High-Accuracy RGB-D Face Recognition via Segmentation-Aware Face Depth Estimation and Mask-Guided Attention Network

Meng-Tzu Chiu[1], Hsun-Ying Cheng[1], Chien-Yi Wang[2], and Shang-Hong Lai[1,2]
[1] Department of Computer Science, National Tsing Hua University, Taiwan
[2] Microsoft AI R&D Center, Taiwan

2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021) | 978-1-6654-3176-7/21/$31.00 ©2021 IEEE | DOI: 10.1109/FG52635.2021.9667000

***Abstract*— Deep learning approaches have achieved highly accurate face recognition by training the models with very large face image datasets. Unlike the availability of large 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 including semantic segmentation information for more accurate face region localization. The other is a novel mask-guided RGB-D face recognition model that contains an RGB recognition branch, a depth map recognition branch, and an auxiliary segmentation mask branch with a spatial attention module. Our DepthNet is used to augment a large 2D face image dataset to a large RGB-D face dataset, which is used for training an accurate RGB-D face recognition model. Furthermore, the proposed mask-guided RGB-D face recognition model can fully exploit the depth map and segmentation mask information and is more robust against pose variation than previous methods. Our experimental results show that DepthNet can produce more reliable depth maps from face images with the segmentation mask. Our mask-guided face recognition model outperforms state-of-the-art methods on several public 3D face datasets.**

I. INTRODUCTION

Face recognition has been a rapidly developing research task in recent years and has been widely used for many different applications, such as video surveillance, biometric identification, security verification, etc. Although 2D face recognition based on deep learning has achieved very high accuracy in most public datasets, face recognition is still very challenging under large pose variations [1]. To overcome this problem, some face-frontalization methods have been proposed to normalize profile face images to frontal pose [11][30], and some focused on RGB-D face recognition. 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 large pose and illumination variations.

The development of 3D or RGB-D face recognition is slower than 2D face recognition. The main reason is the lack of large 3D or RGB-D face datasets available to the public. 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 [39][5][14], whereas existing public 3D face datasets usually contain only hundreds of subjects or at most thousands of images. [40][29][9]. It is easy to fall into the overfitting problem when we only use a limited number of subjects in a 3D dataset to train a face recognition model.

978-1-6654-3176-7/21/$31.00 ©2021 IEEE

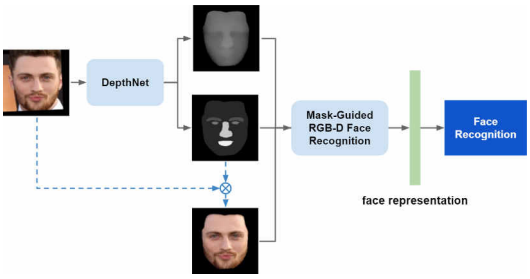

Fig. 1. The pipeline of the proposed RGB-D face recognition system

To address the problem of lacking large 3D face datasets for model training, many works [44][19][4] applied different data augmentation methods to train their face recognition models. [19] changed the values of expression parameters of 3DMM model and randomly generated rigid transformations matrices to the input 3D point cloud to synthesize expression and pose variations. [39] 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.

This paper presents a new method to convert a 2D face dataset to a 3D face dataset to address inadequate numbers of subjects in 3D face datasets for model training. Our system has two major parts, i.e., the depth estimation module (DepthNet) and the mask-guided face recognition module. We include a face semantic segmentation branch into the depth estimation network model as an auxiliary task for the depth estimation module. The module can correctly recognize where the facial features are located to estimate realistic face depth images. The proposed mask-guided face recognition model takes RGB face images, segmentation 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 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 mask-guided face recognition model which contains an RGB recognition branch, a depth map recognition branch, and an auxiliary segmentation mask branch with spatial attention module to overcome challenging variations in expression, pose, and occlusion.
4) Experiments on several public 3D face datasets demonstrate that the proposed mask-guided face recognition model outperforms the state-of-the-art methods for RGB-D face recognition.

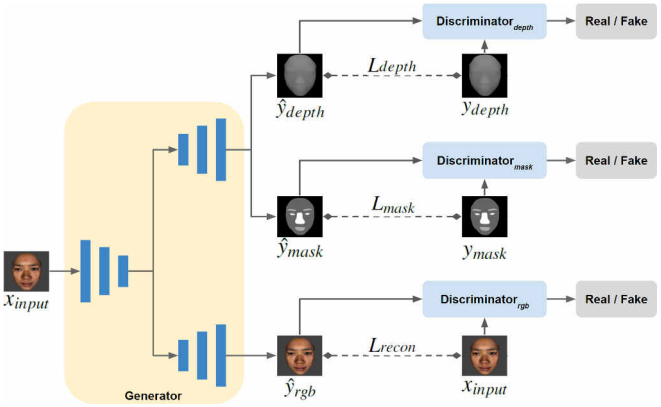

Fig. 2. **Architecture of the proposed DepthNet model.** Each input image (X_{input}) is first fed into the encoder to encode the face feature vector. Then, the two branches of the decoder generate the semantic segmentation mask ($\hat{y}_{mask}$), the depth map ($\hat{y}_{depth}$), and the reconstructed image ($\hat{y}_{rgb}$) from the embedded features.

II. Related Works

A. 3D Data Augmentation

Due to 3D face data scarcity, many 3D face recognition works focused on developing different 3D data augmentation methods. Kim et al. [19] proposed a 3D face augmentation technique that synthesizes several different facial expressions from a single 3D face scan. Each point cloud from FRGCv2 dataset [29] was fitted to a BFM [28] model to produce 25 expressions for each face model by modifying the expression parameters. Gilani et al. [44] generated millions of 3D facial models of different virtual identities by simultaneously interpolating between the facial identity and facial expression spaces. Zhang et al. [42] applied GPMM to generate a large 3D face training dataset and compensated the distribution difference between the generated data and real faces by constraining the face sampling area.

The methods mentioned above proposed to achieve 3D face data augmentation either via sampling from a low-dimensional identity and expression parametric space for a 3D face morphable model, such as GPMM, or interpolating 3D face models from actual 3D face scans. However, it is still not clear how effective the data synthesis of new virtual identities will benefit the training of face recognition models. This paper proposes converting an existing 2D face dataset to an RGB-D face dataset by estimating associated depth maps from 2D face images. A new CNN-based face depth estimation model, called DepthNet, is developed for this specific image-to-image translation problem.

B. RGB-D Face Recognition

Deep learning-based RGB-D face recognition research is not very active compared to 2D face recognition. One reason is that an effective way to pass the 3D data to the neural network is still under research. Additionally, there is a lack of large 3D face datasets available in public, as mentioned above. Therefore, many works [19][20][33][37] employed the CNN that was pre-trained on 2D face images to fine-tune on the relatively small 3D dataset. Gilani et al. [44] took depth, azimuth, and elevation angles of the normal vector as a 3-channel input and proposed the first deep CNN model specifically designed for RGB-D face recognition. Jiang et al. [18] 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. [23] 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. [42] 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.

III. Proposed Method

We aim to build a robust RGB-D face recognition model from the 2D face image dataset. To achieve this goal, we propose a new CNN model for generating the associated depth map and segmentation mask from an input face image. We can then generate a large RGB-D face dataset from a large 2D RGB face dataset to improve the training of RGB-D face recognition models. Our system consists of two modules: (1) the DepthNet and (2) the mask-guided 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 mask-guided RGB-D face recognition model for RGB-D face recognition. Our mask-guided 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.

A. DepthNet

Fig. 2 illustrates the framework of the proposed DepthNet model, which includes a generator and three discriminators. The generator can be divided into three networks, the face encoder, the face decoder, and the auxiliary decoder. This generator is based on the UNet [31] architecture, which is an encoder-decoder model with a skip-connection module. With the skip-connection module, the decoder can directly use the features from the encoder. For a given source face image X_{input}, which passes through the face encoder and the auxiliary decoder to encode image x_{input} information.

We obtain the reconstructed image $\hat{y}_{rgb}$ with the face decoder. To minimize the distance between the reconstructed image $\hat{y}_{rgb}$ and source face image X_{input}, we adopt the L1 loss as follows:

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

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 output the segmentation mask and depth at the same time. To minimize the distance between the generated depth image $\hat{y}_{depth}$ and ground truth depth y_{depth}, we adopt the L1 loss as follows:

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

We adopt the binary cross-entropy loss to train the network to generate the semantic segmentation mask for an input image. This loss enforces the output of the encoder to be similar to the ground-truth semantic segmentation. It is given by

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

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 generative models to learn to reconstruct images to train the RGB discriminator D_{rgb}, depth discriminator D_d, and mask discriminator D_m, given by

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

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

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.

B. Mask-Guided RGB-D Face Recognition

Fig. 3 demonstrates our mask-guided RGB-D face recognition network architecture, which contains an RGB recognition branch, a depth map recognition branch, and an auxiliary segmentation mask branch with spatial attention module proposed in [35]. At the training stage, the RGB recognition branch extracts the face representation feature, f_{RGB} in the figure, by the backbone network SENet [17] network denoted as SENet_RGB. Similarly, the depth map recognition branch extracts corresponding f_D by SENet_D. The auxiliary segmentation mask branch extracts different level of feature map from segmentation mask by SENet_M, and then applies spatial attention module (SAM) on those feature maps to aid RGB and D branches while training. This SAM is shared-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 classifier with ArcFace [8] additive angular margin loss predicts a vector of probabilities with one value for each possible identity.

The proposed mask-guided RGB-D face recognition network is a two-stream-multi-head architecture, and we apply the cross-entropy loss as classification losses L_{cls} on individual branches. We adopt the cross-modal focal loss $L_{CMFL}^{m,n}$ in [12] to learn robust representations jointly, which is defined as

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

α and γ are tunable hyper-parameters.

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

where m_t and n_t denote the classification probabilities after fully connected layer in current branch m and the other branch n, respectively. The CMFL contributed by branch n will reduce when branch n can predict with high confidence.

Although the inputs to the mask-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 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 classification losses between m and n to measure the performance of the network, and it is denoted as $L_{cls}^m - L_{cls}^n$. If the difference is positive, then it means modality m is

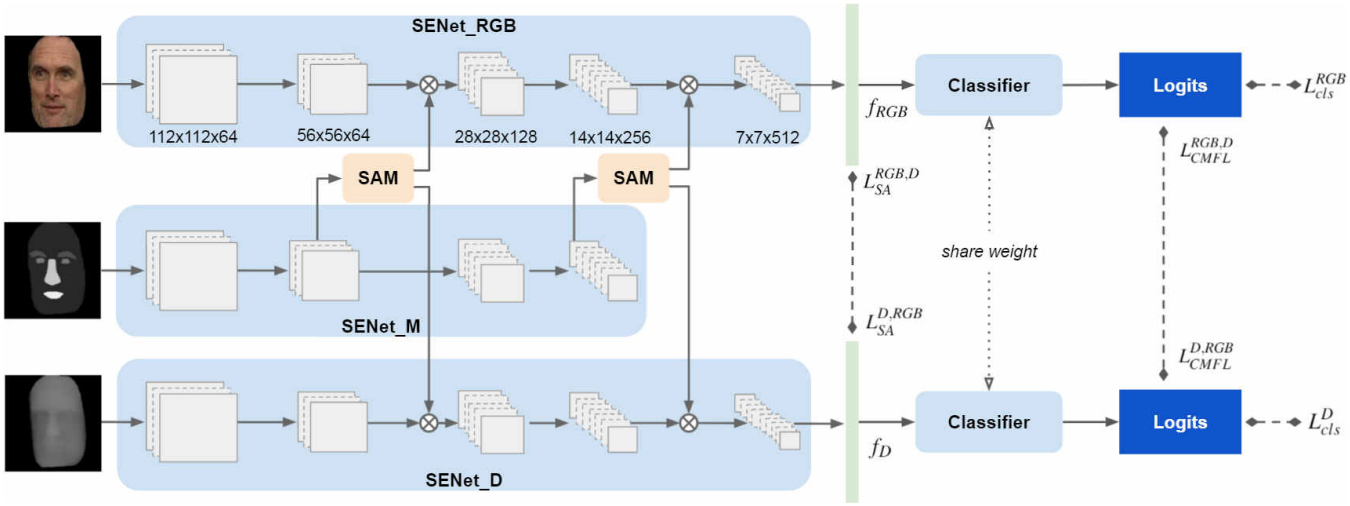

Fig. 3. The proposed mask-guided RGB-D face recognition network architecture.

weaker than modality n. The model will enforce f_m to be similar to f_n. The focal regularization parameter is defined as follows

$$\rho^{m,n} = \begin{cases} e^{\beta(L_{cls}^m - L_{cls}^n)} - 1, & \text{if } L_{cls}^m > L_{cls}^n \\ 0, & \text{if } L_{cls}^m \leq L_{cls}^n \end{cases} \quad (11)$$

where β is a positive focusing parameter.

The overall loss functions in branch RGB and D are given as

$$L_{total}^{RGB} = (1-\lambda_1)L_{cls}^{RGB} + \lambda_1 L_{CMFL}^{RGB,D} + \lambda_2 L_{SA}^{RGB,D} \quad (12)$$

$$L_{total}^{D} = (1-\lambda_1)L_{cls}^{D} + \lambda_1 L_{CMFL}^{D,RGB} + \lambda_2 L_{SA}^{D,RGB} \quad (13)$$

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

At the testing stage, instead of simply concatenating f_{RGB} and f_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 the combination of all modalities provides the most accurate result for face recognition. More details will be provided in the next section.

IV. Experimental Results

A. Experimental Datasets

Here we introduce the datasets that were used in the training and evaluation. The DepthNet model is trained with BU-3DFE 3D database [24] dataset. The VGGFace2 [5] 2D face dataset is augmented to the corresponding RGB-D face dataset by applying the DepthNet model, and then is used to train the mask-guided RGB-D face recognition model. We experiment on public 3D datasets: BU-3DFE [24], Texas FR3D [15], Bosphorus [32], FRGCv2 [29], and Lock3DFace [41] to evaluate the proposed DepthNet and mask-guided RGB-D face recognition model.

1) 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[24] 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. 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 training data. It contains 9,131 subjects and a total of 3.31 million RGB images. The proposed DepthNet model produces the corresponding depth images and segmentation masks. 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 larger RGB datasets. However, due to the memory and training time consideration, we choose to use the VGGFace2 dataset for conversion into the RGB-D dataset to train our face recognition model.

2) 3D Face Datasets:

BU-3DFE 3D database[24] includes 100 subjects with 2,500 scans. Each identity performs seven expressions with four levels of intensity for each expression except for the neutral one. There is no pose variation in this database. The evaluation protocol as [13] is adopted so that we have 100 images in the gallery and 2,400 images in the probe to calculate the identification accuracy.

Texas FR3D database[15] contains 1,149 scans of 118 subjects. All the scans are frontal with different expressions. We select the first images for all the 118 subjects as the gallery and put the remaining 1,031 images as the probe.

Bosphorus database[32] consists of 4,666 scans of 105 subjects in various poses, expressions and occlusions. There are two settings for evaluation. *Setting-1* considers expression variations for recognition. The first neutral image of each subject is selected as the gallery. The other 2,797 images are regarded as probe images. *Setting-2* takes all the remaining 4,561 images in the probe and 105 images in the gallery.

FRGCv2 database[29] 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.

BUAA Lock3DFace database[41] contains 5711 RGB-D face videos of 509 subjects with variations in facial expression, pose, occlusion and time-lapse. We follow the same testing protocols as described in [41]. 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).

B. 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). For the hyper-parameter of the loss function in (12) and (13),we set $\lambda_1 = 0.5$ and $\lambda_2 = 0.05$. We train RGB-D face recognition on 2 Tesla V100 GPUs with batch size 256 and image size 112x112.

C. DepthNet Evaluation

In this section, we demonstrate 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. 4, we depict some examples of applying our DepthNet to VGGFace2 2D face database, which is the training set for our RGB-D face recognition model. The results show that our method can produce well-preserved face contour and face shape features of different expressions.

In Table I, 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 [36] and PRNet [38]. For a fair comparison, we only calculate the MSE in the intersection of ground truth depth and all predicted depth images from our DepthNet, 3DDFA, and PRNet. 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 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. 5. This is because our model includes semantic segmentation together with depth estimation. Our DepthNet is trained with BU3DFE which was 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.

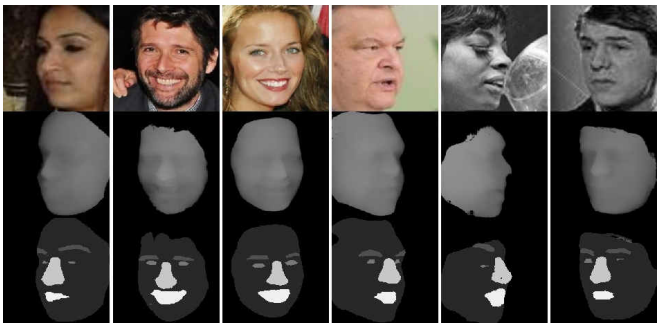

Fig. 4. Generated augmented depth and segmentation images of VGGFace2. Rows from top to bottom: RGB images, augmented depth images, and segmentation images.

TABLE I

QUANTITATIVE COMPARISON OF DEPTH ESTIMATION ERRORS FOR DIFFERENT METHODS: MSE BETWEEN GROUND TRUTH DEPTH AND ESTIMATED DEPTH IMAGES.

Method	BU3DFE	FRGCv2	Bosphorus
3DDFA [43]	125.78	597.17	540.48
PRNet [10]	74.86	**216.29**	615.99
Ours	**16.65**	435.88	**421.46**

In Figure 6, we further illustrate how our DepthNet provides superior depth estimation for face images with large poses. 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 that is consistent with the RGB face image.

D. Mask-Guided RGB-D Face Recognition Evaluation

Our mask-guided RGB-D face recognition model has good generalization ability on other 3D datasets, even though it is trained with a depth-augmented 2D dataset. Our network is trained on VGGFace2 [5] and directly tested on all other 3D face datasets without any fine-tuning. Our mask-guided RGB-D face recognition model takes RGB face image,

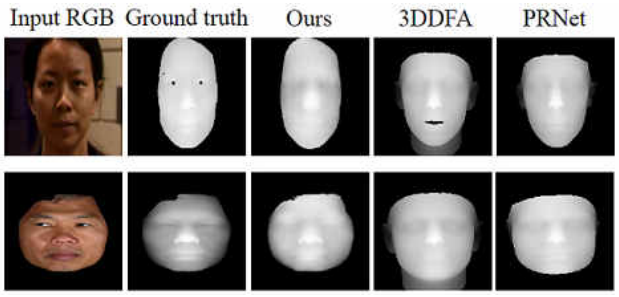

Fig. 5. Depth estimation results by using different methods on some sample images in FRGCv2 dataset (top row) and Texas dataset (bottom row).

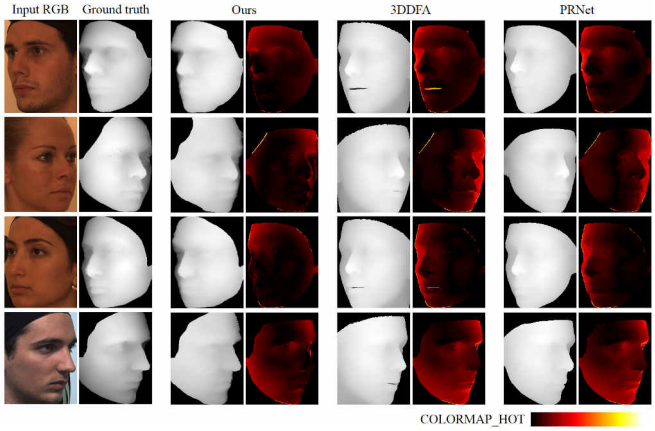

Fig. 6. Depth estimation results by using different methods on the Bosphorus dataset. We utilize hot colormap to illustrate the MSE. The darker the color, the smaller the error.

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 II. For all the datasets, our method provides state-of-the-art verification accuracy. Especially for Bosphorus-2, which has pose variations, the proposed method marginally outperforms the other methods by around 1.7% accuracy.

Table III 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 mask-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.*[18] 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 get an RGB-D database with sufficient subjects from the existing 2D face databases and do not need to collect a new 3D face database.

TABLE II
THE RANK-1 IDENTIFICATION ACCURACY ON PUBLIC 3D FACE DATABASES.

Method	BU3DFE	Texas	Bosphorus-1	Bosphorus-2
Li *et al.*[22]	-	-	98.8	96.6
Lei *et al.*[21]	93.25	-	98.9	-
Mian *et al.*[26]	95.9	98.0	-	96.4
Lin *et al.*[25]	96.2	-	99.71	-
Kim *et al.*[19]	95.0	-	99.2	-
FR3DNet [44]	98.64	100	-	96.18
Ours	**100**	**100**	**100**	**97.94**

The rank-1 identification accuracies for the Lock3DFace dataset are shown in Table IV. Especially in the subset with pose variations, our result achieves a 96.55% accuracy which is significantly better (+25%) than others. For occlusion variations such as covering the face with hand or glasses, we reach an accuracy of 97.31% obtaining about +12% performance gain. In the subset over time scenario, our method also accomplished an accuracy of 92.38%, which exceeds others by +11%. It is worth noting that some other methods include part of the Lock3DFace in their training set; However, our mask-guided directly test on Lock3DFace without any fine-tuning. In general, our mask-guided RGB-D recognition model achieves a much higher (+9%) average accuracy of 96.43% comparing to other state-of-the-art methods. This indicates that our mask-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.

Fig. 7 demonstrates visualization results of the mask-guided spatial attention module on some pulic 3D datasets. The result shows some samples with expression, pose, and occlusion variations. The segmentation mask branch provides auxiliary information to spatial attention module; therefore, we can observe that the attention have selectively focused on the informative parts such as eyes, nose, eyebrows, and lips for RGB-D face recognition.

V. Ablation Study

A. Effect of the Segmentation Mask

In this section, we first analyze the effects of the segmentation mask branch in the proposed DepthNet. Table V demonstrates significant improvement of the depth estimation by including the semantic segmentation into the model. Different from section IV-C, we directly calculate the MSE between the estimated depth image and the ground truth image. We can observe that with the addition of the semantic segmentation branch, it can focus on the face features and provide precise depth estimation. We can easily perceive the expression of both profile images and frontal images with the segmentation mask.

TABLE III

RANK-1 IDENTIFICATION ACCURACY APPLIED TO DIFFERENT MODALITIES. VGGFACE2* DENOTES THE AUGMENTED DATA OF VGGFACE2 THAT PRODUCED BY DEPTHNET. D* AND M* DENOTES THE AUGMENTED DEPTH MAP AND SEGMENTATION MASK GENERATED BY DEPTHNET.

Method	Training data	Subjects	Testing Modality	FRGCv2	Bosphorus-1
VGG-Face[5]	Private[44]	100	RGB	87.92	96.39
Jiang *et al.*[18]	TRAINING-SET-I[18]	60,000	RGB	95.69	96.08
Ours	VGGFace2*[5]	9,131	RGB + M*	**99.07**	**99.75**
Li *et al.*[22]	Part of Bosphorus[32]	105	Depth	96.30	95.40
FR3DNet[44]	Private[44]	100	Depth	97.06	96.18
Jiang *et al.*[18]	TRAINING-SET-I[18]	60,000	Depth	97.45	**99.37**
Ours	VGGFace2*[5]	9,131	D* + M*	**98.42**	98.61
Li *et al.*[22]	Part of FRGCv2[29]	466	RGB + Depth	95.20	99.40
Jiang *et al.*[18]	TRAINING-SET-I[18]	60,000	RGB + Depth	98.52	99.52
Ours	VGGFace2*[5]	9,131	RGB + D* + M*	**99.27**	**100**

TABLE IV

THE RANK-1 IDENTIFICATION ACCURACY ON LOCK3DFACE DATABASES. D* AND M* DENOTES THE AUGMENTED DEPTH MAP AND SEGMENTATION MASK GENERATED BY DEPTHNET.

Method	Input	Accuracy				
		Expression	Pose	Occlusion	Time	Average
He *et al.*[16]	RGB	96.3	58.4	74.7	75.5	76.2
Hu *et al.*[17]	RGB	98.2	60.7	77.9	78.3	78.7
Cui *et al.*[7]	RGB + D	97.3	54.6	69.6	66.1	71.9
Mu *et al.*[27]	RGB + 3D Model	98.2	70.4	78.1	65.3	84.2
Uppal *et al.*[34]	RGB + D	99.4	70.6	85.8	81.1	87.3
Ours	RGB + D* + M*	**99.92**	**96.55**	**97.31**	**92.38**	**96.43**

TABLE V

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**

TABLE VI

IDENTIFICATION ACCURACY ON LOCK3DFACE DATASET OF OUR MASK-GUIDED RGB-D FACE RECOGNITION MODEL AND ITS VARIANTS WITH DIFFERENT MODALITIES.

Method	Lock3DFace				
	Expression	Pose	Occlusion	Time	Average
D*+M*	99.46	81.36	79.08	71.89	83.12
RGB+M*	99.77	93.98	94.22	88.68	94.09
RGB+D*+M*	**99.92**	**96.55**	**97.31**	**92.38**	**96.43**

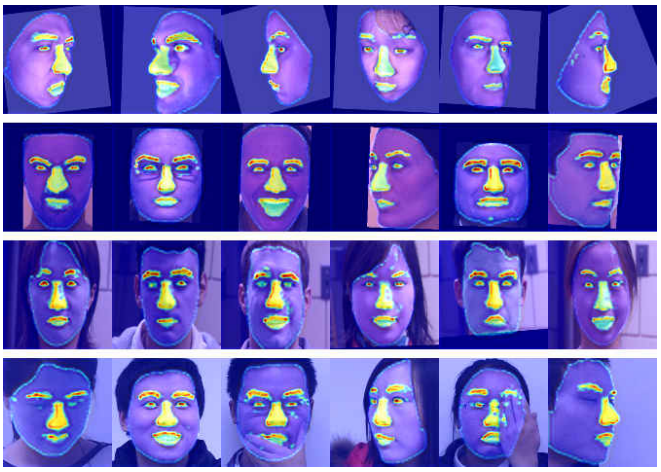

Fig. 7. Spatial attention maps four RGB-D face datasets. First row: BU-3DFE dataset; Second row: Bosphorus datase; Third row: FRGCv2 dataset; and Last row: Lock3DFace dataset.

B. *Effect of the DepthNet*

We report the rank-1 face identification results of the proposed mask-guided RGB-D face recognition module and its variants with different combination of modalities, RGB images or augmented depth images D* or segmentation mask images M*, as the ablation study. The comparison results are presented in Table VI. The first row is the result that we test with augmented depth and the augmented segmentation mask images. The second row is tested with RGB and the augmented segmentation mask images. The third row is tested with RGB, the augmented depth, and the augmented segmentation mask images. We can observe that using both augmented depth and segmentation mask images achieves the highest accuracy, which indicates each component is essential in our mask-guided RGB-D face recognition method.

VI. CONCLUSIONS

In this paper, we propose a novel framework that estimates depth maps from RGB face images by including a semantic segmentation module for more precise face region localization. The estimated depth maps can be combined with 2D images to augment the 2D face image dataset to RGB-D face dataset. This data augmentation approach helps to

improve the accuracy and stability for training the RGB-D face recognition model. Furthermore, we developed a mask-guided RGB-D face recognition model, which includes the auxiliary segmentation attention module to fully exploit the augmented depth and segmentation mask information. Our experiments showed that our DepthNet model provide accurate depth map estimation and the proposed mask-guided RGB-D face recognition model outperforms state-of-the-art face recognition 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] 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.

[8] 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.

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

[10] 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.

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

[12] 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.

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

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

[15] 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.

[16] 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.

[17] 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.

[18] 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.

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

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

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

[34] 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.

[35] 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.

[36] 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.

[37] 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.

[38] 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.

[39] D. Yi, Z. Lei, S. Liao, and S. Z. Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014.

[40] L. Yin, X. Wei, Y. Sun, J. Wang, and M. J. Rosato. A 3d facial expression database for facial behavior research, 2006.

[41] 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.

[42] Z. Zhang, F. Da, and Y. Yu. Data-free point cloud network for 3d face recognition. *arXiv*, pages arXiv–1911, 2019.

[43] 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.

[44] S. Zulqarnain Gilani and A. Mian. Learning from millions of 3d scans for large-scale 3d face recognition, 2018.