FACE RECOGNITION IN THE PRESENCE OF VARIANCE IN POSE, EXPRESSION, AND OCCLUSIONS

A Dissertation Presented to

the Faculty of the Department of Computer Science

University of Houston

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

By

Xiang Xu

May 2019

FACE RECOGNITION IN THE PRESENCE OF VARIANCE IN POSE, EXPRESSION, AND OCCLUSIONS

Xiang Xu

APPROVED:

Ioannis A. Kakadiaris, Chairman
Dept. of Computer Science

Aron Laszka
Dept. of Computer Science

Omprakash Gnawali
Dept. of Computer Science

Baris Coskun
Amazon

Dean, College of Natural Sciences and Mathematics

ACKNOWLEDGMENTS

Many people have helped me during my Ph.D. program. I would like to express my appreciation to my advisor, my dissertation committee members, my colleagues, my family, and my friends.

First, I would like to thank my advisor, Prof. Ioannis A. Kakadiaris, who offered me the great opportunity of pursuing my Ph.D. in the Computational Biomedicine Laboratory (CBL). During my five years in Houston, he has served not only as my academic advisor, but also as a longtime mentor and friend. He has always provided me with insights and suggestions about the research, work, and life. I have enjoyed every discussion with him. While working with him, beyond the research expertise, I learned to take the initiative, to manage my time, work hard, improve my communications with people, and devote the time to my family members.

Besides my advisor, I would like to thank the rest of my dissertation committee members, namely Prof. Aron Laszka, Prof. Omprakash Gnawali, and Dr. Baris Coskun, for their great support and invaluable advice. Their broad knowledge of security and wireless networking, have been a great inspiration to me. I am genuinely proud to have them as committee members for this dissertation.

I want to thank all my previous and current colleagues at CBL. Specifically, I would like to thank Dr. Pengfei Dou for his support and knowledge to guide me in the face recognition area and Dr. Yuhang Wu for his referral to CBL. In addition, I would thank Mr. Nikolaos Sarafianos, Mr. Le Anh Vu Ha, Mr. Lei Shi, Mr. Christos Smailis, and Mr. Charles Livermore for the research discussion and their life support as friends. We had

a lot of cooperations on the research by exchanging the experiences and ideas, collecting data, and producing papers and proposals. I also appreciate the advice provided by Dr. Ioannis Konstantinidis and Prof. Michalis Vrigkas. It was a nice time working with them.

I appreciate the support and understanding of my parents, my fiancée and her family. I would like to thank them for always encouraging me to pursue my dreams and reach my goals. My parents really care about my health and happiness. They tried their best to support me and help to go over challenges happened in my life. My fiancée knew when to cheer me up, when to push me, and when to back off. I thank her for the love and appreciate the time we have together. I am looking forward to all the exciting adventures we will have in the rest of life. Moreover, I also want to give my thanks to my fiancée's family. They helped me to take care of a lot of issues and support me. Without them, I cannot focus on the research as I did.

Thanks also should be given to the institutes that provided funds or computational resources to support my Ph.D. work: the U.S. Department of Homeland Security (under Grant Award Number 2015-ST-061-BSH001 and 2017-ST-BTI-00001-02-01), the US Army Research Lab (under Grant Award Number W911NF-13-1-0127), as well as the University of Houston (Hugh Roy and Lillie Cranz Cullen Endowment Fund). I am also grateful for the support of the Core Facility for Advanced Computing and Data Science at the University of Houston for assisting me with the computations that were required for this dissertation.

FACE RECOGNITION IN THE PRESENCE OF VARIANCE IN POSE, EXPRESSION, AND OCCLUSIONS

An Abstract of a Dissertation

Presented to

the Faculty of the Department of Computer Science

University of Houston

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

By

Xiang Xu

May 2019

Abstract

Face recognition is a technology in which a computing device either classifies a human identity based on a facial image or verifies whether two images belong to the same subject. The recent advances have achieved remarkable performance when comparing images that are both frontal and non-occluded. However, significant challenges remain in the presence of variations in pose, expression, and occlusions. The goal of this dissertation is to achieve statistically significant improvement in the performance of face recognition systems using 2D images that depict individuals with facial expressions and accessories. Four contributions made in this dissertation can be summarized as follows: (i) a 3D-aided 2D face recognition system with additional evaluation package that is modular, easy to use, and easy to install was designed, implemented, and evaluated. This proposed system can work with the facial images that have variations in head pose as large as 90° and improved the face recognition performance by 9% on average when compared with FaceNet on UHDB31 dataset. (ii) two landmark detectors were developed and evaluated on 2D images that are fast and accurate; (iii) feature aggregation learning was proposed for face reconstruction from a single image, which achieved 16% and 10% improvement when compared with the current state-of-the-art on the BU-3DFE and JNU-3D datasets, respectively. and (iv) an occlusion-aware face recognition approach was proposed that improved the generalizability of the facial embedding generator and a graph neural network was designed in an unsupervised manner to adapt the knowledge learned in the image-based scenario to mixed-media set scenario.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Face recognition is a technology in which the computing device either classifies a human identity according to the face (face identification) or verifies whether two images belong to the same subject (face verification). A common face recognition system consists of two stages: enrollment and matching. In the enrollment stage, features are computed from a single facial image or a set of facial images to generate a template for each subject. In the matching stage, these templates are compared to obtain a distance or similarity for the identification or verification problem. Recent advances in face recognition led by the deployment of deep learning can be categorized in the following aspects:

(i) Large-scale image collections: in recent years, several large-scale training and evaluation datasets have been made publicly available for training deep neural networks.

In order to provide a large amount of data on which models can be learned discriminatively, large datasets such as MS1M [28], UMDFace [5], and VGG-Face2 [9] were proposed to cover the large distribution of pose and illuminations of facial images. Since the frontal face verification is almost a solved problem, several challenging datasets have been published to push the frontiers of unconstrained face recognition such as MegaFace [44] and IJB [46, 95] benchmarks. In addition, to evaluate the face recognition algorithm across the pose and illumination, some datasets were recently proposed [83, 48].

(ii) Advanced network architectures: Some advanced network architectures have been deployed to learn the feature extractor for the image classification problem such as VGG [60], ResNet [30], and DenseNet [34]. In particular, ResNet has shown the necessity for identity mapping when training a very deep model, which has been widely deployed in many applications [76, 29].

(iii) Discriminative learning approaches: DeepFace, proposed by Taigman *et al.* [78], introduced 3D alignment and proved that the face recognition performance using 3D alignment achieved superior performance compared to using 2D alignment on LFW dataset by 3%. They also reported on LFW that the deep learning model achieved similar performance with human efforts (97.25% and 97.5%, respectively). FaceNet, proposed by Schroff *et al.* [70], introduced a new loss function named triplet loss, which combines the metric learning concept to guide the training process of deep neural network (DNN). The FaceNet, trained on 200 million private labeled faces with triplet loss, achieved a performance of 99.63% verification accuracy on LFW

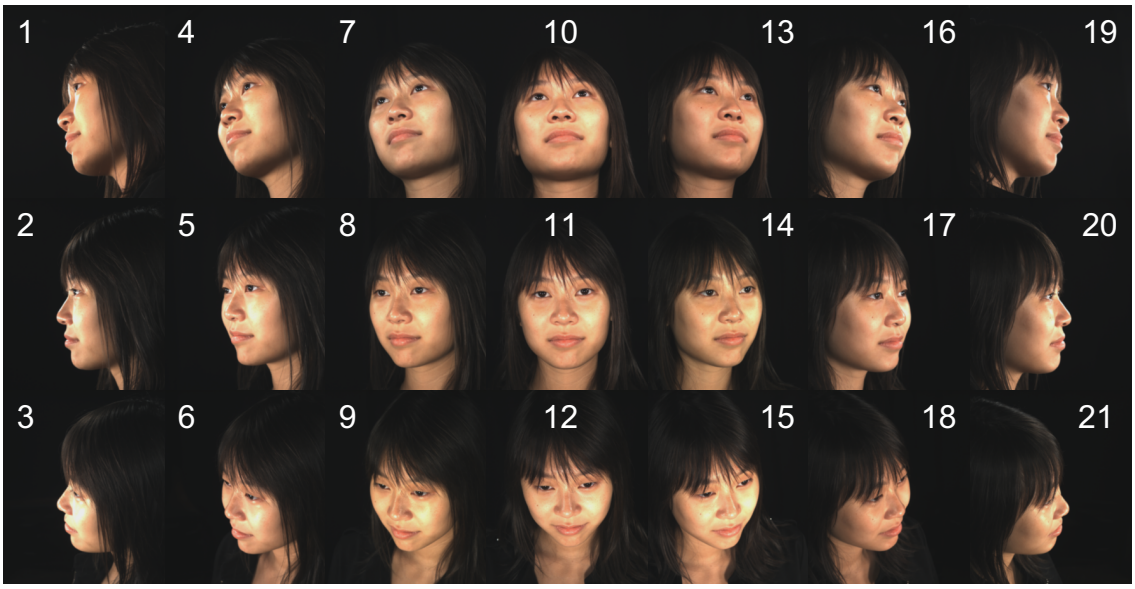
dataset. Various loss functions have also been explored to learn discriminative features for face recognition. Center loss [94] was proposed to learn instance centers with softmax cross-entropy loss. Based on computing the additive angular distance in sphere space with normalized weights, ArcFace [18] was developed and achieved 99.83% on the LFW and 98% identification accuracy on MegaFace benchmark.

1.2 Unsolved Challenges

Face recognition is still not a solved problem in real-world conditions. In unconstrained scenarios, especially using surveillance cameras, there is a plethora of images with large variations in head pose, expression, and occlusions, which limit the performance of face recognition systems:

(i) Pose variations: VGG-Face [60] was performed on UHDB31 dataset [48]. The frontal face was enrolled as the gallery and the other faces in different 20 poses were enrolled as the probes. The cosine distance was computed to evaluate the face identity rank-1 accuracy rate. As depicted in Figure 1.1, the face identification rate significantly drops when the face pose goes to 90°. This evaluation demonstrated the effect of pose variance to the current face feature extractor.

(ii) Expression variations: Facial expression influences the performance of face recognition systems [22, 14], especially in cases where only a single sample per person is available for enrollment.



(a)

14%	69%	94%	99%	95%	79%	19%
22%	88%	100%	-	100%	94%	27%
8%	2%	91%	95%	96%	52%	9%

(b)

Figure 1.1: (a) Depiction of 21 different poses in UHDB31 [48] dataset; (b) Face identification rank-1 accuracy using VGG-Face when the frontal face (pose 11) was used as gallery set and the faces in different poses (pose 1-10, 11-21) were used as probe sets.

(iii) Occlusion existence: Current face recognition systems tend to suffer from occlusions caused by facial accessories such as scarves and sun-glasses [59, 21]. This increases the risk because identity-related information might be excluded when the face is occluded. These occlusions by facial accessories lead to large variations in the final feature representations.

1.3 Limitations of Previous Work

Current facial recognition systems expect those variances of pose, expression, and occlusions can be learned from the existing large amount of data. Existing methods address some of the challenges mentioned in Section 1.2. However, most of them are limited in at least one of the following aspects.

(i) Imbalance problem: Current face recognition research focuses on learning general feature representations from a training set, which prevents such algorithms from learning the characteristics of facial images which might rarely occur in the dataset but actually be common in practice.

(ii) Generalization problem: Some methods use synthetic data to train and evaluate the model, which is not practical in real-life applications. The performance of a model trained from the synthetic data will decrease when applied to the in-the-wild images.

(iii) Scenario variance: Most datasets are designed for image-based face recognition and most face recognition systems require that the input is only a single image. However, in some scenarios, there are sets of images collected from the different sources that describe a single identity, which is not suitable for image-based face recognition systems.

1.4 Goal and Objectives

The goal of this dissertation is to achieve a statistically significant improvement to the performance of face recognition systems using 2D images that depict individuals with open mouth and facial accessories. To realize this goal, the following four objectives are proposed in this dissertation summarized as such:

(i) Design, implement, and evaluate an architecture of a 3D-aided face recognition system that is modular, in which the components are easy to use and easy to install.

(ii) Design, implement, and evaluate an algorithm for landmark detection on 2D images.

(iii) Design, implement, and evaluate an algorithm for 3D face reconstruction from a single image that depicts individuals with variations in pose and expression.

(iv) Design, implement, and evaluate an algorithm for 2D face recognition from a single image or set of images that depict individuals with variations in pose, expression, and occlusions.

1.5 Contributions

1.5.1 Objective 1

A 3D-aided pose-invariant 2D face recognition system named UR2D-E was designed and implemented, which has been demonstrated to be robust to pose variations as large as 90°. This system fills a gap by providing a 3D-aided 2D face recognition system that

has compatible results with 2D face recognition systems using deep learning techniques. UR2D-E consists of several independent modules: face detection, landmark detection, 3D model reconstruction, pose estimation, lifting texture, feature representation, and matching. It provides sufficient tools and interfaces to use different sub-modules designed in the system.

In addition, to efficiently evaluate the performance of the face recognition system, a light-weight, maintainable, scalable, generalizable, and extendable face recognition evaluation toolbox named FaRE was designed, implemented, and evaluated. FaRE supports both online and offline evaluation to provide feedback to algorithm development and accelerate biometrics-related research. It consists of a set of evaluation metric functions and provides various APIs for commonly used face recognition datasets including LFW, CFP, UHDB31, and IJB datasets, which can be easily extended to include other customized datasets.

1.5.2 Objective 2

An ensemble of random ferns (ERF) was proposed to detect landmarks on 2D facial images. As the first step, a classification method was used to obtain a facial shape as initialization for face alignment. Then, an ensemble of local random ferns was learned based on the correlation between the projected regression targets and a local pixel-difference matrix for each landmark, which was used to generate local binary features. Finally, the global projection matrix was learned based on concatenated binary features using ridge regression. Because the learning algorithm and test program were implemented using parallel

programming, the performance of the method was not only accurate but also efficient.

A joint learning framework was proposed that explores both global and local features for learning to estimate head pose and localize landmarks. First, a global network was used to detect the face region to obtain a rough estimate of pose and localize the primary seven landmarks. The most similar shape was selected for initialization from a reference shape pool constructed from the training samples according to the estimated head pose. Starting from the initial pose and shape, a local network was used to learn local CNN features and predict the shape and pose residuals. This framework was designed in a coarse-to-fine manner during which the global network estimates the rough shape and pose but the local network refines the shape in the cascade way.

1.5.3 Objective 3

The feature aggregation network (FR-FAN) was proposed to generate a 3D point cloud from a single image. Features from different layers were aggregated to predict shape and expression parameters of 3D face morphable model. The proposed method resulted in an increased reconstruction precision compared to the baseline with only half the number of the weights. The contributions are two-fold: (i) A feature aggregated network was designed with the principle of memory efficiency and fast speed in mind. (ii) The efficiency and robustness of the designed network were demonstrated with extensive experiments by observing that it improved the state-of-the-art method by 16% on BU-3DFE [108] and 10% on the JNU-3D dataset [47] in terms of reconstruction error.

1.5.4 Objective 4

An occlusion-aware face recognition approach (OREO) was proposed that improved the generalization ability of the facial embedding generator. An attention module was introduced that disentangles the features into global and local parts, resulting in more discriminative representations. In this way, the proposed approach successfully handles occlusions in face recognition without requiring additional supervision (*e.g.*, pose or occlusion labels) and achieves a relative improvement of 1.6% in terms of accuracy on the CFP dataset. An occlusion-balanced sampling strategy, along with a new loss function, was proposed to alleviate the large class imbalance that is prevalent due to non-occluded images. Our experimental results on the Celeb-A dataset [54] indicated that OREO achieved statistically significant improvements of more than 10% in terms of average degradation percentage.

An unsupervised graph-based template adaptation training framework was proposed that adapts the knowledge of the network learned from a still image to a mixed-media set without requiring any ground-truth label in the set domain. This is based on a paired teacher-student learning containing two identical networks. To improve the performance of set-based face recognition, a curriculum was designed for teacher and student networks in two steps: First, a graph-based template adapter was inserted and learned to generate a single feature/template that represented a set considering relationships of all features belonging to the same set. To effectively generate a template from a set, each set was formulated as a subgraph and all sets in the dataset constructed a large graph. Second, the teacher and student networks were updated in an unsupervised manner considering similarities in the set domain. To optimize the unsupervised framework end-to-end, the

teacher network was optimized using the supervision signals while the weights in the student network were updated using the supervision from the source domain and the teacher network's supervision for the target domain. There are two advantages of the proposed method in real-life applications: (i) it aggregated information from all samples within a subgraph to generate the more discriminative, robust, and compact templates by enlarging the similarities of the matched samples and decreasing the similarities of the non-matched samples; (ii) it did not require any modification of the backbone network since the graph-based template adapter is a plug-and-play module. This means that not only it preserved the performance for single image-based face recognition, but also achieved better performance for the mixed set-based face recognition.

1.6 Dissertation outline

The rest of the dissertation is organized as follows: the background and related previous work on face detection, landmark detection, face reconstruction, and face recognition are presented in Chapter 2. The proposed methods for each of the objectives are described, discussed, and evaluated in Chapter 3 to Chapter 6. Finally, Chapter 7 concludes all the works and provides directions for future research.

1.7 Publications

1.7.1 Conferences

1. **X. Xu**, X. Zhou, R. Venkatesan, G. Swaminathan, and O. Majumder. dSNE: domain adaptation using stochastic neighborhood embedding. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, Long Beach, CA, Jun. 16-21, 2019.

2. **X. Xu** and I.A. Kakadiaris. Open source face recognition performance evaluation package, In *Proc. International Conference on Image Processing*, Taipei, Taiwan, Sept. 22-25, 2019.

3. **X. Xu**, H. Le, and I.A. Kakadiaris. On the importance of feature aggregation for face reconstruction, In *Proc. Winter Conference on Applications of Computer Vision*, Waikoloa Village, Hawaii, Jan. 7 - 11, 2019.

4. L. Shi, **X. Xu**, and I. A. Kakadiaris, Smoothed attention network for single stage face detector, In *Proc. International Conference on Biometrics*, Crete, Greece, Jun. 4-7, 2019.

5. L. Shi, **X. Xu**, and I. A. Kakadiaris, A simple and effective single stage face detector, In *Proc. International Conference on Biometrics*, Crete, Greece, Jun. 4-7, 2019.

6. N. Sarafianos, **X. Xu** and I.A. Kakadiaris. Deep imbalanced attribute classification using visual attention aggregation, In *Proc. Europe Conference on Computer Vision*, Munich, Germany, Sept. 8-14, 2018.

7. L. Shi, **X. Xu**, and I. A. Kakadiaris, SSFD: a face detector using a single-scale feature map, In *Proc. IEEE International Conference on Biometrics: Theory, Applications, and Systems*, Los Angeles, CA, Oct. 22-25, 2018.

8. **X. Xu**, H. Le, P. Dou, Y. Wu, and I. A. Kakadiaris. Evaluation of a 3D-aided pose invariant 2D face recognition system, In *Proc. International Joint Conference on Biometrics*, Denver, CO, Oct. 1-4, 2017

9. **X. Xu** and I. A. Kakadiaris. Joint head pose estimation and face alignment framework using global and local CNN features, In *Proc. IEEE Conference on Automatic Face and Gesture Recognition*, Washington, D.C., May 30-June 3, 2017.

10. **X. Xu**, S. K. Shah and I.A. Kakadiaris. Face alignment via an ensemble of random ferns, In *Proc. International Conference on Identity, Security and Behavior Analysis*, Sendai, Japan, Feb. 29 - March. 2, 2016

11. Y. Wu, **X. Xu**, and I. A. Kakadiaris. Towards fitting a 3D dense facial model to 2D image without landmarks, in *Proc. International Conference on Biometrics: Theory, Applications, and Systems*, Arlington, VA, Sept. 8-11, 2015

1.7.2 Conferences Under Review

1. **X. Xu**, N. Sarafianos, and I.A. Kakadiaris. On improving the generalization of face recognition in the presence of occlusions. In *Proc. International Conference on Computer Vision*, Seoul, Korea, Oct. 27-Nov. 2, 2019 (under review).

2. **X. Xu**, and I.A. Kakadiaris. Unsupervised graph template adaptation from single

image-based to mixed-media set-based face recognition. In *Proc. International Conference on Computer Vision*, Seoul, Korea, Oct. 27-Nov. 2, 2019 (under review).

3. N. Sarafianos, **X. Xu**, and I.A. Kakadiaris. Adversarial representation learning for text-to-image matching. In *Proc. International Conference on Computer Vision*, Seoul, Korea, Oct. 27-Nov. 2, 2019 (under review).

1.7.3 Journal Under Review

1. **X. Xu** and I.A. Kakadiaris. FR^2-FAN: feature aggregation network for face reconstruction and recognition, In *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2019 (under review).

2. L. Shi, **X. Xu**, and I.A. Kakadiaris. SSFD$^+$: a two-stage robust face detector, In *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2019 (under review).

Chapter 2

Related Work

2.1 Face Detection

Face detection is the first step in the face recognition domain. Zefeiriou *et al.* [112] presented a comprehensive survey on this topic in which they divided the approaches into two categories: rigid template-based methods and deformable-parts-models-based methods. In addition to the methods summarized in [112], the approaches of object detection under the regions with a convolutional neural network framework (R-CNN) [26] have been well developed. Some techniques can be directly integrated to face detection [38]. Li *et al.* [51] used a 3D mean face model and divided the face into ten parts. The approach proposed by Hu and Ramanan [33] explored context and resolution of images to fine-tune the residual network [30]. Despite the two-stage detectors mentioned above, single-stage detectors [52, 65] have also been developed. Shi *et al.* [73] proposed a simple face detector to localize multi-scale faces [75] using a single feature map [74]. A context anchor and low-level

feature pyramid network were developed in PyramidBox [79] to detect tiny faces.

2.2 Face Alignment

Face alignment refers to aligning the face image to a canonical position. Jin and Tan [39] summarized the categories of popular approaches for this task. Cascaded regression methods learn a mapping function from the feature domain to the target output (*e.g.*, shape residuals in this problem). The features in the cascaded regression are usually shape-indexed [10, 66, 107] and are learned based on the predicted landmarks. In addition, some algorithms chose different regressors for the fitting process (*e.g.*, random ferns [10], random forest [43, 50], linear regressor [99, 66, 104], and neural networks [113, 87]). Specifically, Xiong *et al.* [99] developed a method called Supervised Descent Method (SDM) by investigating linear regression with strong hand-crafted features such as Scale-invariant Feature Transform (SIFT) [55]. An incremental face alignment method [4] incrementally updated the linear regressors in parallel by addressing the re-training problem of sequence learning when the new samples arrive. Ren *et al.* [66] proposed learning local binary features by using random forests and demonstrated that performance of 3,000 frames per second can be achieved. In addition, an ensemble of regression trees was used by Kazemi *et al.* [43] to localize the face landmarks. Zhu *et al.* [122] combined the exemplar searching and regression method together, searching for similar shapes from a shape pool using a probabilistic function. Xu *et al.* [104] proposed an initialization method based on part detection and used random ferns to learn features. Zhu *et al.* [123] trained random forest to choose a homogeneous domain of optimization, each of which was handled by

a regressor. Xu and Kakadiaris [100] proposed to jointly learn head pose estimation and face alignment tasks in a single framework using global and local CNN features.

2.3 Face Reconstruction

Zhu *et al.* [124] used a fitting algorithm to generate ground-truth parameters for 2D images and learned cascaded networks to generate 3DMM parameters for face alignment [104, 100]. Jourabloo *et al.* [40] used multiple neural networks in a cascaded manner to jointly regress camera rotation matrix, shape, and expression parameters of 3DMM. To reconstruct the 3D model in a collection of 2D images, a quality measurement was proposed by Piotraschke *et al.* [63] to select and combine reconstructed face meshes of different facial regions into a single 3D face. A 3DMM fitting algorithm was applied on an image collection by Roth *et al.* [67] to generate personalized template while the fitting results were used to estimate albedo, lighting conditions, and surface normals. Dou *et al.* [20] modified VGG-16 [60] architecture to learn a direct mapping from facial image to the shape and expression parameters of 3DMM. Similarly, Tran *et al.* [84] applied ResNet-101 [30] to predict the shape and texture parameters with the purpose of face recognition. A recurrent neural network was applied by Dou *et al.* [19] to estimate the unique shape and expression parameters from a set of images. By encoding the fitting process in the DNN learning process with a differentiable render layer, Tewari *et al.* [82] proposed an unsupervised learning method to estimate 3DMM parameters on near-frontal facial images. The encoder and decoder networks were presented by Tran *et al.* [85] to generate a bump map and recover the detailed face meshes from occluded facial regions. Similar

to [84], a ResNet-101 network was used by Chang *et al.* [11] to estimate expression parameters from a single 2D image. A trainable generative model along with differentiable image formation model was applied by Tewari *et al.* [81] to update the principal components and expand the variations in 3DMM for in-the-wild images. Following the idea of image-to-image translation, the UV position map was generated by encoding 3D point cloud information in an image [24] and was used as the ground-truth to learn a DNN model for face reconstruction and alignment tasks. Xu *et al.* [103] designed a feature aggregation learning method that improved the face reconstruction performance significantly on several datasets.

2.4 Face Template Generation

An emerging topic in face recognition research is generating a discriminative representation for a human subject. Parkhi *et al.* [60] proposed the VGG-Face template generator. Triplet loss was proposed by Schroff *et al.* [70] to train a deep neural network using 200 million labeled faces from Google. Face recognition techniques that generate 2D frontal images or facial embeddings from a single image have been proposed that: (i) use a 3D model [56, 102], (ii) generative adversarial networks [6, 17, 36, 86, 110, 116, 117], and (iii) various transformations [8, 119, 121]. Additionally, multiple loss functions [18, 53, 91, 93, 94, 118] have been developed to guide the network to learn more discriminative face representations but these usually ignore facial occlusions. Kang *et al.* [42] used a RoI pooling layer to obtain local patches from the image and jointly learned the global and local features. Early methods approached face recognition in occluded scenarios by

using variations of sparse coding [25, 90, 111]. However, such techniques work well only with a limited number of identities, and with frontal facial images in a lab-controlled environment. The works of He *et al.* [32] and Wang *et al.* [92] addressed this limitation by matching face patches under the assumption that the occlusion masks were known beforehand and that the occluded faces from the gallery/probe were also known, which is not realistic. Egger *et al.* [21] combined segmentation with an occlusion-aware 3D morphable model adaptation, which requires an additional segmentation network to generate the occlusion mask. De-occlusion methods [13, 115] have also been introduced but a major limitation is that they assume the input images exhibit same simulated occlusion as the training set. Cheng *et al.* [13] simulated the occlusion with a black rectangle and used an auto-encoder to reconstruct the image whereas Egger *et al.* [21] combined segmentation with an occlusion-aware 3D morphable model adaptation. Zhao *et al.* [115] proposed an LSTM-based auto-encoder to reconstruct the simulated images to original images. They rendered nine types of occlusion objects on gray-scaled images in their training set.

2.5 Face Recognition Systems

OpenCV and OpenBR are some well known open-source computer vision and pattern recognition libraries. However, the eigenface algorithm in OpenCV is out-of-date. OpenBR has not been updated since 9/29/2015. Both libraries only support nearly frontal face recognition, since the face detector can only detect the frontal face. OpenFace is an open-source implementation of FaceNet [70] by Amos *et al.* [1] using Python and Torch, which provides four demos for use. OpenFace applied Dlib face detector and landmark detector

to do the pre-processing, which is an improvement over OpenBR. There is another official Tensorflow implementation of FaceNet in which the authors used MTCNN [114] to detect and align the face, which boosted the performance speed and detection accuracy.

Chapter 3

Objective 1: 3D-aided 2D Pose-invariant Face Recognition System

In this chapter, the objective is to design, implement, and evaluate a 3D-aided face recognition system that is modular, in which the components are easy to use and easy to install. To achieve this goal, a 3D-aided 2D face recognition system (UR2D-E), along with an evaluation package (FaRE), was implemented: (i) UR2D-E is a 3D-aided 2D face recognition system designed for pose-invariant face recognition. Specifically, UR2D-E includes face detection and landmark detection for the localization of faces as well as their attendant landmarks. A 3D model was reconstructed from a 2D image or several 2D images. By estimating the 3D-2D projection matrix, the correspondence between the 3D model and the 2D image was computed and used to render the frontalized images. The different feature extractors were used to generate the template that represented the face. In the matching stage, the computation of the similarity between templates utilized the cosine

similarity metric. In the end, the similarity score matrix was generated from the system by comparing the similarity between the gallery and the probe set. (ii) To evaluate the performance of the implemented system, a light-weight, maintainable, scalable, generalizable, and extendable face recognition evaluation package named FaRE was designed and implemented. Commonly used face recognition datasets were selected and their metrics were analyzed to generalize an evaluation pipeline and evaluate the performance of face recognition algorithms. To support offline evaluation, a file management module was implemented to organize and match the generated template file with meta-data for each dataset. To support online evaluation, data loaders were implemented to feed the data to the neural network and generate a template from a facial image or an image set. The similarity matrix was obtained by computing the similarity of the templates from probe and gallery set based on the evaluating dataset protocol. Based on the similarity matrix and the ground-truth label provided in datasets, different quantitative measurement functions were used according to the protocols provided in the datasets. To visualize the quantitative results, comparison figures were plotted using FaRE. With our evaluation package, the new datasets and protocols were easily extended and evaluated. In addition, new fusion functions were easily added for set-based face recognition.

3.1 3D-aided 2D Face Recognition System

3.1.1 System Design

UR2D-E is a 3D-aided 2D face recognition system designed for pose-invariant face recognition. The algorithm modules were constructed as high-level APIs. The users can directly call these applications and obtain the results. The advantages of this architecture are that it is simple and well-structured. With the full development of libraries, the system easily used CPUs/GPU and other features.

Data Structures: In UR2D-E, the basic element is a `File` on the disk. All operations or algorithms are based on the files. The basic data structure is `Data`, which is a hash table with pairs of keys and values. Both keys and values are stored as a string datatype.

Configuration: The configuration file points to the datasets, input files, output directories, involved modules and their model locations, and evaluations. Attribute `dataset` contains the information for the input dataset including the name and path. Attribute `input` contains the list of galleries and probes. Attribute `output` defines the output directories. Attribute `pipelines` defines the modules used in the pipeline. The `pip` command line application only accepts the argument of the configuration file, which will parse the configuration file, load the models, and run defined modules. The advantages of this approach are simplicity and flexibility. These operations do not require a detailed understanding of the options or require the input of long arguments in the command line. The users only need to change some values in the attributes `dataset` and `input` (*e.g.*, set dataset directory and file to enroll), and program `pip` will generate the output they defined in this

configuration file.

Command Line Interface: To make full use of SDK of UR2D-E, some corresponding applications were created to run each module. All applications accept the file list (text or CSV file by default, which includes a tag at the top line), a folder, or a single image. The `IO` system will load the data into memory and process the data according to the data list. The arguments specify the location of the input file/directory and where the output should be saved. When UR2D-E's enrollment is executed, it generates signatures to the output directory. The path of the signature is recorded in the `Data`. By calling the API from `IO` system, the list of `Data` will be written to the file (default is in `.csv` format).

3.1.2 Enrollment

UR2D-E contains face detection, face alignment, 3D face reconstruction, pose estimation, texture lifting, and signature generation.

Face Detection: To detect the face in multi-view poses, some modern detectors such as Headhunter [57], DDFD [23], and Dlib-DNN [45] face detectors were supported in our system. To support different face detectors for downstream modules, bounding box regression was performed. The first advantage of this approach is that it does not need to re-train or fine-tune the models for downstream modules after switching the face detector. The second advantage is that this approach provides a more robust bounding box for the landmark localization module.

Landmark Detection: To detect face landmarks, GoDP [96] was deployed in the system. Features from shallow and deep layers were aggregated to predict the confidence map.

Each confidence map indicates the possibility of a landmark appearing at a specific location in the original image. Predictions were made by selecting the location that had the maximum response in the confidence map. The landmarks of the corner of eyes, nose tip, and corner of the mouth were selected as output from this module.

3D Shape Reconstruction: To reconstruct the 3D facial shape of the input 2D image, the E2FAR algorithm and DRFAR proposed by Dou *et al.* [20, 19] were integrated into the pipeline. The features were aggregated from two different convolutional layers in VGG-16 [60] and were used to predict the 3D AFM parameter vector from a single 2D image in E2FAR. A recurrent neural network was deployed in DRFAR [20] to support multiple facial images as input. To improve the robustness to illumination variation, the weights trained on real facial images were used for initialization and the model was fine-tuned on the synthetic data. Compared with existing work, it was more efficient due to its end-to-end architecture, which required a single feed-forward operation to predict the model parameters. Both of them do not require landmark detection and predict the shape parameter directly from the facial image.

Pose Estimation: With the assumption of perspective projection between 3D landmarks obtained from a 3D model and corresponding 2D landmarks obtained from landmark detection, the projection matrix can be estimated by solving a least-squares problem.

Texture Lifting: Texture lifting is a technique proposed by Kakadiaris *et al.* [41], which lifts the pixel values from the original 2D images to a UV map. Given the 3D-2D projection matrix, the 3D AFM model, and the original image, it first generates the geometry image, each pixel of which captures the information of an existing or interpolated vertex on the 3D AFM surface. A set of 2D coordinates referring to the pixels on an original 2D

facial image was computed. In this way, the facial appearance was lifted and represented into a new texture image. A 3D model and the Z-Buffer technique [41] were used to estimate the occlusion status for each pixel. This module has the following two advantages: It generates the frontal normalized face images, which is convenient for feature extraction and comparison. Second, it generates occlusion masks, which identify the parts of the facial images that are occluded.

Signature Generation: To improve the performance of face recognition in matching non-frontal facial images, unlike the previous method, deep learning was deployed on the local patches from frontalized texture and self-occlusion mask. The texture was divided equally into eight partially-overlapping regions. A ResNet-18 [30] was trained for each cropped region.

3.1.3 Matching

In the template matching stage, local similarities were computed on non-occluded local patches and the overall similarity was the average of local similarities. Assuming that the template of image I_i is denoted by f_i, the cosine similarity between two templates $\{f_i, f_j\}$ was computed as follows:

$$o(f_i, f_j) = \frac{f_i \cdot f_j}{||f_i|| \cdot ||f_j||}. \tag{3.1}$$

3.1.4 Experiments

In this section, a systematical and numerical analysis on three challenging datasets is provided in both constrained and in-the-wild scenarios.

3.1.4.1 Datasets and Protocols

UHDB31 [48] was created in a controlled lab environment, which allows face-related research on pose and illumination issues. In addition to 2D images, it also provides the corresponding 3D model of subjects. An interesting fact of this dataset is that pose follows a uniform distribution on two dimensions: pitch and yaw. The pitch range is $[-30^{\circ}, +30^{\circ}]$ and the yaw varies in $[-90^{\circ}, +90^{\circ}]$. For each subject, a total of 21 high-resolution 2D images from different views and 3D data were collected at the same time. Then, a 3D model was registered from the 3D data from different poses to generate a unified 3D face mesh. In addition to three illuminations, the resolutions were downsampled to 128, 256, and 512 from the original size. The face identification protocol in this dataset independently used the frontal image (pose 11) as the gallery and the images from the other 20 different poses (poses 1 - 10, 12 - 21) as probes. IJB-A [46] is another challenging dataset which consists of images in the wild. This dataset merged images and frames together and provided evaluations on the template level. A template contains one or several images/frames of a subject. According to the IJB-A search protocol (face identification), galleries and probes were split into 10 folders.

Table 3.1: Comparison of rank-1 identification rate of different systems on UHDB31 dataset.

Pitch \ Yaw	$-90°$	$-60°$	$-30°$	$0°$	$+30°$	$+60°$	$+90°$
$+30°$	14/11/ 58/**82**	69/32/ 95/**99**	94/90/ **100/100**	99/**100/** **100/100**	95/93/ 99/99	79/38/ 92/**99**	19/7/ 60/**75**
$0°$	22/9/ 84/**96**	88/52/ 99/**100**	**100**/99/ **100/100**	-	**100/100/** **100/100**	94/73/ 99/**100**	27/10/ 91/**96**
$-30°$	8/0/ 44/**74**	2/19/ 80/**97**	91/90/ 99/**100**	96/99/ 99/**100**	96/98/ 97/**100**	52/15/ 90/**96**	9/3/ 35/**78**

3.1.4.2 Baselines

To perform a fair comparison with current state-of-the-art face recognition systems, VGG-Face, FaceNet, and COTS v1.9 were chosen as baselines. (i) The VGG-Face descriptor was developed by Parkhi *et al.* [60]. In our implementation, different combinations of descriptor and matching methods were tried. It was observed that embedding features with cosine similarity metrics worked the best for the VGG-Face. (ii) The FaceNet algorithm was proposed by Schroff *et al.* [70]. They first used MTCNN [114] to align the face and extracted 128 dimensions features. They provided pre-trained models that achieve $99.20\% \pm 0.30\%$ accuracy on the LFW dataset. (iii) COTS is a commercial software developed for scalable face recognition.

3.1.4.3 UHDB31: Lab-controlled Pose-invariant Face Recognition

In this experiment, a configuration from the UHDB31 dataset named UHDB31.R0128.I03 was used. This subset was chosen to demonstrate that our system, UR2D-E, is robust to different poses. Therefore, this configuration was used to exclude the other variations such as illumination and expression, but only kept the pose variations.

Table 3.1 depicts the comparison of rank-1 identification rate on UHDB31 dataset. The methods were ordered as VGG-Face, COTS v1.9, FaceNet, and UR2D-E-DPRFS. The index of poses were ordered from the left to right and from the top to bottom (*e.g.*, pose 3 is pitch -30° and yaw -90°, pose 11 is pitch 0° and yaw 0°). The frontal face was gallery while the other poses were probes. The experimental results indicated that UR2D-E was robust to the different poses compared with other systems. It was observed the VGG-Face and COTS v1.9 algorithms cannot generalize to all pose distributions. However, in cases such as pose 3 (-30°, -90°) and pose 21 (-30°, -90°) in Table 3.1, the performance of 2D face recognition pipelines still had significant room for improvement. On the other hand, with the help of the 3D model, our system showed consistent and symmetric performance among different poses. Even in cases with yaw -90° or $+90^\circ$, our system tolerated the pose variations, and achieved around 80% rank-1 identity accuracy with DPRFS features on average.

3.1.4.4 IJB-A: In-the-Wild Face Recognition

A different protocol for face identification experiments was designed based on the original ten splits. Unlike the original template-level comparison, ten closed-set image comparison

Table 3.2: Comparison of rank-1 percentage of different systems on 10 splits of IJB-A.

Method	Split-1	Split-2	Split-3	Split-4	Split-5	Split-6	Split-7	Split-8	Split-9	Split-10	Avg.
VGG-Face	76.18	74.37	24.33	47.67	52.07	47.11	58.31	54.31	47.98	49.06	53.16
COTS v1.9	75.68	76.57	73.66	76.73	76.31	77.21	76.27	74.50	72.52	77.88	75.73
UR2D-E-DPRFS	**78.20**	**76.97**	**77.31**	**79.00**	**78.01**	**79.00**	**81.15**	**78.40**	**74.97**	**78.57**	**78.16**

pairs were generated by removing the samples in the IJB-A splits. The face was cropped according to the annotations. Table 3.2 depicts the rank-1 identification rate with different methods on IJB-A dataset. UR2D-E with DPRFS reported better performance compared with VGG-Face and COTS v1.9. In addition, UR2D-E results were consistent on 10 splits, which indicated that our system was robust.

3.2 Open-source Face Recognition Performance Evaluation Package

3.2.1 System Design

To help researchers obtain quick feedback from evaluation and accelerate research process, a *light-weight*, *extendable*, *generalizable*, *scalable*, and *maintainable* face recognition evaluation package was designed and implemented, which was easily generalized to evaluate other biometric applications.

Considering the generalization for commonly used face verification datasets such as LFW [35] and CFP [72] and face identification datasets such as IJB-A [46], IJB-B [95],

and IJB-C [58], two main protocols were defined in these datasets: a comparison protocol for face verification and a search protocol for face identification. The protocols were abstracted into three parts: comparison protocol, closed-set protocol, and open-set protocol. Each protocol called its intrinsic metric functionality to measure face recognition performance. The datasets consisted of the generated templates from online training or loaded templates in offline mode and the corresponding labels. Therefore, a custom dataset was easily extended by inheriting the existing dataset, which was mainly required to feed the templates and labels into the system. In addition, to fit the set-based face recognition, a template was defined and a custom template fusion function was easily added to generate one template from a set of feature vectors. To organize the files and templates in datasets, some classes were defined for managing the data or meta-data. On top of the system, the users can easily call the dataset wrapper and perform the evaluation.

Metrics: As one of the basic functions defined in FaRE, metrics class defines and manages several commonly used metrics including Receiver Operating Characteristic (ROC) curve, Precision-Recall (PR) curve, Accuracy (ACC), and Equal Error Rate (EER) for face verification comparison protocol, Cumulative Matching Characteristic (CMC) curve and Detection Error Tradeoff (DET) for face identification search protocol.

Protocols: With pre-defined metric functions, in the comparison protocol, the system considers the ground-truth labels and the similarity vectors as input. The search protocol includes both the closed-set protocol and an open-set protocol. In the closed-set protocol, the identities in the probe are assumed to be within the identities set in the gallery, forcing the system to assign a label from the gallery to the testing probe according to similarity ranking. In the open-set protocol, the identities in the probe might be out of the range of

identities in the gallery, which allows the system the ability to reject some samples based on their similarity scores and defined threshold.

Datasets: Some dataset APIs were implemented and provided for users to quickly evaluate their algorithm on commonly used datasets with different purposes. Each dataset supports both offline and online evaluation: In the offline mode, the dataset loads the features from the disk and computes the similarities. To evaluate the training process of the deep neural network, several data loaders were implemented to load the image data and forward them to the trained network to obtain the templates.

Light-weight: Unlike other libraries such as Bob [2], our package was implemented in Python and only requires a few basic dependencies such as numpy for array operation, matplotlib for visualization, scikit-learn [61] for metric computing, and MXNet [12] for deep learning. Therefore, FaRE is a light-weight package because it only requires a few common dependencies, which makes FaRE extremely easy to install.

Extensibility: FaRE features four extensibility aspects: adding new template fusion functions, new metrics, new protocols, and new datasets. In set-based face recognition, a common way is to compute the mean feature vectors or assign different weights to compute weighted average feature vectors as the template for a set, which was implemented in FaRE. In addition, it supports adding new template fusion functions to fuse the features from a set of images and new metrics functions to compute new quantitative measurement. Extending current protocols or datasets is suggested to inherit the corresponding super-class and adjust the protocol process based on customized requirements, which can be quickly extended.

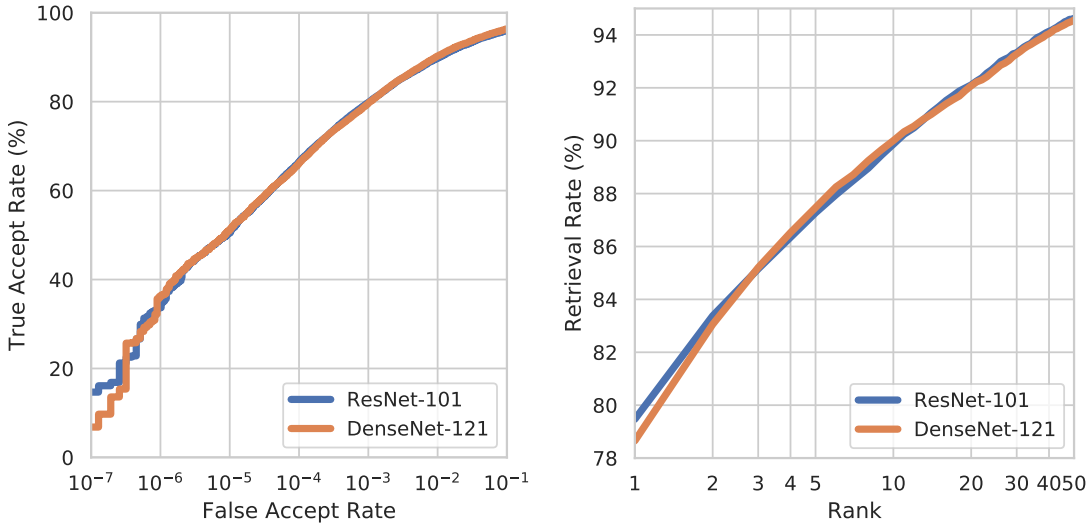
Generalizability: The generalizability is defined that the system can incorporate different datasets, running modes, and template generators. The package is abstracted to fit the requirements of various datasets and template generators. FaRE supports both online and offline performance evaluation. The online evaluation mode can be used in validating the training process while the offline evaluation mode is designed for evaluating existing algorithms.

Scalability: The system can process one image or a set of images at the same time. Several data loaders were designed and implemented to process a batch of images at the same time for online evaluation. The researchers have options to use multiple CPUs and GPUs for evaluation in this package.

Maintenance: Due to the separation of different modules and implemented logger in FaRE, the system can easily track errors and help the developer to quickly update this package.

3.2.2 Experiments

Two baselines using ResNet-101 [30] and DenseNet-121 [34] were trained on VGG-Face2 dataset [9] to generate a facial template from a single image or a set of images. The average of the feature representations generated from a set of images of a subject was computed and treated as the template of that subject. These two baselines were evaluated on the IJB-C [58] dataset for both face verification and identification tasks to present the advantages in two aspects: generalizability and scalability. In the online evaluation mode, it took around one hour to generate the templates and compute the similarity scores for the



(a) (b)

Figure 3.1: Depiction of (a) the ROC curves for the 1:1 face verification protocol, (b) the CMC curves for the 1:N open-set face identification protocol on the IJB-C dataset (best view in color).

mix identification task with a two-fold evaluation according to the IJB-C protocol. The mean feature vectors were computed from the set of features as the final facial template. The average ROC performance across gallery sets for 1:1 mixed verification protocol and average CMC and IET performance across gallery sets for 1: N mixed identification protocol was computed by FaRE. In addition, the corresponding figures generated by FaRE are depicted in Figure 3.1.

To demonstrate scalability, for simplicity, 10-fold evaluation was directly performed using FaRE on LFW dataset in the online evaluation mode. The relation of a number of

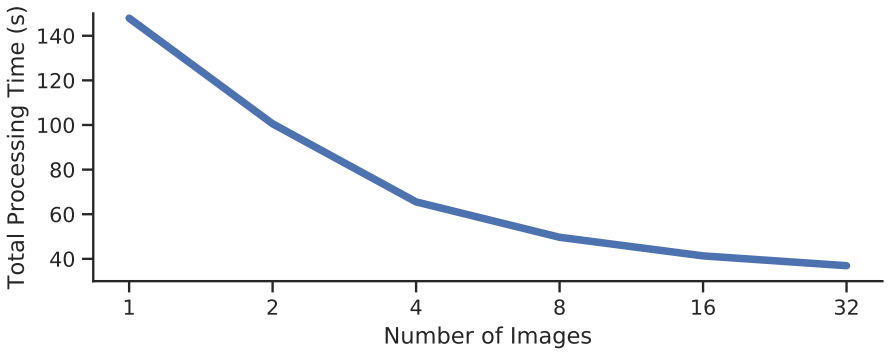


Figure 3.2: Depiction the scalability of FaRE.

images processed at the same time with the total processing time of generating the templates and computing the similarity scores is depicted in Figure 3.2. It took approximately 35 seconds to finish generating templates using DenseNet-121 and comparing all pairs in LFW by processing 32 images at the same time, which provided quick feedback during the development stage.

Chapter 4

Objective 2: 2D Face Landmark Detection

In this chapter, the objective is to develop and evaluate an algorithm for landmark detection on 2D images. In particular, there are two methods developed to achieve this goal. In the first approach, following the coarse-to-fine principle, an ensemble of random ferns (ERF) was applied to progressively learn the shape increments. The shape-indexed features were extracted and progressively adapted to learn the ensemble of ferns, which were simple and computationally efficient. Features and thresholds were learned based on the correlations between the randomly projected regression targets and a local pixel-difference matrix. Then, the intensities were extracted from the training images according to an ensemble of random ferns and compared with the corresponding thresholds to derive the local binary features. The local binary features obtained from the surroundings of each landmark were concatenated to form a global binary feature matrix. The global linear regression matrix

was learned from these global binary features by minimizing the squared loss function with L_2 regularization at the last step. The main contributions of this work are: (i) A probabilistic model was applied to select the initial shape for face alignment; (ii) An ensemble of random ferns was proposed to learn local features. In the second approach, a joint learning framework (JFA) was proposed by jointly learning the head pose estimation and face alignment using the global and local CNN features. JFA first estimated the head pose and primary points on the entire face image and initialized the shape according to the exemplars from the shape pool constructed from the training set. Then, another network was applied to learn the local features from the patches cropped from the current shape. With the local and global deep features, the head pose's residual and shape increments were both learned from the coarse-to-fine regression, which aimed to map features to the shape increment and pose residuals. JFA was designed in a hierarchical way, which analyzed the face from global to local in a cascade manner. It used global CNN features to provide better initialization, which reduced the variation from the realistic bounding box. The local CNNs provided the discriminative features for the cascaded regression. The contributions are summarized as follows: (i) The relationship between head pose and landmarks was used to search for the best shape for initialization; (ii) This was the first work to explore the deep global and local features together via CNNs on the joint head pose estimation and face alignment in a cascaded way.

4.1 Ensemble of Random Ferns

4.1.1 Method

Local Component-based Initialization: The facial patch was defined as a square region around the facial components (eyebrows, eyes, nose, mouth). Given the set of training data, to train several local detectors for each part, the image was scaled to 150×150. Positive and negative facial component patch sets were constructed with the assumption that the facial component in the training dataset and testing dataset obeyed the same distribution. The positive samples were extracted according to a Gaussian distribution centered at the component center. Furthermore, the negative image patches were sampled using a uniform distribution but the negative patches were kept at the distance of 20% of the inter-pupil distance from the component center. For each patch, Histogram of oriented gradients (HOG) features [16] were extracted and assigned a label to the component. After obtaining features, the SVM classifiers [15] were applied on each face to obtain the response maps.

Ensemble of Random Ferns: The overview of our regression approach is illustrated in Figure 4.1. ERF comprised $M \times L$ ferns $\{\mathbb{F}\}_{m=1,l=1}^{M,L}$. For each landmark, there were M ferns as local learners. Each fern was composed of F features and thresholds. To construct each fern, P pixels were uniformly sampled around each landmark. Then, the correlation-based selection method was adopted to choose F pairs of pixels out of P^2 pixel-difference features with the aim of reducing the correlation between features but retaining discriminative power. Finally, ridge regression was applied to learn the projection matrix based on the concatenated binary features learned from the ensemble of ferns.

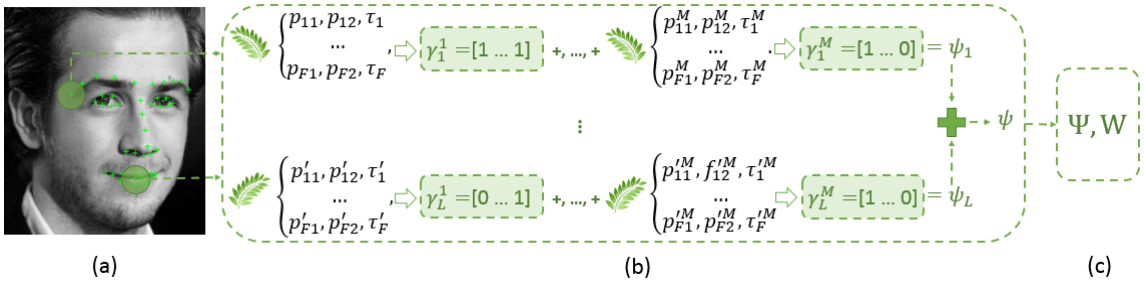


Figure 4.1: Overview of the ensemble of ferns: (a) Pixels randomly sampled around each landmark; (b) An ensemble of random ferns; (c) Features matrix and ridge regression.

Correlation-based Selection: In the training phase, the training set was divided into M subsets randomly with replacement. Then, the intensities on the P pixels were extracted for each image in the same subset.

To form a good fern (*e.g.*, a fern in which features were highly correlated to the shape increment while there was a low correlation between any feature pairs), F features were selected based on the correlation value. First, all of the shape regression targets were divided into L landmark regression targets $\{\Delta \mathbf{s}\}_{i=1,l=1}^{N,L}$. Then, a random direction $\mathbf{v}$ was dot multiplied with each regression target to produce a scalar. These scalars were concatenated to the vector $\mathbf{u}$. Assuming that ρ_m and ρ_n are any pair of pixels from the set of the P pixels, the correlation between the projected regression targets $\mathbf{u}$ and pixel-difference features $\rho_m - \rho_n$ was computed as:

$$\mathbb{C}(\mathbf{u}, \rho_m - \rho_n) = \frac{\mathbb{V}(\mathbf{u}, \rho_m) - \mathbb{V}(\mathbf{u}, \rho_n)}{\sqrt{\sigma(\mathbf{u})\sigma(\rho_m - \rho_n)}}, \tag{4.1}$$

$$\sigma(\rho_m - \rho_n) = \mathbb{V}(\rho_m, \rho_m) + \mathbb{V}(\rho_n, \rho_n) - 2\mathbb{V}(\rho_m, \rho_n), \tag{4.2}$$

where $\mathbb{C}$ denotes the correlation and $\mathbb{V}$ denoted the covariance.

Learning Local Features: To learn the local function, the fern learner used the pixel-difference features. Assuming that two pixel locations $\{\rho_j^1, \rho_j^2\}_{j=1}^F$ were selected (ρ_j^1, ρ_j^2 to represent the j^{th} pair of pixel locations ρ^1, ρ^2), the value of each binary feature $\{f_j\}_{j=1}^F$ depended on the intensities of these pixels:

$$f_j = \begin{cases} 1 & \text{if } I(\rho_j^1) - I(\rho_j^2) \geqslant \tau_j, \\ 0 & \text{otherwise,} \end{cases} \tag{4.3}$$

where $I(\rho_j)$ represented the image intensity on ρ_j, and τ_j was the corresponding threshold. Therefore, each fern generated F binary features, defined as $\gamma = [f_1, f_2, \ldots, f_F]$. The binary features were concatenated to form the global binary shape features $\psi^t = [\gamma_1^1, \gamma_1^2, \ldots, \gamma_1^M, \ldots, \gamma_L^M]$. From the training set, a $N \times (M \times L \times F)$ matrix $\Psi = [\psi_i;], i = 1, \ldots, N$ was obtained.

Global Shape Regression: Ridge regression was used to avoid over-fitting, which is expressed as:

$$\mathbf{W}_*^t = \arg\min_{\mathbf{W}^t} \sum_{i=1}^{N} \|\Delta \mathbf{s}_i^t - \Psi^t(I_i, \hat{\mathbf{s}}_i^{t-1})\mathbf{W}^t\|_2^2 + \lambda \|\mathbf{W}^t\|_2^2, \tag{4.4}$$

where λ controls the regularization strength.

4.1.2 Implementation Details

Initialization: Achieving reasonable model training time required mirroring the data and five initial shapes sampled using our local component-based initialization method. In the testing phase, two different initializations were used. The mean shape was used as the shape initialization in the first method "ERF-mean". The component detectors were used to sample five shapes in the second method "ERF-init".

Measurement: The error was measured using Mean Root Square Error (MRSE) defined as follows:

$$\mathrm{MRSE} = \frac{1}{N * L} \sum_{i=1}^{N} \frac{\|\mathbf{s}_i - \hat{\mathbf{s}}_i\|_2}{d_i}, \tag{4.5}$$

where d_i was the inter-pupil distance.

4.1.3 Experiments

4.1.3.1 Datasets

LFPW [68] is a dataset that contains 811 training images and 224 testing images collected from the Internet. Helen [49] comprises 2,330 high-resolution face images: 2,000 images for training and 330 images for testing. 300-W includes multiple datasets such as AFW [125] and IBUG [68]. AFW dataset was built by collecting the images from Flickr. The total number of images in the AFW is 205. All images in 300-W were annotated with 68 points.

4.1.3.2 Comparison with State-of-the-art Methods

For a full evaluation, the results of a protocol were reported using 51 inner landmarks and 68 full landmarks. The results were summarized in Table 4.1. The results of methods with $*$ were obtained directly from the corresponding paper. The other results were obtained by testing the publically available code with the model the author provided. As depicted in Table 4.1, it was observed that our method achieved the best performance on LFPW and Helen datasets compared with the other algorithms. Compared with ESR, ERF reduced the

Table 4.1: Comparison of different methods on LFPW, HELEN, and 300W datasets.

Method	LFPW		HELEN		300-W		
	49 pts	66 pts	49 pts	66 pts	Common	Challenge	Full set
ESR [10]	4.10	-	4.04	-	-	-	-
RCPR [7] $*$	5.48	6.56	4.64	5.93	6.18	17.26	8.35
DRMF [3] $*$	4.40	5.80	4.60	5.80	-	-	-
SDM [99] $*$	4.47	5.67	4.25	5.50	5.57	15.40	7.50
LBF [66] $*$	-	-	-	-	4.95	11.98	6.32
GNDPM [89] $*$	4.43	5.92	4.06	5.69	5.78	-	-
IFA [4]	6.12	-	5.86	-	-	-	-
POCR [88]	4.08	-	3.90	-	-	-	-
CFSS [122] $*$	3.78	4.87	3.47	**4.63**	**4.73**	**9.98**	**5.76**
ERF-mean	4.05	4.80	3.63	5.22	5.05	17.14	7.43
ERF-init	**3.70**	**4.61**	**3.46**	4.98	4.83	15.05	6.84

error from $4.10\ \%$ to $3.70\ \%$ on the LFPW dataset with smaller training and testing time. In training, ESR augmented the data 20 times, which imposed a computational burden. By comparison, our method only augmented the data five times. Moreover, ESR learned 500 cascade ferns on each iteration, while ERF only learned 15 independent ferns. Therefore, the training time was greatly decreased. Moreover, our method exhibited excellent results on high-resolution images by significantly decreasing the MSRE on Helen dataset

(14.35%) compared with ESR. Our algorithm obtained similar results as CFSS [122], but CFSS required to search similar shapes in a large shape database and use hand-designed features. Our method selected the shape from our pre-detected key points and used the pixel-differences as features, which were much more straightforward.

4.2 Joint Pose Estimation and Face Alignment

4.2.1 Method

Head pose and face alignment exhibit high positive correlation. The head pose distributions from the reference database were used and CNN features were learned to jointly reduce errors on head pose estimation and face alignment tasks in the same framework (JFA). JFA consists of two neural networks: *GNet* and *LNet*. *GNet* estimates the head pose and facial landmarks using global CNN features. With the predicted head pose, the probabilities of reference shapes were computed according to the pose. The initial shape was generated by selecting the reference shape with the highest probability and aligning to the predicted shape. The next step was a coarse-to-fine regression for the pose and landmarks using *LNet*. The patches were extracted according to the current shape and were fed into *LNet* to obtain the non-linear local CNN feature representations for pose and shape. The local CNN features were used to learn the shape and pose residuals by a linear projection. The shape and pose residuals were added to the current shapes for the next iteration. The system was designed in a hierarchical way based on coarse-to-fine principles, which sequentially refined the shape and pose.

Initialization: *GNet* was designed to explore the global information from the whole face image. To estimate the initial head pose and landmarks, a set of images $\{I_i\}$ with ground-truth poses and shapes $\{\bar{q}_i, \bar{L}_i\}$ were used as the training data, where $i = 1, \ldots, N$. The head was treated as a 3D object and its orientation was represented by three angles: pitch, yaw, and roll. For the global prediction, the face area was extracted using the predicted face bounding box to avoid the background.

Seven primary landmarks included the corners of the eyes, the nose tip, and the corners of the mouth. *GNet* consisted of two CNN sequences which shared the first two convolutional layers. The first sequence was used to predict the head pose while the second one was used to localize the initial primary landmarks. An additional convolutional layer was used to extract the features and predict the head pose. The second sequence contained three additional convolutional layers to extract low-level features for localizing the initial primary landmarks. The loss function used in training was defined as follows:

$$l_1 = \frac{1}{N_b} \sum_{i=1}^{N_b} ||q - \bar{q}||_2^2 + \frac{\lambda}{N_b * L} \sum_{i=1}^{N_b} ||L - \bar{L}||_2^2, \tag{4.6}$$

where N_b was the number of mini-batch used in training, and λ denoted the weight to balance the contributions of the two terms.

Feature extraction and Regression: Given an estimated pose q^* and initial shape L^*, the probabilities of shapes were computed in a reference shape pool and the one with the highest probability was chosen as the initialization. Based on the observation that CNN features were more discriminative than conventional features, the *LNet* was designed to obtain the local CNN features from local patches. *LNet* consisted of three convolutional layers, two max-pooling layers, and two fully-connected layers. The feature maps generated by

Table 4.2: Comparison of MRSE from different state-of-the-art approaches and corresponding face detector on 300-W *Common* set, *Challenge* set, and *Full* set.

Method	Face detector	51 landmarks					68 landmarks				
		LFPW	HELEN	*Common*	*Challenge*	*Full*	LFPW	HELEN	*Common*	*Challenge*	*Full*
DRMF [3]	MATLAB	4.95	6.11	5.64	14.82	7.44	5.80	7.26	6.67	16.66	8.63
Chehra [4]	MATLAB	4.10	4.95	4.60	15.83	6.80	-	-	-	-	-
LBF [66]	OpenCV	4.63	5.69	5.26	18.58	7.87	5.58	6.58	6.18	18.94	8.68
ERT [43]	Dlib	**3.81**	**4.04**	**3.94**	12.17	**5.55**	**4.59**	**4.96**	**4.81**	13.66	6.55
3DDFA* [124]	Dlib	66.64	13.03	34.71	28.60	33.51	-	-	-	-	-
JFA	Dlib	4.65	5.26	5.01	**8.98**	5.79	5.08	5.48	5.32	**9.11**	**6.06**

the second and third convolutional layers were concatenated and fed to a fully-connected layer The loss function in *LNet* was defined as follows to minimize the difference between the predicted and ground-truth residuals:

$$l_2 = \frac{1}{N_b}\sum_{i=1}^{N_b}||\Delta q^t - \Delta \bar{q}^t||_2^2 + \frac{\lambda}{N_b * L}\sum_{i=1}^{N_b}||\Delta L^t - \Delta \bar{L}^t||_2^2, \tag{4.7}$$

where Δq^t and ΔL^t denoted the predicted pose and shape residuals in t iteration. Two linear regressions were used to predict the shape increment and pose increment. With the predictions Δq^t and ΔL^t in t iteration, the pose and shape were updated by $q^{t+1} = q^t + \Delta q^t$ and $L^{t+1} = L^t + \Delta L^t$.

4.2.2 Experiments

4.2.2.1 Landmark detection

JFA was compared against state-of-the-art methods in two types of experiments. To provide a fair and intuitive comparison, face alignment task was evaluated on the *full testing*

Table 4.3: Comparison of AME for head pose estimation.

Method	**Pitch**	**Yaw**	**Roll**
Yang *et al.* [106]	5.1	4.2	**2.4**
Random forest	4.7	5.5	4.8
SVR	4.8	7.8	5.3
GNet	3.5	3.3	2.6
JFA	**3.0**	**2.5**	2.6

set from 300-W including the *common testing set* and *challenge testing set*. The baselines included ERT [43], LBF [66], Chehra [4], DRMF [3], and 3DDFA [124]. For a fair comparison, ERT was re-trained using the same bounding box. 3DDFA [124] was tested using the Dlib bounding box. Moreover, the evaluation of inner face alignment was provided, which excluded the outer contour of the face (17 landmarks). Different results were obtained using 51 inner face landmarks and 68 full landmarks, due to the variations of the predicted face contours. The full evaluation was summarized in Table 4.2. Although JFA was not prominent on the common pose face images contained in LFPW and Helen, it outperformed the conventional methods such as ERT by 33% on the face images with a large pose.

4.2.2.2 Head pose estimation

The work of Yang *et al.* [106] was selected as the baseline on head pose estimation. The absolute mean error (AME) of three dimensions: pitch, yaw, and roll were computed as $\mathrm{AME} = \frac{1}{N}\sum_{i=1}^{N} ||q - \bar{q}||_1$. Both methods were trained on the training set of 300-W and tested on the *full set*. To obtain a fair comparison, their result was directly used from the literature. Then, the learned features were compared with the traditional method. Using the same augmented training data with our method, the model was trained using the random forest and super vector regression, respectively, using HoG features. The random forest was set to contain 100 trees. The cell size of the HoG extractor was set to 8×8 with nine cells in the same block. The results are depicted in Table 4.3. Compared with a conventional random forest approach such as random forest, our method boosted the accuracy of 54% in total: 36% on the pitch, 55% on the yaw, and 46% on the roll.

Chapter 5

Objective 3: 3D Face Reconstruction in the presence of Pose and Expression

In this chapter, the objective is to design, implement, and evaluate an algorithm for 3D face reconstruction from a single image that depicts individuals with variations in pose and expression. In particular, the feature aggregation network (FR-FAN) was proposed to join the multi-path information from shallow and deep layers to predict the morphable model parameters. A synthetic dataset was produced with $1,000$ identities, which covered a wide range of pose variations ($[-90^{\circ}, +90^{\circ}]$). This dataset was used to analyze existing state-of-the-art algorithms such as E2FAR [20] and 3DMM-CNN [84]. From extensive experiments and detailed analysis, some intuitive conclusions were drawn about the deep neural network design. Two main contributions are summarized as follows: (i) The E2FAR and a few of its variances were analyzed to obtain the principles for the network design.

(ii) An efficient network was designed for 3D facial reconstruction that improved the performance using existing common backbone networks.

5.1 On the Importance of Feature Aggregation for Face Reconstruction

5.1.1 Method

An efficient network named FR-FAN was proposed based on ResNet-101 architecture. FR-FAN merged the information from shallow, middle, and deep layers, which provided an augmented featureset and improved prediction for the parameters of a morphable model from a 2D facial image, when compared with the features only from the deep layer. Figure 5.1 depicts the feature aggregation network and different aggregated approaches from the different pathways, which are explained in detail below.

Bottom-Up Pathway: The ResNet-101 served as the backbone network architecture in FR-FAN, which consisted of four blocks. The size and connections of these four blockwise feature maps were illustrated in Figure 5.1. Because there were four blocks in the network, it was natural and easy to obtain features from each block for feature aggregation in current deep learning frameworks. The outputs of each successive block were named as $\{\mathrm{B}_1, \mathrm{B}_2, \mathrm{B}_3, \mathrm{B}_4\}$, respectively. The feature aggregation module aggregated the feature maps from these blocks into a single feature vector.

Top-Down Pathway: Based on the observations that feature fusion was of benefit for both

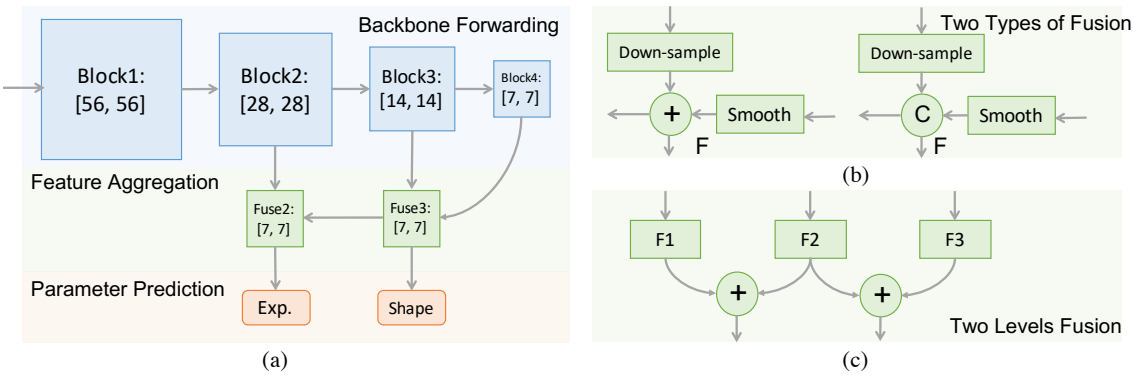


Figure 5.1: Depiction of: (a) Three levels of FR-FAN network architecture; (b) Feature fusion by adding and concatenation operations; and (c) Two level fusion for shape and expression predictions.

shape and expression prediction, and that low-level features resulted in slightly improved performance of expression parameters prediction, it was natural for us to abandon the direct prediction from the last layer features. Instead, the features from the last two deep blocks were used to predict shape parameters and the features from the two middle blocks were used to predict expression parameters. Since the features from $\{\mathrm{B}_2, \mathrm{B}_3, \mathrm{B}_4\}$ have 512, $1,024$, and $2,048$ channels, respectively, a 1×1 convolutional layer was used to reduce the channel dimensions of the output feature maps. To fuse the feature maps, a down-sample layer was designed to reduce the dimension of the feature map and reduce the memory overhead. These down-sampled layers consisted of a single or multiple 5×5 convolutional layers with a stride of two. In this manner, the feature maps were aggregated in the same dimensions.

Feature Aggregation: Regarding the one-level fusion, the features from the upper layer

and bottom layer were fused by element-wise addition or channel-wise concatenation (Figure 5.1 (b)). It was noticed that these two options achieved similar performance in practice. However, the network with adding operation was easier to optimize and had fewer parameters. Therefore, the element-wise addition operation was used by default. The features F_3 from $\{\mathrm{B}_4, \mathrm{B}_3\}$ were fused to predict shape parameters and the features from $\{\mathrm{F}_3, \mathrm{B}_2\}$ were fused to predict expression parameters. One-level fusion manner did not use the feature from B_1, so the second level of features fusion (Figure 5.1 (c)) was designed. Assuming that the features fused using the first level fusion were named $\{\mathrm{F}_3, \mathrm{F}_2, \mathrm{F}_1\}$, then a second fusion operation was applied on $\{\mathrm{F}_3, \mathrm{F}_2\}$ and $\{\mathrm{F}_2, \mathrm{F}_1\}$ to predict shape and expression parameters, respectively. Differently, in second level fusion, the kernel size in the down-sample block was changed to three to reduce redundant parameters and reduce GPU memory utilization when inferencing the model. While FR-FAN had a few more parameters than ResNet-101, our designed network was simpler and had far fewer parameters than E2FAR (approximately 50% less). Therefore, the main advantage of our network design was that it aggregated the features from different pathways and significantly boosted the performance of a pure ResNet.

Optimization: The performance to optimize the whole network was the same as E2FAR, which used the projected mean square error in 3D point cloud space defined as follows:

$$\begin{aligned} l_s &= ||A * \hat{\alpha} - A * \alpha||_2^2, \\ l_e &= ||B * \hat{\beta} - B * \beta||_2^2, \end{aligned} \tag{5.1}$$

where $\hat{\alpha}$ and $\hat{\beta}$ were the shape and expression parameter predicted by the network.

5.1.2 Experiments

5.1.2.1 Datasets, Baseline, and Metrics

Four public available datasets were evaluated for 3D face reconstruction from a single image. UHDB31 [48] is a dataset obtained under controlled lab conditions that consists of 77 subjects with 21 pose variations. For the pose distribution, the yaw angle varies from -90° to 90° with 30° interval and pitch angle varies from -30° to 30°. FRGCv2 [62] validation set is a dataset that consists of 466 subjects and $4,007$ images for evaluating the illumination problem. BU-3DFE [108] is another dataset which consists of frontal faces for evaluating the expression problem. JNU-3D [47] is a part of Surrey-JNU data for 3D face reconstruction evaluation challenge. These 2D images were collected in varying conditions, which exhibit large pose, illumination variation, motion blur, and low resolution. Two state-of-the-art algorithms, 3DMM-CNN [84] and E2FAR [20], were selected as our baseline. To compare the performance of different methods, we used the root mean squared error (RMSE) between the reconstructed 3D face point cloud and the ground truth mesh after rigid alignment and registration using the iterative closest point algorithm to measure the accuracy of 3D face reconstruction. For our method and 3DMM-CNN, the face region was trimmed according to the mean depth before evaluation. All three methods provided a similar number of vertices in the evaluation.

5.1.2.2 Comparison with State-of-the-art

Pose Variation: In the first experiment for evaluating pose variation, the performance of these three methods on the 21 view facial images in the UHD31.R0128 dataset were

Table 5.1: Quantitative comparison on UHDB31.R0128, FRGCv2, BU-3DFE, and JNU-3D datasets.

Method	UHDB31.R0128	FRGCv2	BU-3DFE	JNU-3D
3DMM-CNN [84]	3.74 ± 0.80	4.78 ± 4.31	4.08 ± 0.94	3.18 ± 0.94
E2FAR [20]	$\mathbf{3.03 \pm 0.75}$	4.51 ± 4.51	4.43 ± 1.04	3.46 ± 0.86
FR-FAN-L_1	3.25 ± 0.78	$\mathbf{4.40 \pm 4.31}$	$\mathbf{3.70 \pm 0.91}$	3.14 ± 0.59
FR-FAN-L_2	3.12 ± 0.76	4.52 ± 4.40	3.74 ± 0.87	$\mathbf{3.13 \pm 0.60}$

evaluated. The quantitative results of the mean and standard deviation of RMSE (mm) are presented in the Table 5.1. From the results, E2FAR achieved the best performance in this dataset, which might be because the distribution of the synthetic data that E2FAR was trained on was close to the distribution of UHDB31.R0128.

Illumination Variation: The quantitative results of RMSE mean and standard deviation are presented in Table 5.1. In this dataset, FR-FAN-L_1 achieved the best results. FR-FAN-L_2 exhibited compatible performance to E2FAR, but with less variation. Figure 5.2 depicts the reconstruction results generated by our method and two baselines compared to ground-truth.

Expression Variation: BU-3DFE was used to evaluate the algorithms with expression variation. The quantitative results of RMSE mean and standard deviation are presented in Table 5.1. Both our methods improved the performance approximately by 7% on this dataset compared to the performance of E2FAR.

In-the-wild: JNU-3D was used to evaluate two parametric methods (3DMM-CNN and

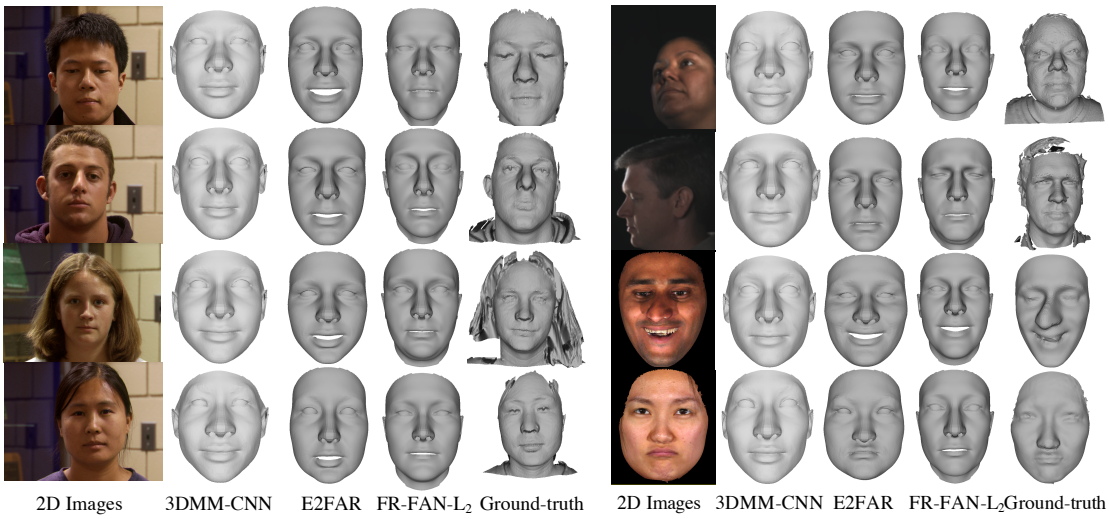


Figure 5.2: Depiction of face reconstruction results on evaluation datasets: FRGC, UHDB31 and BU-3DFE.

E2FAR), two non-parametric algorithms (VRN [37] and Pix2Vertex [71]), and FR-FAN with "in-the-wild" images. The quantitative results of RMSE mean and standard deviation are depicted in Table 5.1. Due to the dataset license constraints, only one individual's face reconstruction results of various methods are presented in Figure 5.3. It was observed that our algorithm captured the expression well while 3DMM-CNN did not provide expression parameters and E2FAR failed in some cases. Comparing the volumes generated by the VRN of Jackson *et al.* [37] and mesh generated by Pix2Vertex [71], our method not only reconstructed the shapes and expressions but also provided more reasonable, detailed, and smooth meshes.

Figure 5.3: Depiction of one sample of face reconstruction results in JNU-3D dataset with the sequence of methods as following: 2D images, 3DMM-CNN [84], E2FAR [20], VRN [37], Pix2Vertex [71], FR-FAN-L_2, and the ground-truth 3D mesh. The 3D meshes generated by our method and the ground-truth were rotated according to the pose of 2D images to visualize the similarity between 2D images and 3D meshes.

Chapter 6

Objective 4: Face Recognition in the Presence of Variance in Pose, Expression, and Occlusions

In this chapter, the objective is to design, implement, and evaluate an algorithm for 2D face recognition from a single image or set of images that depict individuals with variations in pose, expression, and occlusions. Realization of this objective resulted in the development of the following two algorithms. In the first method, an occlusion-aware face recognition (OREO) was proposed to improve the generalization of face recognition in the presence of occlusions. To address the challenge of identity signal degradation due to occlusions, an attention mechanism was introduced which was learned directly from the training data. Since the global feature representation captured information learned from the whole facial image (regardless of whether occlusion occurs), the aim of the attention

mechanism was to disentangle the identity information from the global representation and extract local identity-related features from the non-occluded regions of the image. In this way, global and local features were jointly learned from the facial images and were then aggregated into a single feature embedding. Addressing the challenge of the occlusion imbalance in the training set required the development of an occlusion balancing strategy to train the model with batches drawn as samples from an equally balanced distribution of non-occluded and occluded images. Based on this strategy, an additional learning objective was proposed that improved the discriminative ability of the embeddings learned from our algorithm. The main contributions of this work are: (i) An attention module was introduced that disentangles the features into global and local parts, resulting in more discriminative representations. In this way, the proposed approach successfully handled occlusions in face recognition without requiring additional supervision (*e.g.*, pose or occlusion labels) and achieved a relative improvement of 1.6% in terms of accuracy on the CFP dataset. (ii) An occlusion-balanced sampling strategy, along with a new loss function, was proposed to alleviate the large class imbalance that was prevalent. Our experimental results on the Celeb-A dataset indicated that OREO achieved statistically significant improvements of more than 10% in terms of average degradation percentage. In the second method, an unsupervised graph-based template adaptation (GTA) training framework was proposed to adapt the knowledge of the network learned from a still image to a mixed-media set without requiring any ground-truth label in the set domain. To improve the performance of set-based face recognition, a curriculum was designed for the teacher and student networks in two steps: First, a graph-based template adapter was inserted and

learned to generate a single feature/template that represented a set considering relationships of all features belonging to the same set. Second, the teacher and student networks were updated in an unsupervised manner considering similarities in the set domain. End-to-end optimization of the unsupervised framework required to update the teacher network using supervising signals. Updating the weights in the student network used the supervision from the source domain and the teacher network's inference in the target domain. There are two advantages of the proposed method in real-life applications: (i) information from all samples within a subgraph was aggregated to generate the more discriminative, robust, and compact templates by enlarging the similarities of the matched samples and decreasing the similarities of the non-matched samples; (ii) no modification was required of the backbone network since the graph-based template adapter is a plug-and-play module. This suggested that not only performance was preserved for single image-based face recognition, but also improved performance was obtained for the mixed-media set-based face recognition.

6.1 On improving the generalization of face recognition in the presence of occlusions

6.1.1 Method

Aiming to quantitatively analyze the impact of occlusion, a series of experiments were conducted on the Celeb-A dataset [54], which consists of $10,177$ face identities and 40 facial attributes. Attributes that describe the subject (*e.g.*, `Gender`) were ignored and

only those that might impact the face recognition performance were selected: `Bangs`, `Eyeglasses`, `Mustache`, `Sideburns`, and `Wearing Hat`. For each attribute, the images without this attribute were enrolled as the gallery and images with or without this attribute were enrolled as the probes. In both the gallery and probe, each identity had only a single image. Three face recognition systems were selected to evaluate five close-set face identification experiments:

(i) VGG-Face [60]: a commonly used face template generator baseline in the research community, which was trained on VGG-Face dataset.

(ii) COTS: a commercial off-the-shelf face recognition software which claimed to be the world's most versatile face recognition technology in the industry. The version of this software was v1.18.

(iii) ResNeXt-101: a ResNeXt-101 model [31] trained on the VGG-Face2 dataset [9].

The public state-of-the-art face recognition system, ArcFace [18], was not selected as one of the baselines because the identities between the Celeb-A dataset and the MS-Celeb-1M dataset [94] they used for training overlapped.

For each of the five attributes, the rank-1 accuracy is reported in Table 6.1 for each algorithm in the scenario with and without each attribute, respectively. The occlusion-related attributes were ranked according to the rank-1 identification rate degradation as follows: `Eyeglasses` $>$ `Wearing Hat` $>$ `Bangs` $>$ `Sideburns` $\sim$ `Mustache`. These results demonstrated that occlusion originating from facial accessories (eyeglasses and hat) as well as facial hair (mustache, bangs, and sideburns) was an serious challenge

Table 6.1: Comparison of rank-1 identification rate (%) of different face recognition systems on the Celeb-A dataset w/ and w/o the specified attribute.

Method	Bangs		Eyeglasses		Mustache		Sideburns		Wearing Hat	
	w/o	w/	w/o	w/	w/o	w/	w/o	w/	w/o	w/
VGGFace [60]	50.9	40.6	60.8	34.1	60.8	55.9	62.7	57.8	56.0	37.0
ResNeXt-101 [31]	73.8	66.8	81.4	63.1	83.3	80.9	81.5	79.0	77.7	65.6
COTS v1.18	78.5	75.4	83.0	70.2	84.3	84.0	85.3	84.4	81.8	74.4

that affected the performance of face recognition algorithms. Additionally, it was observed that occlusion due to external accessories affected the performance more than occlusion originating from facial hair.

6.1.1.1 Face Recognition in the Presence of Occlusion

The overall training architecture of OREO is depicted in Figure 6.1 (a), which consists of (i) an occlusion-balanced sampling (OBS) technique to address the occlusion imbalance; (ii) an occlusion-aware attention network (OAN) to jointly learn the global and local features; and (iii) the objective functions to guide the training process. Aiming to balance the occluded and non-occluded images within the batch, random pairs of non-occluded and occluded images were sampled and provided as input to the deep neural network. Then, the proposed attention mechanism was plugged into the backbone architecture to generate the attention mask and to construct a single representation via aggregation of the local

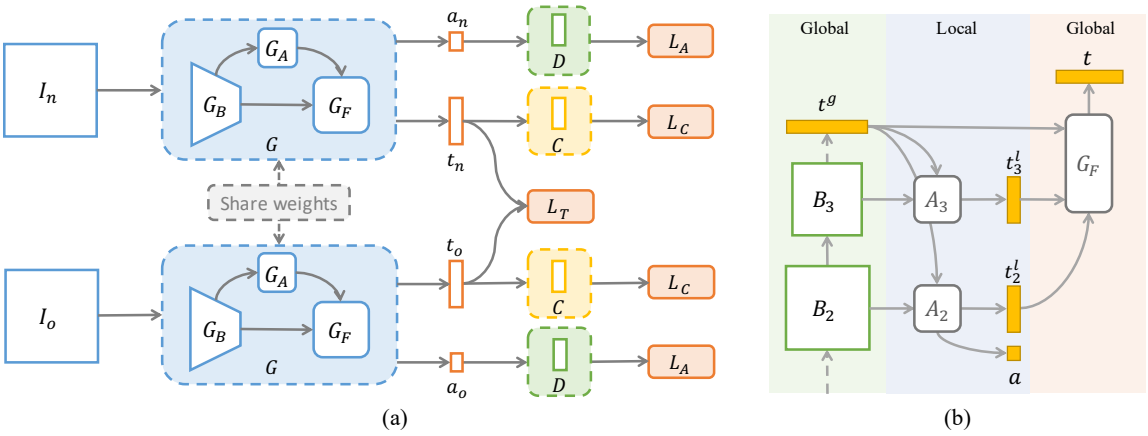


Figure 6.1: Given a pair of non-occluded and occluded images (I_n, I_o), the generator G learned the facial embeddings (t_n, t_o) and the attributes predictions a_n, a_o using loss functions for attribute classification, identity classification and the proposed similarity triplet loss. On the right, the generator is presented in detail, which contains: (i) the output feature maps of the last two blocks (B_2, B_3) of the backbone architecture, (ii) the attention mechanism G_A consisting of masks (A_2, A_3) that learn the local features in two different ways, and (iii) G_F which aggregates the global and local features to the final embedding.

with the global features. The final aggregated embeddings were trained to learn occlusion-robust representations guided by the softmax cross-entropy, sigmoid cross-entropy, and similarity triplet loss (STL) functions.

Occlusion-Balanced Sampling: In a hypothetical scenario in which training data would be accompanied by occlusion ground-truth labels, the training set could easily be split into two groups of occluded and non-occluded images from which balanced batches could be sampled and fed to the neural network. However, this is not the case with existing face recognition training datasets since no such information is provided. Aiming to generate

occlusion labels, facial attributes were selected that contained occlusion information. A state-of-the-art face attribute predictor [69] was trained on the Celeb-A dataset and it was then applied to the training set to generate pseudo-labels. Those attribute pseudo-labels were utilized to facilitate occlusion-balanced sampling during training. By using this strategy, the network G was feed-forwarded with pairs of randomly chosen non-occluded and occluded images denoted by $\{\{I_n, y_n\}, \{I_o, y_o\}\}_i, i \in \mathbb{R}^N$, where y contained the identity and attributes of the facial images and N was the total number of pairs. Since OBS randomly generated pairs of occluded and non-occluded facial images in each epoch, their distribution within the batch was ensured to be balanced.

Occlusion-aware Attention Network: The facial embedding generator G consists of three components: (i) a backbone network G_B, (ii) an attention mechanism G_A, and (iii) a feature aggregation module G_F.

Features were generated from two pathways as depicted in Figure 6.1(b): a bottom-up for the global representations and a top-down for the local features. In the bottom-up pathway, the process of the network to generate global features was described by $t^g = G_B(I)$. In the top-down pathway, since the global features include information from the occluded region of the face, an attention module G_A was proposed that distilled the identity-related features from the global feature maps to the local feature representations t^l. Finally, a feature aggregation module G_F was employed that aggregated the global and local feature representations into a single compact representation t. The goal of attention mechanisms G_A is to help the model identify which areas of the original image contain important information based on the identity and attribute labels. Assuming the feature maps extracted from different blocks of the backbone network were denoted by $\{B_1, B_2, B_3, B_4\}$, then

the two-step attention mechanism was designed as follows. In the first level (denoted by A_3 in Figure 6.1(b)), the feature map B_3 was first broadcasted and then added with the global representation t^g to generate the attention mask A_3. The goal of A_3 is to find the high-response region of the feature map by giving emphasis to the identity-related features and construct the local representation t^{l_3}. Mathematically, this process was described by the following equation:

$$t_3^l = A_3 * B_3 = h_3(t^g, B_3) * B_3, \tag{6.1}$$

where h_3 was a sequence of operations to reduce the channels and generate a single-channel convolutional attention map with spatial normalization and the "$*$" operation corresponded to element-wise multiplication. The final global feature t^g was preserved as part of the final representation of the network so that t^g learned identity-related information. In this way, t^g guided the network to learn local attention maps on features from the previous block and distilled the identity information from the most discriminative region to construct t^l.

Improving the generalization ability of the model required an additional attention mechanism on the feature map B_2 to force the network to focus on other regions of the face. In the second level (denoted by A_2 in Figure 6.1 (b)), the attention map was guided by the facial attribute predictions in a weakly-supervised manner. In the attributes prediction branch, the visual attribute predictions were output and a channel-wise summation of the spatial normalized feature maps was used to predict the attention mask A_2. Thus, the local representations at this level were computed by:

$$t_2^l = (1 - A_2) * B_2 = (1 - h_2(t^g, B_2)) * B_2, \tag{6.2}$$

where h_2 is an attention operation guided by both identity labels and attributes labels. Since the attention map A_2 was guided not only by the identity loss function, but also by the attribute predictions, the network was capable of focusing on image regions related to both the identity and the visual attributes. The global and local features $\{t^g, t_2^l, t_3^l\}$ were concatenated into a single vector and then this vector was projected into a single feature representation t, which enforced both global and local features to preserve semantic identity information.

Loss Functions: The network training comprised three loss functions: (i) the softmax cross-entropy loss L_C for identity classification, (ii) the sigmoid binary cross-entropy loss L_A for attribute prediction, and (iii) a new loss L_T designed for the occlusion-balanced sampling. The identity classification loss was defined as:

$$L_C = -\frac{1}{M}\sum_{i=0}^{M} log\frac{\exp\left(W_{y_i^c}t_i + b_{y_i^c}\right)}{\sum_{j=1}^{n}\exp\left(W_j t_i + b_j\right)}, \tag{6.3}$$

where t_i, y_i^c represented the features and the ground-truth identity labels of the i^{th} sample in the batch, W and b denoted the weights and bias in the classifier, and M and n corresponded to the batch size and the number of identities in the training set, respectively. Following that, the sigmoid binary cross-entropy loss L_A was defined as

$$L_A = -\frac{1}{M}\sum_{i=0}^{M}\log\sigma(a_i)y_i^a + \log(1-\sigma(a_i))(1-a_i), \tag{6.4}$$

where y^a corresponded to the attribute labels and $\sigma(\cdot)$ was the sigmoid activation applied on the attribute predictions a.

In the matching stage, the cosine distance was used to compute the similarity between two feature embeddings. Since images with the same identity had a higher similarity score

than those with a different identity, a similarity triplet loss (STL) was used as regularization to make the final facial embedding more discriminative. During training, each batch comprised pairs of non-occluded and occluded images with each pair having the same identity. Let t_n and t_o be the final feature representations of non-occluded images I_n and occluded images I_o. The similarity matrix $\mathbf{O} \in \mathbb{R}^{M \times M}$ within the batch was computed, where M was the batch size. In the similarity matrix $\mathbf{O}$, it should be identified: (i) hard positives which were pairs of samples that originated from the same identity but had low similarity score $o^p(t_n, t_o)$, and (ii) hard negatives which were pairs of samples with different identities but with high similarity score $o^n(t_n, t_o)$. Then, the objective function was defined as follows:

$$L_T = \sum_{i=1}^{M} [o_i^n(t_n, t_o) - o_i^p(t_n, t_o) + m]_+. \tag{6.5}$$

A margin $m \in \mathbb{R}^+$ was maintained to enforce that small angles (high similarity score) belong to the same identity and large angles (low similarity score) belong to different identities. Finally, the whole network was trained as follows:

$$L = L_C + \beta \cdot L_A + \gamma \cdot L_T, \tag{6.6}$$

where β and γ were the weighting parameters of the losses.

6.1.2 Experiments

Three face evaluation datasets were used to evaluate OREO under different scenarios.

(i) Celeb-A: This dataset contains images of different occlusion-related attributes for closed-set image-based face identification. It was used to analyze face recognition

performance in the presence of different types of occlusions;

(ii) CFP: This dataset [72] that contains images with pose variation. This dataset was selected to demonstrate that the algorithm performance will not degrade for the different poses;

(iii) IJB-C: This is an image-set-based dataset [58] with an emphasis on images with occlusion. It consists of 31,334 still images and 117,542 video clips from 3,531 subjects, which was used to generate 23,124 templates with 19,557 genuine matches and 5,678,932 impostor matches.

The evaluation protocol of each dataset was strictly followed. For face verification experiments, Identification Error Trade-off (IET) was reported with true acceptance rates (TAR) at different false accept rates (FAR). For face identification experiments, CMC including rank-1, rank-5, and rank-10 were reported.

6.1.2.1 Celeb-A: In-the-wild Face Identification

The algorithms were tested on the Celeb-A dataset [54] under various types of occlusion. The Cumulative Match Character (CMC) curves are depicted in Figure 6.2. It was observed that OREO outperformed ResNeXt-101 in all settings (w/ and w/o attributes), which suggested that our algorithm learned robust discriminative feature representations regardless of occlusions. Use of the McNemar's test indicated a statistically significant improvement of all attributes. In addition, OREO demonstrated a lower average degradation percentage than the baseline by 10.17% in terms of relative performance. This indicated

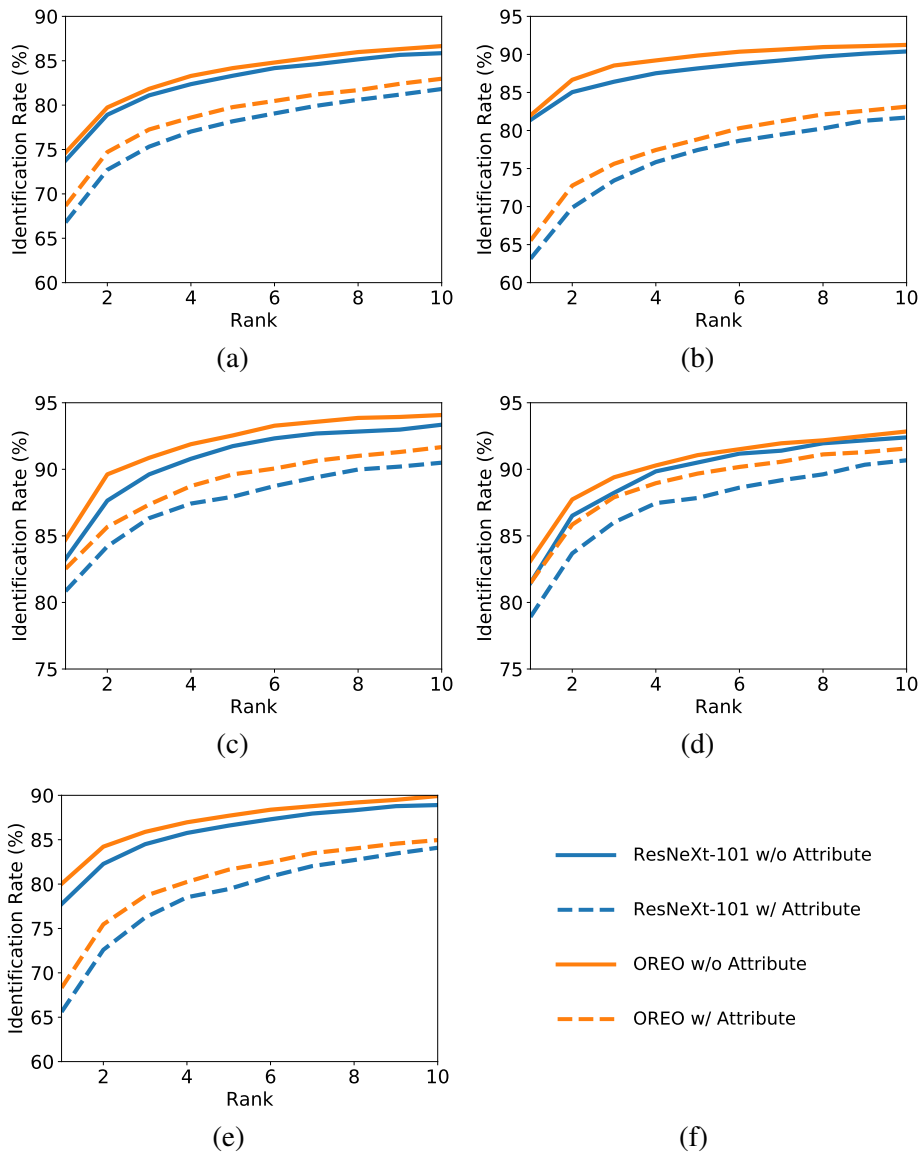


Figure 6.2: Comparison of CMC curves of ResNeXt-101 and OREO with and without the selected occlusion-related attribute: (a) `Bangs`, (b) `Eyeglasses`, (c) `Mustache`, (d) `Sideburns`, and (e) `Wearing Hat`. The last figure (f) depicts the legend.

that our algorithm improved the generalization ability of the facial embedding generator in the presence of occlusions.

Table 6.2: Comparison of the face verification performance with state-of-the-art face recognition techniques on the CFP dataset using CFP-FP protocol.

Method	Acc. (%)	TAR (%) @ FAR=		
		10^{-3}	10^{-2}	10^{-1}
DR-GAN [86]	93.4 ± 1.2	-	-	-
MTL-CNN [109]	94.4 ± 1.2	-	-	-
ArcFace [18]	93.9 ± 0.8	80.2 ± 5.9	86.0 ± 2.8	94.3 ± 1.5
ResNeXt-101 [31]	97.1 ± 0.8	81.9 ± 11.4	92.3 ± 4.1	98.9 ± 0.8
OREO	$\mathbf{97.5 \pm 0.5}$	$\mathbf{85.5 \pm 5.3}$	$\mathbf{94.1 \pm 2.5}$	$\mathbf{99.2 \pm 0.7}$

6.1.2.2 CFP: In-the-wild Face Verification

The CFP [72] was used to evaluate face verification performance on images in-the-wild, which contained variations in pose, occlusion, and age. In this experiment, the CFP-FP protocol was used to quantitatively evaluate the face verification performance and the experimental results are presented in Table 6.2. These results were presented by the average $\pm$ standard deviation over 10 folds. The following metrics were used to evaluate the performance: (i) the verification accuracy, and (ii) the TAR at FAR equal to 10^{-3}, 10^{-2}, and 10^{-1}. Compared to all baselines, OREO achieved state-of-the-art results on this dataset in terms of accuracy and increases the TAR at low FARs. The moderately better accuracy results demonstrated that OREO also improved the performance of general face recognition.

Table 6.3: Comparison of the face verification and identification performance of different methods on the IJB-C dataset.

Method	1:1 Mixed Verification TAR (%) @ FAR=							1:N Mixed Identification TPIR (%) @ FPIR=			Retrieval Rate (%)		
	10^{-7}	10^{-6}	10^{-5}	10^{-4}	10^{-3}	10^{-2}	10^{-1}	10^{-3}	10^{-2}	10^{-1}	Rank-1	Rank-5	Rank-10
GOTS [58]	3.00	3.00	6.61	14.67	33.04	61.99	80.93	2.66	5.78	15.60	37.85	52.50	60.24
FaceNet [70]	15.00	20.95	33.30	48.69	66.45	81.76	92.45	20.58	32.40	50.98	69.22	79.00	81.36
VGGFace [60]	20.00	32.20	43.69	59.75	74.79	87.13	95.64	26.18	45.06	62.75	78.60	86.00	89.20
MN-vc [98]	-	-	-	**86.20**	**92.70**	**96.80**	**98.90**	-	-	-	-	-	-
ArcFace [18]	**60.50**	**73.56**	**81.70**	**87.90**	91.14	95.98	97.92	**70.90**	**81.98**	**87.63**	**92.25**	94.31	95.30
ResNeXt-101 [31]	28.72	58.09	71.19	81.76	90.70	95.75	98.86	53.66	71.50	82.47	91.88	**95.51**	**97.29**
OREO	**51.97**	**62.36**	**75.86**	85.19	**92.81**	**97.11**	**99.37**	**65.47**	**77.11**	**85.92**	**93.76**	**96.68**	**97.74**

6.1.2.3 IJB-C: Set-based Face Identification and Verification

The IJB-C dataset [58] is a mixed media set-based dataset with open-set protocols comprising images with different occlusion variations. Two experiments were performed on this dataset following 1:1 mixed verification protocol and 1:N mixed identification protocol. To generate the facial embedding for the whole set, the corresponding images were fed to the neural networks and the average of the embeddings from all images within a set was computed. The evaluation metrics included the verification metric of ROC, the identification metric of retrieval rate, the true positive identification rate (TPIR) at different false positive identification rates (FPIR). In Table 6.3, top two performance are marked in bold. From the obtained results in Table 6.3, it was observed that OREO outperformed four out of six baselines in all metrics and came second to ArcFace only in some cases under

the mixed verification and identification protocols. ArcFace was trained on the MsCeleb-1M [28] dataset which contains significantly more identities and data than the VGGFace2 dataset. Our method still outperformed ArcFace at high FARs in the verification protocol as well as in the identification retrieval protocol. When compared against the baseline, OREO significantly improved the performance in both verification and identification. For example, when FAR was equal to 10^{-7} the TAR improved from 28.72% to 51.97%. These results demonstrated that OREO successfully learned features that were robust to occlusions.

6.2 Unsupervised Graph Template Adaptation

6.2.1 Method

In this section, an unsupervised domain adaptation framework named GTA (depicted in Figure 6.3) is presented to transfer the knowledge learned from good visual quality still image domain to an unknown set domain comprising both still image and video frames without requiring the ground-truth labels in the set domain.

6.2.1.1 Problem Formulation and Notations

Given access to good visual-quality facial still images I^S and corresponding labels y^S as the source domain $\mathcal{D}^S$, and the unknown mixed image-set I^T as the target domain $\mathcal{D}^T$, the goal is to improve the performance of $G(\cdot)$ trained with the classifier C from domain $\mathcal{D}^S$ on the domain $\mathcal{D}^T$ without knowing the target labels y^T.

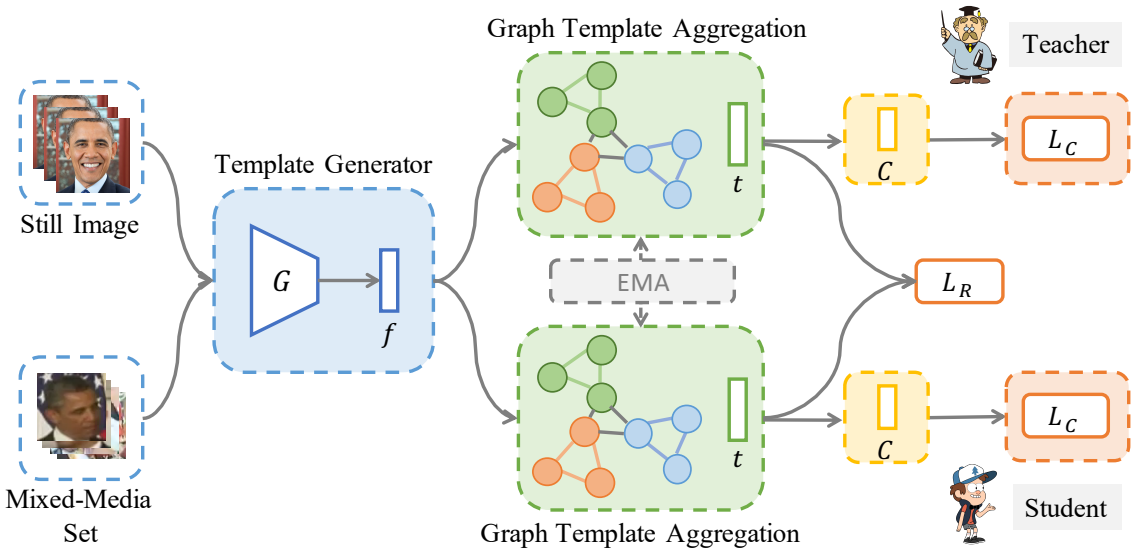


Figure 6.3: Depiction of the overview of the training framework of GTA for unsupervised domain adaptation from image-based to set-based face recognition, which consists of multiple components: graph-based template adapter $\mathcal{G}$ with supervised and unsupervised losses.

6.2.1.2 Curriculum of GTA

Successful adaptation of the image-based template generator to the set-based template generator with minimum modification required the development of a curriculum using a teacher-student learning framework [105] as per the following proposal:

Prerequisites: This self-learning framework required a prerequisite course: Both teacher and student networks should have some prior information on the face recognition from the still image domain. To obtain this prior knowledge, any still image dataset can be used as the training dataset and any network can be trained with a supervision loss serving as

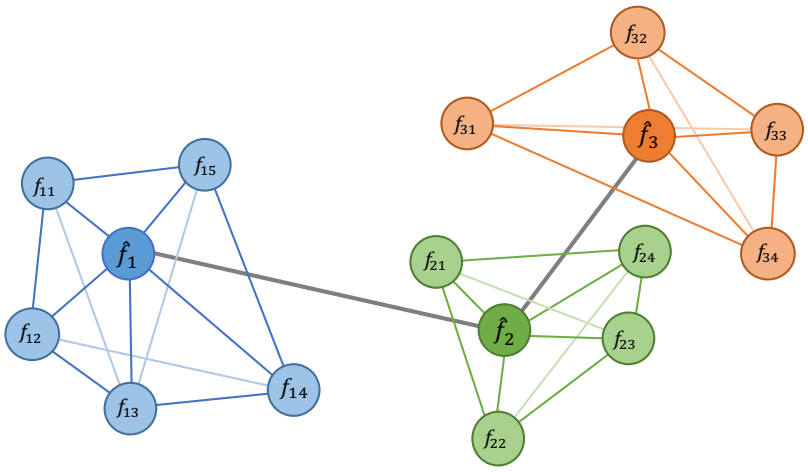


Figure 6.4: Depiction of a template graph consisting of three subgraphs, each of which represents an image-set. The node in the light color represents the features generated from an image. The node in the middle of the subgraph contains the learned higher-level template that represents the whole mixed-media set. The edge connecting each subgraphs denotes the similarity between the template features.

the pre-trained model for both networks. To make the approach more general, the softmax cross-entropy was used as this supervision loss defined below:

$$L_C = -\frac{1}{M}\sum_{i=1}^{M}\log\frac{C(f_i)}{\sum_{j=1}^{c}C(f_j)}, \tag{6.7}$$

where M is the batch size, c is the number of classes, and f is the feature vector generated from a single image. In theory, L_C can be replaced or added with other supervision losses such as ring loss [118] and cosine loss [91, 18] proposed recently and help the network to obtain better performance. This network served as one of our baselines as it only learned from the source domain.

Course 1: Build Knowledge Graph. To help the neural network learners to systematically acquire knowledge, a knowledge graph $\mathcal{G}$ was proposed to learn based on the

prior knowledge from the image domain. This graph consists of multiple local subgraphs $\mathcal{G} = \{\mathcal{G}_1, \mathcal{G}_2, \ldots, \mathcal{G}_M\}$, where M is the number of the mixed-media sets. Two subgraphs were connected with one edge if their similarity was above a threshold, depicted in Figure 6.4. Each subgraph $\mathcal{G}_i = \{\mathcal{V}_{ij}, \mathcal{E}_{ij}\}, i \in [1, M], j \in [1, N]$ represented an image-set consisting of a set of nodes $\mathcal{V}_{ij}$ containing the features generated by image-based template generator G and a corresponding edges $\mathcal{E}_{ij}$, where N represented the images in the set. To aggregate the features within a subgraph, a graph-based attention template aggregation module was used to learn more higher-level discriminative, robust, and compact feature representations by self-attention within the subgraph $\mathcal{G}_i$ (the graph index i is discarded in the following for simplicity):

The input to the aggregation module was a set of nodes $\{\mathcal{V}_j, \mathcal{E}_j\}$ containing the feature embeddings $\{f_j\} \in \mathbb{R}^{N \times l}$, where l is the feature dimension. By applying a linear transformation $W \in \mathbb{R}^{\,l \times l'}$ on the feature representation matrix, the features of the node were projected to lower-dimension l' feature space to generate compact feature representation but with more discriminative power. To interact with other learned feature representations to obtain additional sufficient express power, the concatenation operation $h(\cdot)$ was applied to the two learned representations $\{f'_j, f'_k\}$ in the subgraph $\mathcal{G}_i$. The feature attention mask $\mathbf{A}_1 \in \mathbb{R}^{2l'}$ was multiplied to concatenated feature vectors to learn which feature was important in terms of classification. Another feature attention matrix $\mathbf{A}_2 \in \mathbb{R}^{N \times N}$ was used to learn the importance of the neighborhood nodes $\mathcal{V}_k$ to the current node $\mathcal{V}_j$. The importance coefficient a_j in the attention matrix $\mathbf{A}_2$ was normalized by comparing across the nodes in the subgraph including the node itself and applying the softmax activation to summation of attended features as follows:

$$a_{jk} = \frac{\exp(\sum_{u=1}^{2l'} \mathbf{A}_1 \cdot h(f'_j, f'_k))}{\sum_{k=1}^{N} \exp(\sum_{u=1}^{2l'} \mathbf{A}_1 \cdot h(f'_j, f'_k))}. \tag{6.8}$$

After applying the attention from the neighborhood nodes, the new feature representation $\hat{f}$ was computed by the weighted average of learned features as follows:

$$\hat{f}_j = \frac{1}{N} \sum_{k}^{N} a_{jk} \cdot f'_k. \tag{6.9}$$

To optimize this graph-based template aggregation module, a loss function was proposed considering the current subgraph $\mathcal{G}_i$ and other subgraphs $\mathcal{G}_j$ by enlarging the similarity o of samples within the same graph high enough ($o >= m_1$) while decreasing the similarity of samples in different graphs low enough ($o < m_2$). Usually, the similarity score o was computed as the cosine distance in the literature. Mathematically, in the mini-batch, there are M^p positive samples with the same identity and M^n negative samples with the different identity as the sample I_i, so the objective function can be written as follows:

$$L_R = \frac{1}{M} \sum_{i=1}^{M} L^p_{R_i} + L^n_{R_i} + L^r_{R_i}, \tag{6.10}$$

$$L^p_{R_i} = \frac{1}{M^p} \sum_{j=1, j\neq i}^{M^p} [m_1 - \frac{\hat{f}_i \cdot \hat{f}_j}{||\hat{f}_i|| \cdot ||\hat{f}_j||}]_+, \tag{6.11}$$

$$L^n_{R_i} = \frac{1}{M^n} \sum_{j=1, j\neq i}^{M^n} [\frac{\hat{f}_i \cdot \hat{f}_j}{||\hat{f}_i|| \cdot ||\hat{f}_j||} - m_2]_+, \tag{6.12}$$

$$L^r_{R_i} = \frac{1}{M^p} \sum_{j=1}^{M^p} ||\hat{f}_i - \hat{f}_j||_2^2, \tag{6.13}$$

where $[\cdot]_+$ denotes a ReLU activation. The network was optimized to generate the similarities between the samples within the same set that were larger than m_1 by Equation (6.11)

and to generate the similarity between different sets that were lower than m_2 by Equation (6.12). Due to the fact that the final representation was based on the interactions between the nodes in the set, the features in each node attempted to be discriminative enough to represent that set. Therefore, the mean average of the learned features nodes was simply used to compute the final template that represents the set. To make the graph-based template aggregation module $\mathcal{G}$ work for the set-based face recognition, the weights in the image-based template generator were frozen and the only weights in the $\mathcal{G}$ were updated with the supervision signal in the still-image domain $\mathcal{D}^S$.

Course 2: Refresh with New Knowledge. Assuming there were two networks $\{G_t, G_s\}$, which consisted of the image-based template generator and the graph adapter initialized by the weights learned from the source domain $\mathcal{D}^S$, a paired teacher-student learning was proposed to help both two network learners to deepen what they had learned in the source domain $\mathcal{D}^S$ and explore the new knowledge from the target domain $\mathcal{D}^T$. The key idea in this course was that the teacher network G_t provided the supervision for the student network G_s when the student network G_s tried to explore the knowledge in the new domain.

Assuming that labeled image samples in the source domain and unlabeled images sample in the target domain were $\{I^S, y^S, I^T\}$, the teacher network G_t was updated by the supervision from the source domain $\{I^S, y^S\}$ and the feedback from the student network G_s. The student network G_s was updated by the supervision from the source domain $\mathcal{D}^S$ to keep and deepen the original knowledge and exploring the target domain with the supervision from the teacher network G_t to refresh with new knowledge. Due to unknown classes in the target domain $\mathcal{D}^T$, similarity learning was applied to update the weights in

the G_s: Instead of predicting the labels in the target domain, the similarity matrix $\mathbf{O}$ between the unlabeled data I^T was computed. Because of the fact that the samples within the same set belong to the same identity, Equation (6.11) was applied to optimize the network by the information from the matched samples. To obtain the non-matched samples, the supervision from the teacher network trained in $\mathcal{D}^S$ was used. When the similarity score o of a pair of sets predicted from the teacher network G_t was lower than a threshold m_3, this pair was defined as pseudo-non-matched set while the samples within these set were pseudo-non-matched samples. Similarly, when the similarity of a pair of sets predicted from the teacher network G_t was higher than a threshold m_4, this pair was defined as a pseudo-matched set. When the similarity score o of a pair of sets predicted from G_t was in the range of $[m_3, m_4]$, this pair was simply discarded because the predictions from teacher network G_t were not reliable. Therefore, the network G_s was updated with the objective functions (6.10)-(6.13) when it was exploring the new domain. The feedback from the student network G_s was used to compute the exponential moving average (EMA) [80] and update the knowledge in the teacher network G_t. In this manner, the knowledge in both the teacher network and the student network were refreshed by new knowledge in the target domain.

Final Examination: Template Generator and Matcher. In the testing stage, the classification layer was discarded from the well-trained models. When a set of images (no matter the size of the set) was feed-forwarded to the template generator, a unique feature representation $\hat{f}$ was generated as a template that represented this image-set. To compare

the similarity between two adapted templates $\hat{f}_i$ and $\hat{f}_j$, the cosine similarity was used:

$$o_{ij} = \frac{\hat{f}_i \cdot \hat{f}_j}{||\hat{f}_i|| \cdot ||\hat{f}_j||}. \tag{6.14}$$

This similarity score was used to generate the threshold and decide the identity of the templates in the identification scenario or accept/reject the template in the verification scenario.

6.2.2 Experiments

Two baselines selected are described as follows:

- RX50: a model trained on VGG-Face2 [9] using the ResNeXt-50 [30] architecture with softmax loss without any adaptation;
- IF-R50: a state-of-the-art face template generator [18] trained on the MS1M [94] dataset.

The IJB-A [46] and IJB-C [58] datasets were used to evaluate the set-based face recognition using the model trained with GTA and the baselines. The ROC curve in the mixed verification scenario and IET curve along with the retrieval rate in the open-set mixed identification scenario were reported to evaluate the set-based face recognition. In the mixed verification scenario, the True Acceptance Rates (TARs) at different False Acceptance Rates (FARs) in the ROC curve was reported. In the open-set mixed identification scenario, the True Positive Identification Rates (TPIRs) at different False Positive Identification Rates (FPIRs) were reported. The evaluations were performed using an off-the-shelf toolbox [101].

Table 6.4: Comparison of the face verification and identification performance of different methods on the IJB-A dataset.

Method	Category	Template Size	1:1 Mixed Verification TAR (%) @ FAR=				1:N Mixed Identification TPIR (%) @ FPIR=		Retrieval Rate (%)	
			10^{-4}	10^{-3}	10^{-2}	10^{-1}	10^{-2}	10^{-1}	Rank-1	Rank-5
MN-vc [98]	SU	2,048	-	92.0	96.2	98.9	-	-	-	-
GhostVLAD [120]	SU	128	-	93.5	97.2	99.0	88.4	95.1	97.7	99.1
GA-GANv2 [117]	SU	3,072	**94.6**	**97.3**	**98.9**	**99.5**	**93.9**	**98.2**	**99.0**	**99.5**
IF-R50 [18]	SU	512	78.1	91.0	94.0	97.0	86.5	91.1	93.4	96.1
Sohn *et al.* [77]	US	320	-	64.9	86.4	97.0	-	-	89.5	95.7
RX50	SU	2048	58.1	86.6	94.5	98.3	76.4	86.6	96.7	98.9
GTA	US	512	**80.0**	**92.1**	**96.8**	**99.0**	**86.3**	**94.3**	**97.2**	**98.9**

6.2.2.1 IJB-A: Small-scale Set-based FR

Table 6.4 summarizes the state-of-the-art performance reported in the literature and the performance of our baselines including both supervised and unsupervised methods. The best results in both supervised (SU) and unsupervised (US) methods are marked in bold. The supervised method [117] leveraged the supervision in the target domain with extreme data augmentations and multiple networks. Compared to them, our method was light-weight and did not require multiple feature extraction networks as well as access to labels in the target domain. Compared with the baseline RX50, the proposed method reduced the 2,048 dimensions of the features output from the baselines to 512-dimensions (75% less) but significantly improved the performance in the set-based face recognition (37.7%

Table 6.5: Comparison of the face verification and identification performance of different methods on the IJB-B dataset.

Method	Category	Template Size	1:1 Mixed Verification TAR (%) @ FAR=					1:N Mixed Identification TPIR (%) @ FPIR=		Retrieval Rate (%)	
			10^{-5}	10^{-4}	10^{-3}	10^{-2}	10^{-1}	10^{-2}	10^{-1}	Rank-1	Rank-5
VGG-FACE [9]		2,048	34.2	53.5	71.1	85.0	-	42.9	63.5	75.2	84.3
VGG-FACE2 [9]		2,048	64.7	78.4	87.8	93.8	97.5	70.1	82.4	88.6	93.6
MN-vc [98]	SU	2,048	70.8	83.1	90.9	95.8	98.5	-	-	-	-
CN [97]		2,048	-	84.1	**93.0**	**97.2**	**99.5**	-	-	-	-
Ghost-VLAD [120]		256	**74.1**	**85.3**	92.5	96.3	-	76.4	**88.5**	**92.1**	**95.5**
IF-R50 [18]		512	75.4	84.9	90.1	93.7	96.8	**78.2**	85.8	88.6	92.1
RX50	SU	2,048	53.7	73.5	88.0	95.9	99.1	62.4	80.1	88.1	94.5
GTA	US	512	**66.9**	**82.0**	**91.0**	**95.8**	**98.5**	**70.7**	**84.7**	**90.4**	**94.8**

improved TAR at FAR is 10^{-4} using RX50 backbone).

6.2.2.2 IJB-B: Medium-scaled Set-based FR

Table 6.5 summarizes the state-of-the-art performance reported in the recent literature and the performances of our baselines. The best results in both supervised (SU) and unsupervised (US) methods are marked in bold. Because there are no unsupervised methods reported in the IJB-B dataset, we only compared with supervised methods. Similar to the results obtained in the previous experiment, it was observed that GTA improved the performance of the baseline model RX50 in both 1:1 mixed verification and 1: N mixed identification protocols. Specifically, in the 1:1 mixed verification scenario, GTA improved

Table 6.6: Comparison of the face verification and identification performance of different methods on the IJB-C dataset.

Method	Category	1:1 Mixed Verification TAR (%) @ FAR=							1:N Mixed Identification TPIR (%) @ FPIR=			Retrieval Rate (%)	
		10^{-7}	10^{-6}	10^{-5}	10^{-4}	10^{-3}	10^{-2}	10^{-1}	10^{-3}	10^{-2}	10^{-1}	Rank-1	Rank-5
GOTS [58]		3.0	3.0	6.6	14.7	33.0	62.0	80.9	2.7	5.8	15.6	37.9	52.5
FaceNet [70]		15.0	21.0	33.3	48.7	66.5	81.7	92.5	20.6	32.4	51.0	69.2	79.0
VGGFace [60]	SU	20.0	32.2	43.7	59.8	74.8	87.1	95.6	26.2	45.1	62.8	78.6	86.0
MN-vc [98]		-	-	-	86.2	92.7	96.8	98.9	-	-	-	-	-
CN [97]		-	-	-	**88.5**	**94.7**	**98.3**	**99.8**	-	-	-	-	-
RX50	SU	20.2	42.0	59.2	75.7	89.0	96.4	99.2	42.5	57.9	76.6	**87.9**	**93.9**
GTA	US	**21.4**	**51.1**	**68.0**	**82.3**	**91.7**	**96.6**	**99.0**	**49.6**	**68.8**	**82.0**	**89.6**	**94.7**

TAR of RX50 by 24.6% and 11.6% at FAR equal to 10^{-5} and 10^{-4}, receptively. In the 1: N mixed identification scenario, GTA improved TPIR of RX50 by 13.0% and 9.3% at FPIR are 10^{-2} and 10^{-1}, receptively. In addition, it improved 2.3% in term of Rank-1 accuracy. These results indicated that GTA learned to generate higher-level discriminative templates from a mixed-media set.

6.2.2.3 IJB-C: Large-scaled Set-based FR

Table 6.6 summarizes the state-of-the-art performance reported in recent literature and the performance of our baselines. The best results in both supervised (SU) and unsupervised (US) methods are marked in bold. In the 1:1 Mixed Verification scenario, GTA improved TARs of the baseline RX50 by 21.7%, 14.9%, and 8.7% when FAR equals to 10^{-6}, 10^{-5},

and 10^{-4}, respectively. In the 1: N Mixed Identification scenario, GTA improved TPIR by 16.7% and 18.8% when FPIR are 10^{-3} and 10^{-2}, respectively. In addition, GTA improved the Rank-1 accuracy of RX50 by 1.7%. Using the backbone network RX50, GTA achieved the state-of-the-art Rank-1 accuracy reported on this dataset.

Chapter 7

Conclusions and Future Work

7.1 Conclusions

This dissertation focused on the pipeline of face recognition. The primary contributions were achieved by developing different modules/algorithms to build a robust face recognition system in the presence of variances of pose, expression, and occlusion. Addressing the existing challenges, a series of algorithms were proposed to improve the performance of each objective.

A well-designed 3D-aided 2D face recognition system was developed that was robust to pose variations as large as 90° using deep learning technology. Detailed experiments were conducted on UHDB31 and IJB-A to demonstrate that the proposed system was robust to theses variations of pose, and it outperformed existing 2D face recognition systems such as VGG face descriptor, FaceNet, and a commercial face recognition software. A

light-weight, maintainable, scalable, generalizable, and extendable face recognition evaluation toolbox was designed and implemented in Python that supports both online and offline evaluation to benefit the biometrics research community and to accelerate biometrics-related research. FaRE was designed to evaluate general FR systems, which consisted of commonly used evaluation metrics functions, closed-set, and open-set FR datasets.

ERF demonstrated its efficiency in detecting the landmarks on frontal images, while JFA showed that the pose estimation and face alignment tasks were jointly learnable in the same framework. An initialization method for face alignment and a learning procedure using an ensemble of random ferns to learn local features were proposed. ERF was constructed in a cascade manner to extract the local features. By using the proposed methods, improved performance was achieved compared with the state-of-the-art methods. A joint hierarchical head pose estimation and face alignment learning system was proposed by exploration of the global and local CNN features. Based on the coarse-to-fine manner, the global CNN features were used to estimate face attributes such as head pose and facial components while local CNN features were used to refine the shape in the cascade. The experiments demonstrated that JFA outperformed conventional head pose estimation on the challenging head pose estimation task.

Comparative studies using E2FAR as the baseline were performed, which demonstrated that feature aggregation from different layers was a key-point to train better neural networks for 3D face reconstruction. FR-FAN was proposed and a significant improvement was observed compared to ResNet-101 and E2FAR on our synthetic validation set. Extensive experiments demonstrated that our model exhibited improved performance when compared to the existing state-of-the-art algorithms on BU-3DFE and JNU-3D datasets and was robust to pose, illumination, and expression variations.

Improving the face recognition performance in the presence of variance of the pose, expression, and occlusion required the proposal of OREO which contained an attention mechanism, a balancing sampling strategy, and a similarity-based loss function. Extensive experiments demonstrated that OREO achieved state-of-the-art results on datasets with both image-based and set-based evaluation protocols. Through ablation studies and qualitative results, we demonstrated the impact of individual components to the final performance and provided an effective way to better understand the representations learned by the proposed method. Regarding the mixed-media set-base face recognition, an unsupervised graph-based template adaptation training framework named GTA was developed. A graph-based template aggregation module is proposed as an add-on to the image-based template generator, adapting the knowledge from the still image to the mixed-media set domain. The proposed graph-based template aggregation module helped to generate compact feature embedding considering the relationships within the subgraph. The proposed unsupervised template adaptation helped to explore the new knowledge in the unknown target domain. Extensive experiments indicated that our method achieved state-of-the-art performance in both set-based face identification and verification scenarios.

7.2 Future Work

> As our circle of knowledge expands, so does the circumference of darkness surrounding it.
>
> — Albert Einstein

Objective 1: 3D-aided Face Recognition System. The developed system consists of multiple modules including detection, alignment, reconstruction, and template generation. The limitations along with the future work for this objective are the following:

(i) Computational Acceleration: There are more than ten networks used in the current system to generate the final template, which limits the computational efficiency and inference time. Ranjan *et al.* [64] demonstrated that a single neural network could be designed for multiple tasks, which helps to fully use the neurons in the networks, reduce the complexity, and speed up the processing time. In addition, if all operations were implemented using CUDA, the time copying data between the GPU and CPU would be reduced resulting in a corresponding reduction in the inference time.

(ii) Automatic Module Registration: When adding a new module to the system, the system requires that this module be imported in the main process. If an automatic module registration was used, this step could be avoided as it would allow an administrator to register all operations or modules implemented in the system. This suggestion would help if there are many modules added to the system.

(iii) Module Updates: The multiple modules are developed by different researchers for

different purposes. Updating a single module might influence the downstream modules. One approach to improve the cooperation in the developing team is that the members should learn to use the development tools such as Git and use the Continuous Integration (CI) to integrate code into a shared repository and test it. It would help to improve the code quality and to assess the current face recognition performance in the test datasets.

Objective 2: 2D Landmark Detection. The algorithms designed in this dissertation localize the landmarks on 2D facial images. The limitations along with the future work for this objective are the following:

(i) 3D Supervision: The performance of the landmark detectors developed as part of this thesis cannot match the current performance of heatmap-based landmark detectors. However, in some cases, heatmap-based landmark detectors generate non-reasonable results that lead to face recognition failures. To further improve the current landmark detector, one might consider adding 3D facial shape constraints in the network to ensure that the predictions of the network are not only based on the response from 2D images but are also congruent with the 3D structure.

(ii) Joint Learning: Face detection and landmark detection tasks share the same feature map. Currently, a trend exists to design anchor-free detectors. To obtain the high response of the face and facial landmarks, global features are used to detect the face and local features are used to detect the landmarks and further reduce the false positives of face detection. It is a trend to use points to predict the human and localize the landmarks, which would further accelerate the face recognition processing speed.

(iii) Network Acceleration: The model size of 2D heatmap-based landmark detector usually is more than 200 MB, which limits their deployment on edge devices. Binary or ternary networks can be explored to address this issue. By making the weights binary or ternary, a significant reduction in the computational burden in deployment is possible, yet still maintain comparable performance with the original model.

Objective 3: 3D Face Reconstruction. While the developed algorithm introduced in this thesis achieved state-of-the-art results in four publicly available datasets, there is still a room for improvement in terms of precision and inference time. The limitations along with the future work for this objective are the following:

(i) Non-linear Model: The linear assumption is a very strong constraint in the 3D morphable model, which does not generalize well beyond the underlying model's restricted low-dimensional subspace. A non-linear model can be explored in the future to generate a 3D facial point cloud. This non-linear mapping from 2D images to 3D point cloud can be represented by the multiple networks such as auto-encoder, encoder-decoder to explore the latent space of the face.

(ii) Network Acceleration: Similar to the landmark detection task, the backbone networks used in the 3D face reconstruction can be designed with a computational speed consideration. By designing efficient convolutional neural networks, it would be easier to make a real-time mobile and embedded vision applications.

Objective 4: 2D Face Recognition. The algorithms developed in this dissertation intended to improve the still-image-based face recognition in the presence of occlusion, and

also improve the mixed-media set-based face recognition in an unsupervised manner. The limitations along with the future work for this objective are the following:

(i) Occlusion-aware Generative Adversarial Face Recognition: The attention mechanism used in OREO pointed to the power of finding the occluded facial region on in-the-wild images. Consequently, a follow-up work may focus on the image generation for producing an occluded face region using a generative adversarial network with an attention mechanism. The generative network would learn the distribution mapping from the occluded images to the non-occluded images. An additional identity classifier could be added to ensure that the generated images maintain the same identity.

(ii) Graph Face Modeling: A graph can be used to explore the relationship within a face region. A face can be described by a global feature vector generated from the whole face and several local features from the local parts, which can be further represented by a graph. The graph neural network can be deployed to generate the template for a single facial image. In addition, the template matcher can be further improved by graph-based matching, which requires not only to match the global representations but also to measure the similarity between two local representations.

(iii) Semi-supervised Learning: The improvement offered by GTA can be further improved by exploring the mass of unsupervised data with the graph-based clustering method and using limited annotated data. There is a large unknown data space without labels including still images and videos. One possible solution is to develop a graph-based semi-supervised clustering algorithm to find novel classes and involve

a human operator to provide limited annotations on these novel classes. Using the supervision signal from the identity labels would help to update the parameters in the template generator and improve the discriminative power.

(iv) Adversarial Attack: In the face verification scenario, occlusions by facial accessories would easily fool the system to output false identity when the input pair of images belong to the same identity [27]. In this scenario, OREO might achieve better performance because the system will focus on the non-occluded face region. In the face identification scenario, there is limited literature reporting on how to fool the system to recognize someone with a different identity using occlusion. In a deployment where the system would be designed following an open-set identification setting, the result could be that the system might reject the adversarial sample with accessories when compared with the gallery samples. The reason is that the occlusion will significantly decrease the similarity compared with changing pixels on the face. However, it is still a very interesting topic to improve face recognition system security to be robust to adversarial attacks.

Bibliography

[1] B. Amos, B. Ludwiczuk, and S. Mahadev. OpenFace: A general-purpose face recognition library with mobile applications. Technical Report CMU-CS-16-118, CMU School of Computer Science, Pittsburgh, PA, 2016.

[2] A. Anjos, L. E. Shafey, R. Wallace, M. Gunther, C. McCool, and S. Marcel. Bob: a free signal processing and machine learning toolbox for researchers. In *Proc. ACM Conference on Multimedia Systems*, pages 1–4, Nara, Japan, 2012.

[3] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic. Robust discriminative response map fitting with constrained local models. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 3444–3451, Portland, OR, 2013.

[4] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic. Incremental face alignment in the wild. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 1859–1866, Columbus, OH, 2014.

[5] A. Bansal, A. Nanduri, C. D. Castillo, R. Ranjan, and R. Chellappa. UMDFaces: an annotated face dataset for training deep networks. In *Proc. IEEE International Joint Conference on Biometrics*, pages 464–473, Denver, CO, 2017.

[6] J. Bao, D. Chen, F. Wen, H. Li, and G. Hua. Towards open-set identity preserving face synthesis. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 6713–6722, Salt Lake City, UT, 2018.

[7] X. P. Burgos-Artizzu, P. Perona, and P. Dollár. Robust face landmark estimation under occlusion. In *Proc. IEEE International Conference on Computer Vision*, pages 1513–1520, Sydney, Australia, 2013.

[8] K. Cao, Y. Rong, C. Li, X. Tang, and C. C. Loy. Pose-robust face recognition via deep residual equivariant mapping. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 5187–5196, Salt Lake City, UT, 2018.

[9] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman. VGGFace2: a dataset for recognising faces across pose and age. In *Proc. IEEE Conference on Automatic Face and Gesture Recognition*, pages 67–74, Xi'an, China, 2018.

[10] X. Cao, Y. Wei, F. Wen, and J. Sun. Face alignment by explicit shape regression. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 2887–2894, Providence, RI, 2012.

[11] F. J. Chang, A. Tuan Tran, T. Hassner, I. Masi, R. Nevatia, and G. Medioni. ExpNet: landmark-free, deep, 3D facial expressions. In *Proc. IEEE International Conference on Automatic Face and Gesture Recognition*, pages 122–129, Xi'an, China, 2018.

[12] T. Chen, M. Li, Y. Li, M. Lin, N. Wang, M. Wang, T. Xiao, B. Xu, C. Zhang, and Z. Zhang. MXNet: a flexible and efficient machine learning library for heterogeneous distributed systems. In *Proc. Neural Information Processing Systems, Workshop on Machine Learning Systems*, pages 1–6, Montreal, Canada, 2015.

[13] L. Cheng, J. Wang, Y. Gong, and Q. Hou. Robust deep Auto-Encoder for occluded face recognition. In *Proc. ACM Multimedia Conference*, pages 1099–1102, Queensland, Australia, 2015.

[14] B. Chu, S. Romdhani, and L. Chen. 3D-aided face recognition robust to expression and pose variations. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 1907–1914, Columbus, OH, 2014.

[15] C. Cortes and V. Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[16] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 886–893, San Diego, CA, 2005.

[17] J. Deng, S. Cheng, N. Xue, Y. Zhou, and S. Zafeiriou. UV-GAN: Adversarial facial UV map completion for pose-invariant face recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 7093–7102, Salt Lake City, UT, 2018.

[18] J. Deng, J. Guo, and S. Zafeiriou. ArcFace: Additive angular margin loss for deep face recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–10, Long Beach, CA, 2019.

[19] P. Dou and I. A. Kakadiaris. Multi-view 3D face reconstruction with deep recurrent neural networks. In *Proc. International Joint Conference on Biometrics*, pages 483–492, Denver, CO, 2017.

[20] P. Dou, S. K. Shah, and I. A. Kakadiaris. End-to-end 3D face reconstruction with deep neural networks. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 5908–5917, Honolulu, HI, 2017.

[21] B. Egger, S. Schonborn, A. Schneider, A. Kortylewski, A. Morel-Forster, C. Blumer, and T. Vetter. Occlusion-aware 3D morphable models and an illumination prior for face image analysis. *International Journal of Computer Vision*, 126(12):1269–1287, 2018.

[22] N. Erdogmus and J. Dugelay. 3D assisted face recognition: dealing with expression variations. *IEEE Transactions on Information Forensics and Security*, 9(5):826–838, 2014.

[23] S. Farfade, M. Saberian, and L. Li. Multi-view face detection using deep convolutional neural networks. In *Proc. International Conference on Multimedia Retrieval*, pages 643–650, Shanghai, China, 2015.

[24] Y. Feng, F. Wu, X. Shao, Y. Wang, and X. Zhou. Joint 3D face reconstruction and dense alignment with position map regression network. In *Proc. European Conference in Computer Vision*, pages 1–18, Munich, Germany, 2018.

[25] Y. Fu, X. Wu, Y. Wen, and Y. Xiang. Efficient locality-constrained occlusion coding for face recognition. *Neurocomputing*, 260(1):104–111, 2017.

[26] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 580–587, Columbus, OH, 2014.

[27] G. Goswami, N. Ratha, A. Agarwal, R. Singh, and M. Vatsa. Unravelling robustness of deep learning based face recognition against adversarial attacks. In *Proc. AAAI Conference on Artificial Intelligence*, pages 6829–6836, New Orleans, LA, 2018.

[28] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao. MS-Celeb-1M: A dataset and benchmark for large-scale face recognition. In *Proc. European Conference on Computer Vision*, pages 87–102, Amsterdam, The Netherlands, 2016.

[29] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask R-CNN. In *Proc. IEEE International Conference on Computer Vision*, pages 2980–2988, Venice, Italy, 2017.

[30] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, Las Vegas, NV, 2016.

[31] K. He, X. Zhang, S. Ren, and J. Sun. Identity mappings in deep residual networks. In *Proc. European Conference in Computer Vision*, pages 1–15, Amsterdam, Netherlands, 2016.

[32] L. He, H. Li, Q. Zhang, and Z. Sun. Dynamic feature learning for partial face recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 7054–7063, Salt Lake City, UT, 2018.

[33] P. Hu and D. Ramanan. Finding tiny faces. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 951–959, Honolulu, HI, 2017.

[34] G. Huang, Z. Liu, V. der Maaten Laurens, and K. Q. Weinberger. Densely connected convolutional networks. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 4700–4708, Honolulu, HI, 2017.

[35] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. In *Proc. Workshop on Faces in Real-Life Images: Detection, Alignment, and Recognition*, pages 1–11, Marseille, France, 2008.

[36] R. Huang, S. Zhang, T. Li, and R. He. Beyond face rotation: global and local perception gan for photorealistic and identity preserving frontal view synthesis. In *Proc. IEEE International Conference on Computer Vision*, pages 2439–2448, Venice, Italy, 2017.

[37] A. S. Jackson, A. Bulat, V. Argyriou, and G. Tzimiropoulos. Large pose 3D face reconstruction from a single image via direct volumetric CNN regression. In *Proc. IEEE International Conference on Computer Vision*, pages 1031–1039, Venice, Italy, 2017.

[38] H. Jiang and E. Learned-Miller. Face detection with the faster R-CNN. In *Proc. IEEE International Conference on Automatic Face & Gesture Recognition*, pages 650–657, Washington, DC, 2017.

[39] X. Jin and X. Tan. Face alignment in-the-wild: a survey. *Computer Vision and Image Understanding*, 162:1–22, 2017.

[40] A. Jourabloo and X. Liu. Large-pose face alignment via CNN-based dense 3D model fitting. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 4188–4196, Las Vegas, NV, 2016.

[41] I. A. Kakadiaris, G. Toderici, G. Evangelopoulos, G. Passalis, D. Chu, X. Zhao, S. K. Shah, and T. Theoharis. 3D-2D face recognition with pose-illumination normalization. *Computer Vision and Image Understanding*, 154:137–151, 2017.

[42] B.-N. Kang, Y. Kim, and D. Kim. Pairwise relational networks for face recognition. In *Proc. European Conference in Computer Vision*, pages 1–18, Munich, Germany, 2018.

[43] V. Kazemi and J. Sullivan. One millisecond face alignment with an ensemble of regression trees. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 1867–1874, Columbus, OH, 2014.

[44] I. Kemelmacher-Shlizerman, S. M. Seitz, D. Miller, and E. Brossard. The MegaFace benchmark: 1 million faces for recognition at scale. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–10, Las Vegas, NV, 2016.

[45] D. E. King. Dlib-ml: a machine learning toolkit. *Journal of Machine Learning Research*, 10(1):1755–1758, 2009.

[46] B. F. Klare, B. Klein, E. Taborsky, A. Blanton, J. Cheney, K. Allen, P. Grother, A. Mah, M. Burge, and A. K. Jain. Pushing the frontiers of unconstrained face detection and recognition: IARPA Janus benchmark A. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 1931–1939, Boston, MA, 2015.

[47] P. Koppen, Z.-H. Feng, J. Kittler, M. Awais, W. Christmas, X.-J. Wu, and H.-F. Yin. Gaussian mixture 3D morphable face model. *Pattern Recognition*, 74:617–628, 2018.

[48] H. A. Le and I. A. Kakadiaris. UHDB31: a dataset for better understanding face recognition across pose and illumination variation. In *Proc. IEEE International Conference on Computer Vision Workshops*, pages 2555–2563, Venice, Italy, 2017.

[49] V. Le, J. Brandt, Z. Lin, L. Bourdev, and T. S. Huang. Interactive facial feature localization. In *Proc. European Conference on Computer Vision*, pages 679–692, Firenze, Italy, 2012.

[50] D. Lee, H. Park, and C. D. Yoo. Face alignment using cascade gaussian process regression trees. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 4204–4212, Boston, MA, 2015.

[51] Y. Li, B. Sun, T. Wu, and Y. Wang. Face detection with end-to-end integration of a ConvNet and a 3D model. In *Proc. European Conference on Computer Vision*, pages 420–436, Amsterdam, Netherlands, 2016.

[52] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Fu, and A. C. Berg. SSD: single shot multibox detector. In *Proc. European Conference on Computer Vision*, pages 21–37, Amsterdam, Netherlands, 2016.

[53] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song. SphereFace: deep hypersphere embedding for face recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 212–220, Honolulu, HI, 2017.

[54] Z. Liu, P. Luo, X. Wang, and T. Xiaoou. Deep learning face attributes in the wild. In *Proc. International Conference on Computer Vision*, pages 3730–3738, Santiago, Chile, 2015.

[55] D. G. Lowe. Object recognition from local scale-invariant features. In *Proc. International Conference on Computer Vision*, pages 1–8, Kerkyra, Greece.

[56] I. Masi, S. Rawls, G. Medioni, and P. Natarajan. Pose-aware face recognition in the wild. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 4838–4846, Las Vegas, NV, 2016.

[57] M. Mathias, R. Benenson, M. Pedersoli, and L. V. Gool. Face detection without bells and whistles. In *Proc. European Conference on Computer Vision*, pages 720–735, Zurich, Switzerland, 2014.

[58] B. Maze, J. Adams, J. A. Duncan, N. Kalka, T. Miller, C. Otto, A. K. Jain, W. T. Niggel, J. Anderson, J. Cheney, and P. Grother. IARPA Janus benchmark-C: face dataset and protocol. In *Proc. IEEE International Conference on Biometrics*, pages 158–165, Queensland, Australia, 2018.

[59] R. Min, A. Hadid, and J.-L. Dugelay. Improving the recognition of faces occluded by facial accessories. In *Proc. IEEE International Conference on Automatic Face & Gesture Recognition and Workshops*, pages 442–447, Santa Barbara, CA, 2011.

[60] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *Proc. British Machine Vision Conference*, pages 1–12, Swansea, UK, 2015.

[61] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[62] P. Phillips, W. Scruggs, A. O'Toole, P. Flynn, K. Bowyer, C. Schott, and M. Sharpe. FRVT 2006 and ICE 2006 large-scale experimental results. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(5):831–846, 2010.

[63] M. Piotraschke and V. Blanz. Automated 3D face reconstruction from multiple images using quality measures. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 3418–3427, Las Vegas, NV, USA, 2016.

[64] R. Ranjan, V. M. Patel, and R. Chellappa. HyperFace: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(1):121–135, 2019.

[65] J. Redmon and A. Farhadi. YOLO9000: better, faster, stronger. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 7263–7271, Honolulu, HI, 2017.

[66] S. Ren, X. Cao, Y. Wei, and J. Sun. Face alignment at 3,000 FPS via regressing local binary features. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 1685–1692, Columbus, OH, 2014.

[67] J. Roth, Y. Tong, and X. Liu. Adaptive 3D face reconstruction from unconstrained photo collections. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 4197–4206, Las Vegas, NV, 2016.

[68] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. 300 faces in-the-wild challenge: the first facial landmark localization challenge. In *Proc. IEEE International Conference on Computer Vision Workshops*, pages 397–403, Sydney, Australia, 2013.

[69] N. Sarafianos, X. Xu, and I. A. Kakadiaris. Deep imbalanced attribute classification using visual attention aggregation. In *Proc. European Conference on Computer Vision*, pages 1–18, Munich, Germany, 2018.

[70] F. Schroff, D. Kalenichenko, and J. Philbin. FaceNet: a unified embedding for face recognition and clustering. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 815–823, Boston, MA, 2015.

[71] M. Sela, E. Richardson, and R. Kimmel. Unrestricted facial geometry reconstruction using image-to-image translation. In *Proc. International Conference on Computer Vision*, pages 1576–1585, Venice, Italy, 2017.

[72] 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 *Proc. IEEE Winter Conference on Applications of Computer Vision*, pages 1–9, Lake Placid, NY, USA, 2016.

[73] L. Shi, X. Xu, and I. A. Kakadiaris. SSFD: a face detector via a single-scale feature map. In *Proc. IEEE International Conference on Biometrics: Theory Applications and Systems*, pages 1–8, Los Angeles, CA, USA, 2018.

[74] L. Shi, X. Xu, and I. A. Kakadiaris. A simple and effective single stage face detector. In *Proc. International Conference On Biometrics*, pages 1–8, Crete, Greece, 2019.

[75] L. Shi, X. Xu, and I. A. Kakadiaris. Smoothed attention network for single stage face detector. In *Proc. International Conference On Biometrics*, pages 1–8, Crete, Greece, 2019.

[76] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, Y. Chen, T. Lillicrap, F. Hui, L. Sifre, G. van den Driessche, T. Graepel, and D. Hassabis. Mastering the game of Go without human knowledge. *Nature*, 550:354–359, 2017.

[77] K. Sohn, S. Liu, G. Zhong, X. Yu, M. H. Yang, and M. Chandraker. Unsupervised domain adaptation for face recognition in unlabeled videos. In *Proc. International Conference on Computer Vision*, pages 5917–5925, Venice, Italy, 2017.

[78] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. DeepFace: closing the gap to human-level performance in face verification. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 1701–1708, Columbus, OH, 2014.

[79] X. Tang, D. K. Du, Z. He, and J. Liu. PyramidBox: a context-assisted single shot face detector. In *Proc. European Conference in Computer Vision*, pages 1–21, Munich, Germany, 2018.

[80] A. Tarvainen and H. Valpola. Mean teachers are better role models: weight-averaged consistency targets improve semi-supervised deep learning results. In *Proc. Neural Information Procceessing Systems*, pages 1–10, Long Beach, CA, 2017.

[81] A. Tewari, M. Zollhöfer, P. Garrido, F. Bernard, H. Kim, P. Pérez, and C. Theobalt. Self-supervised multi-level face model learning for monocular reconstruction at over 250 Hz. In *Proc. Computer Vision and Pattern Recognition*, pages 2546–2559, Salt Lake City, UT, 2018.

[82] A. Tewari, M. Zollöfer, H. Kim, P. Garrido, F. Bernard, P. Perez, and T. Christian. MoFA: model-based deep convolutional face autoencoder for unsupervised monocular reconstruction. In *Proc. International Conference on Computer Vision*, pages 3735–3744, Venice, Italy, 2017.

[83] G. Toderici, G. Evangelopoulos, T. Fang, T. Theoharis, and I. A. Kakadiaris. UHDB11 database for 3D-2D face recognition. In *Proc.* 6^{th} *Pacific-Rim Symposium on Image and Video Technology*, pages 73–86, Guanajuato, Mexico, 2013.

[84] A. Tran, T. Hassner, I. Masi, and G. Medioni. Regressing robust and discriminative 3D morphable models with a very deep neural network. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–10, Honolulu, HI, 2017.

[85] A. Tran, T. Hassner, I. Masi, E. Paz, Y. Nirkin, and G. Medioni. Extreme 3D face reconstruction: seeing through occlusions. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 3935–3944, Salt Lake City, UT, 2018.

[86] L. Tran, X. Yin, and X. Liu. Disentangled representation learning GAN for pose-invariant face recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 1415–1424, Honolulu, HI, 2017.

[87] 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 *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 4177–4187, Las Vegas, NV, 2016.

[88] G. Tzimiropoulos. Project-out cascaded regression with an application to face alignment. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 3659–3667, Boston, MA, 2015.

[89] G. Tzimiropoulos and M. Pantic. Gauss-Newton deformable part models for face alignment in-the-wild. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 1851–1858, Columbus, OH, 2014.

[90] A. Wagner, J. Wright, A. Ganesh, Z. Zhou, H. Mobahi, and Y. Ma. Toward a practical face recognition system: Robust alignment and illumination by sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(2):372–386, 2012.

[91] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu. CosFace: Large margin cosine loss for deep face recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–10, Salt Lake City, UT, 2018.

[92] R. Wang, J. Lu, and Y.-P. Tan. Robust point set matching for partial face recognition. *IEEE Transactions on Image Processing*, 25(3):1163–1176, 2016.

[93] Y. Wang, D. Gong, Z. Zhou, X. Ji, H. Wang, Z. Li, W. Liu, and T. Zhang. Orthogonal deep features decomposition for age-invariant face recognition. In *Proc. European Conference in Computer Vision*, pages 1–16, Munich, Germany, 2018.

[94] Y. Wen, K. Zhang, and Zhifeng Li and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *Proc. European Conference on Computer Vision*, pages 499–515, Amsterdam, Netherlands, 2016.

[95] 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 *Proc. IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 90–98, Honolulu, HI, 2017.

[96] Y. Wu, S. K. Shah, and I. A. Kakadiaris. GoDP: Globally optimized dual pathway deep network architecture for facial landmark localization in-the-wild. *Image and Vision Computing*, 73(1):1–16, 2017.

[97] W. Xie, L. Shen, and A. Zisserman. Comparator networks. In *Proc. European Conference on Computer Vision*, pages 1–15, Munich, Germany, 2018.

[98] W. Xie and A. Zisserman. Multicolumn networks for face recognition. In *Proc. British Machine Vision Conference*, pages 1–12, Northumbria University, United Kingdom, 2018.

[99] X. Xiong and F. De la Torre. Supervised descent method and its applications to face alignment. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 532–539, Portland, OR, 2013.

[100] X. Xu and I. A. Kakadiaris. Joint head pose estimation and face alignment framework using global and local CNN features. In *Proc. IEEE Conference on Automatic Face & Gesture Recognition*, pages 642–649, Washington, DC, 2017.

[101] X. Xu and I. A. Kakadiaris. Open source face recognition performance evaluation package. In *Proc. International Conference on Image Processing*, pages 1–5, Taipei, Taiwan, 2019.

[102] X. Xu, H. Le, P. Dou, Y. Wu, and I. A. Kakadiaris. Evaluation of 3D-aided pose invariant 2D face recognition system. In *Proc. International Joint Conference on Biometrics*, pages 446–455, Denver, CO, 2017.

[103] X. Xu, H. Le, and I. A. Kakadiaris. On the importance of feature aggregation for face reconstruction. In *Proc. Winter Conference on Applications of Computer Vision*, pages 922–931, Waikoloa Village, HI, 2019.

[104] X. Xu, S. Shah, and I. A. Kakadiaris. Face alignment via an ensemble of random ferns. In *Proc. IEEE International Conference on Identity, Security and Behavior Analysis*, pages 1–8, Sendai, Japan, 2016.

[105] X. Xu, Z. Xiong, R. Venkatesan, G. Swaminathan, and O. Majumder. dSNE: domain adaptation using stochastic neighborhood embedding. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–12, Long Beach, CA, 2019.

[106] H. Yang, W. Mou, Y. Zhang, I. Patras, H. Gunes, and P. Robinson. Face alignment assisted by head pose estimation. In *Proc. British Machine Vision Conference*, pages 1–13, Swansea, UK, 2015.

[107] H. Yang, R. Zhang, and P. Robinson. Human and sheep facial landmarks localisation by triplet interpolated features. In *Proc. IEEE Winter Conference on Applications of Computer Vision*, pages 1–8, Lake Placid, NY, 2016.

[108] L. Yin, X. Wei, Y. Sun, J. Wang, and M. Rosato. A 3D facial expression database for facial behavior research. In *Proc. IEEE International Conference on Automatic Face and Gesture Recognition*, pages 211–216, Southampton, UK, 2006.

[109] X. Yin and X. Liu. Multi-task convolutional neural network for pose-invariant face recognition. *IEEE Transaction on Image Processing*, 27(2):964–975, 2018.

[110] X. Yin, X. Yu, K. Sohn, X. Liu, and M. Chandraker. Towards large-pose face frontalization in the wild. In *Proc. International Conference on Computer Vision*, pages 4010–4019, Venice, Italy, 2017.

[111] Y. F. Yu, D. Q. Dai, C. X. Ren, and K. K. Huang. Discriminative multi-scale sparse coding for single-sample face recognition with occlusion. *Pattern Recognition*, 66(1):302–312, 2017.

[112] S. Zafeiriou, C. Zhang, and Z. Zhang. A survey on face detection in the wild: past, present and future. *Computer Vision and Image Understanding*, 138(1):1–24, 2015.

[113] J. Zhang, S. Shan, M. Kan, and X. Chen. Coarse-to-fine auto-encoder networks for real-time face alignment. In *Proc. European Conference on Computer Vision*, pages 1–16, Zurich, Switzerland, 2014.

[114] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016.

[115] F. Zhao, J. Feng, J. Zhao, W. Yang, and S. Yan. Robust LSTM-autoencoders for face de-occlusion in the wild. *IEEE Transactions on Image Processing*, 27(2):778–790, 2018.

[116] J. Zhao, Y. Cheng, Y. Xu, L. Xiong, J. Li, F. Zhao, K. Jayashree, S. Pranata, S. Shen, J. Xing, S. Yan, and J. Feng. Towards pose invariant face recognition in the wild. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 2207–2216, Salt Lake City, UT, 2018.

[117] J. Zhao, L. Xiong, J. Li, J. Xing, S. Yan, and J. Feng. 3D-aided dual-agent GANs for unconstrained face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8828(1):1–14, 2018.

[118] Y. Zheng, D. K. Pal, and M. Savvides. Ring loss: Convex feature normalization for face recognition. In *Proc. Conference on Computer Vision and Pattern Recognition*, pages 5089–5097, Salt Lake City, UT, 2018.

[119] Y. Zhong, J. Chen, and B. Huang. Toward end-to-end face recognition through alignment learning. *IEEE Signal Processing Letters*, 24(8):1213–1217, 2017.

[120] Y. Zhong and A. Zisserman. GhostVLAD for set-based face recognition. In *Proc. Asian Conference of Computer Vision*, pages 1–16, Perth Western, Australia, 2018.

[121] E. Zhou, Z. Cao, and J. Sun. GridFace: face rectification via learning local homography transformations. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 3–20, Salt Lake City, UT, 2018.

[122] S. Zhu, C. Li, C. C. Loy, and X. Tang. Face alignment by coarse-to-fine shape searching. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 4998–5006, Boston, MA, 2015.

[123] S. Zhu, C. Li, C. C. Loy, and X. Tang. Unconstrained face alignment via cascaded compositional learning. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 3409–3417, Las Vegas, NV, 2016.

[124] 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, 2019.

[125] X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 2879–2886, Providence, RI, 2012.