2023 IEEE 5th International Conference on Artificial Intelligence Circuits and Systems (AICAS) | 979-8-3503-3267-4/23/$31.00 ©2023 IEEE | DOI: 10.1109/AICAS57966.2023.10168669

Landmark-Based Adversarial Network for RGB-D Pose Invariant Face Recognition

1st Wei-Jyun Chen
Department of Computer Science
National Tsing Hua University
jyunchen@gapp.nthu.edu.tw

2nd Ching-Te Chiu
Department of Computer Science
National Tsing Hua University
chiusms@cs.nthu.edu.tw

3rd Ting-Chun Lin
Department of Computer Science
National Tsing Hua University
tinglin@gapp.nthu.edu.tw

***Abstract*—Even though numerous studies have been conducted, face recognition still suffers from poor performance in pose variance. Besides fine appearance details of the face from RGB images, we use depth images that present the 3D contour of the face to improve recognition performance in large poses. At first, we propose a dual-path RGB-D face recognition model which learns features from separate RGB and depth images and fuses the two features into one identity feature. We add associate loss to strengthen the complementary and improve performance. Second, we proposed a landmark-based adversarial network to help the face recognition model extract the pose-invariant identity feature. Our landmark-based adversarial network contains a feature generator, pose discriminator, and landmark module. After we use 2-stage optimization to optimize the pose discriminator and feature generator, we removed the pose factor in the feature extracted by the generator. We conduct experiments on KinectFaceDB, $\mathrm{RealSense}_{\mathrm{test}}$ and $\mathrm{LiDAR}_{\mathrm{test}}$. On KinectFaceDB, we achieve a recognition accuracy of 99.41%, which is 1.31% higher than other methods. On $\mathrm{RealSense}_{\mathrm{test}}$, we achieve a classification accuracy of 92.57%, which is 30.51% higher than other methods. On $\mathrm{LiDAR}_{\mathrm{test}}$, we achieve 98.21%, which is 21.88% higher than other methods.**

***Index Terms*—face recognition, pose invariant face recognition, pose invariant feature, adversarial network, facial landmark**

I. Introduction

Current RGB-based face recognition approaches have reached great success, but they only rely on appearance information and may be affected by insufficient lighting conditions and large poses. In addition, depth sensors become an important technology for using depth images to face recognition. Depth images captured by RGB-D sensors supply extra geometric information of 3D contour to increase the robustness of face recognition.

We focus on the problem that recognition accuracy decrease when the head poses increase. Some methods introduce margin penalty in loss function, such as ArcFace [1], SphereFace [2], CosFace [3], and CurricularFace [4]. These models trained with these loss functions achieve high recognition accuracy on the LFW [5] dataset which mainly consists of frontal view images, but still make wrong recognition on CPLFW [6] and CFP-FP [7] datasets which consist of not only frontal faces but also profile faces. We notice these models drop the accuracy by about 5% to 10% on CPLFW [6] than LFW [5]. We also observe the performance on Multi-PIE [8] dataset, which contains face images with different yaw angles. We notice that Wang [9] achieves 87.35% at ±90°, which is 12.65% lower than that at ±15°.

Many other approaches focus on attempting to make the model learn identity feature representations of the same person with different poses. These methods for pose-invariant face recognition can be classified into two categories: face frontalization and pose-invariant representation learning. The main difference between face frontalization and pose-invariant representation learning is that face frontalization pays attention to input image pre-processing, but pose-invariant representation learning pays attention to post-processing of feature representation. The general face recognition model extracts a mixed feature representation that contains not only identity information but also poses information. Extracting the pose invariant identity feature representation can enhance performance for large pose face recognition. We proposed a method based on adversarial learning to remove pose information and extract pose invariant identity feature representation.

To deal with pose-invariant face recognition, we are inspired by the Generative Adversarial Networks (GANs). We design a landmark-based adversarial network to finetune a pretrained RGB-D face recognition model to extract pose-invariant identity features. Our landmark-based adversarial network contains three components: an identity feature generator, a pose discriminator, and a landmark module. The pose discriminator learns to discriminate whether the feature maps from the feature generator contain pose information. The feature generator is trained for two tasks simultaneously: extracting identity features and suppressing pose information via adversarial learning with the pose discriminator. After optimizing the two networks, we removed the pose factor in the feature extraction of the generator. In addition, we design a landmark module for our adversarial network to map facial landmarks into pose features with one-to-one mappings.

II. Proposed Methods

There are two steps in this work: training an RGB-D face recognition model to be the feature generator and using a landmark-based adversarial network to finetune the pre-trained model. Our landmark-based adversarial network includes a feature generator, a pose discriminator, and a landmark module. We introduce two-stage optimization in our landmark-based adversarial network.

A. Overall Scenario of Proposed Method

The data flow of our landmark-based adversarial network is shown in Fig. 1. There are a landmark module, a pose discriminator, and a feature generator in the data flow. The input of the feature generator is RGB-D images and the output is the pose-invariant identity feature. The input of the pose discriminator is the feature maps from the feature generator and the pose discriminator extract pose features from the feature maps. The input of the landmark module is the landmark heatmaps produced by the RGB images.

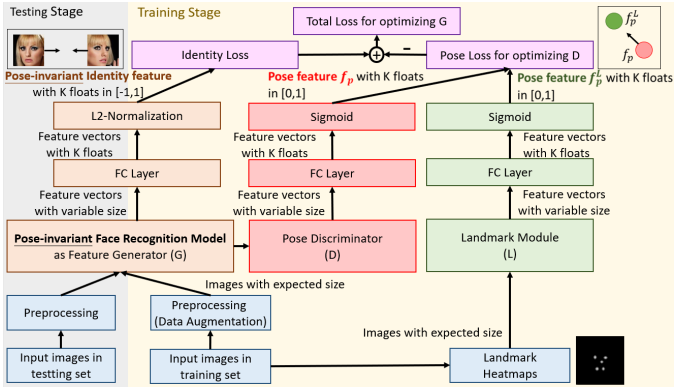

Fig. 1. Data flow of our proposed architecture. In the testing phase, only Feature Generator is passed. In the training phase, we freeze the parameters of the Landmark Module.

B. Network Architecture

1) Landmark Module: We design an Auto-Encoder to embed the landmark heatmaps and use the Encoder part as our landmark module. We generate 68 facial landmarks by SynergyNet [10]. Since some facial landmarks contain identity information, we select 14 of the 68 facial landmarks to generate the heatmap.

Loss Function for Landmark Module We utilize Mean Square Error (MSE) loss shown below:

$$L_{AE} = \frac{1}{D}(H_o - H_i)^2 \tag{1}$$

where H_i and H_o is the inputs and outputs of Auto-Encoder corresponding and D is the number of elements from H_i.

2) Feature Generator: We design a dual path RGB-D face recognition model as the pre-trained model for the feature generator. The model is modified by CvT-13 [11]. We utilize CvT-13 to extract the RGB feature and the depth features separately. After extracting the features of RGB and depth, we add the two features into one feature as the fused feature. We define the output of the final fully connected layer as the classification feature and the output of the penultimate fully connected layer as the identity feature.

Loss Function for Feature Generator

Fig. 2 shows the pipeline of our dual path RGB-D face recognition model. We utilize associate loss and classification loss to assist in training the model. Associate loss between

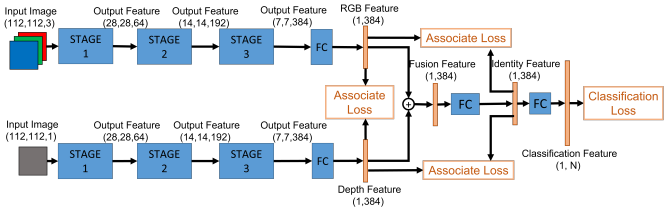

Fig. 2. The pipeline of dual path RGB-D face recognition model. N is the number of identities.

the identity feature and RGB feature and depth feature corresponding is used to make the RGB path and depth path extract more discriminative features. The classification loss function we used is Cross Entropy Loss with label smoothing value α. The total loss of the pre-trained face recognition model is shown below:

$$L_{total,pre} = L_A + L_{CE} \tag{2}$$

$$L_A = \frac{1}{M}\left(\sqrt{\sum_{i=1}^{M}|x_i - y_i|^2} + \sqrt{\sum_{i=1}^{M}|x_i - z_i|^2} + \sqrt{\sum_{i=1}^{M}|y_i - z_i|^2}\right) \tag{3}$$

$$L_{CE} = \frac{1}{NM}\sum_{i=1}^{M}\sum_{i=1}^{N} t_{mn} ln(SoftMax(\hat{t}_{mn})) \tag{4}$$

$$t_{mn} = \begin{cases} 1-\alpha, & m = n \\ \alpha/N, & otherwise \end{cases} \tag{5}$$

where L_A is associate loss function, x_i is the *L2* normalized RGB feature of *i*th image, y_i is the *L2* normalized depth feature of *i*th image, z_i is the *L2* normalized identity feature of *i*th image, L_{CE} is Cross Entropy loss, t_{mn} is target class label with label smoothing value α, $\hat{t}_{mn}$ is the classification feature, M is number of images in a batch and N is number of class.

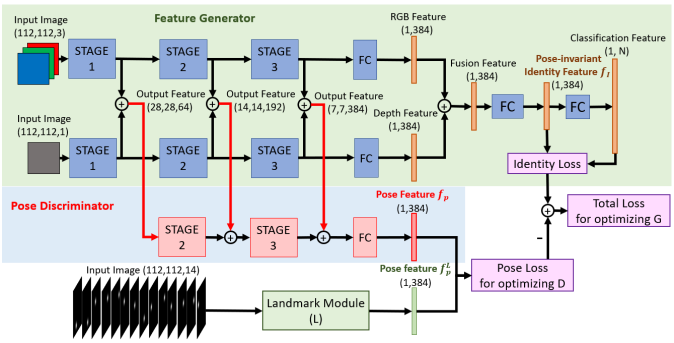

Fig. 3. Details of Landmark-based adversarial network.

3) Pose Discriminator in Landmark-based adversarial network: The architecture of the pose discriminator is combined with stage 2, stage 3, and the first fully connected layer of RGB path. We aim to train the pose discriminator to be sensitive to pose information.

C. Training Loss Function for Landmark-based adversarial network

There are two main loss functions shown in Fig. 3. One is pose loss which helps the pose discriminator extract pose features. The other is total loss in Equation (6) which is identity loss minus pose loss. The identity loss helps the feature generator to extract identity features and the minus of pose loss can help the identity feature to be pose-invariant.

$$L_{total} = L_{Id} - L_{pose} \tag{6}$$

The identity loss includes triplet loss and Cross Entropy loss with label smoothing value α. Triplet loss is used to make the distance of inter-class features far and the distance of intra-class features close. The inputs of the triplet loss are an anchor feature, a positive feature which is the same class as the anchor's class, and a negative class which is a different class from the anchor's class. Since the triplet loss only considers the distance between two classes, the training process is unstable. To deal with this problem, we add Cross Entropy loss with label smoothing value α to assist training. The Cross Entropy loss is denoted in Equation (4). We denote the identity loss and triplet loss below:

$$L_{Id} = L_{triplet} + L_{CE} \tag{7}$$

$$L_{triplet} = max(d(a,p) - d(a,n) + margin, 0) \tag{8}$$

where a is the anchor feature, p is the positive feature which is the same class as the anchor feature, n is the negative feature whose class is different from the anchor feature, $margin$ is a positive value, and $d()$ is the Euclidean distance function.

As mentioned above, the output of the landmark module is a pseudo label for training pose discriminator. The purpose of pose loss is to make the distance between the pose feature from the pose discriminator and pseudo label close. We utilize mean square error loss as pose loss which is shown below:

$$L_{pose} = \frac{1}{d}(f_p^L - f_p)^2 \tag{9}$$

where f_p^L is the pose feature of the landmark module, f_p is the pose feature of the pose discriminator, and d is the dimension of the pose feature.

D. Two-Stage Optimization for Landmark-based Adversarial Network

To train the adversarial network, we propose two-stage optimization to update the parameters. For first stage optimization, we utilize the pose loss in Equation (9) to optimize the pose discriminator. After we optimize the pose discriminator, the pose discriminator can extract pose-related information from the feature generator. Next, we optimize the feature generator with the total loss in Equation (6) for second stage optimization. The minus pose loss tern in the total loss is to make the identity feature to be pose-invariant. Since the pose discriminator cannot extract pose-related information from the feature generator, the pose loss would be high. Therefore, the total loss would decrease. That is to say, the feature generator successfully learns a pose-invariant feature.

III. Experimental Results

A. Implement Details

The proposed methods are implemented in PyTorch 1.10 [12] with python 3.7 interface and CUDA 11.3. The CPU is Intel® Core™ i9-10900K CPU @ 3.70GHz, the main memory is 128GB,and the GPU is NVIDIA GeForce®RTX 3080ti. All the RGBD images are cropped around the face to include the whole face. The input of RGBD face recognition model is a pair of cropped faces, containing a $112 \times 112 \times 3$ RGB image and $112 \times 112 \times 1$ depth map. The input of landmark module is $120 \times 120 \times 14$ landmark heatmaps. The hyper parameter α in Equation (4) is set to 0.1. The hyper parameter $margin$ in Equation (8) is set to 0.4. We use AdamW [13] as our optimizer. The batch size and training epoch are set to 256 and 100 respectively when training the pre-trained dual path RGB-D face recognition model. The batch size and training epoch are set to 64 and 200 respectively when training the landmark-based adversarial network. The learning rate is set to 0.001 with a weight decay 0.0005.

B. Training & Testing Datasets

The tables below are the summary of datasets we used for training models.

TABLE I
THE DATASETS USED FOR TRAINING IN PROPOSED METHODS.

Data Type	Dataset Name	Number of People	Number of Images	Different Views
RGB	MS1MV2	85,742	5,800,000	No
RGB-D	KinectFaceDB	42	756	Yes (0° and 90°)
	$\text{RealSense}_{\text{train}}$	14	35000	Yes (0° ∼90°)
	$\text{LiDAR}_{\text{train}}$	14	49000	Yes (0° ∼90°)

TABLE II
THE DATASETS USED FOR TESTING IN PROPOSED METHODS.

Data Type	Dataset Name	Number of People	Number of Images	Different Views
RGB	LFW	5,749	13,233	No
	CPLFW	5,749	13,233	Yes (0°∼ 90°)
	CFP-FP	500	7,000	Yes (0° ∼90°)
RGB-D	$\text{RealSense}_{\text{test}}$	14	40,320	Yes (0° ∼90°)
	$\text{LiDAR}_{\text{test}}$	11	31,680	Yes (0° ∼90°)
	KinectFaceDB	10	180	Yes (0° and 90°)

C. Each Module in Landmark-based Adversarial Network for RGB-D Pose Invariant Face Recognition

We train and test our method on KinectFaceDB. We use 42 of 52 individuals of KinectFaceDB as our training set and the residual individuals as our testing set. Table III shows the results containing different components including Pose Discriminator, Landmark Module, and Triplet Loss. In this experiment, we add Pose Discrminator with two-stage optimization for training. We do the experiment under two settings, with pre-trained on MS1MV2 or not. The model pre-trained on MS1MV2 can achieve higher recognition performance because the model has more generalization capacity and prevent overfitting on the training dataset. In our initial design, there is no Pose Discriminator. We add Pose Discriminator which discriminates the pose information from feature maps of Feature Generator. But the task of predicting pose angles directly is hard for Pose Discriminator, Pose Discriminator can not have enough ability to discriminate pose-related information.

Therefore, we add Landmark module to assist in training Pose Discriminator. Since we use the penultimate feature to do identity, we not only use Cross Entropy loss on the last classification feature but add Triplet loss on the identity feature.

TABLE III
ABLATION STUDY ON THE PROPOSED MODEL ON EURECOM KINECT FACE DATABASE. POSE DIS.: POSE DISCRIMINATOR.

Pretrained on MS1MV2	Pose Dis.	Landmark Module	Triplet Loss	KinectFaceDB			
				Left Profile	Front	Right Profile	Average
				35.00%	83.08%	35.00%	71.76%
	✓			35.00%	87.69%	40.00%	75.88%
	✓	✓		40.00%	88.46%	40.00%	77.05%
	✓	✓	✓	50.00%	90.00%	50.00%	80.58%
✓				90.00%	97.69%	95.00%	96.47%
✓	✓			95.00%	97.69%	95.00%	97.05%
✓	✓	✓		95.00%	98.46%	95.00%	97.64%
✓	✓	✓	✓	95.00%	100.0%	100.0%	99.41%

D. Effect of 2-Stage Optimization

We do experiments on 1-stage or 2-stage optimization. For 1-stage optimization, we optimize Pose Discriminator and Feature Generator at the same time. As shown in Table IV, we find that model with 2-stage optimization achieves recognition accuracy of 99.41%, which is 2.35% higher than that of with 1-stage optimization.

TABLE IV
EXPERIMENT OF EFFECT OF THE 2-STAGE OPTIMIZATION ON THE PROPOSED MODEL ON EURECOM KINECT FACE DATABASE. 2-STAGE OPT.: 2-STAGE OPTIMIZATION.

Pretrained on MS1MV2	Pose Dis.	Triplet Loss	Landmark Module	2-Stage Optim.	KinectFaceDB			
					Left Profile	Front	Right Profile	Average
✓	✓	✓	✓		91.66%	97.69%	95.00%	97.06%
✓	✓	✓	✓	✓	95.00%	100.0%	100.0%	99.41%

E. Experiment on $\mathrm{RealSense}_{\mathrm{test}}$ and $\mathrm{LiDAR}_{\mathrm{test}}$

We compare our proposed method with eCNN model [14]. For comparison, we use classification accuracy. We train our model with $\mathrm{RealSense}_{\mathrm{train}}$ and $\mathrm{LiDAR}_{\mathrm{train}}$, and test the model on $\mathrm{RealSense}_{\mathrm{test}}$ and $\mathrm{LiDAR}_{\mathrm{test}}$. As shown in Table V, we average the accuracy of frontal view, yaw angle variation, and pitch angle variation under different environments in Table V. We have more improvement under yaw angle variation and pitch angle variation on both datasets.

TABLE V
THE FACE CLASSIFICATION ACCURACY OF OUR PROPOSED METHOD COMPARES WITH ECNN MODEL [14] ON DIFFERENT POSE ANGLES, INCLUDING FRONTAL VIEW, YAW AXIS, AND PITCH AXIS. * MEANS WE PRE-TRAINED OUR MODEL ON MS1MV2 DATASET.

Head Pose	$\mathrm{RealSense}_{\mathrm{test}}$			$\mathrm{LiDAR}_{\mathrm{test}}$		
	eCNN	Proposed	Proposed*	eCNN	Proposed	Proposed*
Front	66.67%	71.58%	94.84%	76.87%	85.48%	97.50%
Yaw	60.32%	64.19%	92.19%	77.12%	82.69%	98.41%
Pitch	59.17%	63.05%	90.68%	75.00%	87.21%	98.73%
Average	62.06%	66.27%	92.57%	76.33%	85.13%	98.21%

F. Comparison with RGB-D face recognition methods

We compare the proposed RGB-D pose invariant face recognition model with the following state-of-the-art methods in Table VI. We achieve recognition accuracy of 99.41%, which is 1.31% higher than [16]. The total number of parameter is 32M, which is 6M lower than [16].

TABLE VI
COMPARISON WITH STATE-OF-THE-ART METHODS ON KINECTFACEDB.

Method	Training Data	Parameters (M)	KinectFaceDB Accuracy
Zhang [15]	Lock3DFace	6.6	96.3%
Zafar [16]	80% KinectFaceBD	38	98.1%
Teng [17]	50% KinectFaceBD	22	97.4%
Two-Level att. [18]	-	328.23	92.0%
Depth-guid att. [19]	-	33.2	93.1%
Proposed Method	80% KinectFaceBD	32	99.41%

TABLE VII
COMPARISON WITH STATE-OF-THE-ART METHODS ON EACH DATASET. MS1M IS MS1MV2 DATASET.

Method	Parameters (M)	LFW	CPLFW	CFP-FP
Deng [21]	17M	99.3%	-	87.11%
An [22]	27M	99.63%	86.95%	-
Co-mining [23]	1M	-	85.70%	-
ArcFace [1]	89M	99.79%	88.78%	-
CurricularFace [4]	248M	99.8%	93.13%	-
Huang [20]	23M	-	89.37%	98.4%
Meng [24]	65M	99.83%	92.87%	98.46%
CvT-13	16M	99.1%	88.5%	92.5%
Proposed Method		99.3%	89.8%	94.5%

G. Comparison with RGB face recognition methods

We use MS1MV2 dataset to train our method and use CvT-13 as our backbone. We train CvT-13 for face recognition task as our baseline for comparison by only using Cross Entropy loss shown in Equation (4). We achieve recognition accuracy of 89.8% on CPLFW, which is 0.43% better than [20]. Our number of model parameters is 16 M, which is 7M lower than [20]. Comparison with the methods also use adversarial learning, we achieve recognition accuracy of 94.5%, which is 7.39% higher than [21] and the number of our model parameter is 1M lower than [21].

IV. CONCLUSION

In this work,we proposed a dual-path RGB-D face recognition model and use our landmark-based adversarial network to improve the recognition accuracy. Our proposed landmark module could embed facial landmarks into pose features, and use the pose feature as a pseudo label for training the pose discriminator. The total number of parameter is 32M, and we evaluation our method on public dataset , KinectFaceDB, $\mathrm{RealSense}_{\mathrm{test}}$ and $\mathrm{LiDAR}_{\mathrm{test}}$. We achieve a recognition accuracy of 99.41% on KinectFaceDB, which is 1.31% higher than [16] with the reduction of 6M parameters. We achieve a classification accuracy of 92.57% on $Realsense_{test}$, which is 30.51% higher than [14] and achieve 98.21% on $LiDAR_{test}$, which is 21.88% higher than [14]. Although our methods need 32 M parameters usage which is 31.46M higher than [14], our proposed method performs well even in different light environments or different head pose angles. We also evaluate RGB face recognition on CPLFW and CPF-FP. On CPLFW, we achieve a recognition accuracy of 89.9%, which is 0.43% higher than [20] with the reduction of 7M parameters. On CFP-FP, we achieve a recognition accuracy of 94.5%, which is 7.39% higher than [21] with the reduction of 1M parameters.

REFERENCES

[1] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4690–4699.

[2] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song, "Sphereface: Deep hypersphere embedding for face recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 212–220.

[3] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu, "Cosface: Large margin cosine loss for deep face recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5265–5274.

[4] Yuge Huang, Yuhan Wang, Ying Tai, Xiaoming Liu, Pengcheng Shen, Shaoxin Li, Jilin Li, and Feiyue Huang, "Curricularface: adaptive curriculum learning loss for deep face recognition," in *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 5901–5910.

[5] Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller, "Labeled faces in the wild: A database forstudying face recognition in unconstrained environments," in *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*, 2008.

[6] Tianyue Zheng and Weihong Deng, "Cross-pose lfw: A database for studying cross-pose face recognition in unconstrained environments," *Beijing University of Posts and Telecommunications, Tech. Rep*, vol. 5, pp. 7, 2018.

[7] Soumyadip Sengupta, Jun-Cheng Chen, Carlos Castillo, Vishal M Patel, Rama Chellappa, and David W Jacobs, "Frontal to profile face verification in the wild," in *2016 IEEE winter conference on applications of computer vision (WACV)*. IEEE, 2016, pp. 1–9.

[8] Ralph Gross, Iain Matthews, Jeffrey Cohn, Takeo Kanade, and Simon Baker, "Multi-pie," *Image and vision computing*, vol. 28, no. 5, pp. 807–813, 2010.

[9] Guoli Wang, Jiaqi Ma, Qian Zhang, Jiwen Lu, and Jie Zhou, "Pseudo facial generation with extreme poses for face recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 1994–2003.

[10] Yeeleng S. Vang, Yingxin Cao, Peter D. Chang, Daniel S. Chow, Alexander U. Brandt, Friedemann Paul, Michael Scheel, and Xiaohui Xie, "Synergynet: A fusion framework for multiple sclerosis brain mri segmentation with local refinement," in *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, 2020, pp. 131–135.

[11] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang, "Cvt: Introducing convolutions to vision transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 22–31.

[12] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al., "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, 2019.

[13] Ilya Loshchilov and Frank Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[14] Ching-Te Chiu, Yu-Chun Ding, Wei-Chen Lin, Wei-Jyun Chen, Shu-Yun Wu, Chao-Tsung Huang, Chun-Yeh Lin, Chia-Yu Chang, Meng-Jui Lee, Shimazu Tatsunori, Tsung Chen, Fan-Yi Lin, and Yuan-Hao Huang, "Chaos lidar based rgb-d face classification system with embedded cnn accelerator on fpgas," *IEEE Transactions on Circuits and Systems I: Regular Papers*, pp. 1–13, 2022.

[15] Hao Zhang, Hu Han, Jiyun Cui, Shiguang Shan, and Xilin Chen, "Rgb-d face recognition via deep complementary and common feature learning," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 2018, pp. 8–15.

[16] Umara Zafar, Mubeen Ghafoor, Tehseen Zia, Ghufran Ahmed, Ahsan Latif, Kaleem Razzaq Malik, and Abdullahi Mohamud Sharif, "Face recognition with bayesian convolutional networks for robust surveillance systems," *EURASIP Journal on Image and Video Processing*, vol. 2019, no. 1, pp. 1–10, 2019.

[17] Wenbin Teng and Chongyang Bai, "Unimodal face classification with multimodal training," in *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*. IEEE, 2021, pp. 1–5.

[18] Hardik Uppal, Alireza Sepas-Moghaddam, Michael Greenspan, and Ali Etemad, "Two-level attention-based fusion learning for rgb-d face recognition," in *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 2021, pp. 10120–10127.

[19] Hardik Uppal, Alireza Sepas-Moghaddam, Michael Greenspan, and Ali Etemad, "Depth as attention for face representation learning," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 2461–2476, 2021.

[20] Junyang Huang and Changxing Ding, "Attention-guided progressive mapping for profile face recognition," in *2021 IEEE International Joint Conference on Biometrics (IJCB)*. IEEE, 2021, pp. 1–8.

[21] Jiankang Deng, Shiyang Cheng, Niannan Xue, Yuxiang Zhou, and Stefanos Zafeiriou, "Uv-gan: Adversarial facial uv map completion for pose-invariant face recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7093–7102.

[22] Zhanfu An, Weihong Deng, Jiani Hu, Yaoyao Zhong, and Yuying Zhao, "Apa: Adaptive pose alignment for pose-invariant face recognition," *IEEE Access*, vol. 7, pp. 14653–14670, 2019.

[23] Xiaobo Wang, Shuo Wang, Jun Wang, Hailin Shi, and Tao Mei, "Co-mining: Deep face recognition with noisy labels," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9358–9367.

[24] Qiang Meng, Shichao Zhao, Zhida Huang, and Feng Zhou, "Magface: A universal representation for face recognition and quality assessment," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 14225–14234.