3D Landmark-based Face Detection and Recognition System for Large Poses

Ching-Tung Tang, Ching-Te Chiu and Wei-Jyun Chen
joycetang@gapp.nthu.edu.tw, ctchiu@cs.nthu.edu.tw, jyunchen@gapp.nthu.edu.tw
Department of Computer Science, National Tsing Hua University, Hsinchu, Taiwan

***Abstract*—Most face recognition algorithms achieve great performances in small poses, but they are unable to extract intact features for large-pose faces. In order to improve large-pose face recognition, we propose a 3D landmark-based face recognition system. We first extract RGB features from a face recognition model. Then we predict 3D landmarks and facial pose degrees via a projected 3DMM vector in the 3D landmark model. To test images, we compute the distance of two RGB features and rotate two 3D 68-point landmarks to the frontal view. If both of two features are smaller than a threshold, we purport these two images are from the same person. Compared to traditional face recognition CNN methods, the proposed method not only consider RGB features but 3D estimated landmarks. With this information, we can achieve higher performance.**

We conduct experiments on large-pose face datasets, CPLFW, CFPFP and IJB-B. The results outperform the state-of-the-art methods. We achieve recognition rate of 94.15% on CPLFW, which is 1.02% higher than the Curricular Face [1], and 98.97% on CFPFP, which is 0.6% higher than Curricular Face [1]. Compared to Curricular Face [1], our model reduces 13M parameters usage and achieves 94.9% on IJB-B.

I. Introduction

The applications of face recognition system are broad. Besides frontal view face recognition system, large-pose face recognition system becomes a trend in real-world applications. For example, immigration checkpoints and criminal identification system. These systems require reliable performance on large-pose face recognition. Due to plenty of face databases, the performance of frontal-view face recognition has achieved great success. However, the accuracy drops when the face poses increase. This problem will affect the stability of face recognition system. The problem of high pose face recognition is the two features from the frontal image and the other pose image are disparate. Current face recognition methods mostly consider the design of loss functions to grapple with the problem of large-scale facial images. For example, the renowned loss functions of SphereFace [2], ArcFace [3] and CurricularFace [1]. But we still notice their improvements are restricted from the experimental results. The main issue is these methods just examine the learned features from 2D images.

We notice that utilize accurate 3D information can improve performance of computer vision applications. Therefore, we surmise the pivotal approach to ameliorate large-scale face recognition is to project 2D facial images to 3D space. 3D mesh estimation from a 2D RGB facial image is needed. We can take an RGB-only image as input then get a 3D mesh of the target face then estimate corresponding 3D landmark. The overall scenario is shown in Fig. 1. [4], [5], [6] proposed CNN-based methods to predict 3D mesh from RGB image. However, these methods lose the details of facial features when estimating high-pose facial images. Therefore, the problem of high-pose face recognition remains an unsolved problem.

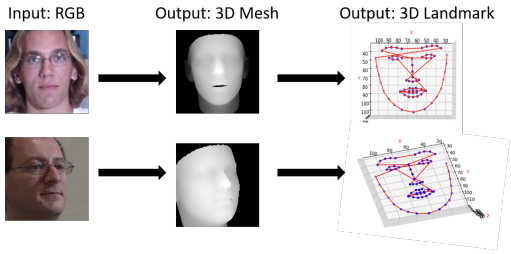

Fig. 1: 3D facial landmark estimation from a 3D face mesh.

In our work, we focus on solving the deficient performance of face recognition in large poses. CNN-based methods for RGB feature extraction are a good way to extract features, but it is not powerful enough when testing on high-pose images. As a result, we propose an innovative evaluation protocol for face recognition. After traditional CNN-based RGB feature extraction, we add another CNN-based model for facial landmark estimation. With the proposed method, we could obtain more information from single face image compared to traditional feature extraction.

II. Proposed Methods

The proposed system consists of three models, 2-stage facial feature and face detection model, RGB face recognition model and 3D facial landmark and reconstruction model.

A. Overall Scenario of Proposed Method

Fig. 2. shows the overview of the proposed architecture. For the first face detection model, it inputs a RGB frame then outputs the coordinates of faces. After cropping faces, Model 2 extracts RGB feature of faces. Meanwhile, we input the cropped faces into Model 3. We can obtain head pose rotation degree and estimated 3D landmark from Model3. Then we acquire two features of the cropped faces, we compare these two features by the proposed 3D landmark-based face recognition evaluation protocol to distinguish whether the two cropped faces are from the same person or not.

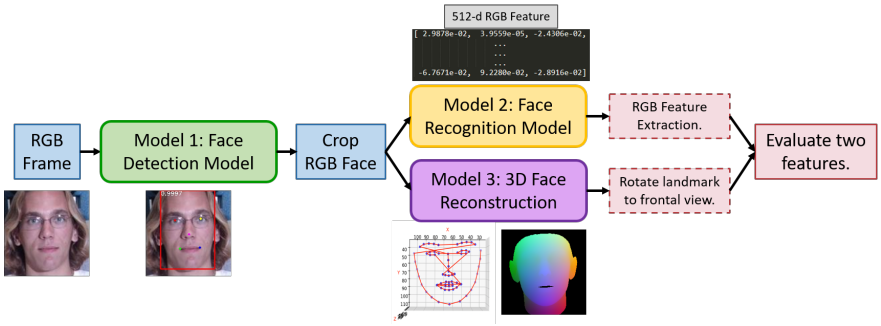

Fig. 2: Overview of the proposed architecture.

B. Network Architecture

1) 2-Stage Face Detection Model: For the proposed face detection model, we separate the scenario into two parts. The first part is 5-point facial landmark extraction, and the second part is bounding boxes detection. The purpose of stage-1 5-point facial landmark extraction is to predict the landmark position of facial features in the faces. When training the proposed model, we use RGB frames as input and 2D 5-point landmark as ground truth. We use AFLW [11] as training dataset. For the learning framework, we use Mobile Net V1. Since AFLW [11] contains 25,993 faces with ground truth annotation of facial landmarks, our model can be trained well to predict accurate facial landmark from unseen images with the aid of the proposed loss function.

Loss Function for Stage-1 Model

$$\arg\min L_{s1}(W) = \sqrt{\frac{1}{n} \times \sum_{i=1}^{n} (y_p - y_g)^2} \tag{1}$$

n is the total training images in a batch, y_p is the prediction of 5-point facial landmark, y_g is the ground truth of 5-point facial landmark of input image and W is all the weight in the architecture.

With this 5-point facial landmark extraction model, we can add 5-point facial landmark as additional ground truth to the existing large face datasets for face detection. As shown in Fig. 3, stage-1 model is trained to generate 5-point facial landmark, then stage-2 can use the additional ground truth to be trained to generate 5-point facial landmark and face bounding boxes.

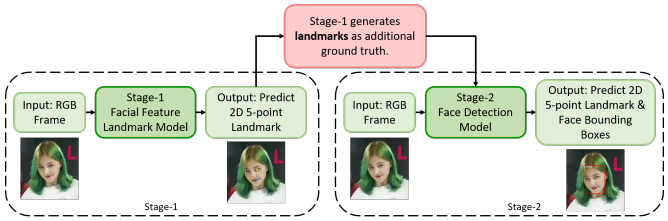

Fig. 3: The dataflow of 2-stage face detection model.

The purpose of stage-2 face detection model is to detect the faces in the input image. We use WIDER FACE Dataset [7] as training dataset. For training, we input RGB frames with the ground truth of the corresponding coordinates of faces from the original dataset. Moreover, with the aid of stage-1 model, we generate 5-point facial landmark of WIDER FACE dataset as an additional ground truth.

Loss Function for Stage-2 Model

$$\begin{aligned}&\arg\min L_{s2}(W)\\ &= \lambda_1 \sqrt{\frac{1}{n} \times \sum_{i=1}^{n} (y_p - y_g)^2} + \lambda_2 \sqrt{\frac{1}{n} \times \sum_{i=1}^{n} (z_p - z_g)^2}\end{aligned} \tag{2}$$

In the L_{s2} loss function, λ_1, λ_2 are weighting constants, n is the total training images in a batch, y_p is the prediction of 5-point facial landmark, y_g is the ground truth of 5-point facial landmarks of input image, z_p is the prediction of the coordinates of bounding boxes, z_g is the ground truth of bounding boxes and W is all the weight in the architecture.

For the output of this model, we examine two conditions to ensure performance. First, we evaluate the coordinates of generated face boxes with ground truth. Second, we examine whether there is a 5-point facial landmark in face boxes. With the second requirement, we can improve the performance of face detection for face in large poses, small scale or dark illumination conditions.

C. RGB Face Recognition Model

The proposed Model 2 is RGB face recognition model, which is used to extract RGB feature from the cropped face from Model 1. The model inputs a cropped facial frame and outputs the n-d feature vector. The output feature vector is an array with many floating numbers in it. We use ResNet 50 as model architectures. Inspired by Arcface loss function [3], we utilize the most widely used loss function for face recognition in the proposed face recognition system. Besides following the Arcface loss function, we survey the best length of dimension for the output vector from model. We do ablation study to test four different lengths of output vector and test on LFW dataset. Then we evaluate accuracy and comparison computational time for each kind of output vectors. Ultimately, we decide to use 512-d as output vector length, which depends on both the performance of accuracy and comparison computational time.

Loss Function for Feature Extraction Model

$$\arg\min L_{Arc}(W) = -\frac{1}{N}\sum_{i=1}^{N} log\frac{e^{s(cos\theta_{y_i}+m)}}{e^{s(cos\theta_{y_i}+m)}+\sum_{j=1,j\neq i}^{N} e^{s(cos\theta_{y_j})}} \quad (3)$$

In the L_{Arc} loss function, N is batch size, s is feature scale, $cos\theta_{y_i}$ means to get the angle between the feature x_i and the ground truth feature, m is the angular margin penalty on the target (ground truth) angle θ_{y_i}, j is j-th class number and W is all the weight in the architecture.

D. 3D Facial Landmark and Reconstruction Model

The purpose of 3D facial landmark and reconstruction model is to estimate the 3D information from 2D RGB images. With the aid of estimated 3D face landmark and face mesh, we can achieve better performance on high-pose face recognition. Fig. 4 shows the proposed network structure for 3D facial landmark prediction.

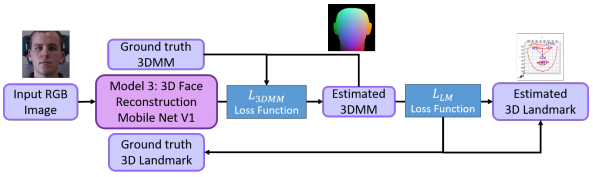

Fig. 4: The dataflow of the proposed 3D face reconstruction architecture.

In this model, we use the cropped RGB face image from proposed 2-stage face detection model as input then get two outputs from the model. The first output is a 3D Morphable Model (3DMM) vector of the corresponding input face. And based on this 3DMM vector we get the estimated 68-point 3D facial landmarks. For fair comparison reason, since most of face landmark estimation researches use 68 points, we decided to use 68-point landmarks, which would not lead to large increase of model parameters or run time. We design the output 3DMM vector as a 233-d vector. The first 199-d is to denote the facial shape parameters. And the next 29-d is for the facial expression parameters. The last 5-d parameters are the estimated rotation degree. The input RGB face image can be projected into a 3D face mesh with the proposed 233-d vector.

Details of 3D Face Mesh

$$S = \bar{S} + A_{id}\alpha_{id} + A_{exp}\alpha_{exp} \quad (4)$$

The above equation is the function to generate a 3D mesh from 2D facial images. S is the generated 3D face reconstruction, $\bar{S}$ is the average 3DMM vector of face. A_{id} is the principal axes trained on the 3D face scans with neutral expression from BFM [8], A_{exp} is the principal axes trained on the offsets between expression scans and neutral scans from FaceWarehouse [9], α_{id} is the output of 199-d shape parameters for facial shape, α_{exp} is the output of 29-d shape parameters for facial expression.

During training stage, the proposed model is trained to fit the generated 3D face mesh with the ground truth 3D face mesh. Therefore, we obtain the rotation degree of the generated 3D face mesh. The last 5-d parameters store the pitch, row, yaw rotation degree of the input face and the scale factor and displacement.

Loss Function for 3DMM Estimation

$$\arg\min L_{3DMM}(W) = \|\triangle M - (M^g - M^0)\|^2 \quad (5)$$

$\triangle M$ is the prediction of 233-d 3DMM vector, M^g is the ground truth of 233-d 3DMM vector, M^0 is the initial 233-d 3DMM vector.

Loss Function for 68-point 3D Landmark Estimation

$$\arg\min L_{LM}(W) = \sqrt{\frac{1}{N}x\sum_{i=1}^{N}\|a^p - a^g\|^2} \quad (6)$$

N is batch size, a^p is the prediction of 68-point landmark, a^g is the ground truth of 68-point landmark and W is all the weight in the architecture. We use Mobile Net V1 as training framework. With the two proposed loss functions, this model can output accurate 3D face reconstruction and 3D facial landmark from RGB images.

E. Rotation Matrix for 3D Facial Landmark

After extracting 3D facial landmark from the proposed model, we rotate 3D facial landmark from profile images to frontal view. There are three angles in Euler angles system. In face and head system, we denote three directions as pitch, roll and yaw angles.

We rotate 3D landmark from Model 3 to frontal view based on the estimated angles on each X, Y, Z-axis. Pitch angle (α) is the head rotation around the horizontal X-axis compare to frontal face plane. Yaw angle (β) is the head rotation around the vertical Y-axis. Roll angle (γ) is the rotation around the Z-axis perpendicular to frontal face plane.

Rotation Matrix for 3D Landmarks on All X, Y, Z-Axis

$$RotationMatrix_{(\alpha,\beta,\gamma)} = R_{Pitch(\alpha)} \cdot R_{Yaw(\beta)} \cdot R_{Roll(\gamma)} \quad (7)$$

$$= \begin{bmatrix} cos\alpha cos\beta & cos\alpha sin\beta - sin\alpha cos\gamma & cos\alpha sin\beta cos\gamma + sin\alpha sin\gamma \\ sin\alpha cos\beta & sin\alpha sin\beta sin\gamma + cos\alpha cos\gamma & sin\alpha sin\beta cos\gamma - cos\alpha cos\gamma \\ -sin\beta & cos\beta sin\gamma & cos\beta cos\gamma \end{bmatrix} \quad (8)$$

Integrated Function to Rotate Angled Image to Frontal View

$$[x, y, z_{(FrontalImage)}] = \quad (9)$$

$$[RotationMatrix_{\alpha,\beta,\gamma}]^{-1} * [x, y, z_{(AngledImage)}] \quad (10)$$

After rotating angled facial landmark into frontal view facial landmark, we need to evaluate difference between two identities. The way we evaluate two landmarks is that we compute the difference of x, y, z value of each 3D landmark from two identities.

Equation for Evaluating Difference between Two 3D Landmarks

$$Different = \frac{\sqrt{(a_i - b_i)^2}}{204} \quad (11)$$

a denotes the 3D facial landmark from the first identity, b is the second identity and i is a length of 204 array. For the first

68 points, they are the coordinates of X in 3D space, 69 to 136 points are the coordinates of Y, 137 to 204 points are the coordinates of Z.

F. 3D Landmark-based Face Recognition Evaluation Protocol

We evaluate the two RGB features from the input RGB testing pair as general methods. Our trait is we examine the distance of two 3D landmarks to check whether these two features are from the same person or not. For the RGB feature comparison part, if the distance of two features is larger than zero, we consider these two images are from two people. For the 3D landmark comparison part, if the distance of two landmark is larger than twenty, we consider these two images are from two people. Otherwise, these two images are from the same person.

III. Experimental Results

A. Implement Details

The proposed face recognition architecture is implemented under the open source Pytorch [10] deep learning framework. We train and evaluate with NVIDIA GTX 1080 GPU with 8GB memory. The CPU is Intel Core i7-7800X 3.5GHz, and the main memory is 32GB DDR4 RAM. In face detection model, the batch size is 32, epoch is 250, learning rate is 0.001. In face recognition model, the batch size is 32, epoch is 30 and learning rate is 0.1. In 3D landmark estimation model, the batch size is 32, epoch is 40 and learning rate is 0.001.

B. Training & Testing Datasets

For training models, in stage-1 face feature extraction model, we use AFLW [11] to train model, which contains 25,993 images with 21-point landmark with faces in different poses. In stage-2 face detection model, we use WIDER FACE [7]. It consists of 32,203 images with 393,703 faces labeled with face bounding boxes in different poses. In RGB face recognition model, we use MS1MV2 [3], which contains 85,742 people and 5,800,000 images with mainly frontal view images. And in 3D landmark estimation model, we use 300W-LP-3D [12]. It includes 61,255 images with faces labeled 3D 68-point facial landmarks in different poses.

For testing models, we test in three random pose face datasets, which are widely used to evaluate face recognition performance. All of them contain face images in 0°to 90°. First, Cross-Pose LFW (CPLFW) [13] dataset, it consists of 5,749 people with 13,133 images. Secondly, CFP FP [14] dataset, it contains 500 people with 7,000 images. Lastly, IJB-B [15] dataset, which has 1,845 people with 61,255 images.

C. Experimental Results on the Overall Face Recognition System

We test our face recognition system, which consists of RGB feature comparison and 3D landmark distance comparison, on CPLFW, CFP FP and IJB-B dataset. There are total 6,000 testing pairs in CPLFW dataset, 7,000 testing pairs in CFP FP dataset and 10,273 testing pairs in IJB-B dataset.

Table I shows the ablation study on the proposed 3D landmark-based face recognition evaluation protocol. We first test the result of face recognition model only. Then we test combination of face recognition model and 3D landmark estimation model without rotation. Finally, we test combination of face recognition model and 3D landmark estimation model with rotating to frontal view.

In CPLFW dataset, we notice that overall face detection and recognition system with rotation can improve 1.97% accuracy. But if we don't use rotation function, the performance only improves 0.02%. Therefore, we prove that the proposed 3D landmark estimation combined with rotation function can improve the performance on face recognition for faces in large poses. In addition, in CFP FP dataset, we notice that without rotation, the accuracy only improves 0.01%. Then we add rotation matrix into face recognition system, the accuracy improves 0.5%. Finally, with the aid of face detection model and 3d landmark estimation with rotation matrix, the final accuracy improves 0.87% compared to RGB feature only. For IJB-B dataset, we notice that after adding rotation matrix into face recognition system, the accuracy improves 0.33%. Finally, with the aid of face detection model and 3d landmark estimation with rotation matrix, the final accuracy improves 2.43% compared to RGB feature only.

D. Comparison with the State-of-the-Art Methods

We compared our method with famous state-of-the-art methods in Table II. The results of the state-of-the-art methods are from the corresponding papers. Our method achieves 1.02% accuracy higher than the best baseline on CPLFW dataset. Compared to the best baseline, we obtain 0.6% higher accuracy on CFP FP dataset. Since this dataset not only contains high-pose facial images but blur images, our method only improve 0.1% accuracy. However, it's worth mentioning that we achieve the same result with 13M parameter reduction.

IV. Conclusion

In this thesis, we proposed a method called 3D landmark-based face detection and recognition system to improve the performance on large-pose face recognition. We combine three CNN models together for face detection, RGB feature extraction and 3D landmark and pose estimation. Besides, we also introduce an innovative 3D landmark-based face recognition evaluation protocol. Compared to the state-of-the-art methods, we achieve 94.15% accuracy on CPLFW, which is 1.02% higher than other methods. And we achieve 98.97% on CFP FP dataset, which is 0.6% better than other works. And for IJB-B dataset, we improve 0.1% and reduce 13M parameters at the same time.

References

[1] Y. Huang, Y. Wang, Y. Tai, X. Liu, P. Shen, S. Li, J. Li, and F. 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.

[2] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "Sphereface: Deep hypersphere embedding for face recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 212–220.

TABLE I: The ablation study results on the proposed 3D landmark-based face recognition evaluation protocol.

Method		Face Detection	RGB Feature	3D Landmark	Rotation Matrix	CPLFW	CFP FP	IJB-B
Without Model 3	Model 2 only		✓			92.18%	98.1%	92.47%
With Model 3	Model 1 + 2 + Model 3 w/o Rotation	✓	✓	✓		92.2%	98.15%	92.5%
	Model 2 + 3 with Rotation		✓	✓	✓	92.7%	98.5%	92.8%
	Model 1 + 2 + 3 with Rotation	✓	✓	✓	✓	94.15%	98.98%	94.9%

TABLE II: Comparison result with the state-of-the-art methods on CPLFW, CFP FP and IJB-B datasets.

Method	Number of Parameters (M)	CPLFW	CFP FP	IJB-B
SphereFace (CVPR'17) [2]	64M	81.4%	-	-
Deng et al. (CVPR'18) [16]	25M	-	94.05%	-
Ranjan Face (CVPR'18) [17]	44M	-	-	90.3%
VGGFace2 (FG'18) [18]	143M	84.00%	-	-
ArcFace (CVPR'19) [3]	44M	92.08%	95.56%	89.8%
Shrink Tea Net (CVPR'19) [19]	48M	-	95.14%	91.5%
Minimum Margin (CVPR'20) [20]	54M	-	-	92.1%
MV-Arc-Softmax (AAAI'20) [21]	44M	92.83%	98.28%	94.8%
Curricular Face (CVPR'20) [1]	44M	93.13%	98.37%	94.8%
Proposed Method	31M	**94.15%**	**98.98%**	**94.9%**

[3] J. Deng, J. Guo, X. Niannan, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *CVPR*, 2019.

[4] Y. Wu, T. Hassner, K. Kim, G. Medioni, and P. Natarajan, "Facial landmark detection with tweaked convolutional neural networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 12, pp. 3067–3074, 2017.

[5] A. Jourabloo and X. Liu, "Large-pose face alignment via cnn-based dense 3d model fitting," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4188–4196.

[6] G. Trigeorgis, P. Snape, M. A. Nicolaou, E. Antonakos, and S. Zafeiriou, "Mnemonic descent method: A recurrent process applied for end-to-end face alignment," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4177–4187.

[7] S. Yang, P. Luo, C.-C. Loy, and X. Tang, "Wider face: A face detection benchmark," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 5525–5533.

[8] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter, "A 3d face model for pose and illumination invariant face recognition," in *2009 Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance*. Ieee, 2009, pp. 296–301.

[9] C. Cao, Y. Weng, S. Zhou, Y. Tong, and K. Zhou, "Facewarehouse: A 3d facial expression database for visual computing," *IEEE Transactions on Visualization and Computer Graphics*, vol. 20, no. 3, pp. 413–425, 2013.

[10] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in neural information processing systems*, 2019, pp. 8026–8037.

[11] M. Koestinger, P. Wohlhart, P. M. Roth, and H. Bischof, "Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization," in *2011 IEEE international conference on computer vision workshops (ICCV workshops)*. IEEE, 2011, pp. 2144–2151.

[12] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li, "Face alignment across large poses: A 3d solution," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 146–155.

[13] T. Zheng and W. 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, 2018.

[14] S. Sengupta, J.-C. Chen, C. Castillo, V. M. Patel, R. Chellappa, and D. 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.

[15] C. Whitelam, E. Taborsky, A. Blanton, B. Maze, J. Adams, T. Miller, N. Kalka, A. K. Jain, J. A. Duncan, K. Allen, J. Cheney, and P. Grother, "Iarpa janus benchmark-b face dataset," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, July 2017.

[16] J. Deng, S. Cheng, N. Xue, Y. Zhou, and S. 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.

[17] R. Ranjan, A. Bansal, J. Zheng, H. Xu, J. Gleason, B. Lu, A. Nanduri, J.-C. Chen, C. D. Castillo, and R. Chellappa, "A fast and accurate system for face detection, identification, and verification," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 1, no. 2, pp. 82–96, 2019.

[18] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "Vggface2: A dataset for recognising faces across pose and age," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 2018, pp. 67–74.

[19] C. N. Duong, K. Luu, K. G. Quach, and N. Le, "Shrinkteanet: Million-scale lightweight face recognition via shrinking teacher-student networks," *arXiv preprint arXiv:1905.10620*, 2019.

[20] X. Wei, H. Wang, B. Scotney, and H. Wan, "Minimum margin loss for deep face recognition," *Pattern Recognition*, vol. 97, p. 107012, 2020.

[21] X. Wang, S. Zhang, S. Wang, T. Fu, H. Shi, and T. Mei, "Mis-classified vector guided softmax loss for face recognition," *arXiv preprint arXiv:1912.00833*, 2019.