Towards Robust Unconscious Face Recognition for Video Surveillance

Meng ZHANG

Abstract

As one of the most promising technologies in the field of computer vision and artificial intelligence, face recognition has been widely considered by academia and industry for many years. At present, face recognition under constrained environment has achieved promising results, and many products with face recognition technology are widely used in our daily life. However, surveillance face recognition is still a challenging problem, especially for unconstrained surveillance scenes. Different from constrained face recognition, where a user is expected to cooperate with the machine to complete the recognition, unconscious surveillance face recognition expects unconstrained scenarios without cooperative users, which can suffer from extremely low quality for each frame, e.g., various occlusions, changing illuminations, dramatic pose variations, especially when a large part of the face is covered by wearing masks, for example, during the COVID-19 pandemic. On the contrary, abundant temporal and multi-view information usually exists between surveillance video frames, which may bring potential to boost accuracy in unconstrained surveillance face recognition.

The main objective of this thesis is to improve the robustness of unconscious face recognition for video surveillance quantitatively and qualitatively. Despite the success of deep learning models under constrained face recognition scenarios, the deep features still demonstrate imperfect invariance to wearing a mask, where the whole face image is not available for description. However, a surveillance video provides us with abundant complementary information across frames compared with a single image. Therefore, this thesis focuses on face recognition with masked faces and feature aggregation-based face recognition between multiple fames. Two methods are proposed. Firstly, a masked face recognition method with mask transfer and self-attention, and secondly, a content-aware contribution estimation feature aggregation for surveillance face recognition.

The first research work presented in this thesis proposes a method used for mitigating the negative effects of mask defects on face recognition. Firstly, a low-cost, accurate

method of masked face synthesis, i.e., mask transfer, is proposed for data augmentation. Secondly, an Attention-aware Masked face recognition Network (AMaskNet) model is proposed to improve the performance of masked face recognition, which includes two modules: a feature extractor and a contribution estimator. Therein, the contribution estimator is employed to learn the contribution of the feature elements, thus achieving refined feature representation by simple matrix multiplications. Meanwhile, the end-to-end training strategy is utilized to optimize the entire model. Finally, a mask-aware similarity Matching Strategy (MS) is adopted to improve the performance in the inference stage. Experiments show that the proposed method consistently outperforms comparative methods on three masked face recognition datasets: RMFRD, COX, and Public-IvS. Meanwhile, qualitative analysis experiments using CAM indicate that the contribution learned by AMaskNet is more conducive to masked face recognition.

The second research work presented in this thesis proposes a content-aware feature aggregation scheme to aggregate complementary information between different frames. The difficulties in video-based face recognition, such as dramatic pose variations and low quality, can be alleviated by leveraging the rich complementary information between the frames. However, limited by the mini-batch training strategy, the current deep learning methods only utilizes the frames in each batch during training, which ignore the content of the entire video. Therefore, firstly, a two-branch structure is designed as the Content-aware feature Aggregation Network (CAN). Secondly, a content-aware training strategy using a content bank is proposed, which alleviates the limitation of minibatch samples by using the content of the entire video or several images belonging to the same identity and thus can estimate the global contribution. Comparative studies on benchmark datasets: IJB-C, YouTube Face (YTF), PaSC and COX, confirm that the proposed approach outperforms comparative methods. Meanwhile, qualitative analysis on Multi-PIE dataset indicates that the contribution learned by the CAN is reasonable and beneficial to video face recognition.

Based on the above research topics, an unconscious access control of a laboratory gate was implementated by setting surveillance cameras and using the trained models

to analyze and verify the feasibility of the proposed methods in practical application scenarios.

In summary, this thesis presents methods towards robust face recognition for video surveillance. Specifically, the prototype also shows good recognition performance for the face with masks caused, for example, by the COVID-19 pandimic. Firstly, Chapter 1 provides introduction, background, research topics and main contributions of this research. Then, Chapter 2 introduces the researches related to this thesis. Furthermore, the proposed methods for masked face recognition and content-aware feature aggregation-based video face recognition are described in detail in Chapter 3 and Chapter 4, respectively. Besides, Chapter 5 concludes the thesis by summarizing the research contributions and providing possible research directions in the future. Finally, a prototype of unconscious face recognition in surveillance scenes is intorduced in the Appendix.

Acknowledgments

This dissertation is formally submitted for fulfilling partial requirements for the degree of Doctor of Intelligent Systems of Graduate School of Informatics, from Nagoya University. This work would not have been possible without the help of many people to whom I owe the sincerest gratitude.

Firstly, I am extremely grateful to Prof. Dr. Hiroshi Murase for accepting me into his lab and giving me many insightful suggestions. I would like to express my gratitude to Prof. Dr. Daisuke Deguchi for his supervision and assistance for this research. He gave me a lot of insightful suggestions, whether about my research or my publications. I am extremely grateful for his contributions and support.

I would like to thank Dr. Satoshi Naoi (Fellow, Fujitsu Laboratories Ltd.) and Dr. Jun Sun (Group manager, Fujitsu R&D Center Co., Ltd.) for their support to my doctor. I would like to express my gratitude to Dr. Rujie liu (Group manager, Fujitsu R&D Center Co., Ltd.) for his comments on my papers and discussion of algorithm details.

I would like to thank Prof. Dr. Hiroshi Murase, Prof. Dr. Kensaku Mori, Prof. Dr. Daisuke Deguchi, Prof. Dr. Ichiro Ide and Prof. Dr. Yu Enokibori for the insightful comments and suggestions during the reviewing process.

I would like to express my gratitude to Fujitsu for its support and for providing experimental equipment and data, especially the two dedicated GPU servers. A GPU server with 8 TITAN RTX, which enabled me to train some state-of-the-art models, for example, I won the 3rd in the world and 1st in Japan (266 participating global teams) on behalf of Fujitsu in FRVT's mask benchmark. A GPU server with 10 1080Ti made me explore more possibilities. It is a pity that some works have not been written into this dissertation because there is no time to write paper or paper is still being reviewed.

Last but not least, I would like to thank my wife and my family for their support and encouragement, which enabled me to overcome difficulties and not be hindered by family chores, so that I could successfully complete my studies.

Contents

List of Figures

List of Tables

Abbreviations

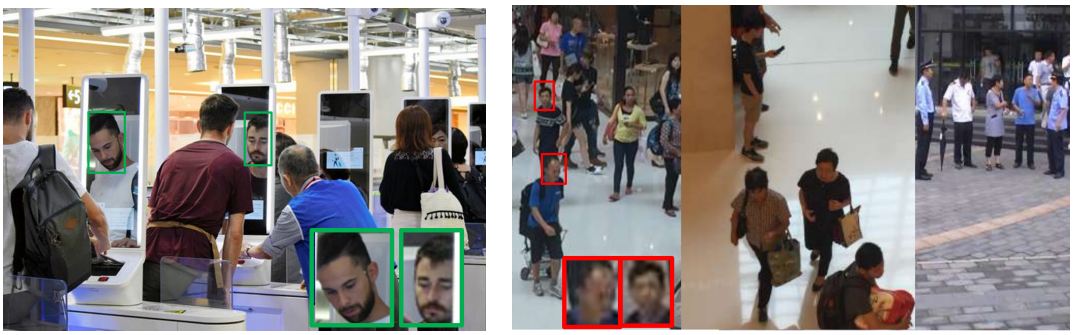
3DMM	**3**-**D**imensional **M**orphable **M**odel
ACC	**ACC**uracy
AMaskNet	**A**ttention-aware **Mask**ed face recognition **Net**work
BS	**B**alance **S**trategy
CAM	**C**lass **A**ctivation **M**ap
CAN	**C**ontent-aware feature **A**ggregation **N**etwork
CCTV	**C**losed **C**ircuit **T**ele**V**ision
CNN	**C**onvolution **N**eural **N**etwork
CS	**C**ontent-aware **S**trategy
CVPR	IEEE Conference on **C**omputer **V**ision and **P**attern **R**ecognition
DA	**D**ata **A**ugmentation
DCNN	**D**eep **C**onvolution **N**eural **N**etwork
DL	**D**eep **L**earning
ECCV	**E**uropean **C**onference on **C**omputer **V**ision
FAR	**F**alse **A**cceptance **R**ate
FR	**F**ace **R**ecognition
GAN	**G**enerative **A**dversarial **N**etwork
ICCV	IEEE **I**nternational **C**onference on **C**omputer **V**ision
ILSVRC	**I**mageNet **L**arge-**S**cale **V**isual **R**ecognition **C**ompetition
JB	**J**oint **B**ayesian
LBP	**L**ocal **B**inary **P**attern
LDA	**L**inear **D**iscriminant **A**nalysis
LFW	**L**abeled **F**aces in the **W**ild

MS	**M**atching **S**trategy
NAN	**N**eural **A**ggregation **N**etwork
PCA	**P**rincipal **C**omponent **A**nalysis
PR	**P**attern **R**ecognition
QAN	**Q**uality **A**ggregation **N**etwork
ReLUs	**Re**ctified **L**inear **U**nit**s**
ROC	**R**eceiver **O**perating **C**haracteristic
SE	**S**queeze-and-**E**xcitation
SGD	**S**tochastic **G**radient **D**escent
SOTA	**S**tate **O**f **T**he **A**rt
SVM	**S**upport **V**ector **M**achine
TAR	**T**rue **A**cceptance **R**ate
TIP	IEEE **T**ransactions on **I**mage **P**rocessing
TPAMI	IEEE **T**ransactions on **P**attern **A**nalysis and **M**achine **I**ntelligence
VFR	**V**ideo **F**ace **R**ecognition
VS	**V**ideo **S**urveillance
YTF	**Y**ou**T**ube **F**ace

Chapter 1

Introduction

With the development of artificial intelligence and computer vision, face recognition has been widely considered by academia and industry. At present, face recognition has achieved promising results, and many products with face recognition technology are widely used in our daily life. However, most current face recognition products assume constrained scenarios with cooperative users, that is, the users need to cooperate with the machine to complete the recognition, and the face image should not contain extreme occlusion, pose, expression, illumination, and so on, as shown in Figure 1.1(a). Face recognition under unconstrained scenarios with uncooperative users for video surveillance, that is, unconscious surveillance face recognition, is still a very challenging problem, as shown in Figure 1.1(b). Different from constrained and cooperative face recognition, the unconscious surveillance face recognition can suffer from extremely low quality for each frame, e.g., various occlusion, changing illumination, dramatic pose, especially when a large part of the face is covered by wearing a mask, for example, during the COVID-19 pandemic. On the contrary, abundant temporal and multi-view information usually exists between surveillance video frames, which may bring potential to boost performance in unconscious surveillance face recognition. This thesis intensively studies the robust unconscious face recognition for video surveillance.



(a) Constrained and cooperative face recognition [12].

(b) Unconscious surveillance face recognition [13].

Figure 1.1: Example comparisons of constrained and cooperative face recognition and unconscious surveillance face recognition in real-world application scenarios. (Best viewed in color)

In this chapter, Section 1.1 explains the background of this thesis, including the history and breakthrough methods of traditional face recognition and deep learning-based face recognition, and the challenges of current face recognition. Then, the description of the research of this thesis followed by the important aspects considered in developing the method, are introduced in Section 1.2. Furthermore, general descriptions of the research overview and proposed solutions are explained in Section 1.3. Lastly, the structure of this thesis is presented in Section 1.4.

1.1 Background

With the development of human society, a variety of science and technology has developed rapidly, especially information technology. Today, information products play a vital role in our life, bringing convenience and rapidity to human life. For example, people can pay online, work remotely, perform online social networking, and so on. However, while information products are convenient for human life, there is also an important problem, that is, security. How to confirm the identity information of users quickly and accurately, has become a widespread concern in academia and industry.

In the past years, passwords have been the most popular authentication mechanism, that is, through the matching of username and password for identity authentication [1]. However, this type of authentication mechanism has drawbacks. First, passwords are easy to be stolen. Once the password is stolen, individuals or units will suffer huge losses. Secondly, password setting often requires a combination of various characters, which is highly complex, making it difficult for users to remember. Especially in today's information society, almost everyone will have various accounts, and each account requires a unique password. If these passwords are set consistently, the security will be reduced, but if the settings are inconsistent, it often brings memory problems, such as often confusing passwords or forgetting passwords. Therefore, it is necessary to find a more convenient, fast, user-friendly, and safe authentication method.

With the development of password authentication, biometric authentication has been widely studied and applied because of its uniqueness, measurability, and lifetime invariance. Biometric identification technology mainly uses the humans' physical or behavioral traits, such as fingerprints, palm veins, iris, face, gait, DNA, voice, and so on [14].

Among these biometric authentication technologies, face recognition has unique advantages over the others especially during the COVID-19 pandemic. First, compared with fingerprint recognition, the face features are relatively complex and less reproducible. Secondly, compared with iris recognition and DNA recognition, face image is easier to collect and can be quickly recognized, even if users do not cooperate at a long distance. Meanwhile, we can obtain the user's tag by image recognition and semantic segmentation, such as identity, race, age, gender, emotion, behavior, and so on [3]. Finally, compared with other biometrics, face recognition is non-contact, highly efficient, user friendly, and so forth, which can eliminate the psychological barriers of users and be easily accepted by them.

Face recognition is a long-standing research topic integrating artificial intelligence, machine learning, video image processing, and other technologies, especially in the

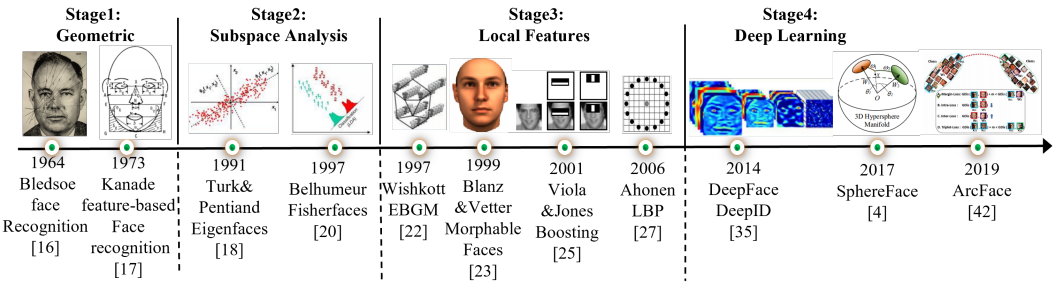


Figure 1.2: Development history of face recognition. (Best viewed in color)

field of computer vision. In the past several decades, face recognition has experienced four climaxes, and its development history is shown in Figure 1.2. Face recognition based on deep learning has achieved a remarkable progress and dramatically improved the state-of-the-art performance. Meanwile, the progress of face recognition in academia has also promoted its application in real-world scenes. So far, face recognition is a mature technology in the field of computer vision and can be widely used in people's lives, as shown in Figure 1.3. However, with the commercial and practical use of the face recognition, many of the ideal assumptions of academic research are being broken, and more and more challenging problems are emerging in real-world applications. Real-world face recognition needs to pursue the ultimate performance under unconstrained scenarios with uncooperative users, for example, financial authentication and watch-list surveillance, demanding the accuracy of matching at very low alarm rates, such as TAR@FAR=10^{-8} [1]. Therefore, face recognition in a real-world application is still a huge challenge even with large-scale training datasets, many well-designed loss functions, and deep learning technology. Thus, in recent years, face recognition has still been a hot research topic, and attracts many researchers to study its challenging problems. Meanwhile, many excellent achievements are published in top journals (TPAMI, TIP, PR) and top conferences (CVPR, ICCV, ECCV) in the computer vision community every year.

In this section, the history and breakthrough methods of conventional face recognition and deep learning-based face recognition are described in Section 1.1.1 and

[1] The true accept rate when the false accept rate equals 10^{-8}.

(a) Check-in at Narita airport (b) Alibaba's Alipay (c) Multi-factor biometric authentication

Figure 1.3: Examples of face recognition application scenarios. (Best viewed in color)

Section 1.1.2, respectively. Then, the challenges of current face recognition are discussed.

1.1.1 Conventional face recognition

Conventional face recognition methods have attempted to recognize faces using one- or two-layer representation learning, such as distribution of the dictionary atoms, filtering responses, or histogram of the feature codes, which can be summed up as face recognition based on shallow representation. They have developed in three-stages, as follows:

(1) Stage 1: Methods based on geometric feature

The earliest research work on face recognition can be traced back at least to the research in psychology in the 1950s. Bledsoe et al. [15] built the first semi-automatic face recognition system at that time. At that time, the focus of face recognition research was mainly to extract the geometric features of the face, such as the distance and ratio between the facial feature points, and the two-dimensional topological structure composed of some feature points on the face, such as the points of the nose and eyes, the corners of the mouth and the eyes, and so on. After that, Kanade [16] developed the first complete face recognition system using computers for his Ph.D dissertation at Kyoto University in 1973, which opened up the research path in the face recognition research community. His thesis described a computer program which performed a complex image processing task, which was to find the

same person in a set of images taken by a TV camera. Generally speaking, the face recognition research at this stage was basically based on the geometric structure of the face and belongs to the primary face recognition.

(2) Stage 2: Methods based on subspace analysis

From 1991 to 1997, in the short period of six years, the research of face recognition reached the second climax, and many representative face recognition algorithms appeared. Firstly, Turk and Pentland [17] proposed the famous Eigenface algorithm, and then many eigenface related face recognition technologies were proposed. In 1993, Brunelli and Poggio [18] showed that the template matching-based method was superior to the features-based methods through experiments. The conclusion of this work basically stopped the research of pure face recognition methods based on structural features and made appearance-based face recognition the mainstream technology. In 1997, Belhumeur et al. [19] proposed the fisher face recognition method to derive the minimum intra-class distance and the maximum inter-class distance, where PCA and LDA were used to reduce and transform the face features. Based on this idea, many researchers have proposed subspace discriminant model, direct LDA discriminant method, and enhancement model.

These methods became the mainstream of face recognition in that period, and even now, they are still one of the mainstream face recognition methods. During this period, the Counterdrug Technology Transfer Program (CTTP) of the U.S. Department of Defense launched another important work as a face recognition technology project called FERET [20]. The main purpose of this project was to promote the research and application of face recognition algorithms, and ultimately provide a reliable practical Automatic Face Recognition (AFR). This project was mainly composed of three parts: The first one was to fund several face recognition studies, which enabled many researchers to have sufficient funding to carry out face recognition research and enable them to quickly join the research field; The second one was to create the FERET face image database [20], which provided a standard platform for the verification of face recognition algorithms; The third one was to organize FERET face recognition

performance evaluation, which provided a standard for face recognition testing. This project had greatly promoted the development of face recognition technology. Meanwhile, the project also pointed out the next research direction of face recognition, that is, face recognition in non-ideal environments such as complex lighting, multi-pose, and complex expression. In conclusion, the research of face recognition became increasingly popular following the idea of the historical Eigenface approaches [17] in the early 1990s, which dominated the face recognition community in the late 1990s and in the early 2000s. But, a well-known problem is that these theoretically plausible holistic approaches cannot tackle the uncontrolled facial changes that deviate from their previous assumptions.

(3) Stage 3: Methods based on local feature

During this period, the main contributions of face recognition technology were the proposal of the 3-Dimensional Morphable Model (3DMM) and the use of Local Binary Pattern (LBP) features. In 1997, Wiskott et al. [21] proposed a face recognition approach based on elastic matching, where a system was used to recognize faces from a single image in a large dataset including one image per person. Based on the Gabor wavelet transform, the face was represented by a marker map. The elastic graph matching process was used to extract the image graphs of new faces, which can be compared by a simple similarity function. This method has derived a series of face recognition algorithms, which has played a certain role in promoting the development of face recognition. In 1999, Blanz and Vetter et al. [22] proposed the 3DMM, which opened the journey of researchers for face recognition under multi-pose and complex lighting conditions. Blanz found through experiments that 3D modeling of a human face using 3DMM achieved a good recognition rate on CMU-PIE [23] and FERET databases. In 2001, Viola and Jones proposed a simple face detector, called V-J face detection [24]. They used rectangle feature and AdaBoost algorithm to realize it. The speed of this method in calibrating the front face could reach more than 1,000 frames per second, which basically achieved real-time processing. This work was published in at ICCV that year. The idea of this work provides a good foundation for back-end face recognition and points out a new way for the development of

face recognition. In the same year, Shashua et al. [25] proposed a face recognition technology based on quotient image, which greatly promoted the development of illumination invariant face recognition. In 2006, Ahonen et al. applied LBP features to face recognition [26]. Since then, various extensions [27, 28] based on LBP features have achieved robust performance in pose invariant face recognition, greatly promoting the development of face recognition. To summarize, in the early 2010s, learning-based local descriptors were widely used in the face recognition community [29, 30, 31], where local filters were built for better distinctive features and the encoding codebook was learned for better compactive features. However, these methods are still considered as shallow representations, which have an unavoidable limitation on robustness of complex non-linear facial appearance changes.

1.1.2 Deep learning-based face recognition

The previous three-stage face recognition methods can be summed up as face recognition based on shallow representation, which attempt to recognize face by one- or two-layer representation. However, the face features of those methods are all based on manual design, and their robustness to non-linear changes in face appearance and external conditions is inevitably limited. Good results can be achieved under limited conditions, but the effect is still not ideal under non-limited conditions. For example, in the Labeled Faces in the Wild (LFW) dataset [32] proposed in 2008, the human recognition rate was 97.5%, but the best algorithm at that time could not reach this recognition rate.

But all that changed in the year of 2012, after AlexNet [33] won the ImageNet competition and showed a large performance gap with the second-place using deep learning [33], which opened a new milestone in computer vision. Deep learning methods, such as Convolutional Neural Network (CNN), use cascaded multi-layer processing units for face feature extraction. They learn multi-level representation corresponding to different abstraction levels, which is more consistent with people's understanding

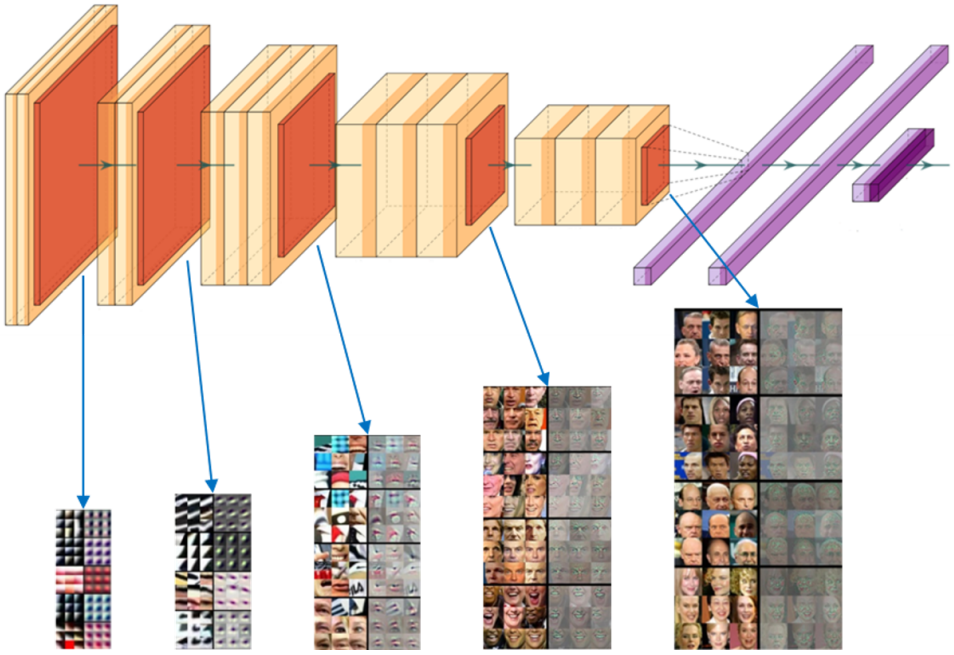


Figure 1.4: Hierarchical architecture of face feature extraction using deep learning-based methods. (Best viewed in color)

of images, as shown in Figure 1.4. The pixels are spliced into a hierarchical structure of facial representation. The model based on deep learning is composed of multi-layer simulated neurons, which convolute and pool inputs. In this process, the receptive field of simulated neurons expands in size to integrate low-level primary elements into various facial attributes, and finally forward the extracted features to one or more fully connected layers at the top of the network. The output is a compressed feature vector representing the face. This kind of deep learning representation is widely used for face recognition, and has achieved good performance. Inspired by this work, most of research focus has shifted to deep-learning-based methods, and the performance was dramatically improved. After that, researchers proposed various network structures and loss functions, which greatly improved the performance of face recognition.

In 2014, Facebook proposed the Deepface [34] model, which introduced the deep learning technology into face recognition for the first time and obtained a high accuracy rate on the challenging and famous LFW dataset. For the first time, it achieved performance close to human recognition under unconstrained conditions (human:

97.53% [35] vs. Deepface [34]: 97.35%). Deepface [34] can be considered as the foundation work of CNN applied to face recognition, which opened a new path for the development of face recognition. In the same year, the Chinese University of Hong Kong (CUHK) proposed DeepID [35], which took an approach of face image segmentation to divide the face image into multiple image blocks and send the segmented image to the deep network for training. This method improved the performance of unrestricted face recognition to a new level. Since then, face recognition has entered the era of face recognition based on deep learning. Most of the research focus has turned to the methods based on deep learning. In 2015, Google proposed FaceNet [36], trained on a massive face dataset with more than 200 million faces collected by themselves and used the triple loss function in the training. Finally, it achieved 99.65% accuracy on the LFW dataset, which greatly exceeded the recognition accuracy of human eyes. In only five years, the face recognition accuracy has increased to more than 99.80% on the LFW dataset, which is nearly perfect.

In the academic community, face recognition has been a long-standing research topic in top journals (TPAMI, TIP, PR) and top conferences (CVPR, ICCV, ECCV) in computer vision community. At present, due to the excellent performance of deep neural network models [37, 38, 39, 40], sophisticated design of loss functions [1, 41, 42], and large-scale training datasets, e.g., MS-Celeb-1M [43], DeepGlit [44], Glint360K [45], and WebFace260M [46], face recognition has made good progress, and under certain conditions, ideal recognition results can be obtained.

In summary, deep learning technology has changed the research field of face recognition in almost all aspects such as algorithms, datasets, and even evaluation protocols, and has brought it to a new climax since 2012.

1.1.3 Is face recognition really a solved problem?

With the development of academic research, face recognition has also been widely used in industry, and more and more products using face recognition technology have entered people's daily life, especially since the COVID-19 pandimic. For example,

travelers can register necessary data through face recognition at a terminal of Narita Airport near Tokyo [47], as shown in Figure 1.3(a); Alipay introduces facial recognition for payments in China [48], as shown in Figure 1.3(b); Fujitsu delivers cashless and hygienic retail experience for masked shoppers using multi-factor face and palm biometric authentication technology [49], as shown in Figure 1.3(c).

One may argue that the most advanced face recognition algorithm, especially with the help of deep learning and large-scale datasets, has reached sufficient maturity and application readiness level, which is proved by its almost saturated performance on large-scale public benchmark challenge, such as MegaFace [50]. Therefore, the face recognition problem should be considered as been tackled, and the remaining work is mainly concentrated in the system production engineering.

However, most current face recognition products need to be under constrained scenarios with cooperative users, that is, the users are expected to cooperate with the machine to complete the recognition and the face images cannot include extreme occlusion, pose, expression, illumination, and so on. For example, the user's face must be kept at a certain distance from the camera and face front to the camera. What is more, the user needs to remove the mask with the risk of spreading the virus under the COVID-19 pandemic, as shown in Figure 1.3(a). If the user's face cannot be placed in the proper position specified by a face box on the screen or is not facing the camera, face payment often fails, as shown in Figure 1.3(b). However, In most real-world applications of face recognition, face images can also come from diverse sources, e.g, surveillance cameras, mobile phone cameras, and have diverse different qualities, e.g, different expressions, posture changes, blur, occlusion. Although remarkable progress for constrained face recognition has been achieved with deep learning and large-scale datasets [1, 41], unconstrained face recognition is still a challenging problem. For example, Cheng et al. [13] have verified that the accuracies on TinyFace [51] and QUML-SurFace [13] are about 25% lower than that on MegaFace [50] or LFW [32] datasets. Therefore, unconstrained and uncooperative face recognition under the surveillance scenario is still difficult.

1.2 Unconscious Face Recognition for Video Surveillance

With the rapid development of video surveillance technology, the number of installed cameras for video surveillance is increasing in public places. In video surveillance, capture conditions typically range from semi-controlled situation with one person in the scene, e.g., terminals at airports and passport inspection lanes, to uncontrolled cluttered and free-flow scenes, e.g., subway stations and airport baggage claim areas. Due to the absence of uncontrolled conditions and user cooperation, face recognition in Video Surveillance (VS) is a less obtrusive technique, which has received more and more attentions.

Although remarkable progress has been achieved in face recognition technology due to the emergence of large-scale datasets, deep learning-based methods [33, 37, 39, 52] and various effective loss functions, e.g. SphereFace [1] or ArcFace [41], most of them are designed for still face recognition. When extending from the still to the video scenario in video surveillance, many approaches tend to ignore the peculiarities of videos compared to still images. However, unconscious surveillance face recognition is evidently more challenging. Images in standard still face recognition datasets are usually captured under good conditions or even framed by professional photographers, e.g., LFW [32] dataset. Different from still face recognition, video face recognition suffers from extremely low quality in each frame, e.g., various occlusion, changing illumination, dramatic pose variations, especially when a large part of the face is covered by masks, for example due to the COVID-19 pandemic. On the contrary, abundant temporal and multi-view information usually exists between surveillance video frames, which may bring potential to boost accuracy in unconstrained surveillance face recognition. Hence, it is necessary to design a method to overcome challenges for effective and robust unconscious face recognition for video surveillance.

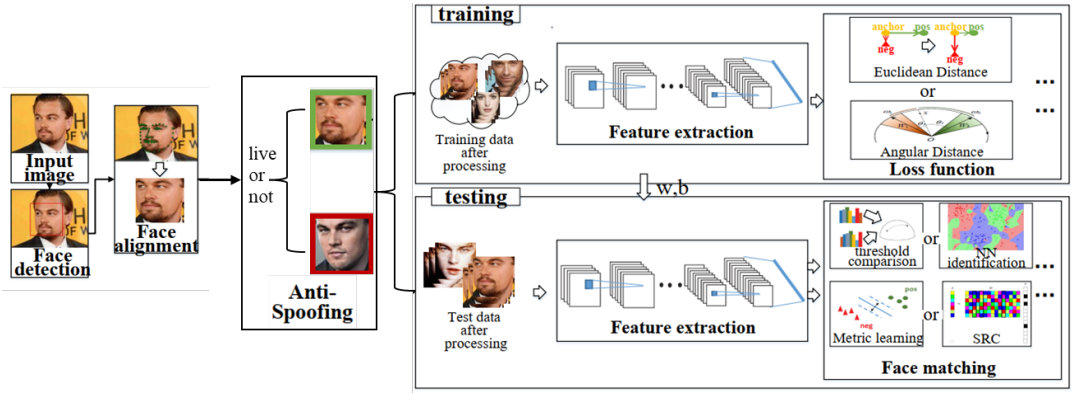


Figure 1.5: Deep face recognition system with face detection, alignment, anti-spoofing. (Best viewed in color)

The research presented in this thesis aims to contribute to the robust unconscious face recognition for video surveillance, which can be deployed on productions with high performance and efficient computation. Unconscious face recognition, also known as unconstrained and uncooperated face recognition, refers to a face recognition model that does not require constrained scenarios with cooperative users, that is, the users are not expected to cooperate with the machines to complete the recognition and the face recognition model does not need to be installed in a fixed scene.

The description of face recognition is introduced below.

Face recognition systems can be mainly divided into four modules: face detection, face alignment, face anti-spoofing recognition, and face recognition, as shown in Figure 1.5. Firstly, the face detection is applied to detect the face bounding box. Secondly, the detected face images are aligned to the normalized canonical coordinates. Thirdly, the face anti-spoofing recognizer is used to recognize whether the detected face is live or spoofed. Finally, the face recognition module is implemented with these aligned face images. Among them, the face recognition module is the most important module, which is the bottleneck for the current research on unconscious surveillance face recognition to move towards practical applications. Therefore, this thesis focuses on the face recognition module.

Furthermore, face recognition can be subdivided into face identification and face verification. In both cases, one set of known subjects is initially enrolled in the system

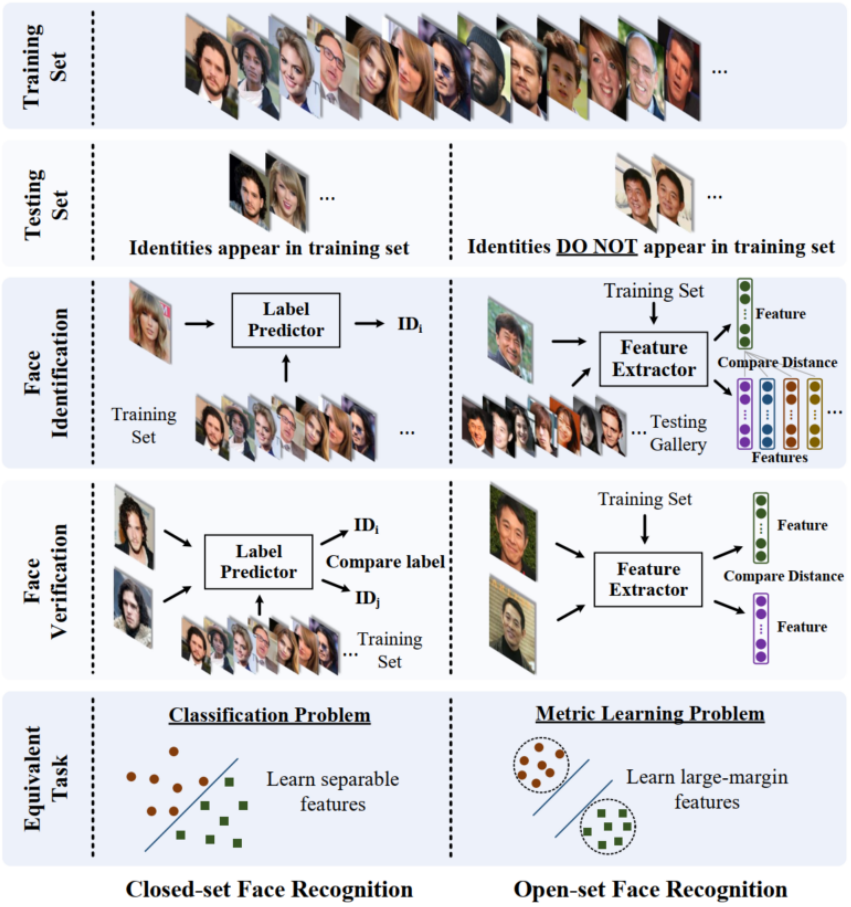


Figure 1.6: Comparison of closed-set and open-set face recognition. [1] (Best viewed in color)

(gallery), whereas a new subject (probe) is presented. Meanwhile, face verification calculates one-to-one similarity between the gallery and the probe to determine if two images belong to the same subject, while face identification calculates one-to-many similarity to determine the specific identity of the probe's face.

Depending on the test protocol, face recognition can be evaluated in either a closed-set or an open-set setting, as shown in Figure 1.6. For closed-set protocols, all test identities are predefined in the training set, which is natural to classify test face images based on given identities. In this scenario, face verification is equivalent to identifying a pair of faces respectively (see the left side of the Figure 1.6). Therefore, closed-set face recognition can be well solved as a classification problem, where features are expected to be separable. Meanwhile, for the open-set protocol, the test

identities are usually disjoint between the test set and the training set, which makes face recognition more challenging but closer to real-world applications. Since it is not possible to classify faces into known identities in the training set, we need to map faces to a distinct feature space. In this scenario, face identification can be considered as performing face verification between the probe face and each identity in the gallery (see the right side of the Figure 1.6). Open-set face recognition is essentially a metric learning problem, where the key is to learn differentiated large edge features.

In most real-world face recognition scenes, it is difficult to obtain user's face image for training. However, the open-set face recognition only needs each user to submit a photo to the gallery. The face photos collected on site will be matched with the photos in the gallery. If it is the person with the highest similarity, then the recognition is correct. This thesis focuses on the open-set face recognition for video surveillance.

The open-set face recognition model is a feature extractor, which is used to extract the features of the probe image and the gallery images, and then uses the matching method to calculate the similarity. It can be described as follows:

$$M\left(F\left(P\left(I_i\right)\right), F\left(P\left(I_j\right)\right)\right), \tag{1.1}$$

where $P(\cdot)$ is a face processor to handle intra-class variations before training and testing, e.g., illuminations, occlusions, poses, and expressions, I_i and I_j represent two face images, respectively, $M(\cdot)$ indicates the face matching algorithm and outputs the similarity scores of face features to identify the specific identity of faces, e.g., cosine similarity for matching. $F(\cdot)$ is a face extractor to extract the discrimination identity features. The feature extractor is trained as a classification task using margin-based loss functions during the training, e.g., Arcface Loss, and is used to extract features of faces when testing.

Unlike object classification, the test identities are usually disjoint with the training data in face recognition, which makes the learned classifier unusable to recognize the test faces. Therefore, the face matching algorithm is an essential part of face

recognition. In these steps, the core step is to obtain a feature extractor, which is expected to have a maximum within-class distance that is smaller than the minimum interclass distance under a suitably chosen metric space.

As explained in the previous descriptions, the main objective of this thesis is to improve the performance of unconscious surveillance face recognition. This research focuses on the most challenging module of face recognition. Several aspects that are of major concern in developing methods for realizing unconscious surveillance face recognition are described in the next subsections.

1.3 Research Overview

As described in the previous section, face recognition has been extensively studied since the 1960s, and increasingly deployed in social applications within the last several years. Compared with fingerprints, iris, gait, and other biometric recognition technologies, face recognition is non-contact, highly efficient, user friendly, and so forth, and thus has been widely applied in access control and security authentication in public places, especially since the COVID-19 pandemic.

However, as described in the previous section, the current face recognition method has poor scalability to real-world surveillance face recognition due to face mask and low-quality images. Typical real-world face images are captured in unconstrained wide-field surveillance video and images, which may be one of the most important face recognition application fields in practice. Specifically, face recognition in unconstrained surveillance images is far from satisfactory, especially in large-scale dataset situations. Different from identifying high-quality Web celebrity images with limited noise, unconscious surveillance face recognition remains extremely challenging and open. This is because the surveillance video data are characterized by low-quality images with heavy noise, subject to poor imaging conditions giving rise to unconstrained pose, expression, occlusion, illumination, and background clutter.

The main objective of this study is to improve the performance of unconscious surveillance face recognition under unconstrained and uncooperative scenes. This thesis mainly focuses on the following two research topics:

[1]. **Masked face recognition.** Face masks bring a new challenge to existing commercial face recognition techniques, especially under the COVID-19 pandemic. Since face recognition becomes more difficult when a large part of the face is covered by a mask, it is essential to study the effect of wearing face masks on the behavior of face recognition systems and design mitigation techniques to offset the inevitable performance loss.

[2]. **Feature aggregation in video face recognition.** Different from still face recognition, video face recognition often suffers from low quality, dramatic pose variations, occlusion, and so on. On the contrary, abundant temporal and multi-view information usually exists in the video, which may bring potential to boost accuracy in video face recognition.

Figure 1.7 shows the relationship between the problems and the proposed solutions discussed in this thesis. In this section, Research Topic 1 is firstly introduced in Section 1.3.1 , and then, Research Topic 2 is introduced in Section 1.3.2.

1.3.1 Research Topic 1: Proposal of a masked face recognition approach with mask transfer and self-attention

The COVID-19 pandemic has caused a global impact: The World Health Organization (WHO) and the U.S. Centers for Disease Control and Prevention (CDC) have suggested everyone should wear a mask in a public setting especially when other social distancing measures are difficult to maintain [53]. Face recognition is non-contact, highly efficient, user friendly, and so forth, and thus has been widely applied in access control and security authentication in public places. However, masks bring

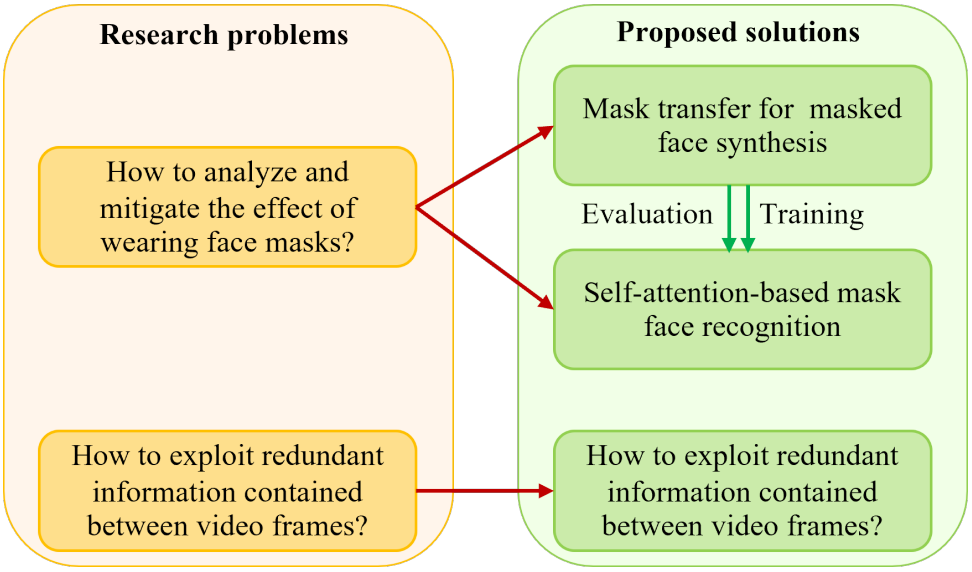


Figure 1.7: Core of the research for developing the robust unconscious face recognition for video surveillance under the COVID-19 pandemic in this thesis. (Best viewed in color)

a new challenge to existing commercial face recognition techniques. Face recognition becomes more difficult when a large part of the face is covered by a mask. Therefore, it is essential to study the effect of masks on the behavior of face recognition systems and design mitigation techniques to offset the inevitable performance loss.

Deep-learning-based approaches predominate in the task of face recognition due to the emergence of advanced CNN, well-designed loss functions [1, 54, 55], and large-scale datasets [41, 56]. Despite the success of deep learning models under general face recognition scenarios, the deep features still demonstrate imperfect invariance to face masks, where the whole face image cannot be used for the description. Therefore, face masks trigger a significant research challenge: Firstly, it is necessary to collect a large-scale training dataset, which includes faces with different types of masks. In order to collect such a large-scale training dataset, on the one hand, it is time consuming and incurs higher labour cost, and on the other hand, maintaining the diversity of data in such datasets is a slow process. Therefore, a low-cost, convenient face data augmentation method is needed as a matter of urgency. Secondly,

it is essential to mitigate the performance loss from the perspective of model design according to the characteristics of masks.

Some methods of simulating a masked face [57, 58, 59, 60] have been proposed for face data augmentation. However, since these methods only utilize affine transformation, the added masks often look unnatural. Furthermore, they ignore pose and illumination consistency, thus leading to biased masked face augmentation. Recently, Generative Adversarial Network (GAN) has become a powerful technique for data augmentation [61]. However, GAN-based methods suffer from mode collapse deeply, which usually manifests that the images generated by the generator tend to be highly similar amongst them, even though their corresponding latent vectors are very different. In addition, GAN-based methods are generally slow and difficult to run online in recognition. On the contrary, the method proposed in this thesis can quickly collect various types of mask images and can transfer them to the face image in run-time for the mask-aware similarity matching strategy in the inference stage.

Based on the above research questions, as Research Topic 1, the effect of wearing a mask on face recognition is qualitatively and quantitatively analyzed, and then a method for mitigating the negative effects of mask defects on face recognition is proposed. Firstly, a low-cost, accurate method of mask transfer is proposed for masked face synthesis by considering pose and illumination consistency. Secondly, the Attention-aware Masked face recognition Network (AMaskNet) model is designed to improve the performance of masked face recognition. This model includes two modules, a feature extractor, and a contribution estimator, wherein the latter is employed to learn the contribution of each spatial region which is then combined with the feature to improve its representation capability. An end-to-end training strategy is adopted to optimize the whole network. Finally, a mask-aware similarity matching strategy is proposed to improve the performance in the inference stage. The experiments show that the proposed method consistently outperforms on three masked face recognition datasets: RMFRD [6], COX [5], and Public-IvS [62]. Meanwhile, qualitative analysis experiments using Class Activation Mapping

(CAM) [7] indicates that the contribution learned by the AMaskNet is beneficial to masked face recognition. The main contributions are summarized as:

[1]. A low-cost, accurate mask transfer method for masked face data augmentation is proposed by considering pose and illumination consistency. This method can add a mask from any face image with a mask to any face image without a mask.

[2]. Qualitative and quantitative experiments are conducted to analyze the effect of wearing face masks on the behavior of face recognition systems.

[3]. AMaskNet is proposed to improve the performance of masked face recognition.

[4]. A mask-aware similarity matching strategy is proposed for the inference stage, which can be applied to any face recognition scene in which one image with a face mask and the other without a face mask are present.

1.3.2 Research Topic 2: Proposal of a content-aware contribution estimation for feature aggregation in video face recognition

Although considerable progress has been achieved in still face recognition owning to the emergence of effective deep learning-based approaches [41, 42, 55, 56, 63, 64, 65, 66, 67], well-designed loss functions, and large-scale datasets, video face recognition remains as a significant research challenge. Different from still face recognition, video face recognition often suffers from low quality, dramatic pose variations, occlusion, and so on. On the other hand, abundant temporal and multi-view information usually exists in the video, which may bring potential to boost accuracy in video face recognition.

To efficiently use more discriminative information in the video, aggregation-based methods [1, 68, 69, 70, 71, 72] have been widely adopted and impressive performance is gained in video face recognition. The basic idea of the aggregation approach is to extract frame-level features at each frame, and then to aggregate them across all frames to form a video-level feature. The most commonly used aggregation technique is average pooling [73], where features of all frames are simply combined with equal importance. However, low-quality frames would deteriorate the quality of features, resulting in degraded performance of face recognition. Another aggregation method is max pooling [74], which only uses the best quality frame feature as video feature. However, the discriminative information contained in low-quality frames is ignored which could be complementary to high-quality frames.

Recent advance has witnessed deep learning network as an adaptive weighting scheme to aggregate all frame-level features together to form a compact and discriminative video-level feature. However, limited by the mini-batch training strategy, the quality prediction in the above methods only utilize video frames in each batch during training, which ignore the content of the entire video as well as all frames corresponding to the subject, thus leading to a biased face quality estimation.

Therefore, it is essential to study video face recognition and design mitigation techniques to alleviate the difficulties by leveraging the rich complementary information between the frames.

Based on the above research questions, as Research Topic 2, a novelty feature aggregation method is proposed for video-based face recognition by considering the content of the entire video. Firstly, a Content-aware feature Aggregation Network (CAN) is designed to learn the contribution for each frame in a video, in which the features coming from multiple frames are adaptively aggregated into a compact video-level feature. The network is composed of two branches; one is a feature extractor to extract face feature from a single frame and the other is a contribution estimator to estimate the image contribution. The video feature is then aggregated by the features and contributions of all frames in a video clip. Secondly, a content-aware

training strategy using a content bank is proposed, where not only the samples in each mini-batch but also the content of the entire video clip are considered, thus achieves a global contribution estimation scheme. In addition, in order to reduce the influence of the long tail problem in the training corpus, i.e., DeepGlint [44] and Glint360K [45] datasets, a balanced batch selection strategy is further carefully designed. The qualitative analysis on the Multi-PIE [11] dataset shows that the contribution learned by the CAN is reasonable in that it is closely related to image quality, and the quantitative experiments on benchmark datasets indicate that the proposed CAN achieves significant perfomance. The main contributions are summarized as:

[1]. CAN is proposed to learn the contribution of each frame in a video, and the features from multiple frames are adaptively aggregated into a compact video-level feature based on their contributions.

[2]. A content-aware training strategy is proposed to achieve a global contribution estimation scheme by leveraging the content of the entire video clip using a content bank.

[3]. A balanced batch selection strategy is carefully designed to reduce the negative impact of the long-tail dataset on performance.

1.4 Thesis Structure

This thesis consists of five chapters and an appendix. The concepts and relationships between these chapters are visualized in Figure 1.8.

Chapter 1 has discussed the background of the research and described the overall problems as well as the proposed solutions in this thesis. Chapter 2 provides the related technologies to the research topics of this thesis, and then reviews the existing studies which are related to this thesis. Chapter 3 describes the first research topic of this thesis in detail: Masked face recognition with mask transfer and self-attention. Chapter 4 presents the second research topic: Content-aware contribution estimation

for feature aggregation for surveillance video face recognition, Finally, Chapter 5 concludes this thesis by summarizing the research contributions and provides possible research directions in the future. The appendix introduces the developed prototype for access control of a laboratory gate using the proposed research works to analyze and verify the feasibility of the proposed methods in practical application scenarios.

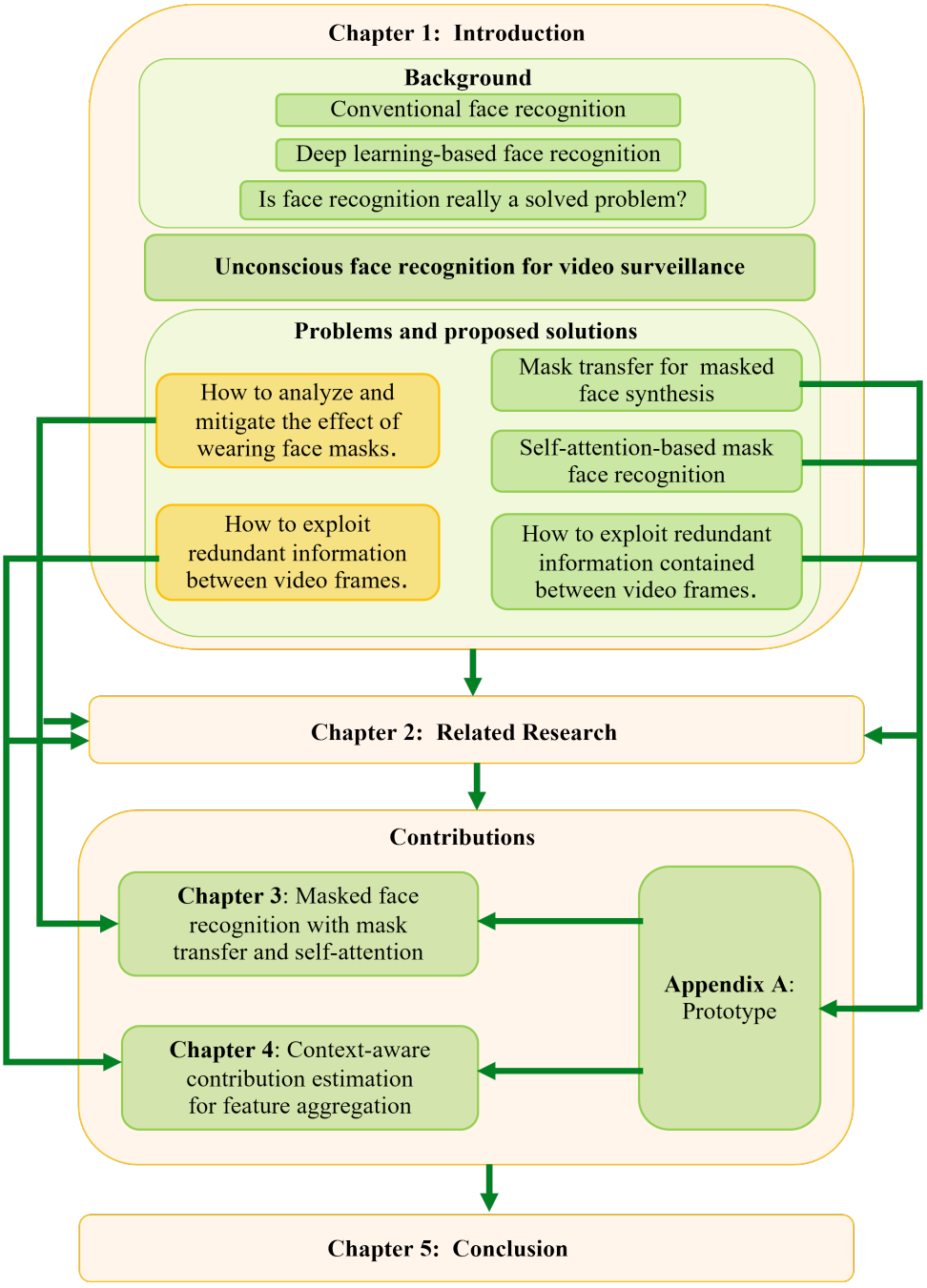


Figure 1.8: Overview of the chapters of this thesis. (Best viewed in color)

Chapter 2

Related Research

In this Chapter, firstly, technologies related to the research topics of this thesis are presented, and then related works to the solutions of the research topics are reviewed.

As described in Chapter 1, the success of current face recognition can be mainly credited to three important reasons: effective deep learning networks, well-designed loss functions, and large-scale datasets. Deep learning has reshaped not only face recognition algorithms, but also face datasets, and even evaluation protocols. Therefore, Section 2.1 presents deep learning related technologies applied to face recognition, including network architecture, attention mechanisms, and training loss. Section 2.2 describes the training and testing datasets used for face recognition. Section 2.3 discusses the protocols of training and testing in this thesis. Section 2.4 introduces related works to the solutions of Research Topic 1, and Section 2.5 introduces related works to the solutions of Research Topic 2.

2.1 Deep Learning Related Technologies

Since 2014, deep learning technology has reshaped the research landscape of face recognition in almost all aspects such as training/testing datasets, algorithm design, application scenarios, and even algorithm evaluation. The success of deep learning

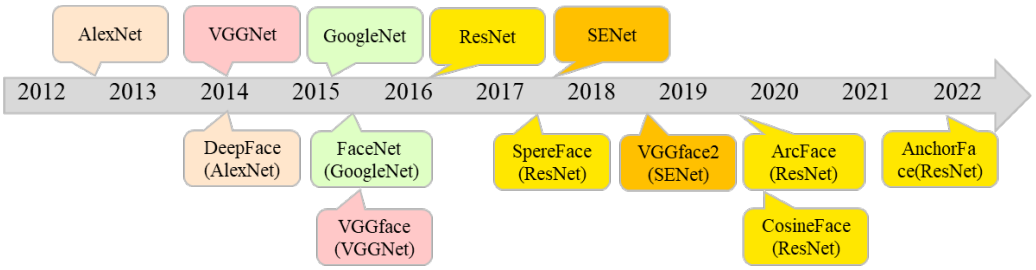


Figure 2.1: Network architectures commonly used in deep face recognition. (Best viewed in color)

for face recognition in recent years can be mainly credited to the following three important reasons: effective neural networks, well-designed loss functions, and large-scale datasets. Firstly, many effective neural network models (e.g., ResNet [37] and SENet [38]) have achieved good results for face recognition. Secondly, many well-designed loss functions are proposed to improve the generalization and discriminative ability of face representation. For example, triplet loss aims to minimize the distances of positive pairs and maximize the distances of negative pairs, and center loss is proposed to reduce the intra-class variations by minimizing the distances within each class. Recently, many margin-based loss functions are proposed by introducing the angular constraints into the cross-entropy loss function. To further increase the feature margin between different classes for enhanced discriminability, CosFace [55] and ArcFace [41] introduce a margin item based on the aforementioned methods. Moreover, CurricularFace [63] and MV-Arc-Softmax [75] are used to introduce the mining-based strategies to emphasize the misclassified samples.

In this section, the network architectures commonly used in deep face recognition are firstly introduced in Section 2.1.1. Then the attention mechanism is described in Section 2.1.2, which is used in research solutions in Chapter 3 and Chapter 4. Finally, many well-designed loss functions are discussed in Section 2.1.3.

2.1.1 Network architectures

Network architectures commonly used in deep face recognition have always followed the architecture of deep object classification and they have rapidly developed

from AlexNet [33]. Here, the most influential deep object recognition and deep face recognition architectures following AlexNet are introduced as in Figure 2.1. The top row shows the mainstream architectures of CNN in object classification, and the bottom row shows famous face recognition models that adopt the main trend CNN. The rectangles with the same color indicate algorithms using the same CNN, where we can easily see that the CNN of deep face recognition have always followed the architecture of object classification and has developed rapidly from AlexNet.

AlexNet has achieved outstanding performance in the ImageNet Large-Scale Visual Recognition Competition (ILSVRC) in 2012, significantly exceeding the state-of-the-art results. Consisting of five convolutional layers and three fully connected layers, AlexNet also integrates various techniques such as data augmentation, dropout, Rectified Linear Units (ReLUs), and so forth. After that, ReLU has been widely regarded as the most important component that makes deep learning possible.

Then, in 2014, a standard network architecture was proposed by VGGNet [39] that used many small 3×3 convolution filters and doubled the number of convolutional neural feature maps after 2×2 pooling, which increased the depth of the CNN to 16 to 19 layers, further enhancing the flexibility of learning asymptotic nonlinear maps through deep architectures.

Next, in 2015, an "inception module" containing a hybrid feature map and two additional intermediate softmax monitoring signals was proposed by GoogleNet [40] that performed multiple convolution in parallel on different receptive fields (5×5, 3×3, and 1×1) and incorporate multi-resolution information by concatenating all feature maps.

More impressively, in 2016, ResNet [37] suggested that a layer learns the remaining mapping with reference to the layer inputs $F(x) = H(x) - x$ to ease the training of very deep CNN (up to 152 layers), instead of directly learning the required underlying mapping $H(x)$. The original mapping is recast to $F(x)+x$, which can be achieved through "shortcut connections". In 2017, a "Squeeze-and-Excitation (SE)" module was proposed and obtained the champion of ILSVRC, which adaptively recalibrated

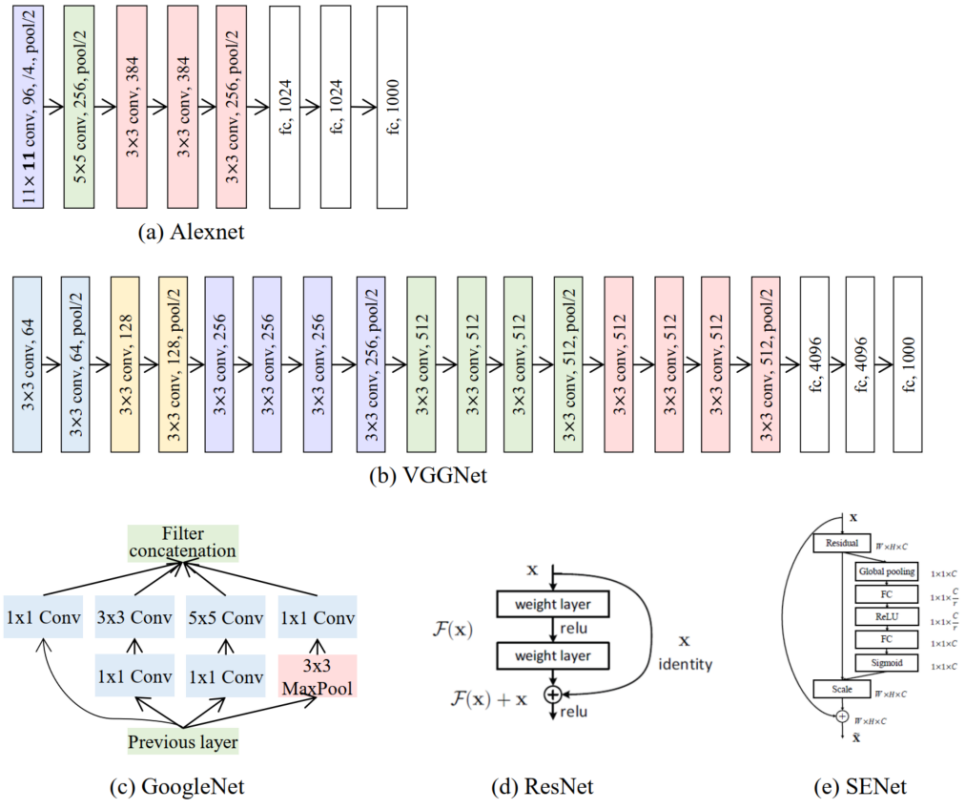


Figure 2.2: Architecture of Alexnet, VGGNet, GoogleNet, ResNet, and SENet. [2] (Best viewed in color)

the channel characteristic response by explicitly modeling the interdependence between channels. These blocks can integrate with modern architectures, e.g., ResNet [37] and ShuffleNet [76], and improve their feature representation capabilities.

With the development of architectures and the cutting edge of training strategies, e.g., Batch Normalization (BN), training became more controllable and the network became deeper. In terms of object classification, following these architectures, the networks in deep face recognition are gradually developing, and the performance of face recognition is also improving.

The mainstream architectures of these deep face recognition methods are shown in Figure 2.2. In 2014, DeepFace [34] was the first proposed deep learning-based face recognition method, which adopted a 9 layers CNN with multiple locally connected layers and has achieved 97.35% accuracy on the LFW dataset. Then, in 2015, FaceNet [36] using GoogleNet and trained by a large private dataset, achieved a good

performance of 99.63%, which employed roughly aligned non-matching/matching patches of triplets generated by the strategy of online triplet mining method and employs a triplet loss function. At exactly the same time, VGGFace [74] proposed a method to collect large-scale datasets from the Internet and then trained the proposed VGGNet using this dataset. Then, the pretrained model was fine-tuned through a triple loss function, which is similar to FaceNet [36]. Finally, the accuracy of VGGFace reached 98.95%. In 2017, SphereFace [1] proposed an Angular softmax (A-softmax) loss and trained a 64-layer ResNet to extract the discriminative face features with angular margins, which has improved the LFW result to 99.42%. At the end of 2017, VGGFace2 [77] was proposed as a new large-scale face dataset, which contains large variations in age, ethnicity, pose, lighting, and occupation. Cao et al. [77] achieved good performance on the IJB-A/B dataset [8] by pre-training SENet [38] firstly using the MS-Celeb-1M dataset [43], and then fine-tuning the model with the VGGFace2 dataset.

2.1.2 Attention mechanism

An attention mechanism is used to mimic human attention, which can concentrate on important information [78, 79]. It makes the neural network focus on important areas of its feature representations. In general, the implementation process of the attention mechanism is divided into two steps: the first step is to calculate the attention distribution on the input information, and the other step is to calculate the context vector according to the attention distribution, as shown in Figure 2.3.

When calculating the attention distribution, neural networks first encode source data features as K, called the key, which can be represented in various representations depending on specific tasks and neural architectures. For example, K can be a feature of a specific area of an image. In addition, it is often necessary to introduce a task-related representation vector q. For example, depending on the particular task, q can also be in the form of a matrix [79] or two vectors [78]. The neural network then

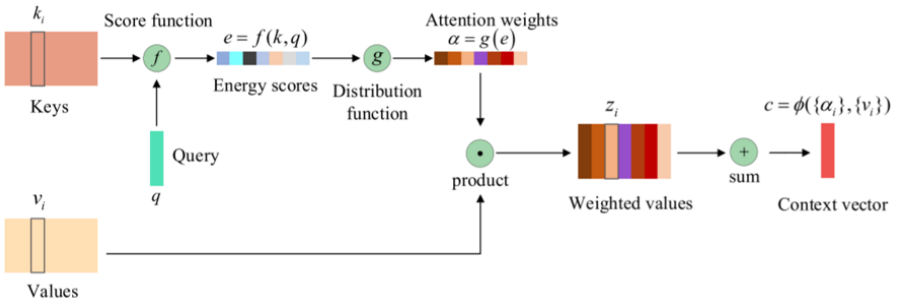


Figure 2.3: Attention mechanism in deep learning. [3] (Best viewed in color)

computes the correlation between the query and the key through the fractional function f (also known as the energy function) to obtain an energy score e that reflects the importance of the query relative to the key when deciding on the next output:

$$e = f(q, K), \tag{2.1}$$

Here, the score function f is the core part of the attention model because it defines how keys and queries are matched and combined.

The attention has been widely used in deep neural networks because of it's advantages, such as sequence-based models, image classification, image localization, and image super-resolution. Residual attention network [80] was proposed as a powerful encoder-decoder style model. A Squeeze-and-Excitation (SE) module was proposed to focus on calculating inter-channel relationships and improve classification performance with a more compact module [38]. Woo et al. [3] extended the SE module and proposed an efficient combination of channel and spatial attention.

In recent years, attention mechanisms have been used in video face recognition. A meta-attention-based aggregation scheme is employed, to fine-grain the weights in an adaptive manner along each feature dimension among all frames to handle the features on a dimensional level. Rao et al. [81] used an attention-aware deep reinforcement learning approach to discard confounding and misleading frames and find the focus of attention in video footage of faces.

2.1.3 Training loss

As described in Section 1.2.1, face recognition is an open-set recognition task for most applications, so we cannot expect to include candidate faces during the training phase, which makes face recognition a "zero-shot" learning task. Fortunately, since all faces have similar shapes and textures, representations learned from a small subset of faces can generalize well to others. Currently, publicly available training database for academic research has only about 10K to 1M subjects. Therefore, academia is devoted to designing effective loss functions and adopting effective architectures to make deep face features more discriminative with training datasets. In this section, research works on different loss functions that have greatly improved deep face recognition methods are reviewed.

Originally, cross-entropy-based softmax loss was firstly adopted for feature learning in face recognition methods, e.g., DeepID [35] and Deepface [34], inheriting from the object classification network, e.g., AlexNet [33], VGGNet [39]. After that, researchers found that the cross-entropy-based softmax losses are not sufficient by itself for learning discriminative face features, and then more and more researchers started to explore novel loss functions for improving the generalization ability, which has become the hottest research direction in the deep face recognition research community. Before 2017, Euclidean distance-based losses played an important role, but since 2017, margin-based losses as well as weight and feature normalization became popular, as illustrated in Figure 2.4. It is noteworthy that, although some loss functions share the similar basic idea, later proposed losses are often designed for facilitating the training procedure by sample selection or easier parameters. Therefore, many well-designed loss functions are presented in this section.

2.1.3.1 Euclidean-distance-based loss

Euclidean-distance-based loss embeds images into Euclidean space using a metric learning approach, in which inter-variance is enlarged and intra-variance is reduced.

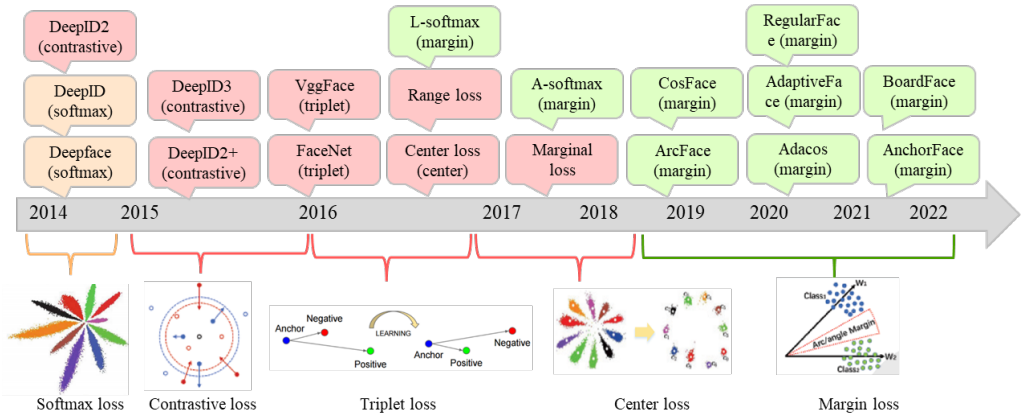


Figure 2.4: Development of training loss functions for deep face recognition. Softmax loss, Euclidean-distance-based loss, and margin-based loss are represented by yellow, red, and blue rectangles, respectively. (Best viewed in color)

Here, triplet loss and contrastive loss are commonly used as the Euclidean-distance-based loss functions. The contrastive loss pushes apart negative pairs and pulls together positive pairs using face image pairs, which can be define as follows:

$$L = y_{ij} \max\left(0, \left\| f(x_i) - f\left(x_j\right)\right\|_2 - \varepsilon^{+}\right) + \left(1 - y_{ij}\right) \max\left(0, \varepsilon^{-} - \left\| f(x_i) - f\left(x_j\right)\right\|_2\right), \tag{2.2}$$

where, $y_{ij} = 1$ means the positive sample pairs and $y_{ij} = 0$ means the negative sample pairs, which can be define as follows:

$$y_{ij} = \begin{cases} 0 & x_i \text{ and } x_j \text{ are from different identities} \\ 1 & x_i \text{ and } x_j \text{ from the same identity} \end{cases} \tag{2.3}$$

$f(\cdot)$ is a face feature extractor, ϵ^{-} and ϵ^{+} control the distance of the positive and negative sample pairs, respectively. DeepID2 [35] combined the face verification (contrastive loss) and identification (softmax loss) supervisory signals to extract a discriminative face representation, and Joint Bayesian (JB) was used to obtain a robust face embedding space. DeepID2+ [82] added supervision to early convolutional layers and increased the dimension of hidden representations by extending from DeepID2 [83], DeepID3 [84] further adopted GoogleNet and VGGNet to their work. Nevertheless, the most serious problem of the contrastive based loss is that the

margin parameter is usually hard to set.

Contrary to the contrastive-based loss that considers the absolute distances of the positive and negative sample pairs, the triplet loss is another idea that considers their relative distances. It was first used by FaceNet [36] proposed by Google, which is widely used in face recognition. The triplet loss first builds face triplets, and then maximizes the distance between a negative sample from a different identity and an anchor, and minimizes the distance between a positive sample from the same identity and the anchor. The triplet loss of FaceNet is defined as follows:

$$\left\| f\left(x_i^a\right) - f\left(x_i^p\right) \right\|_2^2 + \alpha < - \left\| f\left(x_i^a\right) - f\left(x_i^n\right) \right\|_2^2, \tag{2.4}$$

where, x_i^n and x_i^p are negative and positive samples, respectively, x_i^a is the anchor, and α is a margin. $f(\cdot)$ is a feature extractor for embedding a face image into a face feature space. Inspired by FaceNet [36], Triplet Similarity Embedding (TSE) [85] and Triplet Probabilistic Embedding (TPE) [85] construct triplet loss by learning a linear projection W. Meanwhile, some methods optimize deep face recogniton models using both softmax loss and triplet loss [86, 87]. They first pretrain face recognition networks using softmax loss, and then fine-tune it using triplet loss.

However, due to the difficulty of selecting effective training samples, the triplet loss and contrastive loss occasionally encounter training instability, and some researchers started to explore some simple alternatives. Center loss [42] and its variants [88, 89] are good options for reducing intra-variance. Center loss learns the center of each class and penalizes the distance between the deep face feature and its corresponding class center, which is defined as follows:

$$L_C = \frac{1}{2} \sum_{i=1}^{m} \left\| f\left(x_i\right) - C_{y_i} \right\|_2^2, \tag{2.5}$$

where $f\left(x_i\right)$ represents the extracted feature of sample x_i belonging to the y_i-th class. C_{y_i} represents the y_i-th class center of the face feature. To handle long-tail data, range loss [90] is a variant of center loss used to minimize the mean harmonic of

the k largest ranges in a category and maximize the shortest inter-class distance in a batch. Wu et al. [88] proposed a kind of center-invariant loss that penalizes the difference between each center of the class. Deng et al. [91] chose the farthest within-class sample and the nearest between-class sample to calculate the marginal loss. However, the center loss and its variants are affected by the diversity and banlance of training data for each identity, and GPU memory consumes significant size at the classification layer.

2.1.3.2 Margin-based loss

Since 2015, researchers have gradually begun to deeply understand the loss in face recognition. In 2017, researchers have found that samples should be separated more strictly to avoid misclassifying the hard samples. Then, the margin-based loss [1] was proposed to make learned face features potentially separable with a larger margin distance. The decision boundary of the softmax loss is defined as:

$$(W_1 - W_2)\, x + b_1 - b_2 = 0, \tag{2.6}$$

where x denotes the extracted features, W_i and b_i are learned weights and bias of the softmax loss, respectively. Liu et al. [54] firstly proposed the large margin softmax (L-Softmax) loss [54] by reformulating the original softmax loss. When constraining $b_1 = b_2 = 0$, the decision boundaries for class 1 and class 2 can be defined as:

$$\|x\| \left(\|W_1\| \cos\left(m\theta_1\right) - \|W_2\| \cos\left(\theta_2\right)\right) = 0, \tag{2.7}$$

and

$$\|x\| \left(\|W_1\| \, \|W_2\| \cos\left(\theta_1\right) - \cos\left(m\theta_2\right)\right) = 0, \tag{2.8}$$

respectively, where θ_i is the angle between W_i, and x and m are positive integers introducing angular margins. Because of the non-monotonicity of the cosine function, a piece-wise function is used in L-softmax to guarantee the monotonicity. The loss

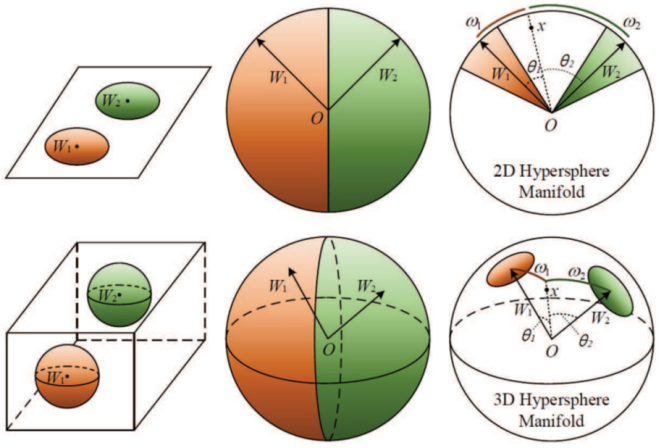


Figure 2.5: Geometry distribution of A-Softmax loss. [1] (Best viewed in color)

function can be defined as follows:

$$L_i = -\log\left(\frac{e^{\|w_{yi}\|\|x_i\|\varphi(\theta_{yi})}}{e^{\|w_{yi}\|\|x_i\|\varphi(\theta_{yi})+\Sigma_{j\neq yi}\|w_{yi}\|\|x_i\|\cos(\theta_j)}}\right), \tag{2.9}$$

where $\varphi(\theta) = (-1)^k \cos(m\theta) - 2k, \theta \in \left[\frac{k\pi}{m}, \frac{(k+1)\pi}{m}\right]$. To facilitate and ensure the model convergence, the L-Softmax loss is always combined with softmax loss since it the L-Softmax [54] is difficult to converge if training directly. So, the loss function is changed into $f_{yi} = \frac{\lambda\|w_{yi}\|\|x_i\|\cos(\theta_{yi})+\|w_{yi}\|\|x_i\|\varphi(\theta_{yi})}{1+\lambda}$, where λ is a dynamic hyper-parameter. Then, Liu et al. [1] proposed A-softmax loss to further normalize the weight W by L_2 norm based on the L-softmax, so that the normalized vector will lie on a hypersphere manifold with an angular margin, as shown in Figure 2.5.

To solve the optimization difficulty of A-Softmax [1] and L-Softmax [54] during training, which combine the angular margin and a multiplicative manner, CosFace [55] and ArcFace [41], respectively introduced an additive cosine/angular margin $\cos(\theta) + m$ and $\cos(\theta + m)$. Here, they are able to converge without the softmax supervision without tricky hyper-parameters λ and are more clear. The decision boundaries of the binary classification case are defined as in Table 2.1.

Based on large margin, AdaptiveFace [92] and FairLoss [93] are further proposed to address the problem of unbalanced data by adaptively adjusting the margins for different classes. Compared to the Euclidean-distance-based loss, the margin-based

Table 2.1: Decision boundaries for class 1 of binary classification case.

Loss Functions	Decision Boundaries
Softmax	$(W_1 - W_2)\, x + b_1 - b_2 = 0$
L-Softmax [54]	$\|\|x\|\| \left(\|\|W_1\|\| \cos\left(m\theta_1\right) - \|\|W_2\|\| \cos\left(\theta_2\right)\right) > 0$
A-Softmax [1]	$\|\|x\|\| \left(\cos\left(m\theta_1\right) - \cos\left(\theta_2\right)\right) == 0$
CosFace [55]	$\|\|x\|\| \left(\cos\left(m\theta_1\right) - m - \cos\left(\theta_2\right)\right) == 0$
ArcFace [41]	$\|\|x\|\| \left(\cos\left(\theta_1 + m\right) - \cos\left(\theta_2\right)\right) == 0$

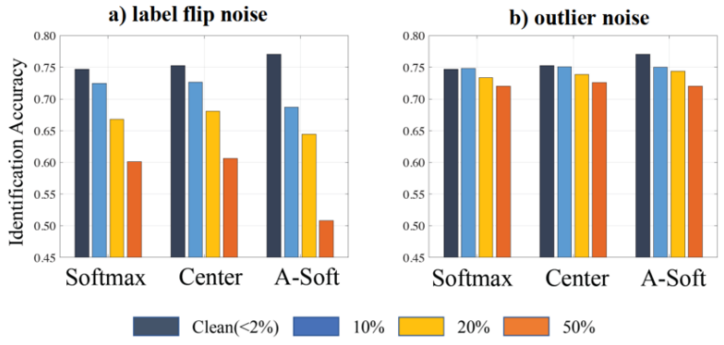


Figure 2.6: Rank-1 identification results on 1:1M MegaFace benchmark. (a) The effect of the label flips in training dataset on performance. (b) The effect of the outliers in training dataset on performance. [4]

loss explicitly adds the discriminative constraints on a hypershpere manifold, which intrinsically matches the prior that a human face lies on a manifold. The recent work Variational Prototype Learning (VPL) [94] first analyzes the limitations of previous methods, which employ sample-to-prototype comparisons during training without considering sample-to-sample comparisons, and then introduces the sample-to-sample comparisons into the classification framework for face recognition. AnchorFace discusses the necessity of the optimization under the Anchor FAR (i.e. Anchor Optimization) for practical face recognition from a new perspective, and introduces a pair of loss functions to reduce the gap of the training and evaluation for FR. However, Wang et al. [4] verified that cosine/angular-margin-based loss can attain better preformance on a clean training dataset, but is liable to noise, and even becomes worse than the center loss and the softmax loss in the high-noise region, as illustrated in Figure 2.6.

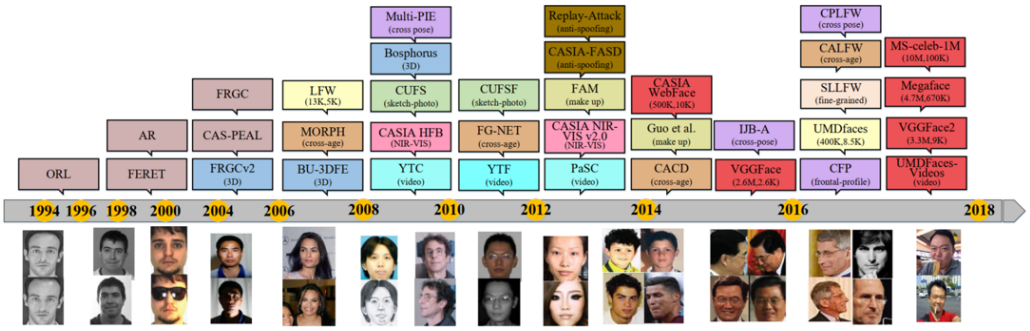


Figure 2.7: Evolution of face recognition datasets. Red rectangles shows face training datasets, and other color rectangles shows face testing datasets with different scenes and task. (Best viewed in color)

2.2 Face Datasets

In the past several decades, many face datasets have been built from single source to diverse sources, from small-scale to large-scale, and from lab-controlled scene to real-world unconstrained scene, as illustrated in Figure 2.7. Prior to 2007, early researches in face recognition focused on small-scale and controlled datasets. In 2007, LFW [32] dataset was constructed which marks the beginning of face recognition under unconstrained scene. As the performance of face recognition on some simple datasets tend to saturate, more and more complex face datasets are continuously constructed to facilitate the face recognition research. Since that, more testing datasets with different scenes and tasks are built. In 2014, CASIA-Webface [95] was the public face training dataset, and is widely used in face recognition community, and then large-scale face training datasets begun to be a hot research topic. It is no exaggeration to say that the development process of face datasets have largely led the direction of face recognition research.

However, publicly available training datasets are mostly celebrity images collected from the Internet, which is far different from images captured in the real-world with diverse scenes. In this section, the development of major face training and testing datasets are reviewed.

2.2.1 Training datasets

The prerequisite for effective deep face recognition is a sufficiently large-scale training dataset. Zhou et al. [87] showed that large amounts of training dataset with deep learning can improve the performance of face recognition. The results of Megaface Challenge [50] also showed that premier deep face recognition methods were typically trained on training datasets larger than 20K subjects and 0.5M images. Early research works of deep face recognition were usually trained on private training datasets. Deepface [34] was trained on a private large-scale training dataset with 4M images of 4K subjects; FaceNet [36] was trained on 200M images of 3M subjects; DeepID serial models [35] were trained on a private large-scale training dataset with 0.2M images of 10K subjects. Although they reported making ground-breaking progresses, we cannot compare their models or accurately reproduce them since they are trained on private datasets.

To overcome this problem, CASIA-Webface [95], which contains 0.5M images of 10K celebrities downloaded from the Internet, was the first public large-scale training face dataset. It is widely used in the face recognition community, and then large-scale face training datasets became a hot research topic. Due to its moderate size and ease of use, it has become an excellent resource for fair comparison of academic deep learning models. Nevertheless, its relatively small-scale training dataset and the number of subjects may not fully utilize the potential ability of many advanced deep learning approaches. Currently, there are more training datasets that provide publicly available large-scale training datasets, as shown in Table 2.2. Especially some datasets are composed of more than 1M images, such as MS-Celeb-1M [43], VGGface2 [77], DeepGlit [44], MegaFace [50], Glint360K [45], and WebFace260M [46].

The unconscious surveillance face recognition task is obviously under-studied in comparison to general face recognition task, which can be seen from Table 2.2. For example, there are 17M celebrity face images from 360K subjects in the Glint360K [45] collected from the Internet, while the largest commonly used video surveillance

Table 2.2: Commonly used publicly available face recognition datasets for training.

Dataset	Published	Celebrity/ commonalty	Still/ Video	# photos/video	# subjects	#of photos per subject
CASIA-WebFace [95]	2014	Celebrity	Still	0.5M	10K	47
CelebFaces [35]	2015	Celebrity	Still	0.2M	10K	20
VGGFace [74]	2015	Celebrity	Still	2.6M	2.6K	1,000
MS-Celeb-1M [43]	2016	Celebrity	Still	10M/3.8M	100K/85K	100/44
MegaFace [50] [50]	2016	Commonalty	Still	4.7M	672K	7
VGGFace2 [77]	2017	Celebrity	Still	3.31M	9K	87
UMDFaces-Still [96]	2017	Celebrity	Still	0.3M	8K	45
UMDFaces-Videos [97]	2017	Celebrity	Video	2.6M	2.6K	1,000
UCCS [98]	2017	Commonalty	Video	0.1M	1.7K	59
QMUL-SurFace [13]	2019	Commonalty	Video	0.22M	5.3K	41
Glint360K [45]	2020	Celebrity	Still	17M	360K	47
WebFace260M [46]	2021	Celebrity	Still	260M	4M	6.5

dataset only has 0.22M from 5.3K subjects in the QMUL-SurFace [13]. This is because still celebrity face images are easier to collect. On the contrary, since data acquisition is greatly limited due to largely restricted data access, it is not feasible to build a large-scale real-world surveillance face image dataset that can be used for training.

2.2.2 Testing datasets

An overview of representative face recognition testing datasets are summarized in Table 2.3. Specifically, early datasets focus on small-scale controlled face recognition scenarios with limited number of images and subjects [19]. The early datasets provides neither adequate inter-variation or diversity for training, nor are effective for solid evaluation. In 2007, the most influential testing dataset LFW [32] was proposed and begun to shift the research community towards recognizing unconstrained faces by providing celebrity face images from the Internet and a standard protocol for evaluation. After that, LFW has greatly promoted the interest and progress of face recognition. This trend towards large datasets is amplified by the creation of even larger face recognition benchmark dataset such as CASIA-WebFace [95], CelebFaces [35], VGGFace2 [77], MS-Celeb-1M [43], and MegaFace [50]. So far, it seems to have solved the problem of the availability of large-scale training and testing benchmark datasets collecting Web celebrity facial images.

Table 2.3: Commonly used public available face recognition datasets for testing.

Datasets	Published	Celebrity/ commonalty	Surveillance	#photos/video	#subjects
Yale [99]	2001	Cooperative	No	10	5,760
CMU [100]	2002	Cooperative	No	68	41,368
Multi-PIE [11]	2010	Cooperative	No	337	750,000
Morph [101]	2006	Celeb(Web)	No	13,618	55,134
LFW [32]	2007	Celeb(Web)	No	5,749	13,233
YouTube Face [10]	2011	Celeb(Web)	No	3,425	1,595
FaceScrub [102]	2014	Celeb(Web)	No	530	100,000
IJB-A [8]	2015	Celeb(Web)	No	500	5,712
VGGFace [74]	2015	Celeb(Web)	No	2,622	2.6M
UMDFaces [96]	2016	Celeb(Web)	No	8,277	367,888
CFP	2016	Celeb(Web)	No	500	7,000
UMDFaces [96]	2016	Celeb(Web)	No	8,277	367,888
IJB-B [103]	2017	Celeb(Web)	No	1,845	11,754
MegaFace2 [50]	2017	Non-Celeb	No	672,057	4,753,320
FERET [20]	1996	Cooperative	No	1,199	14,126
FRGC [104]	2004	Cooperative	No	466+	50,000+
CAS-PEAL [105]	2008	Cooperative	No	1,040	99,594
PaSC [9]	2013	Cooperative	No	293	9,376
SCface [106]	2011	Cooperative	Yes	130	4,160
COX [5]	2015	Cooperative	Yes	1,000	1,000
UCCS [98]	2017	Uncooperative	Yes	1,732	14,016+
QMUL-SurvFace [13]	2018	Uncooperative	Yes	15,573	463,507

With those large benchmark challenges, the performance of face recognition in high-quality face images has achieved an unprecedented level using deep learning, e.g., the performance of face recognition has achieved 99.83% on LFW for face verification and 99.81% on MegaFace [50] for face identification. Nevertheless, this does not scale to real-world surveillance facial images captured in unconstrained environments and uncooperative users. This is due to two reasons: the first one is that existing face recognition challenges have different degrees of bias of data selection (less motion blur, near-frontal pose, and better illumination). The other one is that deep learning methods are often domain-specific for open-set protocols, i.e., only generalize well to face images similar to the training dataset. What is more, there is a big difference in facial images between the celebrity face image collected from the Internet and the real-world surveillance face image in-the-wild, as shown in Figure 1.5.

Since 1996, research on surveillance face recognition has not made much progress, when the well-known FERET challenge [20] was launched. However, it has not been

studied enough and a very few number of benchmark datasets are available. One of the main hurdles is the difficulty of building large-scale real-world surveillance face recognition datasets due to the limited feasibility for such face images. In most cases, only simulated surveillance face images are collected, and the image settings are carefully controlled. Thus, it provides a high face image quality than that can be obtained from local surveillance video.

The UCCS face challenge [98] introduced a notable recent study, which is a large-scale surveillance face benchmark in the public domain. The face images in the challege are not collected based on subjects' cooperation (unconstrained), since they are captured from a long-range distance. These faces images include blur, various poses, and occlusions. This dataset represents a real-world surveillance face recognition scenario in comparison to FERET. Nevertheless, the images in the dataset were captured from a single camera view with high-resolution, so providing obviously more facial details with less viewing angle variations. What is more, the dataset is small in size, particularly in term of the face identity numbers (1,732), statistically limited for evaluating a surveillance face recognition challenge. In 2018, the QMUL-SurvFace benchmark [13] addressed the limitations of the UCCS dataset by building the largest scale natively surveillance face recognition challenge (the QMUL-SurvFace benchmark), which contains 463,507 face images of 15,573 different identities captured from a diverse source of the real-world public spaces.

2.3 Training and Testing protocols

In the past three decades, as face recognition technology moved from laboratories to the real world, not only have face datasets show a clear tendency, e.g., from single-source to diverse-sources, from small-scale to large-scale, and from lab-controlled to real-world unconstrained scene, but also face training and testing protocols have changed dramatically, e.g., from closed-set face recognition to open-set face recognition, from small-scale (10K) to large-scale (10M). In this section, major training and testing protocols are discussed.

2.3.1 Training protocols

According to whether the testing identities are disjoint from the training dataset, face recognition can be divided to the closed-set and open-set protocols, as shown in Figure 1.6.

Closed-set face recognition. In this case, as described in Section 1.2.1, since all testing identities are pre-defined in the training dataset, it is natural to classify testing face images to the given identities. Hence, closed-set face recognition can be well solved as a classification problem, in which features are anticipated to be separable. The closed-set face recognition protocol is mostly used by the early-stage (before 2010) face recognition studies on FERET [20] and AR [107], and only suitable for some small-scale real-world applications.

Open-set face recognition. In this case, the testing identities do not overlap with the training dataset, which makes face recognition more challenging and close to real-world application scenarios. Since it is impossible to classify test face images to known identities in the training dataset, open-set face recognition actually learns a feature extractor to extract the facial features, and then use it for matching. Since human faces exhibit similar within-subject variations, deep models can show transcendent generalization ability when trained with a sufficiently large set of generic subjects, in which the key is to learn discriminative large-margin deep face features. Almost all major face recognition benchmarks, e.g, PaSC [9], IJB-A/B/C [8], LFW [32], and Megaface [50], need testing models to be trained under the open-set evaluation protocol.

2.3.2 Testing protocols

To evaluate whether deep models can address the different problems of face recognition in a real-world application, many testing datasets are built to evaluate the models in different tasks, i.e., face verification and face recognition. In both tasks, a set of known identities is firstly enrolled in the system's gallery, and during testing, a new

subject as a probe is presented. Face verification calculates one-to-one similarity between the gallery and the probe to determine if two images belong to the same subject, while face recognition calculates one-to-many similarity to determine the specific identity of the probe's face. Face verification and face recognition will be discussed in the following part, respectively.

Face verification is relevant to re-identification, access control systems, and application independent evaluations of face recognition algorithms, which is a way of allowing a robot and a computer to confirm that a person is who they claim to be. For example, let us take an example of a user sitting at home. The user wishes to apply for a visa for an upcoming holidays. The user picks up the phone or opens a laptop PC, and then logs into the government visa service. The user then scans the passport with a device-embedded camera to prove his/her identity, then scans the face. Face verification technology can confirm that the user's physical face matches the face in an ID document and that he/she is real and are completing this application. It is typically measured using estimated average accuracy (ACC) and Receiver Operating Characteristics (ROC). Given a threshold (independent variable), ROC analysis measures the True Acceptance Rate (TAR) and the False Acceptance Rate (FAR). TAR is the fraction of genuine comparisons that correctly exceed the threshold, while FAR is the fraction of impostor comparisons that incorrectly exceed the threshold. ACC is a simplified metric introduced by LFW [32] that represents the percentage of correct classification. With the development of deep face recognition, the degree of security is required more and more strictly by testing datasets to match the fact that customers concern more about the TAR when FAR is kept in a very low rate in most security certification scenario. The PaSC dataset [9] evaluates the TAR at a FAR of 10^{-2}; the IJBA dataset [8] increases it to TAR@FAR=10^{-3}; the Megaface dataset [50] focuses on TAR@FAR=10^{-6}; especially, in the Ms-Celeb-1M challenge 3 [43], TAR@FAR=10^{-9} is required.

Meanwhile, face identification (also named face recognition) is relevant to user-driven searches to verify the identity of an individual, which can be widely used in security applications, such as controlling access through gates, doors, or other

physical barriers. It is also used in real-world applications for authentication and identification purposes. For example, let us take an example of a user walking across a shopping mall, or sitting in a seat at a stadium. Face recognition, combined with Closed Circuit TeleVision (CCTV), is scanning the crowds and matching faces against a suspected criminals or a known database. A user may not know if or when face identification is being carried out on him/her. Rank-*N* is a commonly used metric in this scenario. It is based on the percentage of probe searches returning the probe's gallery mate within the top *N* rank ordered results, e.g., the rank-1 and rank-5 recognition rates are adpoted for evaluation.

2.4 Masked Face Recognition

In this section, related works of Research Topic 1 are reviewed, which include methods for simulating masked face images and masked face recognition methods.

2.4.1 Simulating masked face images

Recently, some methods of simulated masked face image have been proposed [57, 58, 59, 60]. MaskTheFace [57] used a Dlib-based [108] face landmark detector to identify the face tilt and six key features of the face necessary for applying a mask. MaskedFace-Net [59] defined a mask-to-face deformable model and applied homographic transformation for mapping mask pixels over the targeted facial areas. Firstly, feature-based cascade classifiers are used to detect a region of interest in the facial image, with which a key-point detector is used to automatically detect 68 landmarks representing the facial structure. Besides, an image of a conventional face mask is selected as a reference image for the mapping where twelve key points are manually annotated for delineating the mask area. Finally, a homographic transformation is applied to map mask pixels over the targeted facial areas relying on the defined point-to-point correspondence of landmarks between mask image and face image.

However, since these methods only utilize affine transformation, the added masks often look unnatural. Furthermore, they ignore pose and illumination consistency thus leading to biased masked face augmentation. Recently, Generative Adversarial Network (GAN) has become a powerful technique used for data augmentation [61]. However, GAN-based methods suffer from mode collapse deeply, which usually manifests that the images generated by the generator tend to be highly similar amongst them, even though their corresponding latent vectors are very different. In addition, GAN-based methods are generally slow and difficult to run online in recognition. On the contrary, the method proposed in this thesis can quickly collect various types of mask images and can transfer them to the face image in run-time for the mask-aware similarity matching strategy in the inference stage.

2.4.2 Occluded face recognition

Many different deep-learning-based approaches have been proposed to solve the occlusion problem. In 2014, Sun et al. found that the features learned by DeepID2+ [82] show certain robustness to image corruption in face verification tasks, and the combination of DeepID2+ features extracted from 25 face patches may further improve the robustness. Daniel et al. [109] used the augmented training data with synthetic occluded faces to tackle the occlusion problem.

Recently, Masked face recognition has attracted much attention during the COVID-19 pandemic. Anwar et al. [57] proposed an open-source tool, named MaskTheFace, to create masked face dataset from a face dataset with extended feature support, and then used this dataset to re-train existing face recognition engines to improve their accuracy. Hariri and Walid [110] developed a reliable method based on occlusion removal and deep learning-based features to address the problem of the masked face recognition process. The first step is to remove the masked face areas. Next, three pre-trained deep CNN namely; VGG-16 [39], AlexNet [33], and ResNet-50 [37], are used to extract deep features from the obtained regions (mostly eyes and forehead regions). The Bag-of-Features (BoF) paradigm is then applied to the feature

maps of the last convolutional layer to quantize them and obtain a slight representation compared to the fully connected layer of a classical CNN. Finally, a Multilayer Perceptron is applied for the classification process. Mundial et al. [111] used a supervised learning approach for masked face recognition together with in-depth neural network-based facial features. First, a CNN model was trained to generate an embedding of features of an image. Then they focused on a dataset which helps in building a classifier for masked face, which consists of three images of a person, two masked face images, and one without a face mask. In the end, the Support Vector Machine (SVM) is used for classification.

The covered facial areas contain many salient features that make it useful for face recognition, such as the nasal region [112], but the extracted features of the covered facial areas are damaged due to occlusions caused by the mask. Therefore, compared with uncovered facial areas such as the eye region, intuitively speaking, the mask area does not contain much discriminative information useful in recognizing a face, which gives us the hint that more attention should be paid to the uncovered region in feature extraction. Recently, attention mechanisms have been introduced to video face recognition systems [81], where an attention mechanism is adopted to mimic human perception to focus on important information.

2.5 Video Face Recognition

In this section, related works of Research Topic 2 are reviewed, which include general video face recognition and feature aggregation for face recognition.

2.5.1 General video face recognition

Video Face Recognition (VFR) has the disadvantages of being in low resolution and includes dramatic pose variations compared with still face recognition. Meanwhile,

it also has the advantages of complementary information in consecutive frames. Existing works on VFR can be categorized into two main categories: one is to exploit complementary information contained in multiple video frames, while the other one is to extract higher quality features from each frame.

With frame sequence as input, person-specific facial dynamics can be extracted from continuous video frames using robust face trackers [113]. Aggregation-based methods [1, 69, 71, 72, 113, 114] aim to obtain a compact and discriminative feature aggregated by all frame-level features in a video using an adaptive weighting scheme. Meanwihle, key frame selection methods [115, 116, 117] attempt to gain only a subset of best-quality frames from video clips using frame quality evaluation for efficient face recognition.

Recent Deep Learning (DL) methods, such as A-softmax [1] , CosFace [55], and ArcFace [41], introduce margin into the softmax loss to extract more discriminative face features. To solve the blurring problem in video caused by the relative motion between the cameras and the subjects, deblur-based methods [113] deblur a blurred image by estimating a blur kernel, and then extract the features. Data uncertainty modeling is another strategy for unconstrained face recognition [118, 119, 120], especially for noisy images. In these works, data uncertainty learning is applied to capture both the feature (mean) and the uncertainty (variance), simultaneously. Inspired by these works, in this thesis, an Attention-aware Masked face recognition Network (AMaskNet) is proposed for masked face recognition, which puts more weight to useful features while (in relative terms) ignoring those corrupted by the face mask by learning a contribution matrix.

2.5.2 Feature aggregation for face recognition

A video provides us with abundant complementary information across frames compared with a still image. Therefore, aggregation of information across frames to obtain more valuable and effective video-level features is a crucial issue for robust recognition against variations. Neural Aggregation Network (NAN) [72] proposed

two attention blocks to adaptively weight the frames. Discriminative Aggregation Network (DAN) [121] proposed a network to aggregate raw video frames directly instead of the features obtained by complex processing. Quality Aware Network (QAN) [122] automatically estimated the quality score for each sample in a set by a quality estimator, and weighted all frames by the predicted quality scores. Region-based Quality Estimation Network (QAN+) [114] further extended the idea of QAN into local regions, which used an ingenious training mechanism to extract the complementary region-based information between different frames. COmpact Second-Order Network (COSONet) [123] proposed a second-order network to extract features from faces with large variations and a mixture loss function to encourage the discrimination and simultaneously regularizes the feature. Multicolumn Network (MN) [71] took entire images in a set as input, and learned to compute a set-level fix-sized feature representation.

Each component of the feature vector may encode different subsets of facial features, thus bias could be caused when we emphasize or suppress all components simultaneously. To alleviate this problem, a meta attention-based aggregation scheme is used in [113], to adaptively fine-grain the weights along each feature dimension among all frames so as to handle the feature on dimension level. Similarly, component-wise feature aggregation scheme is used in C-FAN [69] for video face recognition, where the quality value for each feature component is separately learned. C-FAN automatically learns to suppress features with low-quality scores, while retaining salient face features with high-quality scores.

As a summary, the aim of feature aggregation methods is to automatically learn the weights from frame level or feature component level, and the quality criterion is used therein to represent the importance of each single frame or each feature component. Therefore, these methods are usually called as quality-based feature aggregation methods.

However, limited by the mini-batch training strategy, the existing quality-based feature aggregation methods fail to globally consider the relation among frames in a

video clip or all samples of one identity during training, thus lead to bias or inaccuracy in quality estimation. This motivates us to seek a better solution in this thesis, especially to investigate valuable information in the low-quality images.

Chapter 3

Masked Face Recognition with Mask Transfer and Self-Attention

This chapter is dedicated to the Research Topic 1. As discussed in Chapter 1, face masks bring a new challenge to face recognition systems especially during the COVID-19 pandemic, when it is essential to analyze and mitigate the effect of wearing face masks. Therefore, in this chapter, a method used for mitigating the negative effects of mask defects on face recognition is proposed.

Firstly, a low-cost, accurate method of masked face synthesis, i.e., mask transfer, is proposed for data augmentation. Secondly, an Attention-aware Masked face recognition Network (AMaskNet) is proposed to improve the performance of masked face recognition, which includes two modules: a feature extractor and a contribution estimator. Therein, the contribution estimator is employed to learn the contribution of the feature elements, thus achieving refined feature representation by simple matrix multiplications. Meanwhile, the end-to-end training strategy is taken to optimize the entire model. Finally, a mask-aware similarity Matching Strategy (MS) is adopted to improve the performance in the inference stage. Experiments show that the proposed method consistently outperforms on three masked face recognition datasets:

RMFRD [6], COX [5], and Public-IvS [62]. Meanwhile, a qualitative analysis experiments using CAM [7] indicates that the contribution learned by AMaskNet is more conducive to masked face recognition.

This chapter is structured as follows: Section 3.1 introduces this research topic, including background, motivation, and contributions. Then, Section 3.2 describes the proposed method of mask transfer, AMaskNet, and mask-aware similarity Matching Strategy (MS). Section 3.3 presents the experimental results, discussions are given in Section 3.4, and the summary is presented in Section 3.5.

3.1 Introduction

As discussed in Chapter 1, the COVID-19 pandemic has caused a global impact: the World Health Organization (WHO) and the U.S. Centers for Disease Control and Prevention (CDC) have suggested everyone should wear a mask in a public setting especially when other social distancing measures are difficult to maintain [53]. Face recognition is non-contact, highly efficient, user friendly, and so forth, and thus has been widely applied in access control and security authentication in public places. However, face masks bring a new challenge to existing commercial face recognition techniques. Face recognition becomes more difficult when a large part of the face is covered by a mask. Therefore, it is essential to study the effect of masks on the behavior of face recognition systems and design mitigation techniques to offset the inevitable performance loss.

Deep-learning-based approaches predominate in the task of face recognition due to the emergence of advanced Convolution Neural Network (CNN), well-designed loss functions [1, 54, 55], and large-scale datasets [41, 56]. Despite the success of deep learning models under general face recognition scenarios, the deep features still demonstrate imperfect invariance to face masks, where the whole face image cannot be used for the description. Therefore, face masks trigger a significant research challenge: Firstly, it is necessary to collect a large-scale training dataset, which includes

faces with different types of masks. In order to collect such a large-scale training dataset, on the one hand, it is time consuming and incurs higher labour cost, and on the other hand, maintaining the diversity of data in such datasets is a slow process. Therefore, a low-cost, convenient face data augmentation method is needed as a matter of urgency. Secondly, it is essential to mitigate the performance loss from the perspective of model design according to the characteristics of masks.

Some methods of simulating a masked face [57, 58, 59, 60] have been proposed for face data augmentation. MaskTheFace [57] used a Dlib-based [108] face landmark detector to identify facial tilt and six key features of the face necessary for applying a mask. MaskedFace-Net [59] defined a mask-to-face deformable model and applied homographic transformation to map mask pixels over the targeted facial areas. However, since these methods only utilize affine transformation, the added masks often look unnatural. Furthermore, they ignore pose and illumination consistency thus leading to biased masked face augmentation. Recently, Generative Adversarial Network (GAN) has become a powerful technique for data augmentation [61]. However, GAN-based methods suffer from mode collapse deeply, which usually manifests that the images generated by the generator tend to be highly similar amongst them, even though their corresponding latent vectors are very different. In addition, GAN-based methods are generally slow and difficult to run online in recognition. On the contrary, the method proposed in this thesis can quickly collect various types of mask images and can transfer them to the face image in run-time for the mask-aware similarity matching strategy in the inference stage.

Numerous approaches have been proposed to tackle the problem regarding occlusion which is a common problem in computer vision [5, 109]. Wearing a mask is considered the most difficult facial occlusion challenge since it covers most of the face including the mouth and nose. Anwar et al. [57] developed an open-source tool (MaskTheFace) to create a large dataset of masked faces, and then re-trained existing face recognition systems to improve their accuracies. To reduce the negative influence of masks, Hariri [110] directly discarded the masked region when extracting deep features. Mundial et al. [111] used a supervised arning approach and

an in-depth neural network to recognize masked faces and extract individual facial features, with which an Support Vector Machine (SVM) classifier was established for classification purposes.

The covered facial areas contain many salient features useful for face recognition, such as the nasal region [112], but the extracted features of the covered facial areas are damaged due to occlusions caused by the mask. Therefore, compared with uncovered facial areas such as the eye region, intuitively speaking, the mask area does not contain much discriminative information useful in recognizing a face, which gives us the hint that more attention should be paid to the uncovered region in feature extraction. Recently, attention mechanisms have been introduced to video face recognition systems [81], where an attention mechanism is adopted to mimic human perception to focus on important information.

As Research Topic 1 of this thesis, qualitative and quantitative analysis on the effect of wearing a mask on face recognition is performed here, and then a method for mitigating the negative effects of mask defects on face recognition is proposed. Firstly, a low-cost, accurate method of mask transfer is proposed for masked face synthesis by considering pose and illumination consistency. Secondly, an Attention-aware Masked face recognition Network (AMaskNet) is proposed to improve the performance of masked face recognition, which lends more weight to useful features while (in relative terms) ignoring these corrupted by the face mask by learning a contribution matrix. The AMaskNet includes two modules: a feature extractor and a contribution estimator, wherein the latter is employed to learn the contribution of each spatial region which is then combined with the feature to improve its representation capability. An end-to-end training strategy is adopted to optimize the whole network. Finally, a mask-aware similarity matching strategy is proposed to improve the performance in the inference stage. Experiments show that the proposed method consistently outperforms on three masked face recognition datasets: RMFRD [6], COX [5], and Public-IvS [62]. Meanwhile, qualitative analysis experiments using CAM [7] indicate that the contribution learned by the AMaskNet is beneficial to masked face recognition. Indeed, AMaskNet can localize the salient facial areas and

extract more discriminative features from a non-masked face image and can alleviate the performance degradation of the non-masked scene.

In summary, the main contributions are summarized as:

[1]. A low-cost, accurate mask transfer method for masked face data augmentation is proposed by considering pose and illumination consistency. This method can add a mask from any face image with a mask to any face image without a mask.

[2]. Qualitative and quantitative experiments are conducted to analyze the effect of wearing face masks on the behavior of face recognition systems.

[3]. AMaskNet is proposed to improve the performance of masked face recognition.

[4]. A mask-aware similarity matching strategy is proposed for the inference stage, which can be applied to any face recognition scene in which one image with a face mask and the other without a face mask are present.

3.2 Proposed Method

In this section, the mask transfer method for masked face synthesis is firstly described in two steps: construction of a mask gallery and the generation of synthetic masked face images (Figure 3.1(a) and Figure 3.2), and introduction of the proposed AMaskNet. As shown in Figure 3.1(b), AMaskNet incorporates a feature extractor and a contribution estimator, and the latter further consists of two sub-modules: a self-spatial contribution estimator and a self-channel contribution estimator. The two sub-modules are used to learn the spatial contribution and channel contribution, respectively, and the refined features are obtained by combining the two contributions. Moreover, the entire network is optimized through an end-to-end training procedure.

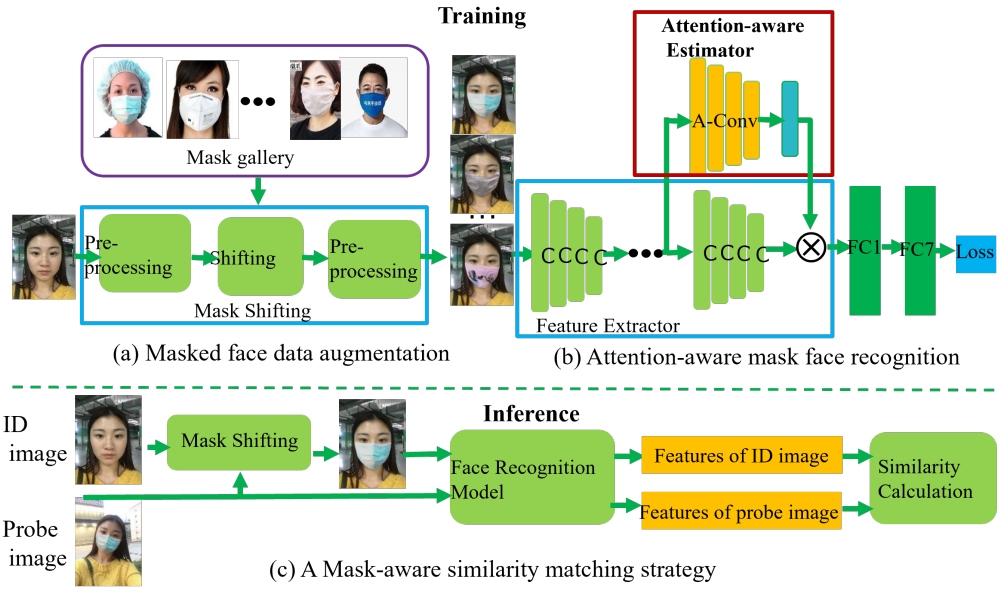


Figure 3.1: Architecture of the proposed masked face recognition method, which includes mask transfer, attention-aware masked face recognition (AMaskNet), and a mask-aware similarity matching strategy for inference. (Best viewed in color)

Finally, a mask-aware similarity matching strategy is introduced for inference purposes, as illustrated in Figure 3.1(c).

3.2.1 Mask Transfer (MT)

To generate the synthetic masked face image, a gallery of different masks should be firstly constructed. Given a non-masked face image and one mask from the gallery, the masked face image is obtained by transferring the mask, including pre-processing, transfer, and post-processing, as illustrated in Figure 3.2. Details are expained below.

3.2.1.1 Collection of mask gallery

The construction of the mask gallery is aimed at obtaining a face image set covering versatile masks, such as different mask colors, shapes, and textures. Since for one mask type, only one face image with this mask is sufficent, it is easy to build this

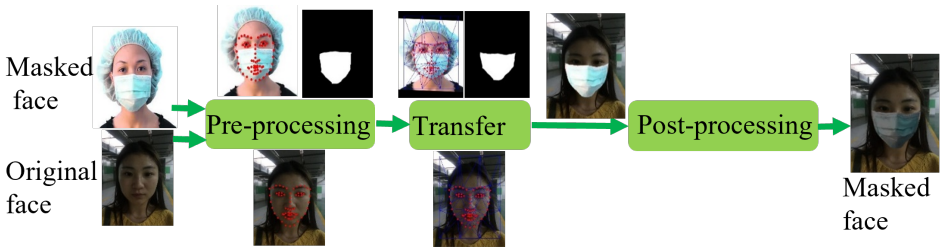


Figure 3.2: Flowchart of Mask Transfer (MT). The masked face is a photo randomly selected from the mask gallery.

dataset through collection from Websites. Generally, any face image with a mask is qualified for the collection. However, to improve the quality of synthetic masked face images, frontal view face images are preferable.

3.2.1.2 Mask transfer for masked face synthesis

Pre-processing. Dlib [108] is used to detect 68 landmark points in both masked and non-masked face images. With these landmarks, a Triangulated Irregular Network [124] is established and the facial area is thus divided into multiple triangular regions. Meanwhile, the grab-cut method [125] is employed to segment mask areas from masked face images.

Transfer. Transfer mask from masked face images to non-masked ones should be implemented based on the geometric relationship bettween the two faces. For each triangular piece in the Triangulated Irregular Network, the affine transformation between the two images is calculated, and the mask region contained in this piece is transformed directly to a non-masked image. The whole mask will be transferred after all triangular pieces have been transformed.

Post-processing. Directly transferring the mask usually leads to inconsistency of illumination in the target image due to the lighting and contrast difference of the two images. For this problem, two post-processing steps are performed after the mask is transferred to the target image: (1) Alpha-matting is utilized to make the boundary more natural in terms of the transition across the boundary; (2) histogram specification is adopted to make the transferred mask region more illumination-consistent

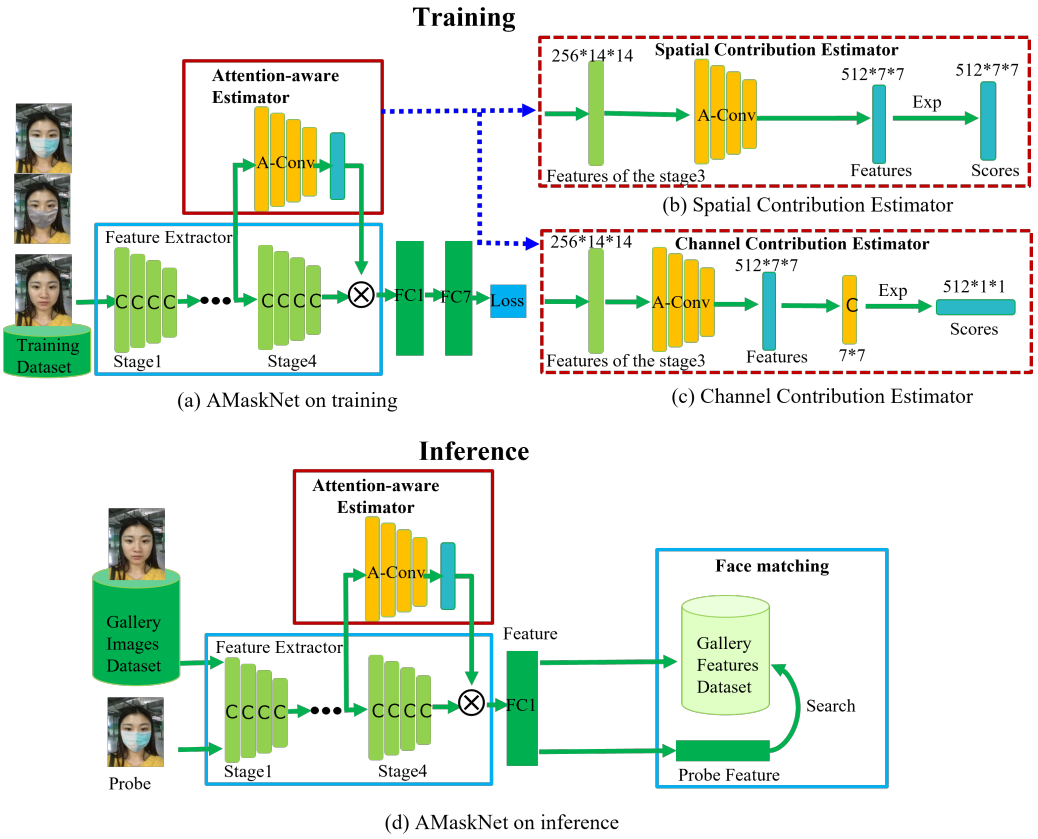


Figure 3.3: Architecture of the proposed AMaskNet. (Best viewed in color)

with the original non-masked face image. More specifically, the gray-scale distribution of the mask region is adjusted according to the gray-scale histogram of the original non-masked image.

3.2.2 Attention-aware Masked face recognition Network (AMaskNet)

The network architecture consists of two modules: a feature extractor and a contribution estimator, as illustrated in Figure 3.3. A conventional face recognition scheme, e.g. ResNet34 [41], may be used for feature extraction, yielding a feature vector as the initial face representation (Figure 3.3(a)). As mentioned previously, the area covered by a face mask does not provide much subject-related information, so the extracted features are less discriminative. For this problem, contribution estimators are designed to learn a contribution matrix to assign more weight to useful features

while (in relative terms) ignoring those corrupted by the face mask. The contribution estimators as an attention scheme is implemented via a branch structure, which includes both spatial (Figure 3.3(b)) and channel components (Figure 3.3(c)).

3.2.2.1 Feature extractor

A conventional face recognition scheme (e.g., ResNet) can be used as a backbone for extraction of features, which is used as the baseline model. Recently, ArcFace [41] has achieved state-of-the-art performance and has been widely used in many papers. Here, ArcFace34 is adopted as the backbone, where BN-Conv-BN-PReLU-Conv-BN module is used as the residual bottleneck and all the convolution kernel size in residual bottlenecks have a size of 3×3. Then, the output feature of all models is fixed to 512-dimensions by a fully connected layer.

3.2.2.2 Attention-aware contribution estimator

The contribution estimators cover both spatial- and channel-wise measurements. Figure 3.3(b) shows the details of the spatial contribution estimator, and Figure 3.3(c) shows the details of the channel contribution estimator. They adaptively aggregate the feature maps in both channel and spatial domains to learn the inter-channel relationship and interspatial relationship matrices. The two matrices are then multiplied with the initial feature representation to produce refined face features. To ensure that the estimated contribution has practical physical significance, the sigmoid function is used instead of the ReLU [33] to map the output onto the interval (0,1), which is used as a contribution coefficient for weighting features. An end-to-end training strategy is adopted to optimize the entire network. After training, the trained feature extractor is applied to extract the feature of gallery images and the feature of probe image, and then carry out face matching to obtain the result of face recognition.

Spatial contribution estimator. In order to estimate the contribution for each component of the feature map, a branch architecture is added to the backbone of the feature extractor model as the spatial contribution estimator, as shown in Figure 3.3(b). In this process, a contribution matrix is learnt which can assign greater weights to useful features while (in relative terms) ignoring those corrupted by the face mask. The contribution matrix and feature map have the same width and height. The structure of the spatial contribution estimator may have different complexities, ranging from one to several convolution layers. A more complex network may result in a better ability to learn, albeit at the cost of the extra computational effort and the risk of overfitting.

Channel contribution estimator. Akin to the spatial contribution estimator, a branch architecture is further added as the channel contribution estimator, in an attempt to estimate the contribution for each channel of the feature map, as shown in Figure 3.3(c). In this module, a contribution matrix will be learnt to put more attention on useful channels. Similarly, different structures of various complexity may be adopted.

3.2.2.3 Feature aggregator

Let the feature extracted by the feature extractor be $F_I \in R^{C*H*W}$, the spatial and channel contribution matrices be $C_S \in R^{C*H*W}$ and $C_C \in R^C$ respectively, the final feature is derived as:

$$F_C = C_C \otimes (C_S \oplus F_I), \tag{3.1}$$

where $\otimes$ respresent matrix multiplication, and $\oplus$ represents element-wise multiplication.

3.2.2.4 Training strategy

In model training, the ArcFace loss [41] is used to penalize identification errors. The bias is fixed as $b_j = 0$, logit is transformed as $W_j^T F_R = \left\|W_j^T\right\| \left\|F_R\right\| \cos\theta_j$, where

θ_j is the angle between the weight W_j and the refined feature F_R. The individual weight is fixed as $\left\|W_j\right\| = 1$ by L_2 normalization. The refined feature is fixed as $\|F_R\|$ by L_2 normalization and is re-scaled to s. This normalization makes the prediction probability only depend on the angle between the feature and the weight. So, the loss is formulated as:

$$L_{id} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{s\cos\left(\theta_{y_i}+m\right)}}{e^{s\cos\left(\theta_{y_i}+m\right)}+\sum_{j=1,j\neq y_i}^{N}e^{s\cos\theta_j}}, \tag{3.2}$$

Here, m is the additive angular margin penalty between F_R and W_{y_i} to enhance the intra-class compactness and inter-class discrepancy simultaneously.

3.2.3 Mask-aware similarity Matching Strategy (MS)

In a real-world application, a given pair of two face photos for matching usually show different styles, that is, the identity or mug-shot photos coming from the gallery are front portrait images without a mask, while the probe images captured in the real-world may be with a mask. Obtaining effective features by focusing more on the non-masked facial region is helpful in such a situation. However, it still cannot eliminate the loss of accuracy caused by the presence of the mask. One straightforward method is to extract features only from upper facial regions when comparing two images with and without masks, but some important information will be neglected, e.g., shape information contained in the mask region. To solve this problem, a mask-aware similarity Matching Strategy (MS) is proposed, as illustrated in Figure 3.1(c). This involves transfer of the mask from the masked face image to a non-masked image, thus mitigating the difference caused by the mask without loss of spatial information. This method is applicable to any face recognition scene in which one image with a mask and the other without a mask is presented, which is especially useful for a 1:1 face verification scene.

3.3 Experiments

In this section, several benchmark datasets and several baseline models are firstly introduced. Then, qualitative and quantitative analyses are undertaken to attain more insight into how face masks affect the performance of face recognition. Finally, the proposed mitigating models are compared with state-of-the-art public models to confirm the effectiveness of the proposed method.

3.3.1 Datasets and protocol

3.3.1.1 Training datasets

The DeepGlint dataset [44] is used as the training corpus. It includes cleaned MS-Celeb-1M [41] and the celebrity Asia [44] datasets, summaring up to a total of 6.6M celebrity images of 172K celebrities therein.

Due to the lack of a large volume of masked face photos to train the model, data augmentation is used for synthesizing masked face images by using the proposed mask-transfer technique. For each training image in DeepGlint, one mask image is randomly selected from the mask gallery and the mask is transferred to this training image. With this manner, the amount of training data is doubled, resulting in around 13M photos.

3.3.1.2 Testing dataset

Several commonly used benchmark datasets, such as RMFRD [6], COX [5], and Public-IvS [62], are used for testing. Details of each dataset are as follow.

COX dataset [5] comprises 1K still images and 3K videos of 1K identities. The video footage is captured using three cameras (Cam1, Cam2, Cam3) set at different locations while the subjects walk in a large gymnasium to simulate a surveillance

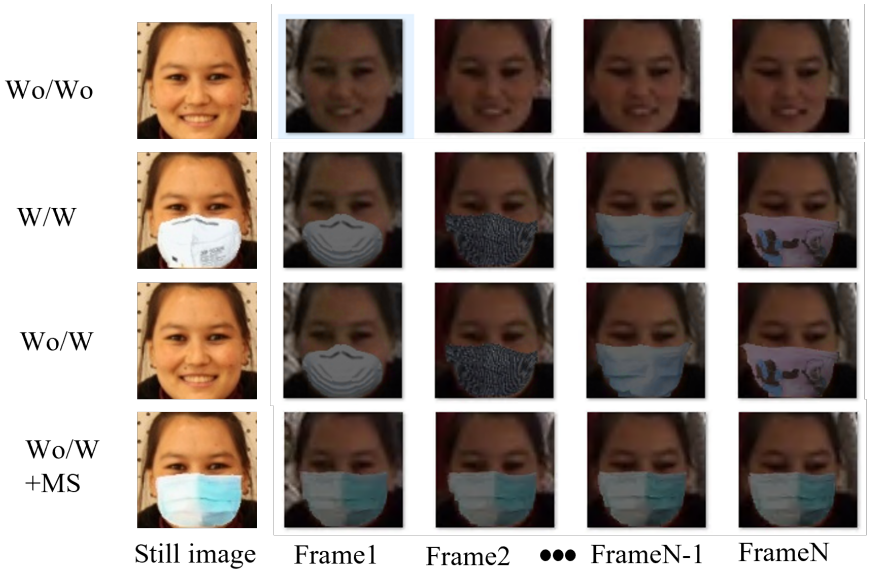


Figure 3.4: Some pairs of face image from the COX dataset [5] and the synthesized COX-mask dataset.

Figure 3.5: Some pairs of face image from the Public-IvS and the synthesized Public-IvS-mask dataset (ID image vs. Spot image).

scenario, as shown in Figure 3.4. A proposed Video-to-Still (V2S) protocol proposed by the author is adopted for performance evaluation, where the true acceptance rate (TAR@FAR = 10^{-4}) is used for the 1:1 verification.

Public-IvS dataset [62] designed for Identity photo versus Spot photo (IvS) face recognition, contains 1,262 identities and 5,503 images, as shown in Figure 3.4. The true acceptance rate (TAR @ FAR = 10^{-5}) for the 1:1 verification protocol is adopted to evaluate its performance.

To analyze the effectiveness of the proposed method on masked face recognition, the following four test conditions are designed to add masks to face photos given the test

image pair in COX and Public-IvS datasets, as shown in Figure 3.4 and Figure 3.5:

[1]. **Wo/Wo**: Original test images without face masks are used.

[2]. **W/W**: Different masks are added to both images in a pair to simulate a scene where both gallery and probe images are of masked faces.

[3]. **Wo/W**: The recognition scene where gallery images do not contain masked faces, but probe images do. To simulate this condition, a mask is added to COX video frames (Public-IvS spot image) while keeping the COX still images (Public-IvS ID image) unchanged.

[4]. **Wo/W+MS**: To improve the accuracy of Wo/W, the mask of the COX video frames (Public-IvS spot image) is transferred to COX still images (Public-IvS ID image), i.e., the same masks are guaranteed to appear in each image pair.

RMFRD dataset [6] is crawled from the Internet, including 5K pictures of 525 people wearing masks, and 90K images of the same 525 subjects without masks, which is mainly devoted to evaluate the existing face recognition system on masked images during the COVID-19 pandemic. Some sample images are illustrated in Figure 3.6. The COX and private-IvS datasets are masked by the proposed mask transfer method, which is convenient to analyze the impact of masks on the performance of the model. However, to evaluate the performance of the proposed model on the real-world masked face recognition, experiments are conducted on the RMFRD dataset and compared with other state-of-the-art methods.

3.3.2 Face recognition model and implementation

3.3.2.1 Face recognition model

The ArcFace model [41], which is the state-of-the-art against several face recognition benchmarks such as LFW [32] and YTF [10], is selected for comparisons.

Figure 3.6: Some pairs of face image from the RMFRD dataset [6]: Face images without a mask (up) and with a mask (down).

Since ArcFace introduced additive angular margin loss to enhance the discriminative power of the face recognition model, it is robust to the condition of wearing a mask. Four publicly available models, MobileFaceNet (ArcFaceM), LResNet34E-IR (ArcFace34), LResNet50E-IR (ArcFace50), and LResNet100E-IR (ArcFace100) are adopted [1]. As introduced previously, the ResNet34 is selected in the proposed method as the backbone for feature extraction, with which the DeepGlint dataset [44] without mask data augmentation is used to construct a baseline model to study the effect of face masks. To mitigate the negative effects of mask, two variants based on ResNet34 are established including data augmentation and the proposed attention scheme, as follows:

[1]. **R34-Baseline**. A ResNet-34 model is trained with the original DeepGlint dataset, to quantify the loss of accuracy that the mask may induce.

[2]. **R34-Mask**. A ResNet-34 model is trained by only mask dataset augmented from the DeepGlint dataset using the proposed masked face synthesis. This model can significantly improve the performance of masked-face recognition, but will degrade the performance of non-masked face recognition.

[3]. **R34-AMaskNet**. This is the proposed AMaskNet model trained with combination of the mask augmented DeepGlint and the original DeepGlint dataset,

[1]Downloaded from `https://github.com/deepinsight/insightface/tree/master/model_zoo` (Accessed 2023-01-26).

which is expected to improve the performance of masked-face recognition while reducing the loss of accuracy inherent to non-masked face recognition.

3.3.2.2 Implementation

The ArcFace loss [41] is used to train the model, where the feature scale s is set to 64 and the arccos margin m is set to 0.5. In training, stochastic gradient descent is adopted with momentum and weight decay values of 0.9 and 0.0005, respectively. The training begins with a learning rate 0.1 for seven epochs, which is then decreased every five epochs by a factor of 10. A total of 25 epochs are taken for the training with the augmented DeepGlint dataset, and 760 images are used in each mini-batch.

3.3.3 Effectiveness of the proposed method

3.3.3.1 Effectiveness of the proposed mask transfer for masked face synthesis

Figure 3.7 illustrates some mask transfer examples. Figure 3.7(a) and Figure 3.7(b) show the mask gallery of the traditional method and the proposed method. The proposed method is simply a face image with a mask, which is easy to obtain and does not need manual annotation. It is low-cost, rapid, and convenient for model development. Figure 3.7(c) and Figure 3.7(d) compare, respectively, the synthesized masked face images of the traditional method and the proposed method. From the results, we can observed that the proposed mask transfer method is effective and can maintain consistency of illumination.

3.3.3.2 Effectiveness of the proposed AMaskNet

Results on COX dataset. Figure 3.8 shows the result of the proposed AMaskNet on the COX dataset. We can see that by re-training the R34-Baseline using a masked augmented dataset, the R34-Mask model achieves a significant improvement on

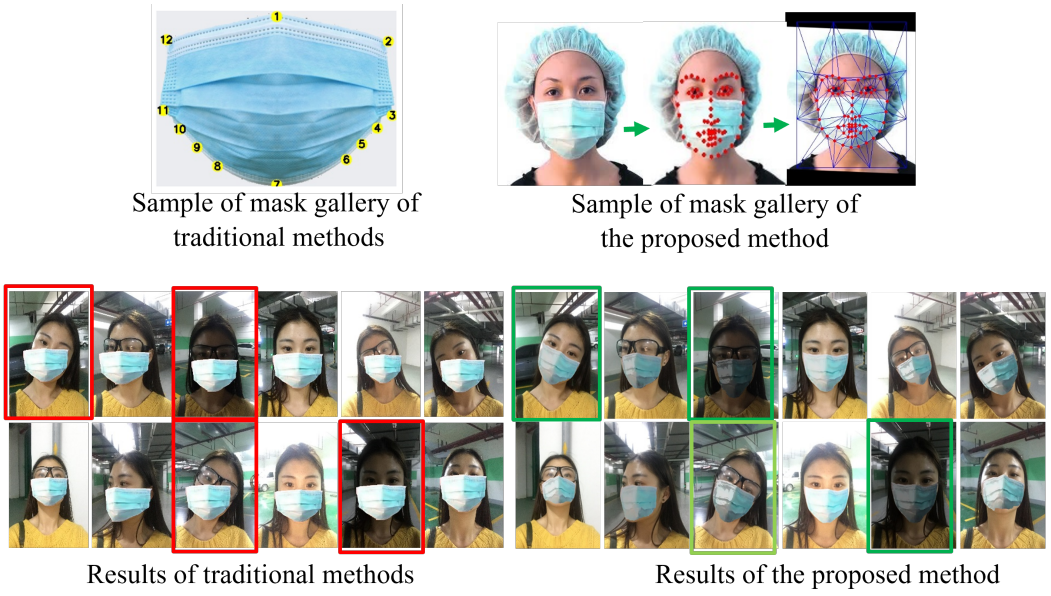


Figure 3.7: Comparation of the MaskedFace-Net [7] and the proposed method. MaskedFace-Net requires manually labeling several key points on the mask boundary (a), while the proposed method automatically extracted the mask region from masked face image (b). Exemplar results when adding a mask are shown in (c) and (d) respectively.

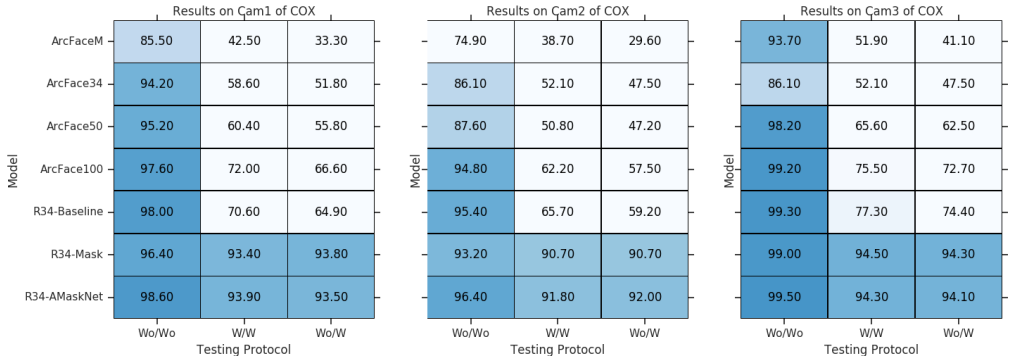


Figure 3.8: Results on the COX dataset with 1:1 verification protocol at TAR@FAR=10^{-4}. From the leftmost to the right are the results of Cam1, Cam2, Cam3, respectively.

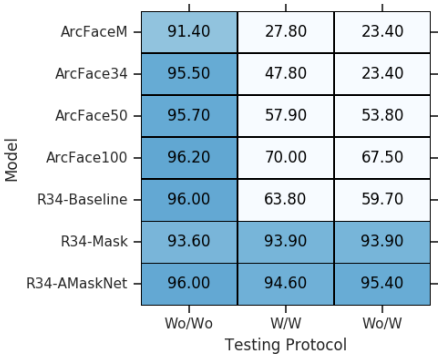


Figure 3.9: Results on the Public-IvS dataset with 1:1 verification protocol at TAR@FAR=10^{-5}.

masked face images, e.g., 28.9 percent improvement in Wo/W on Cam1 of COX. However, this method is likely to cause performance degradation in terms of general face recognition, e.g., 1.6 percent decline in the case of Wo/Wo for the Cam1 of COX. The comparison between R34-AMaskNet and R34-Mask shows that AMaskNet is able to improve the performance, especially for masked face recognition, e.g., 1.1 percent improvement in the case of W/W on the Cam2 of COX, which indicates that the proposed contribution estimator can learn an effective contribution matrix and automatically assign higher weights to the feature map activated by the non-masked facial parts and lower weights to those that are activated by masked facial parts. Meanwhile, the performance of R34-AMaskNet undergoes no significant decline and may even be slightly improved in the case of Wo/Wo. This is because COX is a low-quality video face recognition dataset with dramatic illumination and motion blur. However, AMaskNet can localize the salient facial areas and put more weight to discriminative features, thus improving the performance of wearing masks while minimizing the effect of general face recognition on existing face recognition systems.

Results on Public-IvS dataset. Figure 3.9 shows the results on Public-IvS. We can obtain a similar conclusion to the COX. Here, Wo/Wo: no mask, original test images are used. W/W: different masks are added to both images in a pair. Wo/W: the recognition scene where ID images do not contain masked faces, but spot images do. Although R34-Mask can improve the performance of masks in the case of Wo/W or

Table 3.1: Results on the COX dataset with 1:1 verification protocol at TAR@FAR=10^{-4}.

Model	Cam1		Cam2		Cam3	
	×	✓	×	✓	×	✓
ArcFaceM	33.3	41.4	29.6	40.9	41.1	55.8
ArcFace34	51.8	56.9	47.2	49.8	47.5	49.8
ArcFace50	55.8	59.8	47.5	50.7	62.5	68.2
ArcFace100	66.6	71.0	57.5	65.4	72.7	81.0
R34-Baseline	64.9	73.1	59.2	68.4	74.4	83.2
R34-Mask	93.8	92.9	90.7	92.2	94.3	98.6
R34-AMaskNet	93.5	93.0	92.0	92.7	94.1	98.5

✓means using the proposed mask-aware similarity matching strategy (Wo/W+MS), while × means not applicable (Wo/W).

W/W, it will degrade the performance in the case of Wo/Wo. On the contrary, the R34-AMaskNet improves masked face recognition performance with a slight cost in terms of performance decrease in general face recognition.

3.3.3.3 Effectiveness of the proposed mask-aware similarity Matching Strategy (MS)

Results on COX dataset. Table 3.1 shows the recognition result of without and with mask transfer on the Wo/W conditions (Wo/W vs. Wo/W+MS), where one contains a mask, while the other does not. In addition to R34-Mask and R34-AMaskNet models, the performance of the models is greatly improved after using the proposed mask-aware similarity matching strategy, e.g., an 11.3 percent improvement for ArcFaceM on the Cam2 of COX between images treated without mask-transfer to those with mask-transfer (because there is no mask strategy used therewith). Meanwhile, for R34-Mask and R34-AMaskNet, although the data augmentation strategy has been adopted, it is still improved in most cases, e.g., a 0.7 percent improvement for the R34-AMaskNet on the Cam2 of COX between images treated without mask-transfer to those with mask-transfer.

Table 3.2: Results on the Public-IvS dataset with 1:1 verification protocol at TAR@FAR=10^{-5}.

Model	Public-IvS	
	×	✓
ArcFaceM	23.4	60.3
ArcFace34	23.4	60.3
ArcFace50	53.8	65.3
ArcFace100	67.5	75.7
R34-Baseline	59.7	70.8
R34-Mask	93.9	94.3
R34-AMaskNet	95.4	95.2

✓means using the proposed mask-aware similarity matching strategy (Wo/W+MS), while × means not applicable (Wo/W).

Results on Public-IvS dataset. Table 3.2 shows the comparison between without mask transfer (Wo/W) and with mask transfer (Wo/W+MS) on the Wo/W (one is with mask, while the other is without mask) in Table 3.2 shows that the recognition performance is improved by using the mask-aware similarity matching strategy, especially for the general face recognition model, e.g., 0.4 percent improvement for the R34-Mask, compared to 14 percent improvement for ArcFaceM on the Public-IvS dataset from without mask transfer to with mask transfer. The result shows that transferring the mask from masked image to non-masked image in the face pairs can mitigate the difference caused by the mask without loss of spatial information in the inference stage.

3.3.4 Comparison of state-of-the-art methods on RWMFD dataset

To further verify the effectiveness of the proposed method on real mask data, the proposed method is compared with public models and other literature models on the RMFRD dataset [6]. The results are shown in Table 3.3, where those of literature methods are taken from corresponding papers. The R34-AMaskNet outperforms the R34-Baseline with an improvement of up to 12.5 percent. Meanwhile, R34-AMaskNet outperforms the other methods by a significant margin, which indicates

Table 3.3: Results on the RWMFD dataset.

Method	Accuracy
J. Luttrel et al.[126]	85.7
Hariri et al. [127]	84.6
Almabdy et al. [57]	87.0
Walid Hariri. [110]	91.3
ArcFaceM	34.6
ArcFace34	43.2
ArcFace50	49.3
ArcFace100	61.7
R34-Baseline	81.9
R34-Mask	92.5
R34-AMaskNet	**94.3**

the proposed method is efficient when applied to a real masked face recognition scenario.

3.4 Discussion and Analysis

This section firstly provides a discussion about the the effect of masked faces on the behavior of face recognition model, which is important for design mitigation techniques to offset the inevitable performance loss. Then, qualitative analysis experiments using Class Activation Map (CAM) [7] are used to explore the role of contribution learning by the proposed AMaskNet for masked face recognition.

3.4.1 Effect of mask on performance

Four publicly available models are evaluated under three test conditions, i.e., Wo/Wo, W/W, and Wo/W, to study the effect of masked faces on the recognition accuracy.

Effect on the COX dataset. As shown in Figure 3.10, compared to a scene without a mask, the recognition accuracy is largely decreased when a mask is present. For example, for Cam1 of the COX dataset, ArcFace50 decreases from 95.2% on Wo/Wo

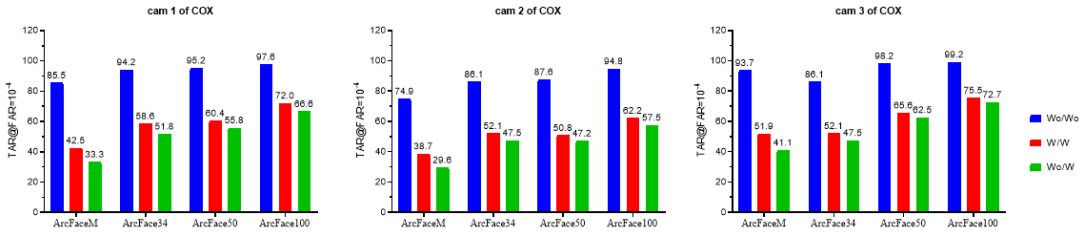


Figure 3.10: Results on COX dataset with a 1:1 verification protocol at TAR@FAR=10^{-4}.

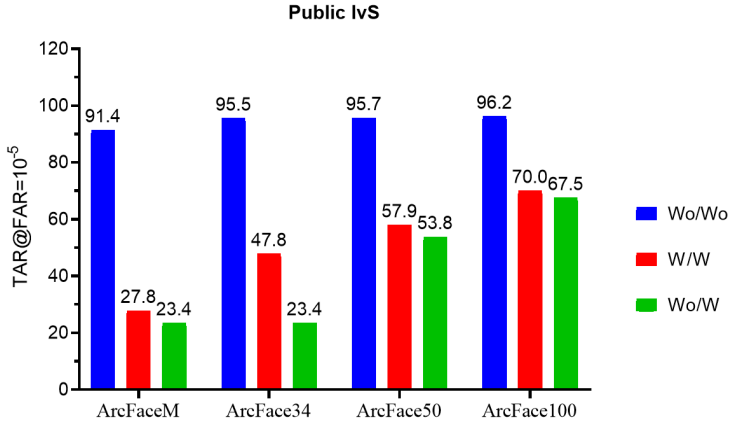


Figure 3.11: Results on Public-IvS dataset with 1:1 verification protocol at TAR@FAR=10^{-5}.

to 55.8% on Wo/W, resulting in 39.4 percent loss of accuracy. This magnitude of decrease in accuracy affects all camera scenes of the COX dataset. Moreover, the poorer the model performance, the greater the loss of accuracy, e.g. 52.2 and 31.0 percent losses in ArcFaceM and ArcFace100, respectively, from Wo/Wo to Wo/W. It is noteworthy that having one image with a mask while the other image without a mask has a greater adverse effect on the performance than both having masks, e.g. the loss of accuracy of ArcFaceM from Wo/Wo to Wo/W is 52.2 percent, but only 40.0 percent from Wo/Wo to W/W.

Effect on the Public-IvS dataset. The recognition results on Public-IvS are shown in Figure 3.11. Again, similar findings as for the COX dataset are obtained. The recognition accuracy is largely decreased when a mask is present. For example, ArcFace50 decreases from 95.7 percent on Wo/Wo to 53.8 percent on Wo/W, resulting

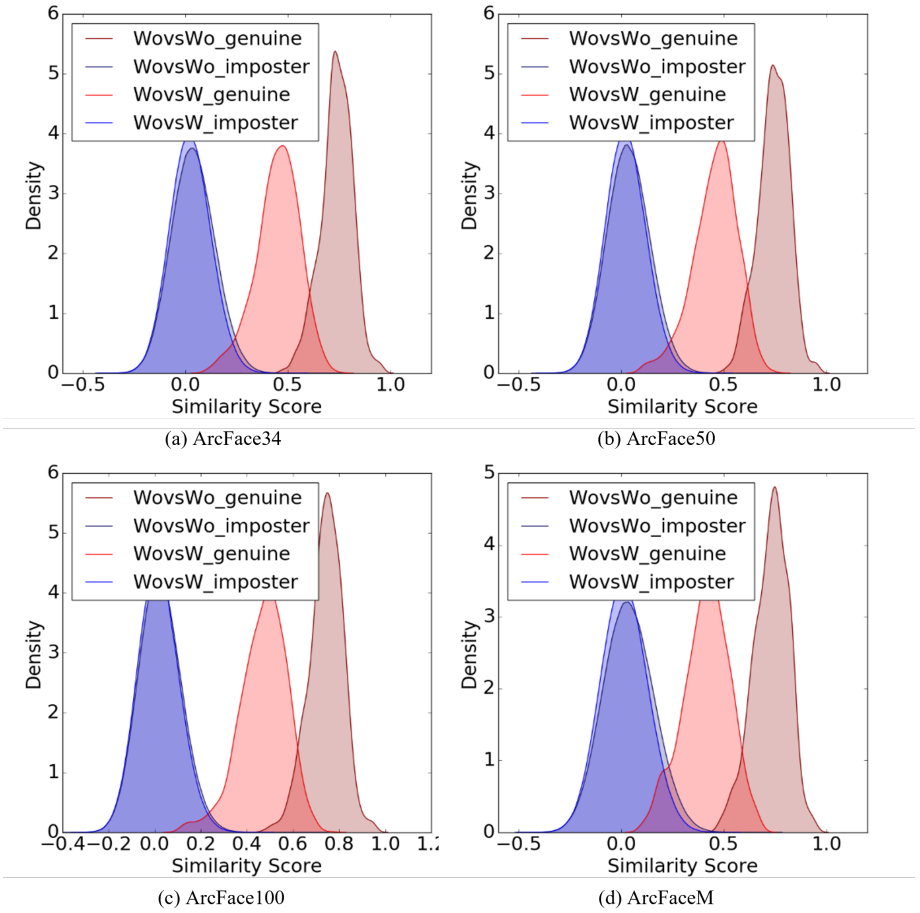
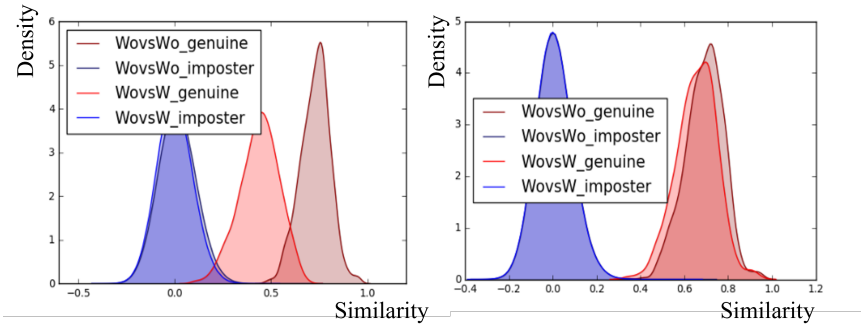


Figure 3.12: Distribution comparison of similarity scores on public models. Here, "Wo" means without wearing a mask, and "W" means wearing a mask.

in a 41.9 percentage point loss of accuracy.

Effect of the similarity distributions. To understand the reason of performance improvment, the similarity score distributions of genuine and imposter pairs in Wo/Wo and Wo/W test conditions are analyzed, with the result on Public-IvS dataset shown in Figure 3.12. The choice of False Acceptance Rate (FAR) determines the score threshold, and then affects the results of True Acceptance Rate (TAR). In comparison with the Wo/Wo condition, the scores of genuine pairs strongly transfer towards the imposters when one image is with a mask (Wo/Wo vs. Wo/W). That means, the scores of genuine pairs become smaller and are nearer to imposter pairs due to the influence of masks, which will cause the TAR to become smaller at the same FAR, making them less recognizable leading to performance degradation.



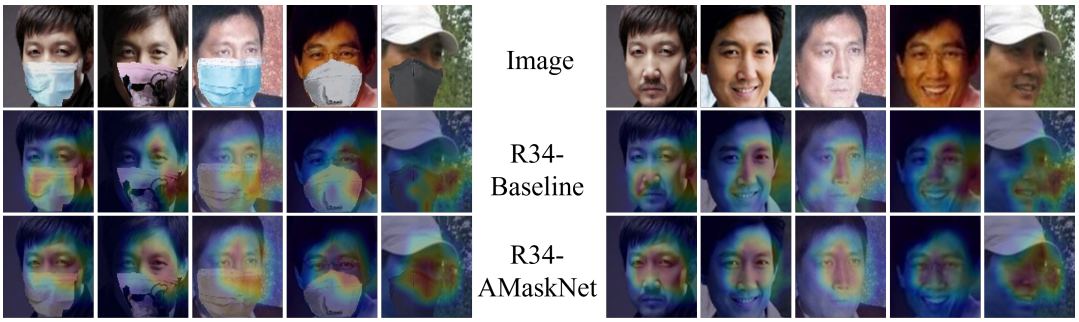
(a) R34-Baseline (b) R34-AMaskNet

Figure 3.13: Distribution comparison of similarity scores.

3.4.2 Qualitative analysis

The distributions of similarity score. The matching score distributions of genuine and imposter pairs in Wo/Wo and Wo/W test conditions are presented in Figure 3.13. We can see from the Figure 3.13(a) that in the baseline model, the genuine scores significantly shift towards the imposter scores when the image is with a mask. Meanwhile, we can see from the Figure 3.13(b) that in R34-AMaskNet, the shift of similarity scores is largely mitigated. In comparison with the Wo/Wo condition, the scores of genuine pairs of the baseline model shift towards the imposter counterparts when one image is of a masked face (Wo/Wo vs. Wo/W), implying that the scores of genuine pairs decrease and are closer to those of imposter pairs due to the influence of masks, which makes them less recognizable. With the R34-AMaskNet, however, the score distribution of the genuine pairs shifts only slightly toward imposter ones, which clearly confirms that a stronger recognition capability is obtained in R34-AMaskNet.

Contribution estimation. Some samples are randomly selected from the testing dataset for visual analysis. Figure 3.14 qualitatively shows the contribution estimation result using CAM [7] for the purpose of intuitive understanding. Figure 3.14(a) shows images wearing a mask and its visualization results on R34-Baseline and R34-AMaskNet, respectively. Figure 3.14(b) shows images without wearing a mask and visualization results on R34-Baseline and R34-AMaskNet, respectively. From the first line to the third line are the original images, the attention results of



(a) With wearing a mask (b) Without wearing a mask

Figure 3.14: Visualization of attention result with CAM [7]. We can see that the model with the contribution module can be successfully able to localize the discriminative regions for face recognition.

R34-Baseline, and R34-AMaskNet, respectively. The maps highlight the discriminative image regions used for face recognition. We can see from this result that R34-AMaskNet can focus on the non-masked regions and exclude the background for most samples, suggesting that discriminative regions for face recognition are obtained. Even in images without a mask, this attention scheme can also localize the facial area and eliminate background interference. This is why the model performs slightly better under Wo/Wo conditions.

3.5 Summary

In this chapter, an effective method was proposed with which to mitigate the effect of mask defects in face recognition. Firstly, a low-cost, accurate method of masked face synthesis was proposed for use in data augmentation and a mask-aware similarity matching strategy was developed, which is low-cost, rapid, and convenient for model development. Secondly, AMaskNet was proposed to improve the performance of masked face recognition, which includes two modules: a feature extractor and a contribution estimator, where the latter is adopted to learn the contribution matrix, thus outputting refined features by successive matrix multiplication. This method can learn an effective contribution matrix and automatically assign higher weights to the feature map activated by the non-masked facial parts and lower weights to those

that are activated by masked facial parts. Finally, a mask-aware similarity method was proposed for use in the inference stage to mitigate the difference caused by the mask without loss of spatial information. Both qualitative and quantitative analysis results showed that the proposed model can mitigate the effects of mask defects in face recognition. While the method is designed for masked face recognition, it can also be applied in other computer vision tasks, especially for other face-related applications such as facial attribute recognition.

Chapter 4

Content-Aware Contribution Estimation for Feature Aggregation

This chapter is dedicated to the Research Topic 2. As discussed in Chapter 1, the difficulties in video-based face recognition, such as dramatic pose variations and low quality, can be alleviated by leveraging the rich complementary information between frames. However, limited by the mini-batch training strategy, the current deep learning methods only utilize the frames in each batch during training, ignoring the content of the entire video. In this chapter, a content-aware feature aggregation scheme is proposed, that aggregates complementary information between different frames.

Firstly, a two-branch structure is designed as the Content-aware feature Aggregation Network (CAN). Secondly, a content-aware training strategy using a content bank is proposed, which alleviates the limitation of mini-batch samples by using the content of the entire video or several images belonging to the same identity and thus estimates the global contribution. Comparative studies on benchmark datasets, such as COX [5], IJB-C [8], PaSC [9], and YTF [10], confirm that the proposed approach exceeds state-of-the-art performance. Meanwhile, qualitative analysis on the Multi-PIE dataset [11] indicates that the contribution learned by the CAN is reasonable and beneficial to video face recognition.

This chapter is structured as follows: Section 4.1 introduces this research topic, including background, motivation, and contributions. Then Section 4.2 describes the proposed CAN framework and the contribution-aware training strategy. Section 4.3 presents the experimental results and discussions, and Section 4.4 summarizes this chapter.

4.1 Introduction

As discussed in Chapter 1, video face recognition has received increasing interests in both academia and industry, and has been widely used in applications such as security authentication and video surveillance. Although considerable progress has been achieved in still face recognition owning to the emergence of effective deep learning-based approaches [41, 42, 55, 56, 63, 64, 65, 66, 67], well-designed loss functions, and large-scale datasets, video face recognition remains as a significant research challenge. Different from still face recognition, video face recognition often suffers from low quality, dramatic pose variations, occlusion, and so on. On the other hand, abundant temporal and multi-view information usually exists in the video, which may bring potential to boost accuracy in video face recognition.

To efficiently use more discriminative information in the video, aggregation-based methods [1, 68, 69, 70, 71, 72] have been widely adopted and impressive performance is gained in video face recognition. The basic idea of the aggregation approach is to extract frame-level features at each frame, and then to aggregate them across all frames to form a video-level feature. The most commonly used aggregation technique is average pooling [73], where features of all frames are simply combined with equal importance. However, low-quality frames would deteriorate the quality of features, resulting in degraded performance of face recognition. Another aggregation method is max pooling [74], which only uses the best quality frame feature as video feature. However, the discriminative information contained in low-quality frames is ignored which could be complementary to high-quality frames.

Recent advance has witnessed deep learning network as an adaptive weighting scheme to aggregate all frame-level features together to form a compact and discriminative video-level feature. Neural Aggregation Network (NAN) was proposed in [72] for feature combination. It has two modules: one is the CNN feature embedding module to extract the feature representation of each face frame, the other is the neural aggregation module to aggregate the video-level feature from face video using two attention blocks. Quality Aggregation Network (QAN) [1] adopted a two branches scheme, where one branch is used to extract face feature of each image and the other branch is adopted to predict the quality score of each image. Then, the final set-level features are obtained by aggregating the features and the quality scores of all images in a set. C-FAN [69] was proposed to learn the quality score of each feature component by adding an aggregation module to the base network, and then to gain the video-level face feature in a video using a single vector aggregated from deep feature vectors. However, limited by the mini-batch training strategy, the quality prediction in the above methods only utilize video frames in each batch during training, which ignore the content of the entire video as well as all frames corresponding to the subject, thus leading to a biased face quality estimation.

As Research Topic 2 of this thesis, a novelty feature aggregation method is proposed here for video-based face recognition by considering the content of the entire video. Firstly, a Content-aware feature Aggregation Network (CAN) is designed to learn the contribution for each frame in a video, in which the features coming from multiple frames are adaptively aggregated into a compact video-level feature. The network is composed of two branches; one is a feature extractor to extract face feature from a single frame and the other branch is a contribution estimator to estimate the image contribution. The video feature is then aggregated by the features and contributions of all frames in a video clip. Secondly, a content-aware training strategy using a content bank is proposed, where not only the samples in each mini-batch but also the content of the entire video clip are considered, thus achieves a global contribution estimation scheme. In addition, in order to reduce the influence of the long tail problem in the training corpus, i.e., DeepGlint [44] and Glint360K [45] datasets, a

balanced batch selection strategy is further carefully designed. The qualitative analysis on the Multi-PIE [11] dataset shows that the contribution learned by the CAN is reasonable in that it is closely related to image quality, and the quantitative experiments on benchmark datasets indicate that the proposed CAN achieves significant performance. The main contribution are summarized as follows:

[1]. CAN is proposed to learn the contribution of each frame in a video, and the features from multiple frames are adaptively aggregated into a compact video-level feature based on their contributions.

[2]. A content-aware training strategy is proposed to achieve a global contribution estimation scheme by leveraging the content of the entire video clip using a content bank.

[3]. A balanced batch selection strategy is carefully designed to reduce the negative impact of the long-tail dataset on performance.

4.2 The Proposed Approach

In this section, the proposed Content-aware feature Aggregation Network (CAN) is described, which incorporates feature extractor network and contribution estimator network to obtain, respectively, the feature and the contribution of a single frame. Then, the content-aware training strategy is introduced, where not only the samples in each mini-batch but also the content of the entire video is considered.

4.2.1 Content-aware feature Aggregation Network (CAN)

The CAN architecture consists of three modules: feature extractor, contribution estimator, and feature aggregator, as illustrated in Figure 4.1. The input of the CAN is a video clip or several images belonging to the same identity. The feature extractor is a base model, which is used to extract each frame feature of the video clip. The

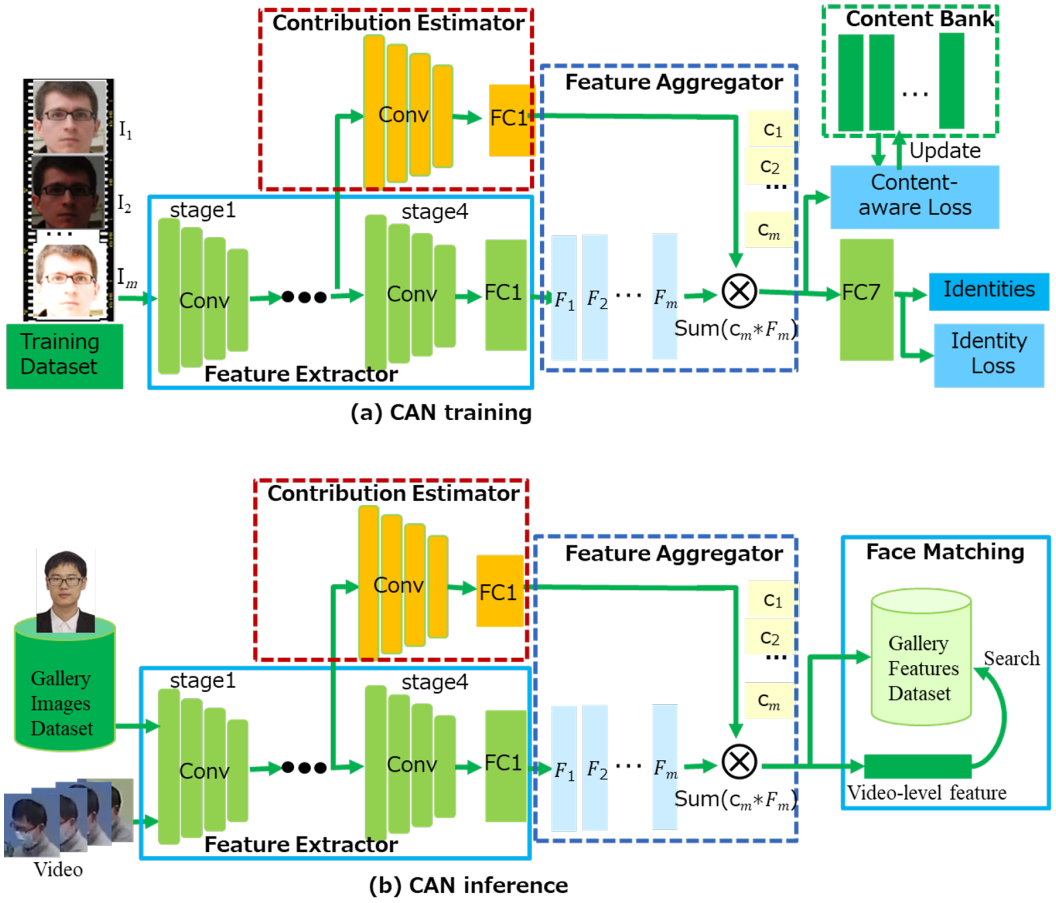


Figure 4.1: Architecture of the proposed Content-aware feature Aggregation Network (CAN). (Best viewed in color)

contribution estimator is added to the base model using several convolution layers and one-node fully connected layer, which is used to estimate the contribution of each frame to their video clip. The feature aggregator is used to aggregate the contribution scores and features of all frames in the video clip. The final video-like feature is thus directly obtained by the feature aggregator. The content bank is maintained to memorize the global features. Here, video-level identity loss and content-aware loss are used to supervise the training. Details of each component are introduced below.

Feature extractor: Most popular deep neural networks (e.g., ResNet34 in ArcFace [41]) can be adopted as a backbone to extract the feature from each frame [128, 129]. Once built, the extractor is kept fixed during the training of the CAN network.

Contribution estimator: A new branch is adopted as the contribution estimator by connecting the feature extractor, aiming to obtain the contribution of each single frame, which is then used as the weight later in feature aggregation stage. The structure may have different complexity, ranging from one-node fully connected layer to one or several convolution layers. More complex network may bring higher learning ability, but certainly with the cost of extra computation and the risk of over-fitting.

Feature aggregator: Let us define $V = \{I_1, I_2, \ldots, I_m\}$ as a video clip, with I_i indicating the i-th frame. Let $F(\cdot)$ and $C(\cdot)$, respectively, denote the feature extractor and contribution estimator, which output the feature vector f_i and attribution value c_i for each frame, i.e., $f_i = F(I_i)$ and $c_i = C(I_i)$. The final face representation of a video is thus directly obtained by a weighted average of the features of the video frames, as follows:

$$F_V = \emptyset(f_1, f_2, \ldots f_m) = \frac{\sum_{i=1}^{m} c_i * f_i}{\sum_{i=1}^{m} c_i}, \tag{4.1}$$

Optionally, the frame feature with the highest contribution value may be used as the video feature, i.e., feature selection scheme.

4.2.2 Content-aware training

Two kinds of losses are proposed to train the model; one is video-level identity loss, while the other is content-aware contribution loss. With the former loss, the ground-truth of contribution or quality value of each frame is not necessary during training, which largely reduces the cost of building training data. With the latter loss, a content-aware features memory bank is introduced to store more information (beyond the information in each training batch as in traditional method) during training stage, in order to increase accuracy. Details of each loss are introduced below.

Video-level identity loss. Video-level feature is firstly obtained by the contribution weighted aggregation of the features of all frames in a video clip, and an ArcFace [41]-like loss is chosen to penalize video-level identification error. In such case, no

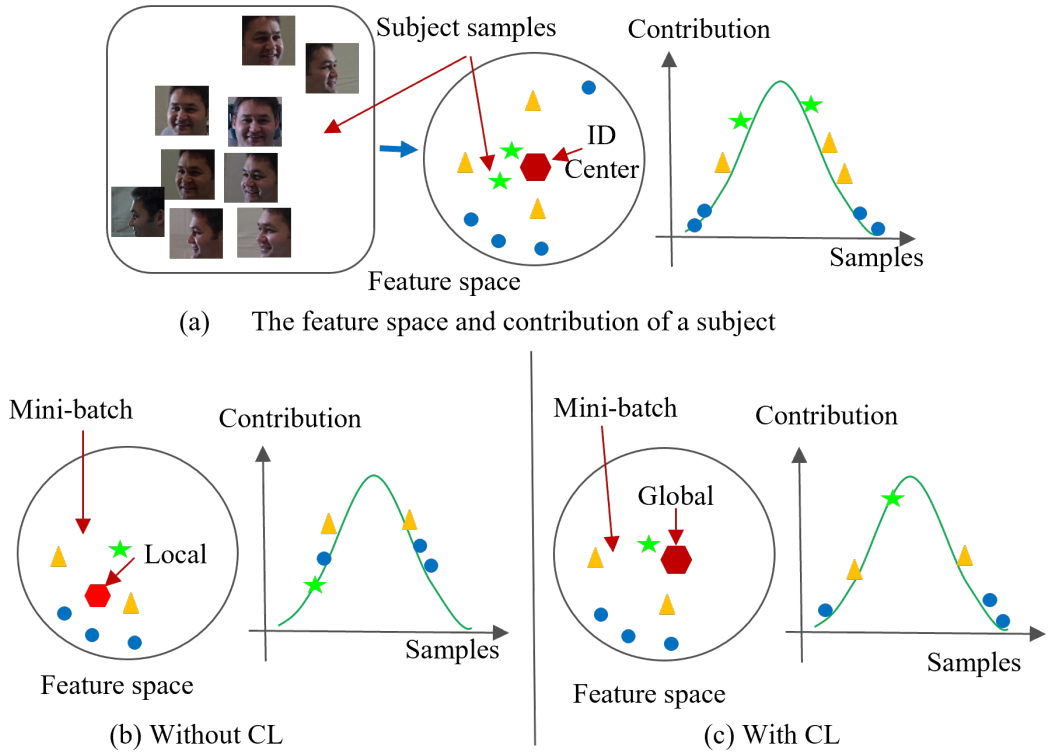


Figure 4.2: Contribution estimation with the proposed content-aware Contribution Loss (CL). (Best viewed in color)

contribution or quality value of each frame is provided as ground-truth for supervision, so, the training can be seemed as unsupervised. The video-level identity loss is defined as:

$$L_{\text{ID}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{s\cos(\theta_{y_i}+m)}}{e^{s\cos(\theta_{ji}+m)}+\sum_{j=1,j\neq y_i}^{n}e^{s\cos\theta_j}}, \tag{4.2}$$

where θ is the angle between a video-level feature and the corresponding weight, N is the number of video clips in a mini-batch, m is a marginal factor, and s is a scale factor.

Content-aware Contribution Loss (CL). Given traditional DL model training strategy where the calculation is limited to mini-batch samples, contribution prediction is achieved only based on the video frames in each mini-batch, while ignoring content information in the entire video or even the whole corpus, as illustrated in Figure 4.2.

Figure 4.2(a) shows the feature space and contribution value of the samples belonging to a subject. The nearer a sample is to the identity center, the higher (better) the contribution (quality). Figure 4.2(b) shows the feature space and contributions of mini-batch samples. The ID center represents the feature space center of all face images of a subject. During training, face recognition is regarded as a classification task, that is, a subject is trained as a class. Therefore, the ID center is the feature space center of a class for classification tasks. The samples close to the mini-batch center are assigned higher contribution value even they are far from the ID center. However, the samples close to the ID center but far from the mini-batch center are assigned a low contribution value, such as the estimated contribution value of the sample represented by the pentagram, thus leading to a biased face contribution estimation. Figure 4.2(c) shows the contribution estimation result under the CL loss, where the samples far from the identity center are assigned lower contribution values even when they are close to the mini-batch center. Meanwhile, the samples close to the ID center are assigned higher contribution values even when they are close to the mini-batch center, such as the estimated contribution value of the sample represented by the pentagram.

To alleviate the above limitations, a renewed global representation is always kept for each identity using a memory bank, and to force the local representation calculated from each mini-batch to be close to the global representation. The global representation is obtained from the entire video (or the whole corpus), thus content information is introduced to the mini-batch based training. Let $\mathrm{B} = \left\{F_{g_1}, F_{g_2}, \ldots, F_{g_i}\right\}$ be the memory bank, with F_{g_i} representing the global representation feature of the i-th class. The content-aware loss can thus be defined as follows:

$$L_C = \sum_{i=0}^{n} \left\| F_{v_i} - F_{g_i} \right\| = \sum_{i=0}^{n} \left\| \frac{\sum_{j=0}^{m} c_j * f_j}{\sum_{j=0}^{m} c_j} - F_{g_i} \right\|, \tag{4.3}$$

where F_{v_i} is the video-level feature as introduced in Eq. 4.1, F_{g_i} is the global video-level feature of identity which is obtained by simply averaging all the video-level features belonging to the i-th class. F_{g_i} is updated by calculating the average of the features of the same class in the mini-batch on each iteration. If the mini-batch does

not include the set features of some classes, their corresponding global features will not be updated.

The gradients of L_C with respect to F_{v_i} and F_{g_i} are given by:

$$\frac{\partial L_C}{\partial F_V} = \frac{1}{n}\left(F_{v_i} - F_{g_i}\right), \tag{4.4}$$

$$\Delta F_{g_i} = \frac{\sum_{i=1}^{n} \delta\left(y_i = j\right) \cdot \left(F_{g_i} - F_{v_i}\right)}{\varepsilon + \sum_{i=1}^{n} \delta\left(y_i = j\right)}, \tag{4.5}$$

where δ is Kronecker's Delta function, and ε is a small positive number to avoid zero denominator, which is set to 10^{-5}.

Combination of the two losses. Finally, the video-level identity loss and contribution loss are combined to jointly train the model, as follows:

$$L = L_{\mathrm{ID}} + \lambda L_{\mathrm{C}}, \tag{4.6}$$

where λ is adopted for balancing the two loss functions.

4.3 Experiment and Discussions

In this section, the proposed method is evaluated through comparison with state-of-the-art approaches to confirm its effectiveness. Firstly, several commonly used benchmark datasets are introduced. Then, the implementation details of the proposed method are presented. To gain more insight into the behavior of the proposed method, ablation study and qualitative analysis are presented.

4.3.1 Datasets

Both feature extractor and the contribution estimator are trained on DeepGlint [41] and Glint360k [45] datasets, while the accuracy is evaluated on three benchmarks, i.e., COX [5], IJB-C [8], PaSC [9], and YTF [10]. In addition, the Multi-PIE dataset [11] is used for qualitative illustration of the effect of contribution estimation.

DeepGlint dataset [44] includes cleaned MsCeleb1M [41] and Asian celebrity [44] datasets with a total of 6.6M celebrity images of 172K identities. Therefore, it is adopted for the analysis of the proposed method and ablation studies with a relatively small dataset for training.

Glint360K dataset [45] is a widely adopted large-scale dataset for training a face recognition model, which includes more than 17M images from 360K identities merged from clean Celeb-500K [130] and MS1M-Retinaface [41] datasets. Therefore, it is adopted for the verification of the effectiveness of the proposed method with large-scale datasets for training, to compare with state-of-the-art methods.

COX dataset [5] contains 1K identities, including 1 still image and 3 videos for each identity; a total of 1K still images and 3K videos with natural variations in pose, expression, lighting, blur, and face resolution. Most of the videos have more than 100 frames each identity and especially videos from Cam3 mostly have 170 frames. The videos captured each identity walking in a large gym to simulate the surveillance scenario from three cameras (Cam1, Cam2, Cam3) at different locations. Three standard matching protocols were also proposed by the author for face identification testing, i.e., Video-to-Still (V2S), as shown in Figure 4.3.

IJB-C dataset [8] includes 3,531 identities, a total of 31,334 (21,294 face and 10,040 non-face) still images, averaging to 6 images per subject, and 117,542 frames from 11,779 full-motion videos, averaging to 33 frames per subject and 3 videos per subject, which is an extension of the IJB-B [103] dataset. All subjects in the dataset are ensured to appear in at least two still images and one video. In the 1:1 verification, there are 23,124 templates with 15,639K impostor pairs and 19,557 genuine pairs.

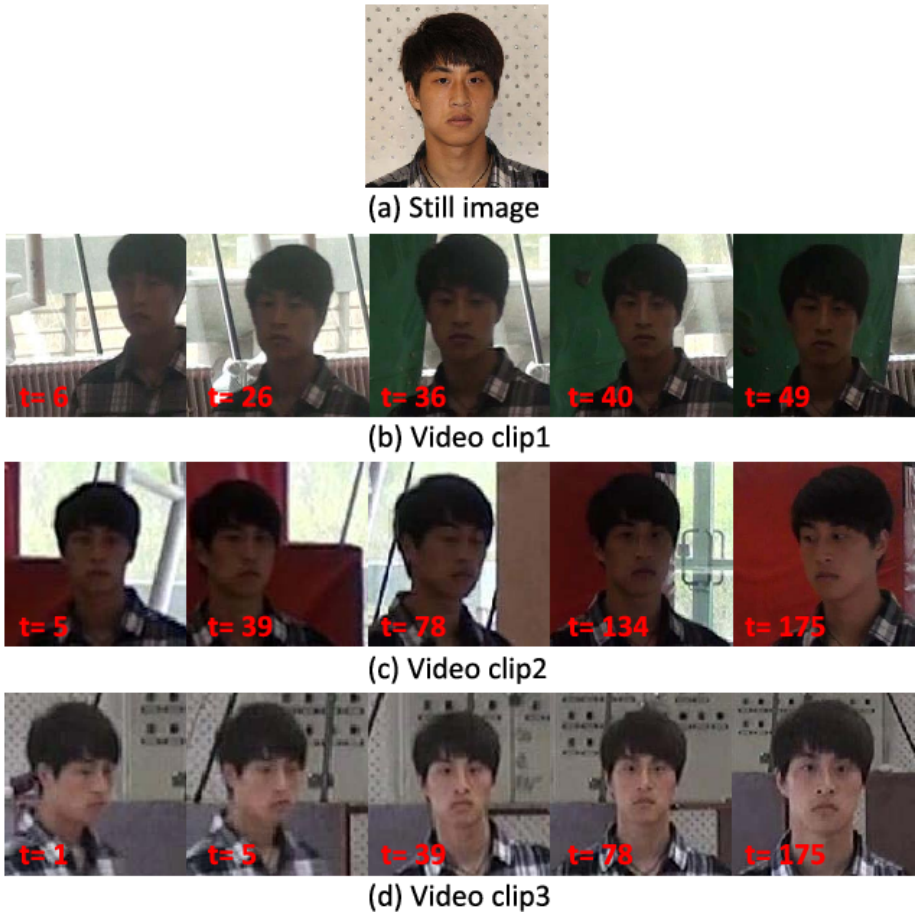


Figure 4.3: Sample images from the COX dataset [5].

The verification protocol of IJB-C contains more impostor pairs, thus the True Accept Rates (TAR) at lower False Accept Rates (FAR) is used in here, as shown in Figure 4.4.

PaSC dataset [9] contains 265 identities and 2,802 videos. Half of its videos are captured by controlled camera (denoted as PaSC-C), while the rest are captured by hand-held camera (denoted as PaSC-H), and each identity is asked to perform some predefined actions. Therefore, the face photos cover serious video-type noises and large pose variations. In total, there are 334,879 and 328,967 frames of video in the 1,401 control and 1,401 hand-held videos, respectively. The evaluation in this chapter on PaSC totally follows the predefined face verification protocol, as shown in Figure 4.5. Note that pose, distance to camera and sensor were varied within sessions, while locations were varied between sessions.

Figure 4.4: Sample images from the IJB-C dataset [8].

Figure 4.5: Sample images from the PaSC dataset [9] from four sessions.

YTF dataset [10] contains videos downloaded from YouTube, which includes 3,425 videos of 1,595 identities with an average of 2.15 videos of each subject. This dataset is designed for studying the problem of unconstrained face recognition in videos. The average length of a video clip is 181.3 frames, of which the longest clip is 6,070 frames and the shortest clip duration is 48 frames, as shown in Figure 4.6.

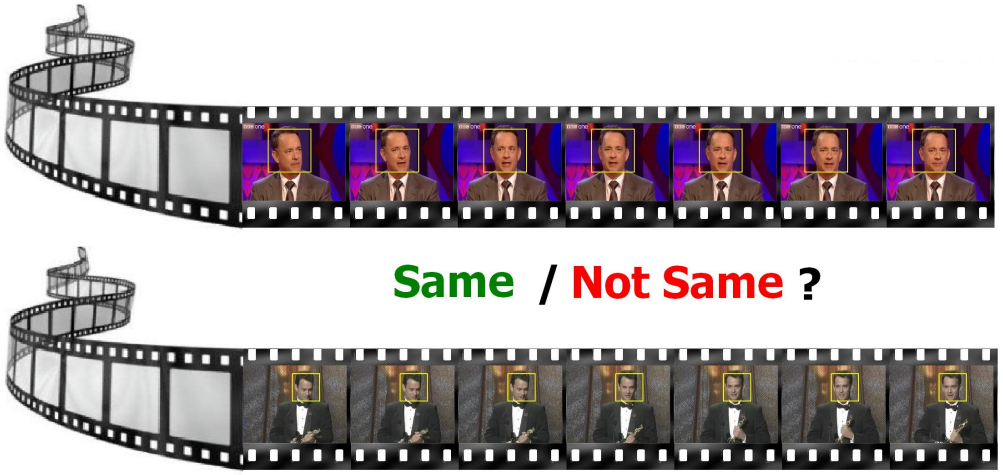


Figure 4.6: Sample images from the YTF dataset [10] downloaded from YouTube, which is designed for studying the problem of unconstrained face recognition in videos.

Figure 4.7: Sample images from the Multi-PIE dataset [11].

Multi-PIE dataset [11] contains face images of 337 identities with comprehensive variety in illumination, expression and pose, by carefully designing the configuration of 15 cameras and 18 flashes. Thirteen cameras were located at head height, spaced at 15° intervals, and two additional cameras were located above the subject, simulating a typical surveillance view. One image without any flash illumination, 18 images with each flash firing individually, and then another image without any flash. Therefore, it is very suitable to confirm the correlation of contribution estimation with image quality, as shown in Figure 4.7.

4.3.2 Implementation details

In this section, firstly, Data augmentation (DA) is introduced. Then, the balance strategy (BS) is discussed. Finally, training parameters are described.

Data Augmentation (DA). Blur is imposed to the Deepglint and Glint360K datasets to simulate video-like training data, where one-dimensional local averaging of neighboring pixels is applied to generate motion blur, while a Gaussian kernel is adopted to simulate the out-of-focus blur. Besides this, the images are split into 5×5 blocks and some blocks are randomly replaced with a black mask for synthesizing occlusion data. Illumination variance is achieved by simply adjusting the brightness of the training images.

Balance Strategy (BS). Long tail distribution of the training data; the fact that a small number of entities appear frequently while most of others remain relatively rare, usually poses great impact on the feature learning process and feature extraction ability. To solve this problem, two strategies are designed in this chapter. One straightforward strategy is to remove the very head and tail identies. More specifically, subjects with more than 500 or less than 10 samples are removed. Given that there are sufficent number of identities remaining, this strategy brings some improvement on the accuracy. The second strategy is to select samples for each mini-batch in training based on identities but not individual images. In image based mini-batch solution, identities with more samples will have a higher probability to be selected. However, in identities based solution. n identities are randomly selected from the identity list and then m images are obtained for each identity to generate the mini-batch of $n*m$ samples, thus avoiding the bias to head subjects.

Training parameters. The ResNet50 (R50) and ResNet100 (R100) models are pre-built as the feature extractor of a small model and a large model, respectively. After that, the feature extractor is fixed and the contribution estimator is then trained on the same dataset with the above data augmentation scheme. Stochastic Gradient Descent (SGD) [131] is used with momentum and weight decay value being 0.9 and 0.005.

Table 4.1: Results on the IJB-C dataset with 1:1 verification protocol (TAR@FAR=10^{-3}, 10^{-4}, 10^{-5}). "CAN" means the proposed content-aware feature aggregation Network.

Method	10^{-5}	10^{-4}	10^{-3}
Multicolumn [71]	77.10	86.20	92.70
ArcFace [41]	87.28	92.13	95.55
PFE [119]	89.64	93.25	95.49
DUL [118]	87.22	92.43	95.38
GroupFace [56]	94.53	96.26	-
VPL [56]	-	96.76	-
R100, CAN, DeepGlint	95.44	96.88	97.87
R100, CAN, Glint360k	**96.29**	**97.62**	**98.52**

The value of λ is set to 0.1, and the size of mini-batch is set to 100 including 20 randomly selected subjects and 5 images per subject.

4.3.3 Evaluation through comparison with state-of-the-art methods

Following the standard evaluation protocols, the proposed model is compared with the state-of-the-art methods on several benchmarks, i.e., IJB-C, PaSC, YTF, and COX.

Evaluation on the IJB-C dataset. The commonly used criterion of true acceptance rate at different false acceptance rate (TAR@FAR=10^{-3}, 10^{-4}, 10^{-5}) is used for the evaluation on the IJB-C dataset. The proposed method is compared with several state-of-the-art face recognition methods including both feature aggregation and non-aggregation methods. The results are shown in Table 4.1. For the purpose of fair comparison, both DeepGlint and Glint360k datasets are used for training. We can clearly see from the results that the proposed model outperforms the non-aggregation methods with a large margin, i.e., 8.12% and 9.01% better than ArcFace at FAR=10^{-5} trained on DeepGlint and Glint360k datasets, respectively.

Table 4.2: Results on the PaSC dataset with 1:1 verification protocol (TAR@FAR=10^{-2}) and YTF dataset (Accuracy(%)). "PaSC-C" means videos captured by controlled camera. "PaSC-H" means videos captured by hand held camera. "CAN" means the proposed content-aware feature aggregation Network.

Method	PaSC-C	PaSC-H	YTF
NAN [72]	—	—	95.72
QAN [122]	—	—	96.17
DAN [121]	92.00	80.30	94.28
ADRL [81]	95.67	93.78	96.52
TBE-CNN [132]	96.20	95.80	94.96
COSONet [123]	97.40	96.00	—
C-FAN [69]	—	—	96.50
R100, CAN, DeepGlint	97.67	96.83	97.18
R100, CAN, Glint360k	**98.46**	**97.62**	**97.53**

Evaluation on the PaSC dataset. The proposed method is further evaluated on surveillance scenes, by using the PaSC dataset. The results are shown in Table 4.2. Similar to that in the IJB-C dataset, the proposed method again behaves consistently better than the literature methods. The video content in PaSC, especially in the hand-held case, suffers from more severe conditions due to camera shaking, pose, blur, etc. Therefore, most of the face images in each video clip are of low quality. The simple average method, such as average pooling aggregation, assigns equal weights to each frame. In this way, the low quality frames with improper features would degrade the performance of the final recognition. On the contrary, the proposed contribution estimator obtains a contribution value closely related to image quality, thus depressing the low quality frames and strengthening the contribution of high quality frames. The proposed CAN outperforms ADRL aggregation by 2.79% at FAR=10^{-2} in controlled scenes, while this superiority increases to 3.84% at FAR=10^{-2} in hand-held scenes, which implies the robustness of the proposed method to deteriorated image quality.

Evaluation on the YTF dataset. The YTF dataset is a widly used benchmark video face dataset, which is designed for analyzing the problem of unconstrained video face recognition. The result of a 10-fold cross-validation is calculated on the YTF dataset as in Table 4.2. Compared with other aggregation methods, no video face

Table 4.3: Rank-1 Identification Rates (%) under the V2S setting for different methods on the COX dataset. "CAN" is the proposed content-aware feature aggregation Network.

Model	V2S_1	V2S_2	V2S_3
PSCL [5]	38.60 ± 1.39	33.20 ± 1.77	53.26 ± 0.80
LERM [133]	45.71 ± 2.05	42.80 ± 1.86	58.37 ± 3.31
VGG Face [74]	88.36 ± 1.02	80.46 ± 0.76	90.93 ± 1.02
TBE-CNN [132]	93.57 ± 0.65	93.69 ± 0.51	98.96 ± 0.17
R100, CAN, DeepGlint	96.14 ± 0.49	94.69 ± 0.25	99.68 ± 0.09
R100, CAN, Glint360k	**98.21 ± 0.28**	**95.18 ± 0.19**	**99.86 ± 0.07**

datasets was used in training the proposed contribution estimator. However, better performance was still achieved, which confirms the superiority of the proposed method among these state-of-the-arts, i.e., 1.03% better than C-FAN.

Evaluation on the COX dataset. The Rank-1 identification rates on COX is listed in Table 4.3. Again, the proposed method outperforms literature methods. Especially on Cam1 and Cam2, more than 4% improvement is achieved.

4.3.4 Ablation studies

In this section, ablation studies are conducted to understand the contribution of each component of the proposed method, ranging from the model structure and the loss to the implementation such as data augmentation and balance strategy. ResNet50 with DeepGlint dataset is adopted for the studies with the relatively small data set for training. TAR@FAR on PaSC dataset is used here for the comparison. The results are shown in Table 4.4.

Model structure. Three contribution estimators of different complexity are firstly compared: (1) Baseline, where only feature extractor is adopted and the features from all frames are aggregated by average pooling, (2) FC1 (Model A), which is a simplified contribution estimator where only a one-node fully connected module is

Table 4.4: Ablation study on PaSC dataset with 1:1 verification protocol (TAR@FAR=10^{-2}). "Conv" is the Convolution Module. "CS" is the Content-aware Strategy. "DA" is the Data Augmentation. "BS" is the Balance Strategy. "PaSC-C" is the videos captured by the control-held camera. "PaSC-H" is the videos captured by the hand-held camera.

Model	Conv	CS	DA	BS	PaSC-C	PaSC-H
Baseline	—	—	—	—	93.43	80.34
A	—	✓	✓	✓	93.41	84.18
B	✓	✓	✓	✓	**97.41**	**97.05**
C	✓	—	✓	✓	95.79	94.03
D	✓	✓	—	✓	93.39	89.16
E	✓	✓	✓	—	96.15	93.89

used, (3) Conv+FC1 (Model B), which is the proposed solution comprising a convolution module and a one-node fully connected module. We can see that introducing the contribution estimator evidently boosts the recognition accuracy. This is more obvious in the hand-held case, e.g., 4% improvement on hand-held data. As mentioned earlier, the images in the hand-held PaSC suffer severe quality degradation, thus it is even crucial to select the most informative and discriminative frames for correct recognition. However, since the structure of one node FC is too simple, thus the learning capability may be not sufficient. By adding additional convolution layers, the learning capability is enhanced and much bigger improvement can be further obtained, e.g., 12.87% improvement on hand-held data.

Content-aware Strategy (CS): Different learning strategies are compared, i.e., with or without CS (Model B or Model C). It is obvious that CS brings further benefits to the accuracy, especially in the hand-held case, i.e., 3.02% improvement in comparison without using this strategy. This result once again confirms the effectiveness of the proposed contribution estimation scheme in aggregating low quality video frames for recognition.

Data Augmentation (DA). Data augmentation plays an important role in the proposed video face recognition solution. The photos in the DeepGlint dataset, which

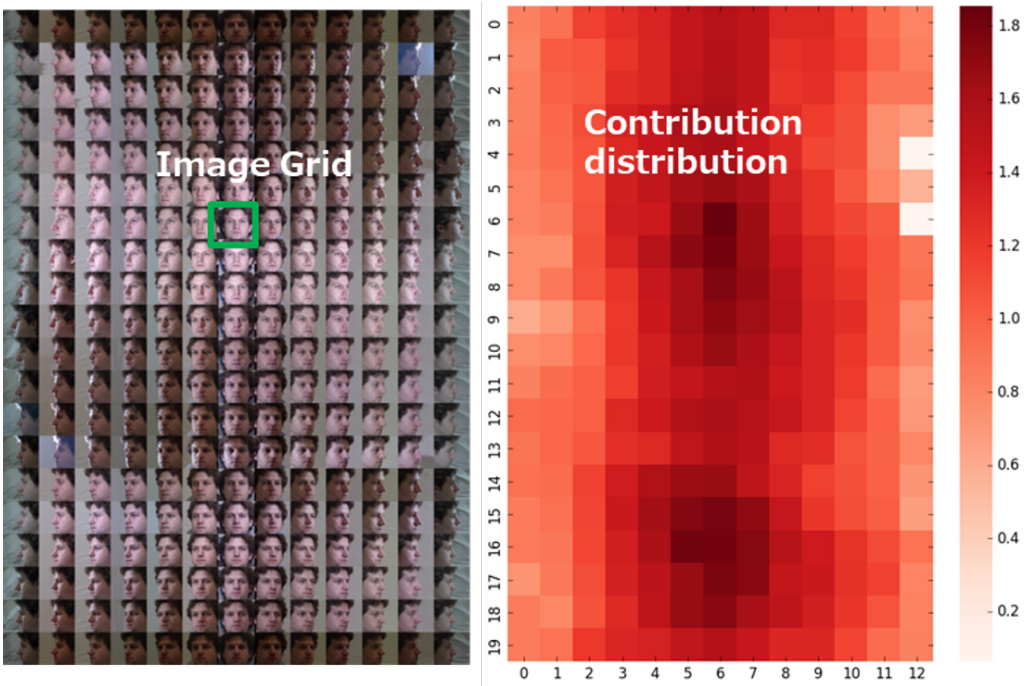
are usually captured under good conditions or even captured by professional photographers, are much different from the surveillance scenario, e.g., the photos are usually in high resolution and are not blured. The feature extractor and estimator built on DeepGlint thus cannot behave well on the PaSC video data. By introducing data augmentation to the training corpus, the data becomes more consistent with the video scene, and better accuracy can be achieved (Model B or Model D). This can be confirmed by the result, where 4.02% and 7.89% improvement are obtained for controled and hand-held cases, respectively.

Balance Strategy (BS). As introduced in Section 4.1, the long tail problem still exists in the DeepGlint corpus. Therefore, the balance strategy may contribute to the accuracy without any surprise. The balance strategy itself may not be a part of the proposed content-aware feature aggregation algorithm, but, it is a helpful training scheme toward better accuracy. Compared with Model B and Model E, it is obvious that BS brings further benefits to the accuracy, i.e., 3.16% improvement for PaSC-H in comparison to not taking this strategy.

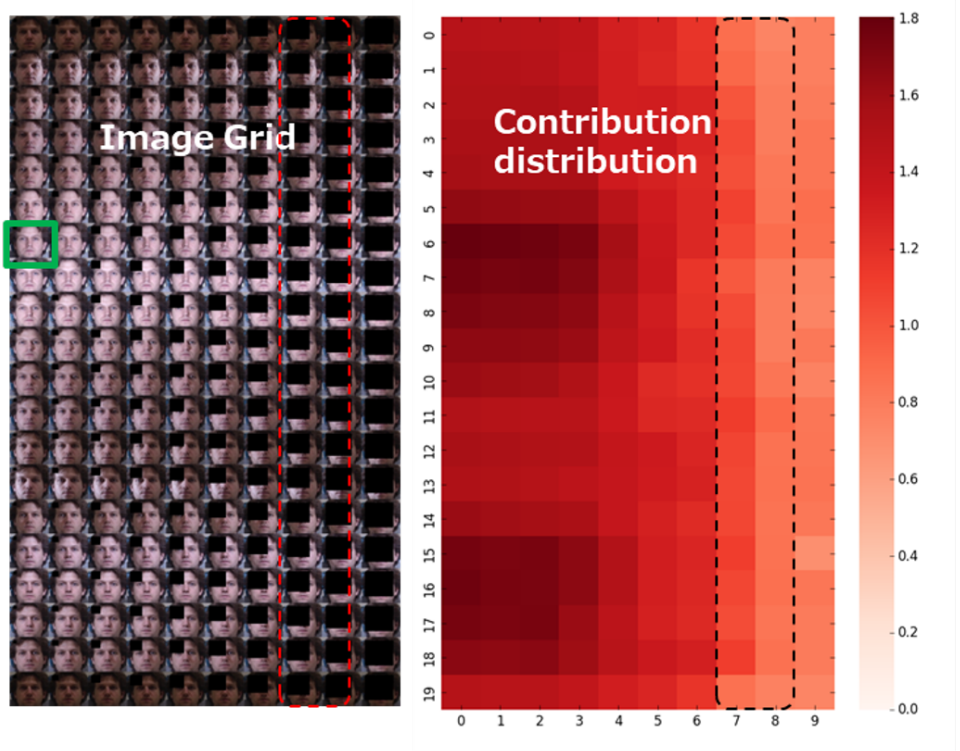
4.3.5 Qualitative analysis

The result of contribution estimation on Multi-PIE dataset [11] is qualitatively illustrated in Figure 4.8 and Figure 4.9.

Figure 4.8(a) shows samples of a subject across varied pose, illumination, and corresponding contributions predicted by the proposed contribution estimator. From left to right, faces with different poses are represented, spaced in 15 degree intervals. We can clearly see that the estimated contribution value is closely related to face pose and illumination from Figure 4.8(a). For example, the frontal face image with normal light shows high contribution. With the increase of pose and the decrease of illumination, the contribution value decreases, and the images in extreme illumination condition and large pose tend to obtain very low scores, which indicates positive correlation between the contribution estimation value and the image quality.



(a) Pose (Yaw) and Lighting variation



(b) Occlusion variation

Figure 4.8: The contribution distribution across varied pose, illumination and occlusion on the Multil-PIE dataset. (Best viewed in color)

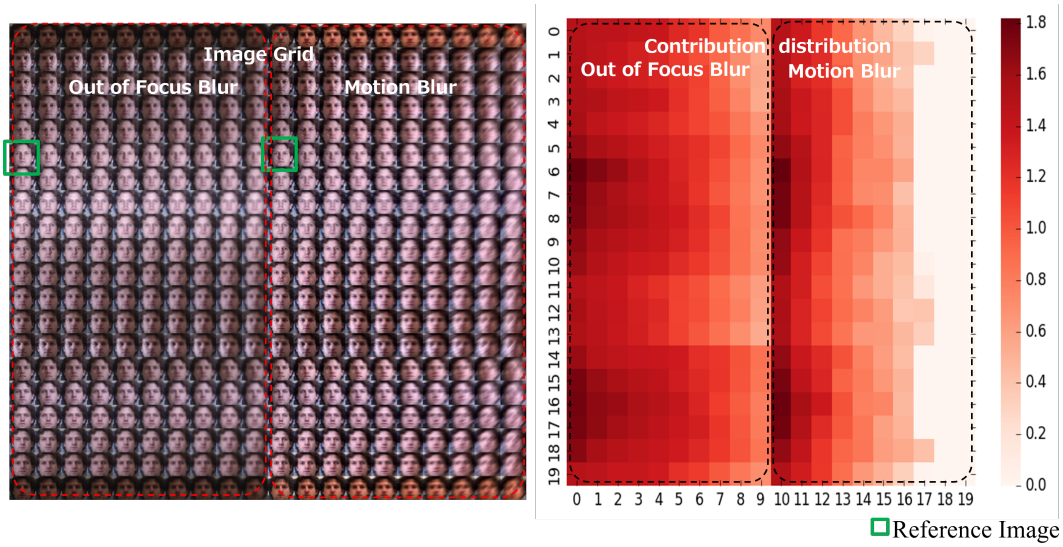


Figure 4.9: Contribution distribution on different motion blur and out of focus blur on the Multil-PIE dataset, where the left is an artificially adding motion and out of focus blur to the face image to simulate the different types of blur and the right is the corresponding predicted contribution of the contribution estimator. (Best viewed in color)

To conduct qualitative analysis on the effectiveness of the contribution estimator on occlusion, an occlusion is artificially added to the face image to simulate different occlusions. Figure 4.8(b) shows samples of a subject with varied occlusion and corresponding contributions predicted by the proposed contribution estimator. We can be see that by adding occlusion to the original images, the contribution gradually decreases, which prove the effectiveness of the contribution estimation for occluded situations.

To further conduct qualitative analysis on the effectiveness on blur, motion and out of focus blur are artificially added to the face image to simulate the different types of blur. The results are shown in Figure 4.9. We can see that by adding more blur into the original images, contribution gradually decreases, which proves the effectiveness of contribution estimation for blured image.

4.4 Summary

In this chapter, a new feature aggregation method was proposed for video-based face recognition by considering the content of the entire video, i.e., a content-aware feature aggregation scheme to aggregate complementary information between different frames in the video. Several innovative ideas were presented including: (1) a two-branch DL network for content-aware feature aggregation, (2) a content-aware training strategy for global contribution estimation by utilizing the content of the entire video clip using a content bank, (3) a balanced batch selection strategy for better accuracy by reducing the negative impact of the long-tail training dataset. While the proposed approach is aimed for video face recognition, it can also be applied to other computer vision tasks, especially for object recognition tasks in video.

Chapter 5

Conclusion

This chapter provides the conclusion of this thesis. The overall summary is introduced in Section 5.1. Section 5.2 discusses remaining challenges in the field of face recognition and potential future research directions. Section 5.3 provides some closing remarks to conclude the thesis.

5.1 Summary

The research presented in this thesis aimed to achieve robut face recognition for video surveillance. At present, face recognition under constrained environment has achieved promising results, and many products with face recognition technology are widely used in our daily life. However, surveillance face recognition is still a challenging problem, especially for unconstrained surveillance scenes. Different from constrained face recognition, unconstrained surveillance face recognition suffers from extremely low quality of each frame, e.g., various occlusions, changing illuminations, dramatic pose variations, especially when a large part of the face is covered by wearing a face mask, for example, during the COVID-19 pandemic. On the contrary, abundant temporal and multi-view information usually exists between

surveillance video frames, which may bring potential to boost accuracy in unconstrained surveillance face recognition. Therefore, this requires us to improve the performance of unconstrained surveillance face recognition from each frame and between multiple frames. This thesis intensively studied two issues of face recognition under unconstrained surveillance scenes and proposed two approaches with sufficient experiments and comparative analysis.

The first research topic described in this thesis proposed a method used for mitigating the negative effects of mask defects on face recognition. Firstly, a low-cost, accurate method of masked face synthesis, i.e., mask transfer, was proposed for data augmentation. Secondly, an Attention-aware Masked face recognition Network (AMaskNet) was proposed to improve the performance of masked face recognition, which includes two modules: a feature extractor and a contribution estimator. Therein, the contribution estimator was employed to learn the contribution of the feature elements, thus achieving refined feature representation by simple matrix multiplications. Meanwhile, the end-to-end training strategy was utilized to optimize the entire model. Finally, a mask-aware similarity Matching Strategy (MS) was adopted to improve the performance in the inference stage. Experiments showed that the proposed method consistently outperformed comparative methods on three masked face recognition datasets: RMFRD, COX, and Public-IvS. Meanwhile, qualitative analysis experiments using CAM indicated that the contribution learned by AMaskNet was more conducive to masked face recognition. The proposed solution was submitted to the National Institute of Standards and Technology's (NIST) Face Recognition Vendor Test in 2021, and achieved the third place in the World and the top ranking among the Japanese vendors for the mask-wearing category.

The second research topic described in this thesis proposed a content-aware feature aggregation scheme to aggregate complementary information between different frames. The difficulties in video-based face recognition, such as dramatic pose variations and low quality, can be alleviated by leveraging the rich complementary information between the frames. However, limited by the mini-batch training strategy, the current deep learning methods only utilizes the frames in each batch during training,

which ignore the content of the entire video. Therefore, firstly, a two-branch structure was designed as the Content-aware feature Aggregation Network (CAN). Secondly, a content-aware training strategy using a content bank was proposed, which alleviate the limitation of minibatch samples by using the content of the entire video or several images belonging to the same identity and thus could achieve global contribution estimation. Comparative studies on benchmark datasets: IJB-C, YouTube Face (YTF), PaSC, and COX, confirmed that the proposed approach could outperform comparative methods. Meanwhile, qualitative analysis on the Multi-PIE dataset indicated that the contribution learned by the CAN was reasonable and beneficial to video face recognition.

Based on the above research topics, a prototype of unconscious face recognition in surveillance scenes was designed to analyze and verify the feasibility of the proposed methods in practical application scenarios. They have also been applied to many practical products and services of Fujitsu.

5.2 Remaining Challenges and Future Directions

Thanks to the publication of large-scale labeled face recognition datasets and the rapid development of deep learning technology, face recognition has continuously achieved significant results on testing datasets. The target has changed from a constrained scene to an actual video surveillance scene where the scenario is completely unconstrained with uncooperative users. However, with the practical and commercial applications of face recognition, many of the ideal assumptions of academic research are being broken, and more and more real-world problems are emerging. Therefore, when results for a testing dataset saturates, some newer datasets that are more challenging and closer to the actual scenario will appear. The remaining challenges and possible future directions are as follows:

[1]. **Pursuit of extreme performance and efficiency**. Many killer-applications, e.g., financial authentication and watch-list surveillance, demand the accuracy

of matching at very low alarm rates, such as FAR=10^{-8}. Even with deep learning and massive training data, this is still a huge challenge. At the same time, the deployment of deep face recognition applied to mobile devices pursues the smallest size feature representation with extreme high accuracy using deep learning. Exploring this extreme face recognition performance beyond human imagination is of great significance to both industry and academia. After the algorithm has surpassed humans, it is also exciting to continuously improve the performance limit of the algorithm.

[2]. **Across different application scenarios**. In real-world applications, it is hard to collect and label sufficiant number of samples for countless real-world scenarios. A promising solution is firstly learning a general model and then transferring it to real-world application-specific scenarios. Although deep domain adaptation [134] has recently been applied to reduce algorithmic bias on different scenarios [135], a general solution for transferring face recognition is still an open problem.

[3]. **Privacy-preservation issues of face recognition**. Nowdays, privacy concerns are becoming increasingly prominent with the leakage of biological data. Facial images can predict not only demographic information such as age, ethnicity or gender, but even genetic information [136]. More recently, pioneering works such as the Semi-Adversarial network [137, 138, 139, 140] have explored generating recognizable biometric templates that can hide some private information from facial images. Further research into the principles of visual cryptography, signal mixing, and image perturbation to protect stored user privacy, such as face templates, is essential to address public concerns about privacy.

[4]. **Multi-modal fusion problems**. Face recognition by itself is inadequate to cover all the cases in biometric authentication tasks, such as matching faces before and after surgery, and biometric recognition of crowded people. Face recognition is also sensitive to occlusion, blur, pose, etc. Moreover, in many

video frames, some faces can even be not visible, for example, due to occlusion. Intuitively, it will be beneficial to combine all biometric features to make full use of them to obtain satisfactory results. However, these information sources can correspond to different biometrics, e.g., combination of face and palm vein for identification, sensors, e.g., combination of 3D and 2D cameras, feature extraction and feature matching techniques. Performing information fusion at the decision-level, rank-level, score-level, feature-level, and data-level is beneficial for facial biometric authentication applications.

In the future, with the development of face recognition technology, face recognition application scenarios will gradually change from semi-cooperative to completely unconstrained and non-cooperative surveillance scenarios. It will not only greatly improve access control, financial security, retail transactions, speed up transport check-ins, and even overhaul national security processes. In the future, face recognition combined with other biometrics will achieve an even more super-smart society according to its specific application field. Let us take the retail industry as an example. Face recognition will help the retail industry establish feasible relationships with customers. The main application of face recognition in the retail industry is self-service shopping. Without a cashier, customers can fill the shopping cart and pay by verifying the face ID connected to their E-wallets. Another application is emotion analysis. Store managers can use artificial intelligence and face recognition to capture customers' emotions about products, find the most attractive or least attractive products, and set their product portfolio correctly. Similarly, retail managers can identify loyal customers and offer them rewards, because facial recognition cameras can accurately detect ordinary shoppers or loyal customers.

However, with the wide application of face recognition in our daily life, more and more people begin to worry about the ethical issues of face recognition, such as, privacy-preservation and biases. In the future, on the one hand, more and more privacy preservation technologies will be used to protect users' privacy from anyone, even users and developers, such as, visual cryptography, signal mixing, and image

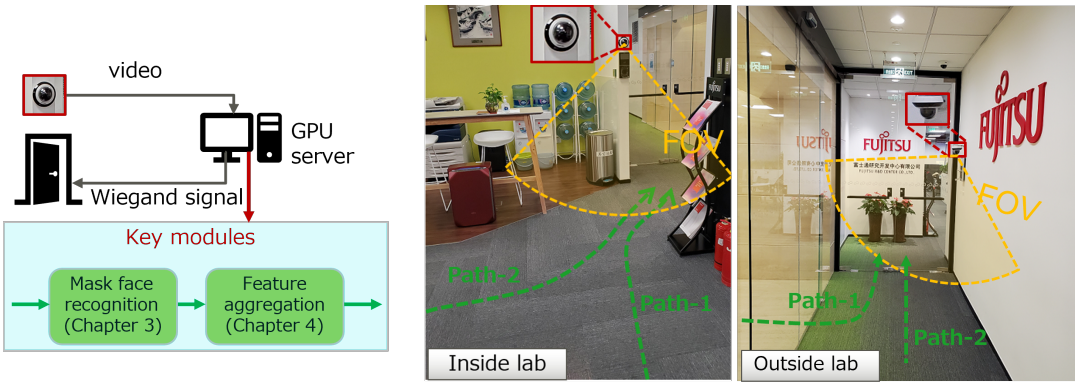
perturbation. On the other hand, many effective strategies to avoid bias will be proposed and used for face recognition, for example, building a more representative large-scale dataset that consists of a large demographically balanced set of faces and developing more advanced debiasing methods that treats all people equally, whether black or white, men or women.

Appendix A

Prototype of Unconscious Surveillance Face Recognition

The main objective of this thesis is to improve the robustness of unconscious face recognition for video surveillance. Despite the success of deep learning models under constrained face recognition scenarios, the deep features still demonstrate imperfect invariance to wearing a mask, where the whole face image can not be provided for description. However, the surveillance video provides us with abundant complementary information across frames compared with a single image. Therefore, this thesis focused on masked face recognition and feature aggregation-based face recognition between multi-fames. Based on these researches and the trained models, I developed a prototype of unconscious surveillance face recognition for access control of a laboratory gate to analyze and verify the feasibility of the proposed methods in practical application scenarios. The prototype recognizes an unconscious laboratory member and opens the door when he/she walks toward the gate.

The process of the prototype is shown schematically in Figure A.1. As Step 1, two surveillance cameras are installed on the inside and outside of the laboratory entrance to collect face images of the subject without any cooperation, and send them to the server. As Step 2, the process of face detection, face tracking, face alignment, feature extraction, and face matching is completed in the server using the face images captured by the surveillance cameras. The core module is the feature extraction, which

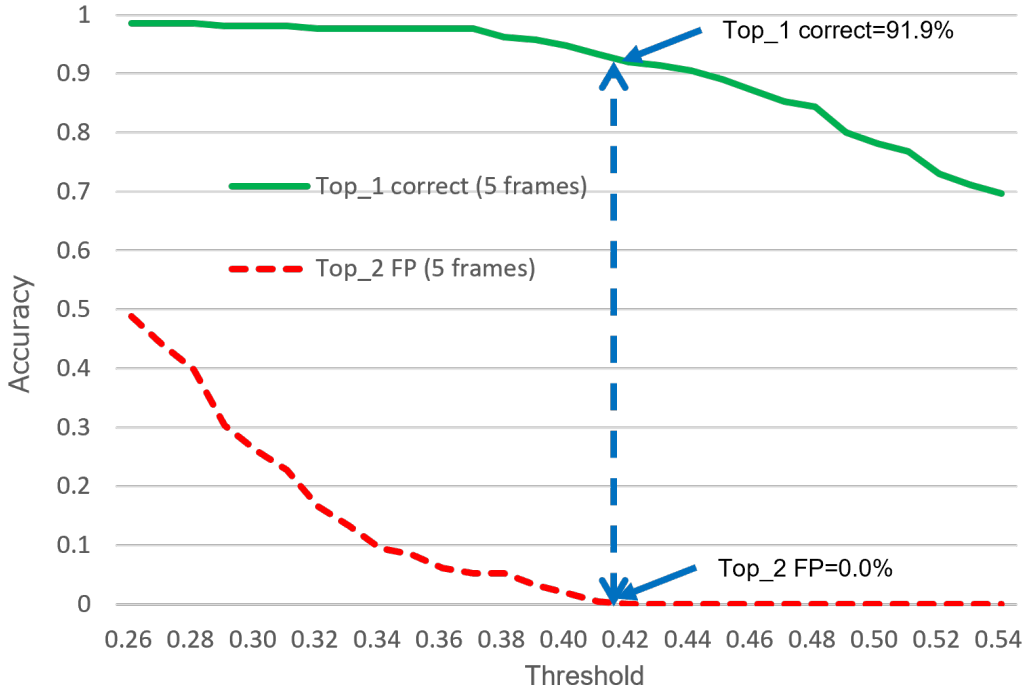


A.1. The process of the prototype

A.2. Sample images of surveillance cameras.

is mainly composed of two parts: The masked face recognition model described in Chapter 3 is used to extract the frame-level face features of each frame, and then the content-aware contribution estimation for feature aggregation method in Chapter 4 is used to extract the video-level face features. As Step 3, the result of face matching in the previous step is sent to the access control system in the form of Wiegand signal, which is the most common communication method used by access control devices [141]

To evaluate the prototype's performance, 211 video clips were collected and labeled manually. The gallery included 89 identities. The criteria for a "correct recognition" of each clip is that the target identity is at the first rank in the result compared with all identities and the similarity score is larger than a given threshold. Note that if the similarity score of the second rank identity is also larger than the threshold, the recognition is considered as failed (False Positive; FP). The threshold is set to when the second rank FP is 0. As shown in Figure A.3, the performance of the prototype can reach 91.9% when FP is set to 0.



A.3. Experimental results of the prototype system. Here, the Top-1 correct indicates that the first rank is the target identity and the score is larger than a threshold. Top-2 FP (false positive) incicates that the second rank is larger than a threshold. Imposter FP incicatess that the score of the 1st rank is larger than the threshold.

In order to avoid false recognition, the FP of this prototype is set to 0, which is the highest level. However, even with highest FP, the performance of the prototype still reached 91.9%. The system has been running steadily and continuously for more than one year, and its recognition results are directly used for the laboratory attendance. Although the performance of the prototype has not reached 100%, the 8.1% of the user who are rejected are mainly from the path-2 of the inside laboratory camera and the path-2 of the outside laboratory camera. In practice, if a user walks to the door, the access control does not open automatically, the user only needs to look at the camera to perform face recognition, and then the gate will be opened automatically by the access control system.

The proposed models in this thesis have been successfully applied to different Fujitsu products, but we cannot use user's face data due to privacy-preservation issues. So this prototype is build to analyze and verify the feasibility of the proposed methods in

practical application scenarios. However, in order to ensure the privacy of face data of laboratory members, we will add a privacy-preservation algorithm to this prototype in the future. In addition, we will also evaluate the face expression recognition on this prototype.

Bibliography

[1] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. SphereFace: Deep hypersphere embedding for face recognition. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6738–6746, June 2017.

[2] Mei Wang and Weihong Deng. Deep face recognition: A survey. *Neurocomputing*, 429:215–244, March 2021.

[3] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and InSo Kweon. CBAM: Convolutional Block Attention Module. In *Computer Vision — ECCV 2018, 15th European Conference, Munich, Germany, September 8–14, 2018, Proceedings, Part VII*, volume 11211 of *Lecture Notes in Computer Science*, pages 3–19. Springer, November 2018.

[4] Fei Wang, Liren Chen, Cheng Li, Shiyao Huang, Yanjie Chen, Chen Qian, and Chen Change Loy. The devil of face recognition is in the noise. In *Computer Vision — ECCV 2018, 15th European Conference, Munich, Germany, September 8–14, 2018, Proceedings, Part IX*, volume 11213 of *Lecture Notes in Computer Science*, pages 780–795. Springer, September 2018.

[5] Zhiwu Huang, Shiguang Shan, Ruiping Wang, Haihong Zhang, Shihong Lao, Alifu Kuerban, and Xilin Chen. A benchmark and comparative study of video-based face recognition on COX face database. *IEEE Transactions on Image Processing (TIP)*, 24(12):5967–5981, October 2015.

[6] Zhongyuan Wang, Guangcheng Wang, Baojin Huang, Zhangyang Xiong, Qi Hong, Hao Wu, Peng Yi, Kui Jiang, Nanxi Wang, Yingjiao Pei, Heling Chen, Yu Miao, Zhibing Huang, and Jinbi Liang. Masked face recognition dataset and application. *Computing Research Repository arXiv Preprints*, art. arXiv:2003.09093, February 2020.

[7] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2921–2929, June 2016.

[8] Brianna Maze, Jocelyn C Adams, James A Duncan, Nathan D Kalka, Tim Miller, Charles Otto, Anil K Jain, W Tyler Niggel, Janet Anderson, Jordan Cheney, and Patrick Grother. IARPA Janus Benchmark-C: Face dataset and protocol. In *Proceedings of the 11th IAPR International Conference on Biometrics (ICB)*, pages 158–165, November 2018.

[9] J Ross Beveridge, P Jonathon Phillips, David S Bolme, Bruce A Draper, Geof H Givens, Yui Man Lui, Mohammad Nayeem Teli, Hao Zhang, W Todd Scruggs, Kevin W Bowyer, Patrick J Flynn, and Su Cheng. The challenge of face recognition from digital point-and-shoot cameras. In *Proceedings of the 2013 IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, pages 1–8, September 2013.

[10] Lior Wolf, Tal Hassner, and Itay Maoz. Face recognition in unconstrained videos with matched background similarity. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 529–534, June 2011.

[11] Ralph Gross, Iain Matthews, Jeffrey Cohn, Takeo Kanade, and Simon Baker. Multi-PIE. *Image and Vision Computing (IVC)*, 28(5):807–813, May 2010.

[12] Asashi Shimbun. Foreign passenger face recognition exits and customs electronic declaration service at Narita Airport. `https://www.asahi.com/articles/ASM8W4WHZM8WUDCB00B.html`, 2019-08-28.

[13] Zhiyi Cheng, Xiatian Zhu, and Shaogang Gong. Surveillance face recognition challenge. *Computing Research Repository arXiv Preprints*, art. arXiv:1804.09691, August 2018.

[14] Zengpeng Li, Ding Wang, and Eduardo Morais. Quantum-safe round-optimal password authentication for mobile devices. *IEEE Transactions on Dependable and Secure Computing (TDSC)*, 19(3):1885–1899, November 2020.

[15] Woodrow Wilson Bledsoe. Some results on multicategory pattern recognition. *Journal of the ACM (JACM)*, 13(2):304–316, April 1966.

[16] Takeo Kanade. Picture processing system by computer complex and recognition of human faces. *PhD Thesis, Graduate school of Engineering, Kyoto University*, May 1974.

[17] Matthew Turk and Alex Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience (JoCN)*, 3(1):71–86, January 1991.

[18] Roberto Brunelli and Tomaso Poggio. Face recognition: Features versus templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 15(10):1042–1052, October 1993.

[19] Peter N Belhumeur, Joao P Hespanha, and David J Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 19(7):711–720, July 1997.

[20] P Jonathon Phillips, Harry Wechsler, Jeffery Huang, and Patrick J Rauss. The FERET database and evaluation procedure for face-recognition algorithms. *Image and Vision Computing (IVC)*, 16(5):295–306, April 1998.

[21] Laurenz Wiskott, Norbert Krüger, N Kuiger, and Christoph von der Malsburg. Face recognition by elastic bunch graph matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 19(7):775–779, July 1997.

[22] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3D faces. In *Proceedings of the 26th ACM Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*, pages 187–194, August 1999.

[23] Terence Sim, Simon Baker, and Maan Bsat. The CMU Pose, Illumination, and Expression (PIE) database. In *Proceedings of the 5th IEEE International Conference on Automatic Face Gesture Recognition (FGR)*, pages 53–58, October 2002.

[24] Paul Viola and Michael J Jones. Robust real-time face detection. *International Journal of Computer Vision (IJCV)*, 57(2):137–154, May 2004.

[25] Amnon Shashua and Tammy Riklin-Raviv. The quotient image: Class-based re-rendering and recognition with varying illuminations. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 23(2):129–139, February 2001.

[26] Timo Ahonen, Abdenour Hadid, and Matti Pietikainen. Face description with local binary patterns: Application to face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 28(12):2037–2041, December 2006.

[27] Weihong Deng, Jiani Hu, and Jun Guo. Compressive binary patterns: Designing a robust binary face descriptor with random-field eigenfilters. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 41(3): 758–767, January 2018.

[28] Dong Chen, Xudong Cao, Fang Wen, and Jian Sun. Blessing of dimensionality: High-dimensional feature and its efficient compression for face verification. In *Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3025–3032, June 2013.

[29] Zhimin Cao, Qi Yin, Xiaoou Tang, and Jian Sun. Face recognition with learning-based descriptor. In *Proceedings of the 2010 IEEE Conference on*

Computer Vision and Pattern Recognition (CVPR), pages 2707–2714, June 2010.

[30] Tsung-Han Chan, Kui Jia, Shenghua Gao, Jiwen Lu, Zinan Zeng, and Yi Ma. PCANet: A simple deep learning baseline for image classification. *IEEE Transactions on Image Processing (TIP)*, 24(12):5017–5032, December 2015.

[31] Zhen Lei, Matti Pietikäinen, and Stan Z Li. Learning discriminant face descriptor. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 36(2):289–302, 2013.

[32] Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. In *University of Massachusetts Amherst, Technical Report, 07-49*, October 2007.

[33] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet classification with deep convolutional neural networks. *Communications of the ACM (CACM)*, 60(6):84–90, June 2017.

[34] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. DeepFace: Closing the gap to human-level performance in face verification. In *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1701–1708, June 2014.

[35] Yi Sun, Xiaogang Wang, and Xiaoou Tang. Deep learning face representation from predicting 10,000 classes. In *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1891–1898, June 2014.

[36] Florian Schroff, Dmitry Kalenichenko, and James Philbin. FaceNet: A unified embedding for face recognition and clustering. In *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–823, June 2015.

[37] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, June 2016.

[38] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7132–7141, June 2018.

[39] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *Computing Research Repository arXiv Preprints*, art. arXiv:1409.1556, September 2014.

[40] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, June 2015.

[41] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. ArcFace: Additive angular margin loss for deep face recognition. In *Proceedings of the 2019 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4690–4699, June 2019.

[42] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *Computer Vision — ECCV 2016, 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VII*, volume 9911 of *Lecture Notes in Computer Science*, pages 499–515. Springer, August 2016.

[43] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. MS-Celeb-1M: A dataset and benchmark for large-scale face recognition. In *Computer Vision — ECCV 2016, 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part III*, volume 9907 of *Lecture Notes in Computer Science*, pages 87–102. Springer, November 2016.

[44] DeepGlint. Challenge 3: Face feature test/trillion pairs. http://trillionpairs.deepglint.com/overview, 2018-09-21.

[45] Xiang An, Xuhan Zhu, Yuan Gao, Yang Xiao, Yongle Zhao, Ziyong Feng, Lan Wu, Bin Qin, Ming Zhang, Debing Zhang, and Ying Fu. Partial FC: Training 10 million identities on a single machine. In *Proceedings of the 18th IEEE International Conference on Computer Vision Workshops (ICCVW)*, pages 1445–1449, October 2021.

[46] Zheng Zhu, Guan Huang, Jiankang Deng, Yun Ye, Junjie Huang, Xinze Chen, Jiagang Zhu, Tian Yang, Jiwen Lu, Dalong Du, and Jie Zhou. WebFace260M: A benchmark unveiling the power of million-scale deep face recognition. In *Proceedings of the 2021 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10492–10502, June 2021.

[47] Kyodo Press. Japan's Narita, Haneda airports start facial recognition in full scale. https://english.kyodonews.net/news/2021/07/f104681f01b8-narita-haneda-airports-start-facial-recognition-in-full-scale.html, 2021-07-19.

[48] Hexun.com. People can pay on Alibaba mobile Alipay using face recognition. http://money.hexun.com/2019-06-24/197629409.html, 2019-06-24.

[49] Fujitsu. Fujitsu delivers cashless, contactless retail experience for masked shoppers with hygienic, multi-factor biometric authentication technology. https://www.fujitsu.com/global/about/resources/news/press-releases/2021/0121-01.html, 2021-01-21.

[50] Ira Kemelmacher-Shlizerman, Steven M Seitz, Daniel Miller, and Evan Brossard. The MegaFace benchmark: 1 million faces for recognition at scale. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4873–4882, June 2016.

[51] Zhiyi Cheng, Xiatian Zhu, and Shaogang Gong. Low-resolution face recognition. In *Computer Vision — ACCV 2018, 14th Asian Conference on Computer*

Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part III, volume 11363 of *Lecture Notes in Computer Science*, pages 605–621. Springer, December 2018.

[52] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1251–1258, June 2017.

[53] National Center for Immunization and Respiratory Diseases (NCIRD). When you've been fully Vaccinated — How to protect yourself and others. November 2021.

[54] Weiyang Liu, Yandong Wen, Zhiding Yu, and Meng Yang. Large-margin softmax loss for convolutional neural networks. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, pages 507–516, January 2016.

[55] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. CosFace: Large margin cosine loss for deep face recognition. In *Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5265–5274, June 2018.

[56] Yonghyun Kim, Wonpyo Park, Myung-Cheol Roh, and Jongju Shin. GroupFace: Learning latent groups and constructing group-based representations for face recognition. In *Proceedings of the 2020 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5621–5630, June 2020.

[57] Aqeel Anwar and Arijit Raychowdhury. Masked face recognition for secure authentication. *Computing Research Repository arXiv Preprints*, art. arXiv:2008.11104, October 2020.

[58] Satya Mallick. learnopencv. `https://github.com/spmallick/learnopencv`, 2020-06-24.

[59] Adnane Cabani, Karim Hammoudi, Halim Benhabiles, and Mahmoud Melkemi. MaskedFace-Net—A dataset of correctly/incorrectly masked face images in the context of COVID-19. *Smart Health*, 19:100144, March 2021.

[60] Ash368. face mask. `https://github.com/ash368/face_mask`, 2019-06-24.

[61] Yujun Shen, Ping Luo, Junjie Yan, Xiaogang Wang, and Xiaoou Tang. FaceID-GAN: Learning a symmetry three-player GAN for identity-preserving face synthesis. In *Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 821–830, June 2018.

[62] Xiangyu Zhu, Hao Liu, Zhen Lei, Hailin Shi, Fan Yang, Dong Yi, Guojun Qi, and Stan Z Li. Large-scale bisample learning on ID versus spot face recognition. *International Journal of Computer Vision (IJCV)*, 127(6):684–700, February 2019.

[63] Yuge Huang, Yuhan Wang, Ying Tai, Xiaoming Liu, Pengcheng Shen, Shaoxin Li, Jilin Li, and Feiyue Huang. CurricularFace: Adaptive curriculum learning loss for deep face recognition. In *Proceedings of the 2020 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5900–5909, June 2020.

[64] Xinxin Shan, Yue Lu, Qingli Li, and Ying Wen. Model-based transfer learning and sparse coding for partial face recognition. *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, 31(11):4347–4356, November 2020.

[65] Meng Zhang, Rujie Liu, Hajime Nada, Hidetsugu Uchida, Tomoaki Matsunami, and Narishige Abe. A pairwise learning strategy for video-based face recognition. In *Proceedings of the 2019 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 38–44, June 2019.

[66] Shanming Yang, Weihong Deng, Mei Wang, Junping Du, and Jiani Hu. Orthogonality loss: Learning discriminative representations for face recognition.

IEEE Transactions on Circuits and Systems for Video Technology (TCSVT), 31 (6):2301–2314, September 2020.

[67] Xiao Zhang, Rui Zhao, Yu Qiao, Xiaogang Wang, and Hongsheng Li. AdaCos: Adaptively scaling cosine logits for effectively learning deep face representations. In *Proceedings of the 2019 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10823–10832, June 2019.

[68] Hakan Cevikalp, Hasan Serhan Yavuz, and Bill Triggs. Face recognition based on videos by using convex hulls. *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, 30(12):4481–4495, July 2019.

[69] Sixue Gong, Yichun Shi, and Anil K Jain. Video face recognition: Component-wise Feature Aggregation Network (C-FAN). *Computing Research Repository arXiv Preprints*, art. arXiv:1902.07327, February 2019.

[70] Wen Heng, Tingting Jiang, and Wen Gao. How to assess the quality of compressed surveillance videos using face recognition. *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, 29(8):2229–2243, August 2018.

[71] Weidi Xie and Andrew Zisserman. Multicolumn networks for face recognition. *Computing Research Repository arXiv Preprints*, art. arXiv:1902.07327, August 2018.

[72] Jiaolong, Yang and Peiran, Ren and Dongqing, Zhang and Dong, Chen and Fang, Wen and Hongdong, Li and Gang, Hua. Neural aggregation network for video face recognition. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5216–5225, June 2017.

[73] Yi-Chen Chen, Vishal M Patel, P Jonathon Phillips, and Rama Chellappa. Dictionary-based face recognition from video. In *Computer Vision — ECCV*

2012, 12th European Conference on Computer Vision, Florence, Italy, October 7–13, 2012, Proceedings, Part VI, volume 7577 of *Lecture Notes in Computer Science*, pages 766–779. Springer, October 2012.

[74] Omkar M Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. In *Proceedings of the 26th British Machine Vision Conference (BMVC)*, pages 41.1–41.12, September 2015.

[75] Xiaobo Wang, Shifeng Zhang, Shuo Wang, Tianyu Fu, Hailin Shi, and Tao Mei. Mis-classified vector guided softmax loss for face recognition. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence (AAAI)*, pages 12241–12248, February 2020.

[76] Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. ShuffleNet: An extremely efficient convolutional neural network for mobile devices. In *Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6848–6856, June 2018.

[77] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. VGGFace2: A dataset for recognising faces across pose and age. In *Proceedings of the 13th IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, pages 67–74, October 2018.

[78] Shikhar Vashishth, Shyam Upadhyay, Gaurav Singh Tomar, and Manaal Faruqui. Attention interpretability across NLP tasks. *Computing Research Repository arXiv Preprints*, art. arXiv:1909.11218, September 2019.

[79] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30:5998–6008, December 2017.

[80] Hefei Ling, Jiyang Wu, Lei Wu, Junrui Huang, Jiazhong Chen, and Ping Li. Self residual attention network for deep face recognition. *IEEE Access*, 7: 55159–55168, 2019.

[81] Yongming Rao, Jiwen Lu, and Jie Zhou. Attention-aware deep reinforcement learning for video face recognition. In *Proceedings of the 16th IEEE International Conference on Computer Vision (ICCV)*, pages 3931–3940, October 2017.

[82] Yi Sun, Xiaogang Wang, and Xiaoou Tang. Deeply learned face representations are sparse, selective, and robust. In *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2892–2900, June 2015.

[83] Yi Sun, Yuheng Chen, Xiaogang Wang, and Xiaoou Tang. Deep learning face representation by joint identification-verification. *Advances in Neural Information Processing Systems*, 27:1988–1996, December 2014.

[84] Yi Sun, Ding Liang, Xiaogang Wang, and Xiaoou Tang. DeepID3: Face recognition with very deep neural networks. *Computing Research Repository arXiv Preprints*, art. arXiv:1502.00873, December 2015.

[85] Swami Sankaranarayanan, Azadeh Alavi, Carlos D Castillo, and Rama Chellappa. Triplet probabilistic embedding for face verification and clustering. In *Proceedings of the IEEE 8th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pages 1–8, November 2016.

[86] Changxing Ding and Dacheng Tao. Robust face recognition via multimodal deep face representation. *IEEE Transactions on Multimedia (TMM)*, 17(11): 2049–2058, October 2015.

[87] Erjin Zhou and Zhimin Cao and Qi Yin. Naive-deep face recognition: Touching the limit of LFW benchmark or not? *Computing Research Repository arXiv Preprints*, art. arXiv:1501.04690, August 2015.

[88] Yue Wu, Hongfu Liu, Jun Li, and Yun Fu. Deep face recognition with center invariant loss. In *Proceedings of the 25th Conference on Thematic Workshops of ACM Multimedia (ACM MM workshop)*, pages 408–414, August 2017.

[89] Yu Liu, Hongyang Li, and Xiaogang Wang. Rethinking feature discrimination and polymerization for large-scale recognition. *Computing Research Repository arXiv Preprints*, art. arXiv:1710.00870, August 2017.

[90] Xiao Zhang, Zhiyuan Fang, Yandong Wen, Zhifeng Li, and Yu Qiao. Range loss for deep face recognition with long-tailed training data. In *Proceedings of the 16th IEEE International Conference on Computer Vision (ICCV)*, pages 5409–5418, October 2017.

[91] Jiankang Deng, Yuxiang Zhou, and Stefanos Zafeiriou. Marginal loss for deep face recognition. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition workshops (CVPRW)*, pages 60–68, June 2017.

[92] Hao Liu, Xiangyu Zhu, Zhen Lei, and Stan Z Li. AdaptiveFace: Adaptive margin and sampling for face recognition. In *Proceedings of the 2019 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11947–11956, June 2019.

[93] Bingyu Liu, Weihong Deng, Yaoyao Zhong, Mei Wang, Jiani Hu, Xunqiang Tao, and Yaohai Huang. Fair Loss: Margin-aware reinforcement learning for deep face recognition. In *Proceedings of the 17th IEEE International Conference on Computer Vision (ICCV)*, pages 10051–10060, October 2019.

[94] Jiankang Deng, Jia Guo, Jing Yang, Alexandros Lattas, and Stefanos Zafeiriou. Variational prototype learning for deep face recognition. In *Proceedings of the 2021 IEEE Computer Vision and Pattern Recognition (CVPR)*, pages 11906–11915, June 2021.

[95] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. Learning face representation from scratch. *Computing Research Repository arXiv Preprints*, art. arXiv:1411.7923, April 2014.

[96] Ankan Bansal, Anirudh Nanduri, Carlos D Castillo, Rajeev Ranjan, and Rama Chellappa. UMDFaces: An annotated face dataset for training deep networks.

In *Proceedings of the 10th IEEE International Joint Conference on Biometrics (IJCB)*, pages 464–473, October 2017.

[97] Ankan Bansal, Carlos Domingo Castillo, Rajeev Ranjan, and Rama Chellappa. The do's and don'ts for CNN-based face verification. In *Proceedings of the 16th IEEE International Conference on Computer Vision Workshops (ICCVW)*, pages 2545–2554, October 2017.

[98] Manuel Günther, Peiyun Hu, Christian Herrmann, Chi-Ho Chan, Min Jiang, Shufan Yang, Akshay Raj Dhamija, Deva Ramanan, Jürgen Beyerer, Josef Kittler, Mohamad Al Jazaery, Mohammad Iqbal Nouyed, Guodong Guo, Cezary Stankiewicz, and Terrance E Boult. Unconstrained face detection and open-set face recognition challenge. In *Proceedings of the 10th IEEE International Joint Conference on Biometrics (IJCB)*, pages 697–706, October 2017.

[99] Athinodoros S Georghiades, Peter N Belhumeur, and David J Kriegman. From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 23(6):643–660, June 2001.

[100] Tom M Mitchell. *Machine Learning*, volume 1 of *McGraw-Hill Series in Computer Science*. McGraw-Hill, New York NY, USA, July 1997.

[101] Karl Ricanek and Tamirat Tesafaye. Morph: A longitudinal image database of normal adult age-progression. In *Proceedings of the 7th IEEE International Conference on Automatic Face and Gesture Recognition (FGR)*, pages 341–345, April 2006.

[102] Hong-Wei Ng and Stefan Winkler. A data-driven approach to cleaning large face datasets. In *Proceedings of the 21st IEEE International Conference on Image Processing (ICIP)*, pages 343–347, October 2014.

[103] Brendan F Klare, Ben Klein, Emma Taborsky, Austin Blanton, Jordan Cheney, Kristen Allen, Patrick Grother, Alan Mah, and Anil K Jain. Pushing the frontiers of unconstrained face detection and recognition: IARPA Janus

Benchmark-A. In *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1931–1939, June 2015.

[104] P Jonathon Phillips, Patrick J Flynn, Todd Scruggs, Kevin W Bowyer, Jin Chang, Kevin Hoffman, Joe Marques, Jaesik Min, and William Worek. Overview of the face recognition grand challenge. In *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 947–954, June 2005.

[105] Wen Gao, Bo Cao, Shiguang Shan, Xilin Chen, Delong Zhou, Xiaohua Zhang, and Debin Zhao. The CAS-PEAL large-scale Chinese face database and baseline evaluations. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans (TSMC)*, 38(1):149–161, December 2007.

[106] Mislav Grgic, Kresimir Delac, and Sonja Grgic. SCface—Surveillance cameras face database. *Multimedia Tools and Applications (MTA)*, 51(3):863–879, October 2011.

[107] Aleix Martinez and Robert Benavente. The AR face database. *CVC Technical Report, 24*, January 1998.

[108] Davisking. dlib. `https://github.com/davisking/dlib`, 2018-07-13.

[109] Daniel Sáez Trigueros, Li Meng, and Margaret Hartnett. Enhancing convolutional neural networks for face recognition with occlusion maps and batch triplet loss. *Image and Vision Computing (IVC)*, 79:99–108, November 2018.

[110] Walid Hariri. Efficient masked face recognition method during the COVID-19 pandemic. *Signal, Image and Video Processing (SIVP)*, 16(3):605–612, November 2022.

[111] Imran Qayyum Mundial, M Sohaib Ul Hassan, M Islam Tiwana, Waqar Shahid Qureshi, and Eisa Alanazi. Towards facial recognition problem in COVID-19 pandemic. In *Proceedings of the 4th International Conference on Electrical, Telecommunication and Computer Engineering (ELTICOM)*, pages 210–214, September 2020.

[112] Mehryar Emambakhsh and Adrian Evans. Nasal patches and curves for expression-robust 3D face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 39(5):995–1007, May 2016.

[113] Nishant Sankaran, Sergey Tulyakov, Srirangaraj Setlur, and Venu Govindaraju. Metadata-based feature aggregation network for face recognition. In *Proceedings of the 11th IAPR International Conference on Biometrics (ICB)*, pages 118–123, February 2018.

[114] Guanglu Song, Biao Leng, Yu Liu, Congrui Hetang, and Shaofan Cai. Region-based quality estimation network for large-scale person re-identification. *Computing Research Repository arXiv Preprints*, art. arXiv:1711.08766, November 2017.

[115] Tejas I Dhamecha, Gaurav Goswami, Richa Singh, and Mayank Vatsa. On frame selection for video face recognition. In *Advances in Face Detection and Facial Image Analysis, cham switzerland*, pages 279–297. Springer, Cham, Switzerland, April 2016.

[116] Kaneswaran Anantharajah, Simon Denman, Sridha Sridharan, Clinton Fookes, and Dian Tjondronegoro. Quality based frame selection for video face recognition. In *Proceedings of the 6th International Conference on Signal Processing and Communication Systems (ICSPCS)*, pages 1–5, December 2012.

[117] Angelina Kharchevnikova and Andrey V Savchenko. Efficient video face recognition based on frame selection and quality assessment. *PeerJ Computer Science*, 7:e391, February 2021.

[118] Jie Chang, Zhonghao Lan, Changmao Cheng, and Yichen Wei. Data uncertainty learning in face recognition. In *Proceedings of the 2020 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5710–5719, June 2020.

[119] Yichun Shi and Anil K Jain. Probabilistic face embeddings. In *Proceedings of the 17th IEEE International Conference on Computer Vision (ICCV)*, pages 6901–6910, October 2019.

[120] Yichun Shi, Xiang Yu, Kihyuk Sohn, Manmohan Chandraker, and Anil K Jain. Towards universal representation learning for deep face recognition. In *Proceedings of the 2020 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6816–6825, June 2020.

[121] Yongming Rao, Ji Lin, Jiwen Lu, and Jie Zhou. Learning discriminative aggregation network for video-based face recognition. In *Proceedings of the 16th IEEE International Conference on Computer Vision (ICCV)*, pages 3801–3810, October 2017.

[122] Yu Liu, Junjie Yan, and Wanli Ouyang. Quality aware network for set to set recognition. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5790–5799, June 2017.

[123] Yirong Mao, Ruiping Wang, Shiguang Shan, and Xilin Chen. COSONet: COmpact Second-Order Network for video face recognition. In *Computer Vision — ACCV 2018, 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part III*, volume 11363 of *Lecture Notes in Computer Science*, pages 51–67. Springer, September 2019.

[124] William Evans, David Kirkpatrick, and Gregg Townsend. Right-triangulated irregular networks. *Algorithmica*, 30(2):264–286, January 2001.

[125] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. "GrabCut" interactive foreground extraction using iterated graph cuts. *ACM Transactions on Graphics (TOG)*, 23(3):309–314, August 2004.

[126] Joseph Luttrell, Zhaoxian Zhou, Yuanyuan Zhang, Chaoyang Zhang, Ping Gong, Bei Yang, and Runzhi Li. A deep transfer learning approach to fine-tuning facial recognition models. In *Proceedings of the 13th IEEE Conference*

on Industrial Electronics and Applications (ICIEA), pages 2671–2676, May 2018.

[127] Walid Hariri, Hedi Tabia, Nadir Farah, Abdallah Benouareth, and David Declercq. 3D face recognition using covariance based descriptors. *Pattern Recognition Letters (PRL)*, 78:1–7, July 2016.

[128] Dulari Bhatt, Chirag Patel, Hardik Talsania, Jigar Patel, Rasmika Vaghela, Sharnil Pandya, Kirit Modi, and Hemant Ghayvat. CNN variants for computer vision: History, architecture, application, challenges and future scope. *Electronics*, 10(20):2470, October 2021.

[129] Chirag I Patel, Dulari Bhatt, Urvashi Sharma, Radhika Patel, Sharnil Pandya, Kirit Modi, Nagaraj G Cholli, Akash Patel, Urvi Bhatt, Muhammad Ahmed Khan, Shubhankar Majumdar, Mohd Zuhair, Khushi Patel, Syed Aziz Shah, and Hemant Ghayvat. DBGC: Dimension-Based Generic Convolution block for object recognition. *Sensors*, 22(5):1780, February 2022.

[130] Jiajiong Cao, Yingming Li, and Zhongfei Zhang. Celeb-500k: A large training dataset for face recognition. In *Proceedings of the IEEE 25th International Conference on Image Processing (ICIP)*, pages 2406–2410, November 2018.

[131] Léon Bottou. Stochastic gradient descent tricks. In *Neural networks: Tricks of the trade*, volume 7700 of *Lecture Notes in Computer Science*, pages 421–436. Springer-Verlag, Berlin Heidelberg, September 2012.

[132] Changxing Ding and Dacheng Tao. Trunk-branch ensemble convolutional neural networks for video-based face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 40(4):1002–1014, April 2017.

[133] Zhiwu Huang, Ruiping Wang, Shiguang Shan, and Xilin Chen. Learning Euclidean-to-Riemannian metric for point-to-set classification. In *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1677–1684, June 2014.

[134] Mei Wang and Weihong Deng. Deep visual domain adaptation: A survey. *Neurocomputing*, 312:135–153, October 2018.

[135] Zimeng Luo, Jiani Hu, Weihong Deng, and Haifeng Shen. Deep unsupervised domain adaptation for face recognition. In *Proceedings of the 13th IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, pages 453–457, May 2018.

[136] Yaron Gurovich, Yair Hanani1, Omri Bar1, Guy Nadav1, Nicole Fleischer1, Dekel Gelbman1, Lina Basel-Salmon, Peter M Krawitz, B Kamphausen Susanne, Martin Zenker, Lynne M Bird, and Karen W Gripp. Identifying facial phenotypes of genetic disorders using deep learning. *Nature Medicine*, 25(1): 60–64, January 2019.

[137] Xin Sun, Chengliang Tian, Changhui Hu, Weizhong Tian, Hanlin Zhang, and Jia Yu. Privacy-preserving and verifiable SRC-based face recognition with cloud/edge server assistance. *Computers and Security (COMPUSEC)*, 118: 102740, July 2022.

[138] Vahid Mirjalili, Sebastian Raschka, and Arun Ross. Gender privacy: An ensemble of semi adversarial networks for confounding arbitrary gender classifiers. In *Proceedings of the 9th IEEE International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pages 1–10, October 2018.

[139] Vahid Mirjalili, Sebastian Raschka, Anoop Namboodiri, and Arun Ross. Semi-adversarial networks: Convolutional autoencoders for imparting privacy to face images. In *Proceedings of the 11th IAPR International Conference on Biometrics (ICB)*, pages 82–89, February 2018.

[140] Vahid Mirjalili and Arun Ross. Soft biometric privacy: Retaining biometric utility of face images while perturbing gender. In *Proceedings of the 10th IEEE International Joint Conference on Biometrics (IJCB)*, pages 564–573, October 2017.

[141] Luis Velasco, Luis Miguel Contreras, Giuseppe Ferraris, Alexandros Stavdas, Filippo Cugini, Manfred Wiegand, and Juan Pedro Fernandez-Palacios. A service-oriented hybrid access network and clouds architecture. *IEEE Communications Magazine (COM-M)*, 53(4):159–165, April 2015.

Publication list

Peer-reviewed Journal

[1] Meng Zhang, Rujie Liu, Daisuke Deguchi, and Hiroshi Murase. Masked face recognition with mask transfer and self-attention under the COVID-19 pandemic. *IEEE Access* 10: 20527–20538, February 2022.

[2] Meng Zhang, Rujie Liu, Daisuke Deguchi, and Hiroshi Murase. Content-aware contribution estimation for feature aggregation in video face recognition. *IEEE Access* 10: 79301–79310, July 2022.

[3] Song Guo, Rujie Liu, Mengjiao Wang, Meng Zhang, Shijie Nie, Septiana Lina, and Narishige Abe. Exploiting the tail data for long-tailed face ecognition. *IEEE Access* 10: 97945–97953, September 2022.

International Conference

[1] Meng Zhang, Rujie Liu, Hajime Nada, Hidetsugu Uchida, Tomoaki Matsunami, and Narishige Abe. "A pairwise learning strategy for video-based face recognition." *In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 38–44, June 2019.

Press Release

[1] Fujitsu. "Fujitsu delivers cashless, contactless retail experience for masked shoppers with hygienic, multi-factor biometric authentication technology". https://www.fujitsu.com/global/about/resources/news/press-releases/2021/0121-01.html. Jan. 2021.

[2] Fujitsu. "Top-level world ranking for Fujitsu's mask-wearing face recognition technology in latest U.S. NIST Evaluation". https://www.fujitsu.com/global/about/research/article/202110-frvt.html. Oct. 2021.

International Patent

[1] Meng Zhang and Rujie Liu. "Device and method for classification using classification model and computer readable storage medium." U.S. Patent Application No. 17/460,316.

[2] Meng Zhang and Rujie Liu. "Apparatus and method for training classifying model." U.S. Patent No. 11,270,139. 8 Mar. 2022.

[3] Meng Zhang, and Rujie Liu. "Apparatus and method for training classification model and apparatus for classifying with classification model." U.S. Patent No. 11,113,513. 7 Sep. 2021.

[4] Meng Zhang and Rujie Liu. "Information processing method and information processing apparatus." U.S. Patent No. 11,113,581. 7 Sep. 2021.

[5] Meng Zhang, Fei Li, and Rujie Liu. "Method and apparatus for training classification model, and classification method." U.S. Patent Application No. 17/076,320.

[6] Meng Zhang, Fei Li, and Rujie Liu. "Information processing device and method, and device for classifying with model." U.S. Patent Application No. 17/090,032.

[7] Meng Zhang, Rujie Liu, and Jun Sun. "Method and apparatus for training face recognition model." U.S. Patent No. 10,769,499. 8 Sep. 2020.

[8] Meng Zhang and Rujie Liu. "Apparatus and method for training classification model and apparatus for performing classification by using classification model." U.S. Patent Application No. 16/736,180.