in a +1% performance gain than the baseline + L_{SA} with generated depth images and a +2% accuracy gain than the baseline with ground truth depth images. The disentanglement loss ensures the extracted identity feature f_{id} and style feature f_{style} are independent of each other so that the identity feature contains only identity-related information. With the disentanglement loss, our model could learn a better facial feature representation that achieves a higher (+0.4% and +3%) average accuracy of 95.73% and 94.74% compared to the baseline + L_{SA} + L_{CMFL} for the generated depth images and ground truth depth images as input, respectively.

What is worth mentioning is the mask-guided RGB-D face recognition network in our previous work [7] is the Baseline + L_{SA} + L_{CMFL} model in Table 8. The results indicate an accuracy decrease (from 95.38% to 91.56% on average) when using the ground truth depth images as the model input due to the domain gap between the augmented and the ground truth depth map. This paper includes the disentanglement component in the RGB-D face recognition model to extract the identity feature that contains

TABLE 7
Identification accuracy on Lock3DFace dataset of our segmentation-guided RGB-D face recognition model and its variants with different training modalities.

Method	Training Modality	Lock3DFace Expression	Pose	Occlusion	Time	Average
Baseline D	D*	83.68	29.29	41.04	25.00	45.61
Baseline RGB	RGB	99.53	82.94	78.78	68.71	82.50
Baseline RGB-D	RGB+D*	99.38	73.96	64.74	75.74	79.51
Baseline D w/ mask branch	D*+M*	93.86	63.31	54.38	47.63	65.28
Baseline RGB w/ mask branch	RGB+M*	99.61	86.39	82.37	87.35	89.43
Baseline RGB-D w/ mask branch	RGB+D*+M*	98.60	90.24	80.38	77.88	86.84

TABLE 8
Identification accuracy on Lock3DFace dataset of our segmentation-guided RGB-D face recognition model and its variants with different training loss functions.

Method	Input	Lock3DFace Expression	Pose	Occlusion	Time	Average
Baseline	RGB + D* + M*	98.60	90.24	80.38	77.88	86.84
Baseline + L_{SA}	RGB + D* + M*	99.22	93.80	93.63	90.31	94.25
Baseline + L_{SA} + L_{CMFL} [7]	RGB + D* + M*	99.69	94.48	94.42	92.83	95.38
Baseline + L_{SA} + L_{CMFL} + L_{dis}	RGB + D* + M*	100	95.46	95.22	92.23	95.73
Baseline	RGB + Dgt + M*	98.14	70.71	61.25	55.10	71.72
Baseline + L_{SA}	RGB + Dgt + M*	99.46	86.19	82.37	87.94	89.54
Baseline + L_{SA} + L_{CMFL} [7]	RGB + Dgt + M*	99.54	90.14	91.14	85.65	91.56
Baseline + L_{SA} + L_{CMFL} + L_{dis}	RGB + Dgt + M*	99.69	93.89	93.13	91.86	94.74

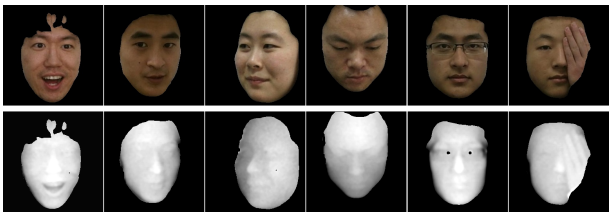

Fig. 11. The cases of the Lock3DFace dataset that our previous work [7] failed to identify, but correctly identified by the proposed method. Top: RGB images; Bottom: ground truth depth images.

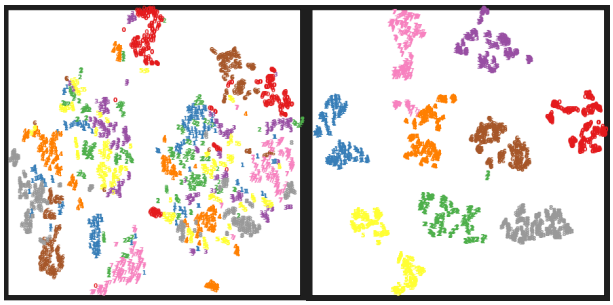

Fig. 12. The facial representations of Lock3DFace dataset visualized by t-SNE. Left: the identity features extracted by the baseline model; Right: the identity feature extracted by the proposed model.

the subject-related information only and minimizes such domain shift. We notice that the proposed disentanglement loss can mitigate the accuracy decrease (around -1% on average). Especially in the subsets containing pose, occlusion, and time variations, only a slight accuracy decrease (up to -2% in all subsets) occurs. We further demonstrate the cases that are miss-classified by our previous work but identified correctly by the proposed method in Fig. 11.

Fig. 12 illustrates how effectively the proposed RGB-D face recognition model disentangles the identity representation from input face images. We randomly select nine subjects that cover all variations, expression, pose, occlusion, and time, from the Lock3DFace dataset and map the extracted identity feature to a two-dimension space by t-SNE [40]. The left side of the figure depicts the distribution of identity feature representations extracted by the baseline model for the nine subjects. The right side of the figure shows the distribution of all the identity feature representations extracted by the model trained with all three proposed loss functions. The three proposed loss functions enforce our segmentation-guided RGB-D face recognition model to extract a much more compact and discriminative face feature representation.

6 CONCLUSIONS

This paper proposed a novel segmentation-guided RGB-D face recognition model and a new face depth estimation model. The face depth estimation model estimates the depth map from an RGB face image by including a semantic segmentation module for more precise face region localization. The estimated depth maps can be combined with 2D images to convert a 2D face image dataset to a RGB-D face dataset. This data augmentation approach helps to improve the accuracy and stability of training the RGB-D face recognition model. Furthermore, we developed a segmentation-guided RGB-D face recognition model, which includes the auxiliary segmentation attention module to fully exploit the augmented depth and semantic segmentation information. The proposed cross-modal focal loss, the semantic alignment loss, and the feature disentanglement loss enforce the proposed model to provide a compact and discriminative face identity representation. In our previous work [7], the domain shift between the augmented depth map and the ground truth depth map causes a significant decrease in

RGB-D face recognition accuracy. The novel identity-style disentanglement can mitigate the decrease in RGB-D face recognition accuracy caused by such domain shift. Our experiments showed that our DepthNet model provides accurate depth map estimation. The proposed RGB-D face recognition model outperforms state-of-the-art methods on several public 3D face datasets.

REFERENCES

[1] A. F. Abate, M. Nappi, D. Riccio, and G. Sabatino. 2d and 3d face recognition: A survey. *Pattern recognition letters*, 28(14):1885–1906, 2007.

[2] M. Abavisani, H. R. V. Joze, and V. M. Patel. Improving the performance of unimodal dynamic hand-gesture recognition with multimodal training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[3] A. Bulat and G. Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks). In *International Conference on Computer Vision*, volume 1, page 8, 2017.

[4] Y. Cai, Y. Lei, M. Yang, Z. You, and S. Shan. A fast and robust 3d face recognition approach based on deeply learned face representation. *Neurocomputing*, 363:375–397, 2019.

[5] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman. Vggface2: A dataset for recognising faces across pose and age, 2018.

[6] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 334–349, 2018.

[7] M. Chiu, H. Cheng, C. Wang, and S. Lai. High-accuracy rgb-d face recognition via segmentation-aware face depth estimation and mask-guided attention network. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pages 1–8, Los Alamitos, CA, USA, dec 2021. IEEE Computer Society.

[8] A. Chowdhury, S. Ghosh, R. Singh, and M. Vatsa. Rgb-d face recognition via learning-based reconstruction. In *2016 IEEE 8th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pages 1–7, 2016.

[9] 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)*, pages 140–147, 2018.

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

[11] T. C. Faltemier, K. W. Bowyer, and P. J. Flynn. Using a multi-instance enrollment representation to improve 3d face recognition, 2007.

[12] Y. Feng, F. Wu, X. Shao, Y. Wang, and X. Zhou. Joint 3d face reconstruction and dense alignment with position map regression network. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 534–551, 2018.

[13] H. Gao, H. K. Ekenel, and R. Stiefelhagen. Pose normalization for local appearance-based face recognition, 2009.

[14] A. George and S. Marcel. Cross modal focal loss for rgbd face anti-spoofing. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[15] S. Ghosh, R. Singh, M. Vatsa, and A. Noore. Rgb-d face recognition using reconstruction based shared representation. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pages 1–8, 2021.

[16] S. Z. Gilani and A. Mian. Learning from millions of 3d scans for large-scale 3d face recognition. *CoRR*, abs/1711.05942, 2017.

[17] S. Gong, X. Liu, and A. Jain. Jointly de-biasing face recognition and demographic attribute estimation. In *2020 European Conference on Computer Vision (ECCV 2020)*, pages 330–347, 10 2020.

[18] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition, 2016.

[19] S. Gupta, K. R. Castleman, M. K. Markey, and A. C. Bovik. Texas 3d face recognition database. In *2010 IEEE Southwest Symposium on Image Analysis Interpretation (SSIAI)*, pages 97–100, 2010.

[20] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[21] X. Hou, Y. Li, and S. Wang. Disentangled representation for age-invariant face recognition: A mutual information minimization perspective. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3692–3701, October 2021.

[22] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[23] L. Jiang, J. Zhang, and B. Deng. Robust rgb-d face recognition using attribute-aware loss. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[24] D. Kim, M. Hernandez, J. Choi, and G. Medioni. Deep 3d face identification, 2017.

[25] 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*, volume 1, page 3, 2016.

[26] Y. Lei, Y. Guo, M. Hayat, M. Bennamoun, and X. Zhou. A two-phase weighted collaborative representation for 3d partial face recognition with single sample. *Pattern Recognition*, 52:218–237, 2016.

[27] H. Li, D. Huang, J.-M. Morvan, Y. Wang, and L. Chen. Towards 3d face recognition in the real: a registration-free approach using fine-grained matching of 3d keypoint descriptors. *International Journal of Computer Vision*, 113(2):128–142, 2015.

[28] H. Li, J. Sun, Z. Xu, and L. Chen. Multimodal 2d+ 3d facial expression recognition with deep fusion convolutional neural network. *IEEE Transactions on Multimedia*, 19(12):2816–2831, 2017.

[29] Lijun Yin, Xiaozhou Wei, Yi Sun, Jun 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)*, pages 211–216, 2006.

[30] S. Lin, F. Liu, Y. Liu, and L. Shen. Local feature tensor based deep learning for 3d face recognition. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, pages 1–5. IEEE, 2019.

[31] A. Mian, M. Bennamoun, and R. Owens. An efficient multimodal 2d-3d hybrid approach to automatic face recognition. *IEEE transactions on pattern analysis and machine intelligence*, 29(11):1927–1943, 2007.

[32] G. Mu, D. Huang, G. Hu, J. Sun, and Y. Wang. Led3d: A lightweight and efficient deep approach to recognizing low-quality 3d faces. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5766–5775, 2019.

[33] 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*, pages 296–301. Ieee, 2009.

[34] P. J. Phillips, P. J. Flynn, T. Scruggs, K. W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek. Overview of the face recognition grand challenge, 2005.

[35] Y. Qian, W. Deng, and J. Hu. Unsupervised face normalization with extreme pose and expression in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9851–9858, 2019.

[36] 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*, pages 234–241. Springer, 2015.

[37] A. Savran, N. Alyuz, H. Dibeklioglu, O. Celiktutan, B. Gokberk, B. Sankur, and L. Akarun. Bosphorus database for 3d face analysis. *Workshop on Biometrics and Identity Management*, pages 47–56, 01 2008.

[38] H. Uppal, A. Sepas-Moghaddam, M. Greenspan, and A. Etemad. Attention-aware fusion rgb-d face recognition. *arXiv preprint arXiv:2003.00168*, 2020.

[39] H. Uppal, A. Sepas-Moghaddam, M. Greenspan, and A. Etemad. Depth as attention for face representation learning. *IEEE Transactions on Information Forensics and Security*, 16:2461–2476, 2021.

[40] L. van der Maaten and G. Hinton. Viualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605, 11 2008.

[41] G. Wang, H. Han, S. Shan, and X. Chen. Cross-domain face presentation attack detection via multi-domain disentangled representation learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6677–6686, 06 2020.

[42] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[43] Xiangyu Zhu, Xiaoming Liu, Zhen Lei, and Stan Z. Li. Face alignment in full pose range: A 3d total solution. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[44] X. Xiong, X. Wen, and C. Huang. Improving rgb-d face recognition

via transfer learning from a pretrained 2d network. In *International Symposium on Benchmarking, Measuring and Optimization*, pages 141–148. Springer, 2019.
[45] Yao Feng, Fan Wu, Xiaohu Shao, Yanfeng Wang, and Xi Zhou. Joint 3d face reconstruction and dense alignment with position map regression network. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.
[46] D. Yi, Z. Lei, S. Liao, and S. Z. Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014.
[47] L. Yin, X. Wei, Y. Sun, J. Wang, and M. J. Rosato. A 3d facial expression database for facial behavior research, 2006.
[48] J. Zhang, D. Huang, Y. Wang, and J. Sun. Lock3dface: A large-scale database of low-cost kinect 3d faces. In *2016 International Conference on Biometrics (ICB)*, pages 1–8, 2016.
[49] K.-Y. Zhang, T. Yao, J. Zhang, Y. Tai, S. Ding, J. Li, F. Huang, H. Song, and L. Ma. Face anti-spoofing via disentangled representation learning. In *2020 European Conference on Computer Vision (ECCV 2020)*, 2020.
[50] W. Zhang, X. Ji, K. Chen, Y. Ding, and C. Fan. Learning a facial expression embedding disentangled from identity. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6755–6764, 2021.
[51] Z. Zhang, F. Da, and Y. Yu. Data-free point cloud network for 3d face recognition. *arXiv*, pages arXiv–1911, 2019.
[52] X. Zhu, X. Liu, Z. Lei, and S. Z. Li. Face alignment in full pose range: A 3d total solution. *IEEE transactions on pattern analysis and machine intelligence*, 41(1):78–92, 2017.
[53] S. Zulqarnain Gilani and A. Mian. Learning from millions of 3d scans for large-scale 3d face recognition, 2018.

Meng-Tzu Chiu is a PhD candidate in the Computer Vision Science Department of National Tsing Hua University, supervised by Prof. Dr. Shang-Hong Lai. She holds a BSc degree in EE from the National Chung Hsing University, and an MSc degree in EE from National Chiao Tung University. Her research focuses on Computer Vision, including Generative Adversarial Neural Networks, Multi-modal Face Recondition and Facial Representation Learning.

Hsun-Ying Cheng is a master student in the Computer Vision Science Department of National Tsing Hua University, supervised by Prof. Dr. Shang-Hong Lai.

Chien-Yi Wang is a senior research engineer at Microsoft, working on face recognition and face anti-spoofing. He received the B.S. degree from the EE department at National Taiwan University in 2013 and M.S. degree from the EE department at University of Southern California in 2016.

Shang-Hong Lai received the Ph.D. degree in electrical and computer engineering from the University of Florida, Gainesville, in 1995. Then, he joined Siemens Corporate Research, Princeton, New Jersey, as a member of technical staff. Since 1999, he became a faculty member with Department of Computer Science, National Tsing Hua University (NTHU), Taiwan, where he is currently a Professor. Since 2018, he has been on leave from NTHU to join Microsoft AI RD Center, Taipei, as a Principal Researcher. He has authored more than 300 articles published in the related international journals and conferences. In addition, he has been awarded around 30 patents for his researches on computer vision and medical imaging. His research interests include computer vision, image processing, and machine learning. He has been a program committee member of several international conferences, including CVPR, ICCV, ECCV, AAAI, ACCV, ICPR, and ICME. He has been an Associate Editor for several international journals, including Pattern Recognition and Journal of Signal Processing Systems.