1 2 9 0

UNIVERSIDADE Đ
COIMBRA

Jin Bo

# Pseudo RGB-D Facial Image Processing

## Towards Face Recognition and Facial Diagnosis

**Tese no âmbito do Doutoramento em Engenharia Electrotécnica e de Computadores, ramo de especialização em Computadores e Electrónica, orientada pelo Professor Nuno Miguel Mendonça da Silva Gonçalves e Doutor Leandro Moraes Valle Cruz e apresentada ao Departamento de Engenharia Electrotécnica e de Computadores da Faculdade de Ciências e Tecnologia da Universidade de Coimbra.**

Julho de 2023

TESE

1 2 90

UNIVERSIDADE D
COIMBRA

Jin Bo

# Pseudo RGB-D Facial Image Processing
## Towards Face Recognition and Facial Diagnosis

PhD Thesis in Electrical and Computer Engineering, Specialization in Computers and Electronics, supervised by Professor Nuno Miguel Mendonça da Silva Gonçalves and Doctor Leandro Moraes Valle Cruz, and presented to the Department of Electrical and Computer Engineering of the Faculty of Sciences and Technology of the University of Coimbra.

July of 2023

THESIS

# Resumo

Hoje em dia, aplicações baseadas em imagens faciais tornaram-se generalizadas em campos como segurança, medicina e entretenimento. Fatores como iluminação, pose e expressões faciais podem impactar o desempenho dessas aplicações. Na última década, o desenvolvimento e a acessibilidade de sensores RGB-D de baixo custo tornaram possível obter informações de profundidade de objetos, levando os pesquisadores a abordar problemas de reconhecimento facial capturando imagens faciais RGB-D. No entanto, devido a restrições de privacidade, a obtenção de dados de profundidade de rostos humanos permanece um desafio, e as imagens faciais RGB 2D ainda são predominantes.

Seres inteligentes, como os humanos, podem usar sua vasta experiência para derivar informações espaciais 3D de cenas 2D. As metodologias de aprendizado de máquina visam resolver tais problemas treinando computadores para gerar respostas precisas. O objetivo de nossa pesquisa é melhorar o desempenho das tarefas de processamento facial subsequentes, como reconhecimento facial e diagnóstico facial, obtendo mapas de profundidade diretamente das imagens RGB correspondentes. Propomos uma estrutura de processamento de imagem facial pseudo RGB-D que substitui sensores de profundidade com mapas pseudo-profundidade gerados e oferece métodos orientados a dados para criar mapas de profundidade a partir de imagens faciais 2D.

Especificamente, projetamos e implementamos um modelo de rede adversarial generativa chamado 'D+GAN' para tradução de imagem para imagem multi-condicional com atributos faciais. Validamos a abordagem de processamento de imagem facial pseudo RGB-D através de experimentos em reconhecimento facial e diagnóstico facial usando vários conjuntos de dados. A estrutura de processamento de imagem facial pseudo RGB-D trabalha em conjunto com algoritmos de fusão de imagens para melhorar o desempenho do reconhecimento facial e diagnóstico facial.

Para explorar ainda mais as características pseudo-profundidade, propomos finalmente uma estrutura de processamento de imagem facial multimodal simulada que melhora significativamente o desempenho com uma probabilidade mais alta.

**Palavras-chave:** Reconhecimento facial pseudo RGB-D; diagnóstico facial profundo; Rede Adversarial Generativa com Profundidade Adicional (D+GAN); estimativa de profundidade facial monocular; rostos específicos de doenças

# Abstract

Today, face image-based applications have become widespread in fields such as security, medicine, and entertainment. Factors like lighting, pose, and facial expressions can impact the performance of these applications. Over the past decade, the development and affordability of low-cost RGB-D sensors have made it possible to obtain depth information of objects, leading researchers to tackle face recognition problems by capturing RGB-D face images. However, due to privacy restrictions, acquiring depth data from human faces remains challenging, and 2D RGB face images are still prevalent.

Intelligent beings, such as humans, can use their vast experience to derive 3D spatial information from 2D scenes. Machine learning methodologies aim to solve such problems by training computers to generate accurate answers. Our research's objective is to enhance the performance of subsequent face processing tasks, such as face recognition and facial diagnosis, by obtaining depth maps directly from corresponding RGB images. We propose a pseudo RGB-D facial image processing framework that replaces depth sensors with generated pseudo-depth maps and offers data-driven methods to create depth maps from 2D face images.

Specifically, we design and implement a generative adversarial network model named 'D+GAN' for multi-conditional image-to-image translation with facial attributes. We validate the pseudo RGB-D facial image processing approach through experiments on face recognition and facial diagnosis using various datasets. The pseudo RGB-D facial image processing framework works in conjunction with image fusion algorithms to enhance face recognition and facial diagnosis performance.

To further exploit pseudo-depth features, we ultimately propose a simulated multimodal facial image processing framework that significantly improves performance with a higher probability.

**Keywords:** Pseudo RGB-D face recognition; deep facial diagnosis; Depth Plus Generative Adversarial Network (D+GAN); monocular face depth estimation; disease-specific faces

# Acknowledgements

Here I would like to thank my supervisor Professor Nuno Miguel Gonçalves. He is a caring, tolerant, and experienced professor, who has benefited me a lot.

The Institute of Systems and Robotics (ISR), based in the University of Coimbra is an advanced research institution with hardworking and intelligent researchers. Thanks to all my ISR colleagues who helped me.

The University of Coimbra has a history of more than 700 years and has cultivated numerous outstanding talents, for which I have the utmost respect. I am honored to study here.

Special thanks to my daughter Jin Shuiyi. I hope you could understand that Dad did not accompany you at home.

This is life. Thank you.

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

In recent decades, biometrics has gained significant attention from researchers due to its uniqueness, stability, versatility, and difficulty to counterfeit. Among various biometric features, facial recognition has become increasingly popular. Face image-based applications can be found in a wide range of fields, such as security, entertainment, and healthcare [8], [9], [10]. Numerous factors, including illumination, posture, and expression, can impact the performance of facial image applications. Thus, it is crucial to address these factors to enhance application performance.

Machine learning algorithms are commonly used in facial image applications. Machine learning (see Figure 1.1) is a process in which computers derive models from input data through training to make decisions. In the field of digital image processing, feature extraction for traditional machine learning methods relies on hand-crafted engineering, which can be challenging. Most hand-crafted features need to be designed by specialists to reduce data complexity. In contrast, deep learning automates feature extraction without depending on hand-crafted engineering. Traditional machine learning methods for image processing mainly involve hand-crafted local descriptors combined with classifiers.

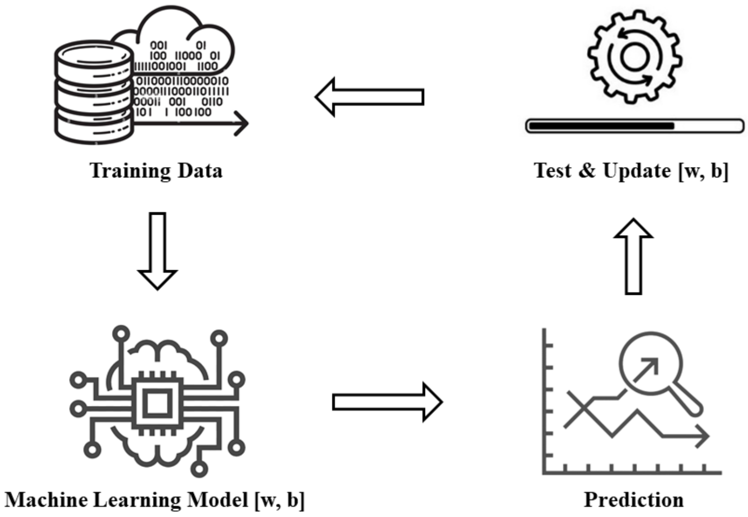


Figure 1.1: Brief process of machine learning

In recent years, deep learning technology has significantly improved state-of-the-art performance in many areas, particularly in computer vision, due to its powerful reasoning capabilities [11], [12]. Deep learning has demonstrated its best performance in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [13] since 2012. In the ILSVRC, numerous research experiments have shown that features learned by deep learning methods can represent the inherent information of the data more effectively than hand-crafted features [14], [15]. Some classic deep neural network models have emerged, such as AlexNet, GoogLeNet, VGGNet, ResNet, and Inception-ResNet. The mainstream approach has shifted toward deep learning with big data.

The most important difference between deep learning and traditional machine learning is that its performance increases as the amount of training data increases. If the dataset is small, the deep learning algorithm does not perform well because the deep learning algorithm requires a large amount of data to understand the patterns implied.

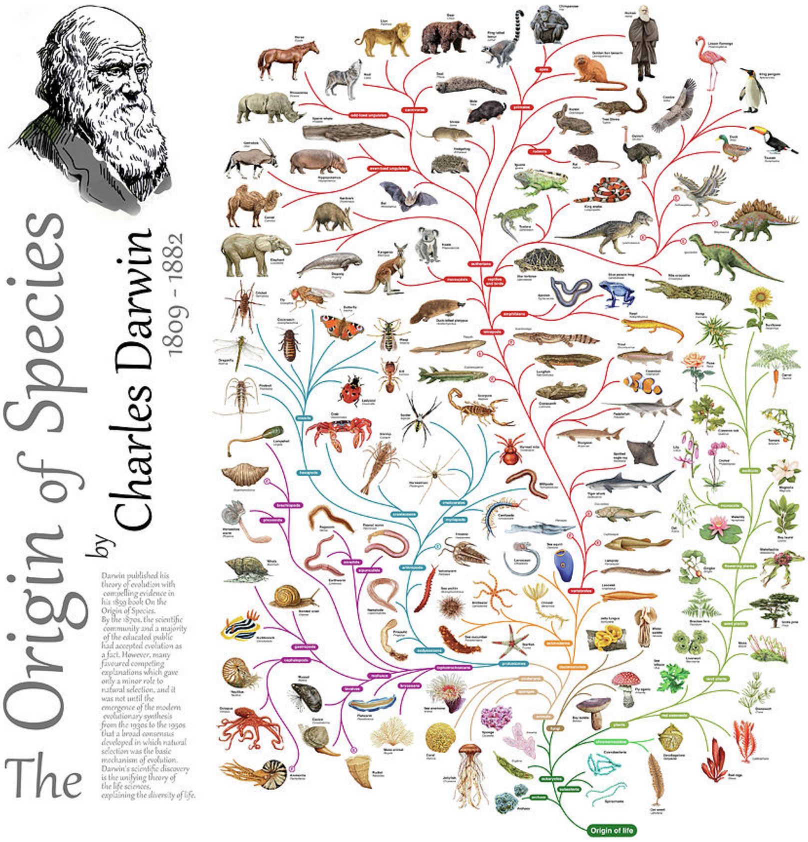
In this PhD work, we focus on two emerging research topics: face recognition and facial diagnosis.

## 1.1 Face Recognition

Face recognition refers to the technology of identifying or verifying the identity of individuals from facial images or videos. Due to its non-invasiveness, face recognition has become one of the most user-friendly biometric methods, leading to a wide range of applications. Face identification involves matching a given face image to one in a database of faces, representing a one-to-many mapping.

Darwin's theory of evolution proposes natural selection, which is the process of survival of the fittest and the elimination of those less adapted [16]. The genetic characteristics of organisms that adapt well to their environment are preserved through natural selection, a concept that is supported by abundant evidence and has had a profound impact on academic research [17]. Nowadays, all higher living creatures have two eyes for three-dimensional positioning, which is crucial for foraging. In contrast, most one-eyed creatures have gone extinct. Humans can still perform 3D positioning with one eye for a short period due to their extensive prior experience. While human vision is three-dimensional, most commonly encountered 2D face images lack facial spatial information. The importance of facial spatial information is undeniable.

In the last decade, advances in and the popularity of inexpensive RGB-D sensors have enabled us to utilize three-dimensional information. Compared to RGB face recognition, RGB-D face recognition requires depth images captured by depth sensors, such as Kinect [18] and PrimeSense [19]. There is some evidence, yet to be confirmed, that RGB-D face recognition performs better in terms of accuracy due to the effective use of spatial features [20], [21], [22].



Darwin published his theory of evolution with compelling evidence in his 1859 book On the Origin of Species. By the 1870s, the scientific community and a majority of the educated public had accepted evolution as a fact. However, many favoured competing explanations which gave only a minor role to natural selection, and it was not until the emergence of the modern evolutionary synthesis from the 1930s to the 1950s that a broad consensus developed in which natural selection was the basic mechanism of evolution. Darwin's scientific discovery is the unifying theory of the life sciences, explaining the diversity of life.

Figure 1.2: A Mixed Media by Gina Dsgn: The Origin of Species by Charles Darwin (1809-1882)

In modern society, although facial recognition systems are very convenient, they also give rise to many information security and privacy issues. In addition, there are no popular file formats for RGB-D data, and RGB-D cameras are not as common as RGB cameras. Therefore, RGB-D face images are not easy to collect and are much less common than RGB face images.

The emergence of machine learning enables computers to imitate the human learning process by learning from historical experiences to make predictions. It occurs to us that using machine learning algorithms might enable us to create models that effectively predict depth maps from their corresponding RGB images.

Prior to 2014, the main approach for image processing involved hand-crafted local descriptors with classifiers. With the development of big data and improvements in computer hardware performance, deep learning technology has become widely used in science and industry, offering more powerful reasoning capabilities than traditional machine learning models. Monocular depth estimation has inspired us to acquire 3D information from 2D face images using deep learning. Synthesizing the above, the idea behind this section is to generate corresponding depth maps solely from RGB face images, replacing depth maps collected by depth sensors, to perform the pseudo RGB-D face recognition.

## 1.2 Facial Diagnosis

The relationship between face and disease has been discussed from thousands years ago, which leads to the occurrence of facial diagnosis. Thousands years ago, Huangdi Neijing [23], the fundamental doctrinal source for Chinese medicine, recorded ”Qi and blood in the twelve Channels and three hundred and sixty-five Collaterals all flow to the face and infuse into the Kongqiao (the seven orifices on the face).” It indicates the pathological changes of the internal organs can be reflected in the face of the relevant areas. In China, one experienced doctor can observe the patient’s facial features to know the patient’s whole and local lesions, which is called “facial diagnosis”. Similar theories also existed in ancient India and ancient Greece. Nowadays, facial diagnosis refers to that practitioners perform disease diagnosis by observing facial features. The shortcoming of facial diagnosis is that for getting a high accuracy facial diagnosis requires doctors to have a large amount of practical experience. Modern medical researches [24], [25], [26] indicate that, indeed, many diseases will express corresponding specific features on human faces.

Nowadays, it is still difficult for people to take a medical examination in many rural and underdeveloped areas because of the limited medical resources, which leads to delays in treatment in many cases. Even in metropolises, limitations including the high cost, long queuing time in hospital and the doctor-patient contradiction which leads to medical disputes still exist. Computer-aided facial diagnosis enables us to carry out non-invasive screening and detection of diseases quickly and easily. Therefore, if facial diagnosis can be proved effective with an acceptable error rate, it will be with great potential. With the help of artificial intelligence, we could explore the relationship between face and disease with a quantitative approach.

With the help of a large amount of face images with labels from public face recognition datasets [27], [28], [29], CNN models are trained for learning most suitable face representations automatically for computer understanding and discrimination, and they get a high accuracy when testing on some specific datasets [30], [31].

The success of deep learning in the face recognition area motivates this project initially. However, the labelled data in the area of facial diagnosis is seriously insufficient. If we train a deep neural network from scratch, it will inevitably lead to overfitting. Apparently, face recognition and facial diagnosis are related. Since there is much more labeled data in the area of face recognition, transfer learning technology comes into view. In traditional learning, separate isolated models are trained on specific datasets for different tasks. Transfer learning, on the other hand, involves applying the knowledge gained while solving one problem to a different but related problem. According to whether the feature spaces of two domains are same or not, transfer learning can be divided into homogeneous transfer learning and heterogeneous transfer learning [32]. In our task of facial diagnosis, it belongs to homogeneous transfer learning. Deep transfer learning refers to transfer knowledge by deep neural networks. Thus, transfer learning makes it possible that identifying diseases from 2D face images by deep learning technique to provide a non-invasive and convenient way to realize early diagnosis and disease screening.

In the experiments, the following six diseases and their corresponding healthy controls are selected for validation purposes. Prevalence and incidence are two important indicators used to describe the epidemiology of diseases. Prevalence primarily focuses on the extent to which a disease is present in a population, while Incidence concentrates on the number of new cases that occur. Both of these indicators play a significant role in epidemiological research and the development of public health policies.

Prevalence is the proportion of total cases of a disease in a population at a specific time. It is calculated by dividing the number of cases of a disease or condition in a population by the total number of individuals in that population, as demonstrated in Equation (1.1). Prevalence provides a snapshot of how common a disease or condition is in a population at a given time, which includes both new and existing cases of a disease or condition.

$$\text{Prevalence} = \frac{\text{Number of existing cases of a disease}}{\text{Total population}} \tag{1.1}$$

Incidence refers to the number of new cases of a disease or condition that develop in a population over a specific period of time. It is calculated by dividing the number of new cases of a disease or condition in a population by the total number of individuals at risk in that population, as demonstrated in Equation (1.2). Incidence provides information on how quickly a disease or condition is spreading in a population. Incidence includes only new cases of a disease or condition that occurred during the specific period of time and does not include existing cases.

$$\text{Incidence} = \frac{\text{Number of new cases of a disease}}{\text{Number of individuals at risk}} \tag{1.2}$$

### 1.2.1 Acromegaly

Acromegaly is a hormone disorder caused by excessive secretion of growth hormone by the pituitary gland in adulthood, which will lead to abnormal hyperplasia or hypertrophy of organs. A survey shows that the prevalence rate of acromegaly ranges from 2.8 to 13.7 per 100,000 individuals approximately, and the annual incidence rate ranges from 0.2 to 1.1 per 100,000 individuals approximately [33]. Acromegaly is not easily noticed by patients for a short period of time, and is often mistaken for a phenomenon of weight gain or normal aging. Acromegaly and related complications such as high blood pressure, diabetes, and heart disease seriously affect patient health, quality of life and longevity. Studies show that if the patients with acromegaly do not receive treatment, the average remaining life is only about 10 years; However, if they can receive treatment, their life expectancy will be no different from that of ordinary people [34]. Therefore, early diagnosis and treatment are necessary. Acromegaly could cause gradual facial changes. Symptoms of acromegaly that probably appear on the patients' face include a prominent lower jaw, prominent brow bones, an enlarged nose, thickened lips, and wider spacing between teeth, which is shown as Figure 1.3.

### 1.2.2 Facial nerve paralysis

Facial nerve paralysis, caused by a dysfunction of the facial nerve, results in an inability to control facial muscles for smiling, blinking, and other facial movements on the affected side. Common causes of facial paralysis include facial nerve infection or inflammation, head trauma, and head or neck tumors. The prevalence of

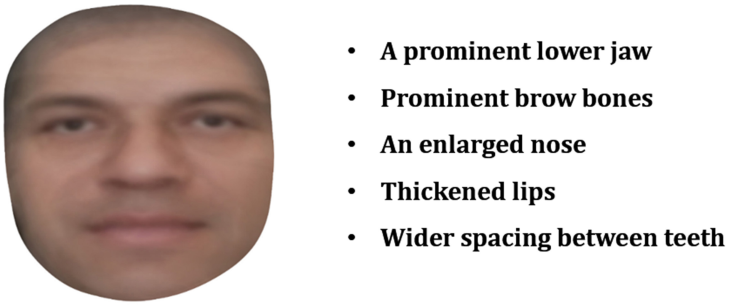


Figure 1.3: Acromegaly-specific face

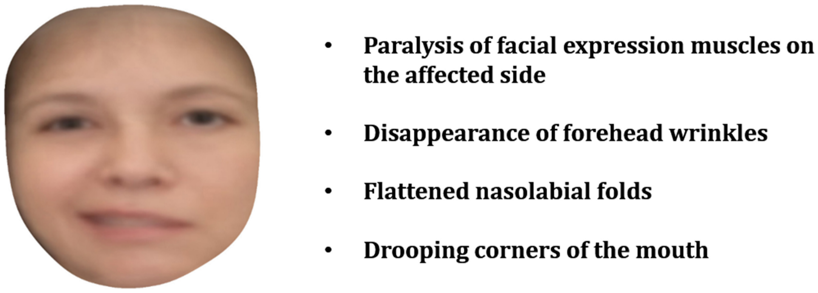

Paralysis of facial expression muscles on the affected side
Disappearance of forehead wrinkles
Flattened nasolabial folds
Drooping corners of the mouth

Figure 1.4: Facial nerve paralysis-specific face

facial nerve paralysis ranges from 11.5 to 40.2 per 100,000 individuals [35], and the annual incidence of facial paralysis ranges from 15 to 30 per 100,000 individuals approximately [36]. Facial nerve paralysis may cause numerous complications, including irreversible facial nerve damage, abnormal regeneration of nerve fibers, and partial or complete blindness in eyes that cannot be closed [37]. Symptoms of facial nerve paralysis, which likely appear on the patients' face, include paralysis of facial expression muscles on the affected side, disappearance of forehead wrinkles, flattened nasolabial folds, and drooping corners of the mouth, as illustrated in Figure 1.4.

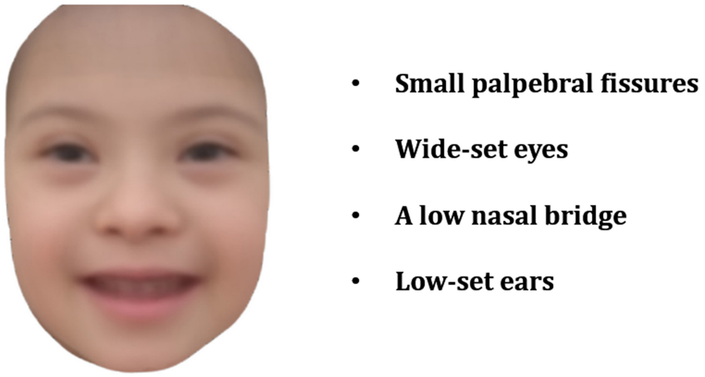


Figure 1.5: Down syndrome-specific face

### 1.2.3 Down syndrome

Down syndrome (DS) is a genetic disorder caused by trisomy of chromosome 21. Most patients with Down syndrome have physical and intellectual disabilities. Proper care can improve the quality of life for patients with Down syndrome. The estimated prevalence of DS approximately ranges from 136.6 to 142.9 per 100,000 individuals[38], [39]. According to the World Health Organization [40], the incidence of DS approximately ranges from 90.9 to 100 per 100,000 live births worldwide. Symptoms of Down syndrome, which may appear on the patients' face, include small palpebral fissures, wide-set eyes, a low nasal bridge, low-set ears, and more, as illustrated in Figure 1.5.

### 1.2.4 Leprosy

Leprosy, also known as Hansen's disease, is an infectious disease caused by a slow-growing type of bacteria called Mycobacterium leprae. If a patient with leprosy doesn't receive timely treatment, the disease can cause a loss of pain sensation, weakness, and poor eyesight. According to the World Health Organization, the incidence of leprosy approximately ranges from 2.5 to 3.2 per 100,000 individuals,

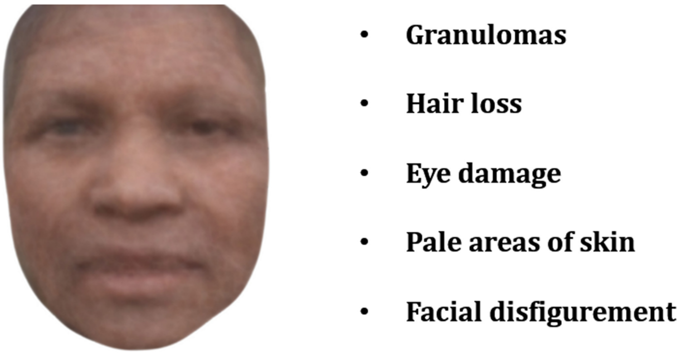


Figure 1.6: Leprosy-specific face

and the prevalence of leprosy approximately ranges from 2.2 to 2.7 per 100,000 individuals [41], [42]. Symptoms of leprosy, which may appear on the patients' face, include granulomas, hair loss, eye damage, pale areas of skin, and facial disfigurement (e.g., loss of nose), as illustrated in Figure 1.6.

### 1.2.5 Thalassemia

Thalassemia is a genetic blood disorder caused by abnormal hemoglobin production and is one of the most common inherited blood disorders worldwide. Hemoglobin is composed of two alpha and two beta chains. Different types of globin gene deletions or defects result in the corresponding inhibition of globin chain synthesis. Based on this fact, thalassemia is primarily divided into two types: $\alpha$ and $\beta$. The global incidence of thalassemia approximately ranges from 0.74 to 39.79 per 100,000 individuals [43], [44], while the prevalence of thalassemia varies from 2,500 to 15,000 per 100,000 individuals, approximately [45].

Since thalassemia can be fatal in early childhood without ongoing treatment, early diagnosis is vital. According to medical research [46], thalassemia can result in bone deformities, especially in the face. Symptoms of thalassemia that may appear on the face include small eye openings, epicanthal folds, a low nasal

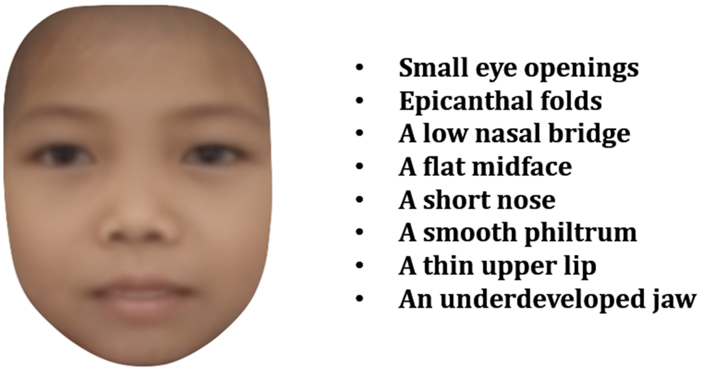


Figure 1.7: Thalassemia-specific face

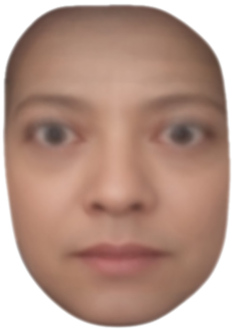
bridge, a flat midface, a short nose, a smooth philtrum, a thin upper lip, and an underdeveloped jaw, as illustrated in Figure 1.7.

### 1.2.6 Hyperthyroidism

Hyperthyroidism is a common endocrine disease caused by excessive amounts of the thyroid hormones T3 and T4 which can regulate the body's metabolism by various causes. The incidence of hyperthyroidism approximately ranges from 50 to 1300 per 100,000 individuals [47], and the average prevalence of hyperthyroidism approximately ranges from 800 to 1300 per 100,000 individuals [48].

If not treated early, hyperthyroidism can cause a series of serious complications and even threaten the patient's life. The typical facial characteristics of hyperthyroidism include thinning hair, shiny and protruding or staring eyes, increased ocular fissure, less frequent blinking, nervousness, consternation, and fatigue, as illustrated in Figure 1.8.

Figure 1.9 summarizes the prevalence of the six aforementioned condition categories. Figure 1.10 summarizes the incidence of the six aforementioned condition categories.



- Thinning hair
- Shiny and protruding or staring eyes
- Increased ocular fissure
- Less frequent blinking
- Nervousness
- Consternation
- Fatigue

Figure 1.8: Hyperthyroidism-specific face

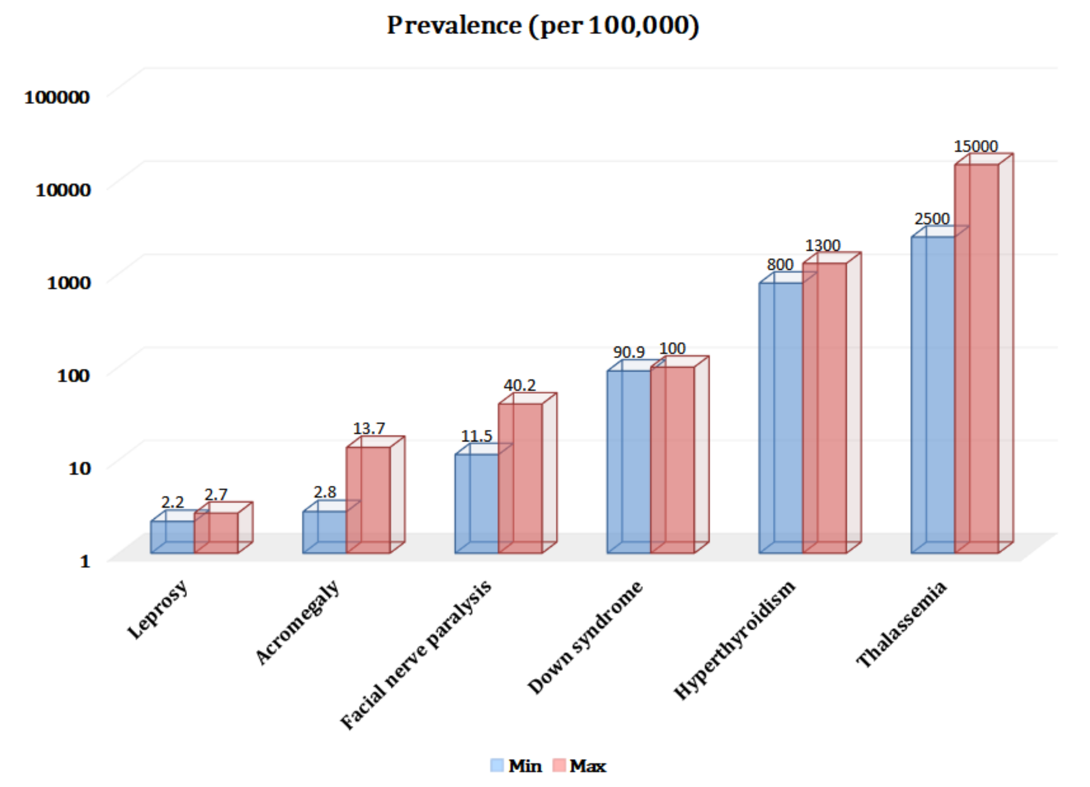


Figure 1.9: Prevalence of the six conditions used for the study

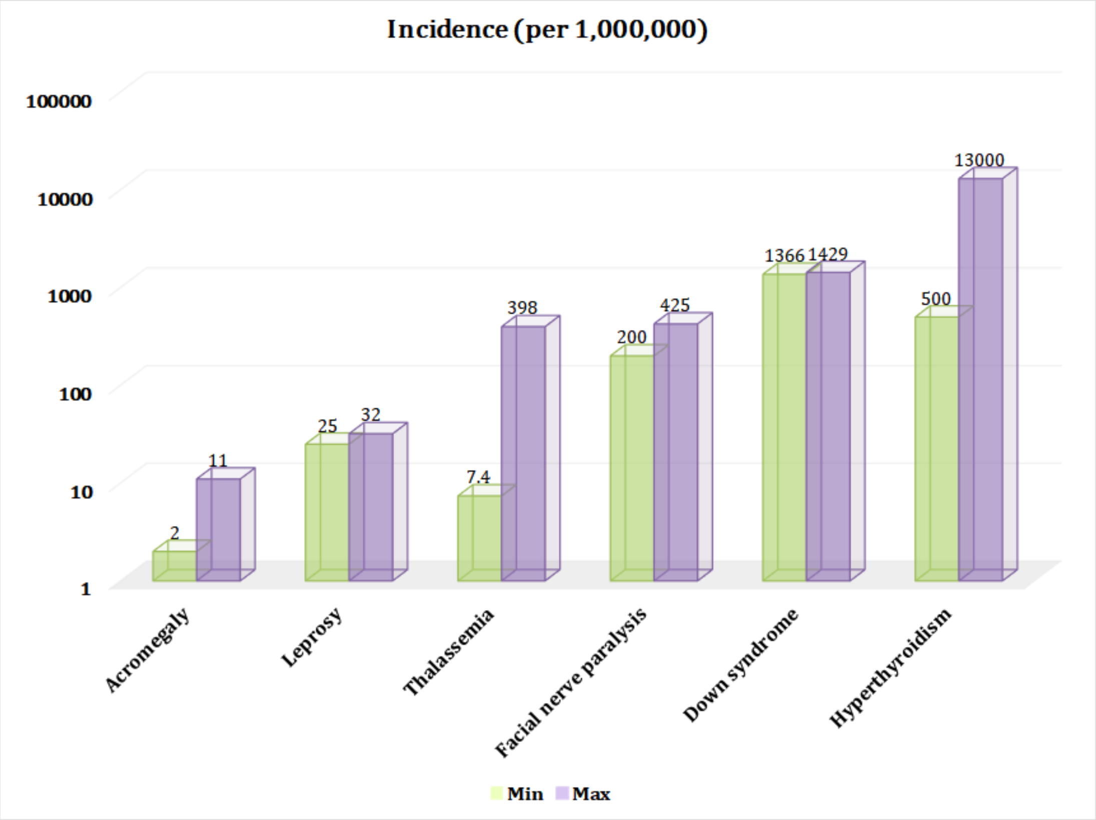


Figure 1.10: Incidence of the six conditions used for the study

## 1.3 Contributions

In the PhD research, our contributions to **face recognition** can be summarized as follows:

1. We propose and validate a pseudo RGB-D face recognition framework, as illustrated in Figure 1.11. Figure 1.11 presents a modular process, where algorithms within the module lists can be selected for preprocessing, depth generation, image fusion, and feature extraction, and then combined for face recognition. The best embodiment discovered is provided.

2. In order to fully utilize facial attributes, we specifically propose a GAN-based model, D+GAN, which performs multi-conditional image-to-image translation with facial attribute labels, transforming RGB face images into corresponding depth maps.

3. Based on the obtained depth maps, we improve the face recognition performance in cooperation with image fusion technologies, especially the Non-subsampled Shearlet Transform (NSST) [49].

In the PhD research, our contributions to **facial diagnosis** could be summarized as follows:

1. We definitely propose using deep transfer learning from face recognition to perform the computer-aided facial diagnosis on four conditions.

2. We apply the pseudo RGB-D facial image processing framework on the facial diagnosis on six conditions.

3. In order to make more effective use of pseudo-depth features, at the end of this dissertation, we propose an improved pseudo RGB-D facial image processing framework, *simulated multimodal framework*, to further improve the facial diagnosis performance.

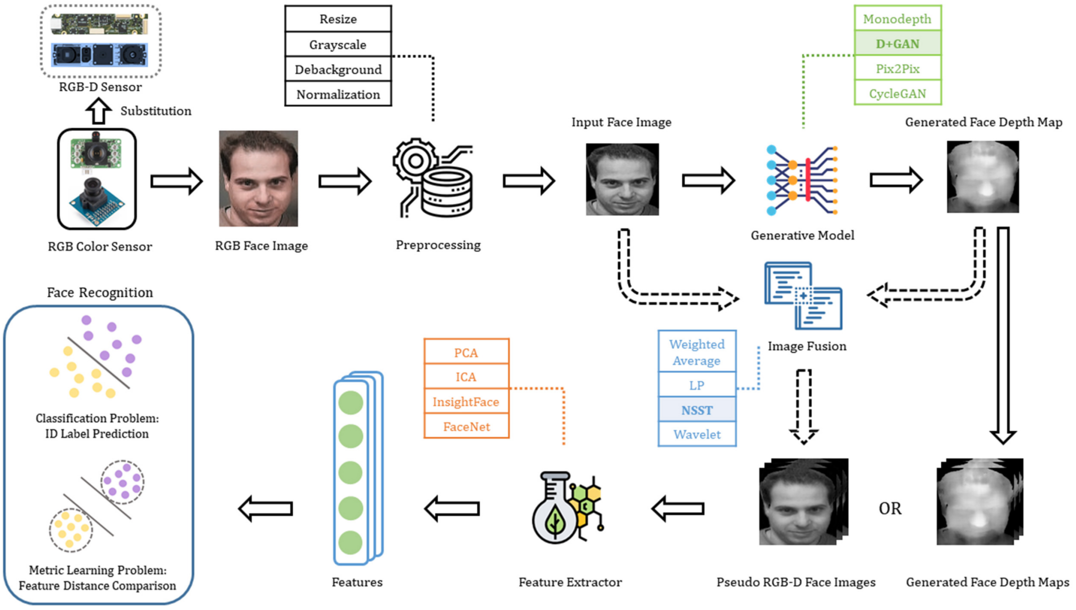


Figure 1.11: Pseudo RGB-D face recognition framework

The aforementioned partial research findings have been published as follows:

1. B. Jin, L. Cruz, and N. Gonçalves, ”Pseudo RGB-D Face Recognition”, ***IEEE Sensors Journal***, vol. 22, no. 22, pp. 21780–21794, 2022, DOI: 10.1109/JSEN.2022.3197235. (Google Scholar citations: 35+)

2. B. Jin, L. Cruz, and N. Gonçalves, ”Deep facial diagnosis: deep transfer learning from face recognition to facial diagnosis”, ***IEEE Access***, vol. 8, pp. 123649–123661, 2020, DOI: 10.1109/ACCESS.2020.3005687. (Google Scholar citations: 105+)

3. B. Jin, L. Cruz, and N. Gonçalves, ”Face Depth Prediction by the Scene Depth”, ***IEEE/ACIS 19th International Conference on Computer and Information Science (ICIS)***, pp. 42–48, 2021.
DOI: 10.1109/ICIS51600.2021.9516598. (Google Scholar citations: 10+)

4. B. Jin, ”Deep learning facial diagnosis system, CN”, ***National Invention***

***Patent***, Priority Date: 03.05.2017, Granted (2022), ZL201711255031.1, Publication of CN108806792B.

The aforementioned partial research findings are in the process of being prepared for publication:

1. B. Jin and N. Gonçalves, ”Pseudo RGB-D Face Recognition, CN”, ***National Invention Patent***, Priority Date: 04.08.2022, Preliminary Examination Passed, AN:202210959034.8.

2. ”Simulated Multimodal Deep Facial Diagnosis”, ***IEEE Journal***

## 1.4 Dissertation Structure

The dissertation is organized as follows: In Chapter 1, we introduce the relevant background of the PhD study. In Chapter 2, we review the related work in the fields of face recognition and facial diagnosis. In Chapter 3, we describe our proposed methods and their implementations. Our experimental results for face recognition are analyzed in Chapter 4. In Chapter 5, we analyze our experimental results for facial diagnosis. In Chapter 6, we engage in a discussion about the methodology and results presented throughout the thesis. Finally, in Chapter 7, we draw conclusions from our research and outline potential future research directions.

# Chapter 2

# Related Works

## 2.1 Face Recognition

Face recognition refers to the technology of identifying or verifying the identity of subjects from faces in images or videos. The history of face recognition algorithms can be traced back to the 1970s. Traditional machine learning methods involve extracting hand-crafted features, which are designed by specialists to reduce the complexity of input data, and training a model from the input to discover patterns for decision-making. Matthew Turk and Alex Pentland proposed the Eigenfaces method for face recognition, which uses a smaller set of face image features approximating the set of known face images [50]. Marian Stewart Bartlett et al. proposed using the Independent Component Analysis (ICA) method for face recognition, demonstrating that ICA representations were superior to Principal Components Analysis (PCA) based representations for face recognition across changes in certain conditions [51]. P. Jonathon Phillips developed a Support Vector Machine (SVM) based algorithm to generate the decision surface for face recognition [52].

In the past ten years, traditional machine learning methods have increas-

Figure 2.1: Thumbnails from ILSVRC

ingly been replaced by deep learning methods based on the convolutional neural network (CNN) in face recognition. The CNN structures mainly used in face recognition are basically consistent with the ones for classification tasks in ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [13], which is shown as Figure 2.1. In order to adapt to the task of face recognition, researchers mainly focus on discovering better training loss functions.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton proposed AlexNet, a classic CNN framework for classifying a large number of images in ILSVRC-2010 [1].

Matthew D. Zeiler and Rob Fergus discovered that the lower layers of a convolutional neural network capture generic features, while the higher ones learn source task-specific features through the deconvolution method. This forms the basis of deep transfer learning [2]. Figure 2.3 shows an example for describing this conclusion.

Yaniv Taigman et al. presented the DeepFace system, which can achieve human-level performance in face recognition [53]. The backbone network of DeepFace is based on AlexNet, and the loss function used is Softmax.

Karen Simonyan and Andrew Zisserman proposed the VGG model, which

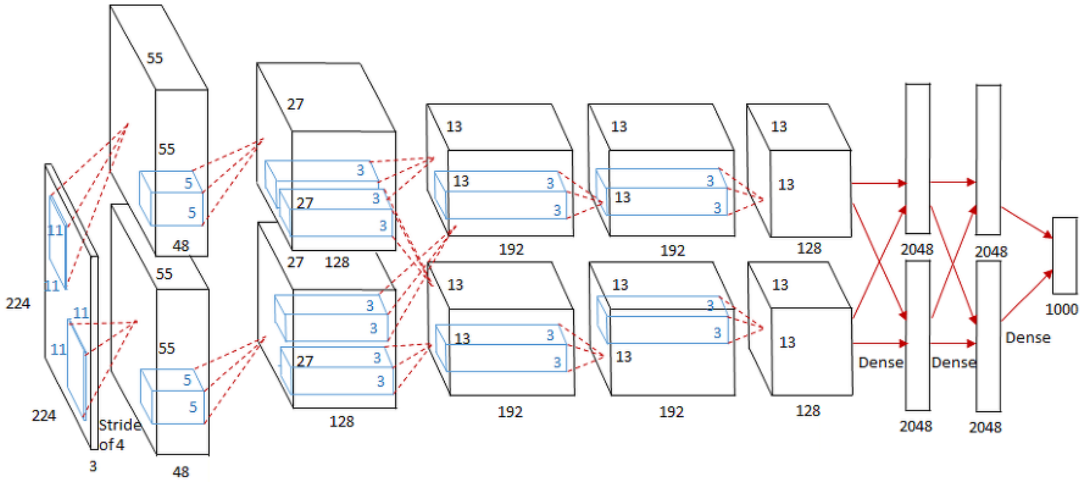


Figure 2.2: An illustration of the architecture of AlexNet from [1]

was developed to increase the depth of Convolutional Neural Networks (CNNs) in order to improve model performance [3].

Omkar M. Parkhi, Andrea Vedaldi, and Andrew Zisserman proposed the VGG-Face model, which is based on the VGG-16 CNN architecture and is designed to represent a face image as a robust vector of scores. The VGG-Face model was trained using 2.6 million images of 2.6 thousand people for face recognition and verification [27].

Christian Szegedy et al. proposed GoogLeNet, a 22-layer deep convolutional neural network that is a variant of the Inception Network [4].

Florian Schroff et al. presented FaceNet, which uses GoogLeNet as its backbone network and employs the triplet loss function for training, to directly map face images to Euclidean space [54].

Kaiming He et al. proposed ResNet, which increases the network depth to 152 layers by using residual blocks [5]. To address the vanishing gradient problem, ResNet adds skip connections, which combine the input x with the output after several weight layers, as shown in the Figure 2.6.

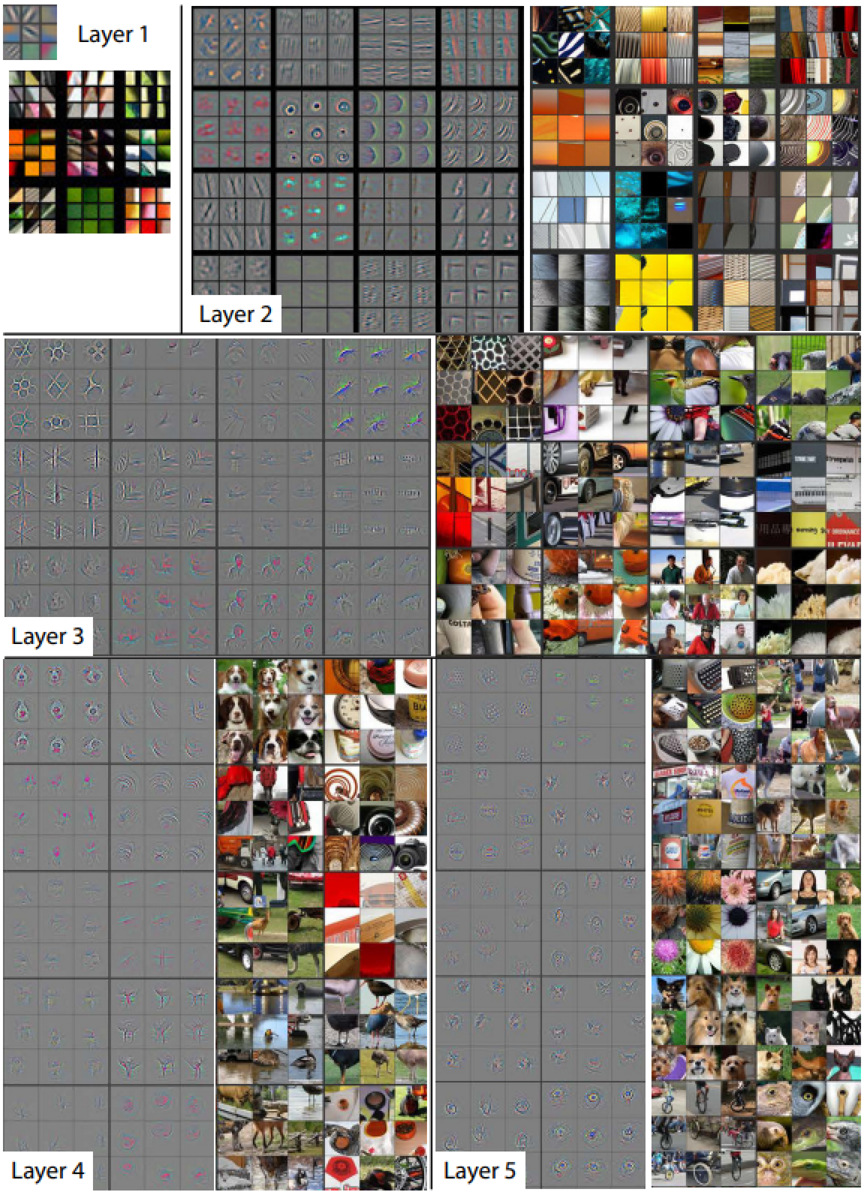


Figure 2.3: Visualization of features in a fully trained CNN model from [2]

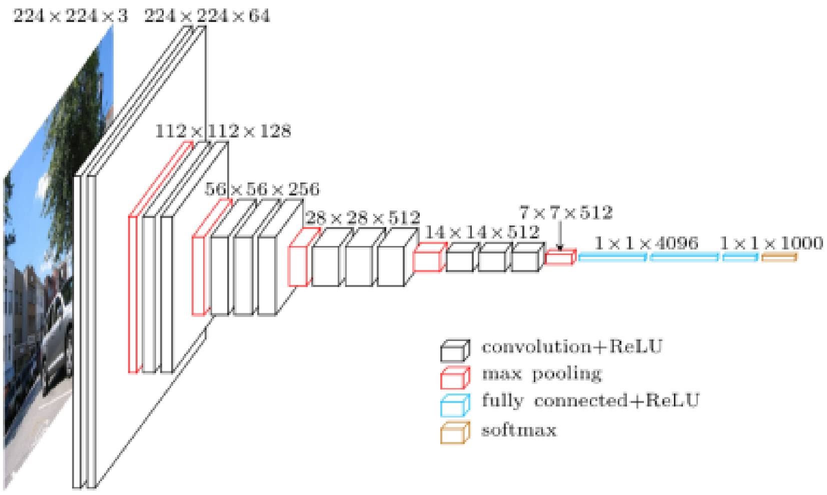


Figure 2.4: An illustration of VGG Net Structure from [3]

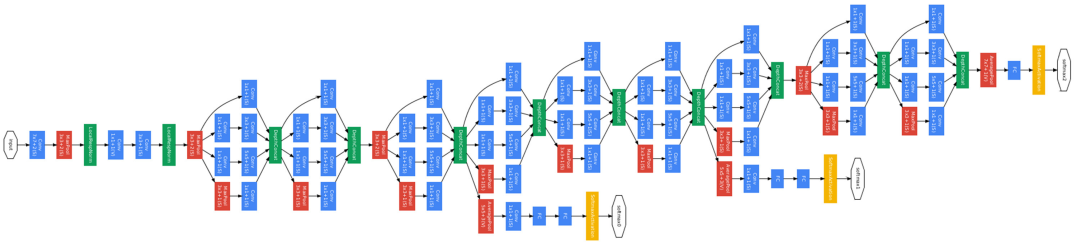


Figure 2.5: GoogLeNet Structure from [4]

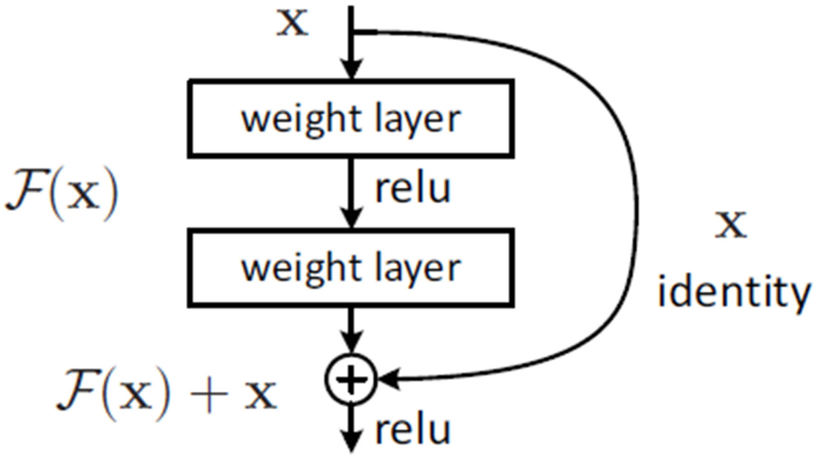

x
weight layer
$\mathcal{F}(\mathrm{x})$
relu
weight layer
x
identity
$\mathcal{F}(\mathrm{x}) + \mathrm{x}$
relu

Figure 2.6: A Building Block of Residual Network from [5]

Jiankang Deng et al. introduced the Additive Angular Margin Loss function, which aims to enhance the discriminative power of learned feature embeddings. When combined with ResNet, this function enables the model to achieve state-of-the-art results in the field of face recognition [30].

Similarly, in the field of RGB-D face recognition research, in the past decade, RGB-D face recognition technology has made significant progress, with researchers achieving breakthroughs in various aspects. The following is a review of RGB-D face recognition research in recent years:

Construction and expansion of RGB-D face datasets: With the popularization of RGB-D sensors, such as Microsoft Kinect and Intel RealSense, researchers have collected a large number of face images with depth information. These datasets, including CASIA-3D [55], BU-3DFE [56], and Bosphorus [57], have laid the foundation for the development of RGB-D face recognition algorithms.

RGB-D Feature Extraction Methods: Traditional feature extraction methods, such as Local Binary Pattern (LBP) and Principal Component Analysis (PCA), have been applied to RGB-D face recognition tasks [58]. Researchers have used deep neural networks with CNN structures to extract features from face depth maps. Yuancheng Lee et al. employed a 12-layer deep neural network, initially trained with a color face dataset, and later fine-tuned on depth face images for feature extraction, to perform joint classification [21]. Donghyun Kim et al. utilized a fine-tuned DCNN to extract features from 2D depth maps converted from 3D point clouds for calculating distances in face matching [59]. Hao Zhang et al. applied convolutional neural network models to learn complementary features between RGB and depth images, thereby enhancing the accuracy of RGB face recognition [60]. Furthermore, Luo Jiang, Juyong Zhang, and Bailin Deng proposed an attribute-aware loss function for RGB-D facial data [61].

### 2.1.1 Depth Estimation

Depth estimation to obtain a representation of the spatial structure of objects plays a crucial role in navigation, robotics, and augmented reality for inferring scene geometry from 2D images [62]. Suppose that there is a 2D image $I$, and we need a function $F$ to calculate its corresponding depth $D$. This process can be written as:

$$D = F(I) \tag{2.1}$$

There is no doubt that $F$ is a very complex function.

Because obtaining specific depth from a single image is equivalent to inferring three-dimensional space from a two-dimensional image, traditional depth estimation methods do not perform well in monocular depth estimation. Consequently, people have focused more on studying stereo vision, which involves obtaining depth information from multiple images. We can obtain the change of disparity between two pictures according to the change of viewing angle, so as to achieve the purpose of obtaining the depth. David Eigen, Christian Puhrsch and Rob Fergus used a multi-scale convolutional network architecture to predict the depth map from a single image on both NYU Depth and KITTI datasets [63]. Iro Laina et al. proposed a fully convolutional architecture encompassing residual learning to model the mapping between monocular scene images and corresponding depth maps [64]. Alhashim and Wonka used a standard encoder-decoder architecture with features extracted using pre-trained networks to get the depth [65]. The encoder part consists of a truncated DenseNet-169 pretrained by ImageNet without any additional modifications. The decoder is composed of basic blocks of convolutional layers, which are applied to the concatenation of the 2x bilinear upsampling from the previous block and the block in the encoder with the same spatial size after upsampling. For the above methods, it is necessary to know in advance the reference standard of the depth value corresponding to a large number of input pictures as training constraints, so as to back-propagate in the deep neural network, and train the neural network to perform depth prediction for scenes. It is referred to as supervised learning. In practice, it is not easy to obtain the depth values corresponding to a scene. At present, the commonly used method is to obtain the depth from the infrared sensor such as Kinect [18] or with the help of a laser LIDAR. Though the infrared sensor is relatively cheap, the collected depth range and accuracy are limited. In contrast, the cost of LIDAR is high. Using unsupervised learning for training is able to get a deep neural network model without knowing the depth before. Clement Godard, Oisin Mac Aodha and Gabriel J. Brostow used unsupervised learning method which is without ground truth to estimate the depth. The basic idea is to match the pixels of the left and right views to get the disparity map so as to calculate and optimize the depth map by Left-Right Consistency [66]. For getting a better performance, Clement Godard et al. used self-supervised learning with a standard, fully convolutional, U-Net to predict the depth [67].

Researchers have applied machine learning methods to estimate the depth of human faces from monocular images since the 1990s. Shang-Hong Lai, Chang-Wu Fu and Shyang Chang estimated the depth from defocus by using the raw image data in the vicinity of the edge [68]. Zhan-Li Sun, and Kin-Man Lam converted depth estimation into an independent component analysis (ICA) problem by incorporating a prior from the CANDIDE 3-D face model [69]. Zhan-Li Sun, Kin-Man Lam, and Qing-Wei Gao employed the nonlinear least-squares model to estimate the depth values of facial feature points and the pose of the 2D face image [70]. Since 2014, with the development of deep learning, researchers have successively used deep learning methods to perform monocular face depth estimation, which is similar with face recognition. Jiyun Cui et al. presented a deep neural network with a cascaded FCN and CNN architecture to estimate depth information of RGB face images [71]. Stefano Pini et al. applied a conditional Generative Adversarial Network (cGAN) for learning to translate intensity face images into their corresponding depth maps [72]. Abdullah Taha Arslan and Erol Seke applied a conditional Wasserstein GAN to perform face depth estimation [73]. Bo Jin, Leandro Cruz and Nuno Gonçalves predicted face depth maps by using pretrained models for scene depth estimation directly [74], which is also within the scope of the dissertation.

## 2.2 Facial Diagnosis

In this section, we primarily review some of the classic studies in computer-aided facial diagnosis, which are relatively limited in number. Since transfer learning serves as the core processing algorithm in our facial diagnosis research, we have dedicated a subsection in this section to introduce the related work on transfer learning.

Schneider et al. applied texture and geometry to compare graphs for similarity in order to detect acromegaly through face classification [75]. Their dataset includes face images of 57 patients with acromegaly. They claimed to have achieved an accuracy of 81.9%.

Zhao et al. proposed using ICA to locate the anatomical facial landmarks to discriminate between Down syndrome and healthy populations [76]. Their dataset includes 50 face images of patients with Down syndrome. They claimed to have achieved an accuracy of 0.967 and a F1 score of 0.956.

Zhao et al. proposed identifying patients with Down syndrome by ensembling the outputs of multiple different classifiers [77]. Their dataset includes 50 face images of patients with Down syndrome. They claimed to have achieved an accuracy of 0.967.

Kong et al. performed detection of acromegaly from facial photographs using a voting method to combine the predictions of basic estimators, including a Generalized Linear Model (GLM), a k-Nearest Neighbors (KNN) model, a Support Vector Machines (SVM) model, a CNN model, and a Random Forests (RF) model [78]. Their dataset includes 641 face images of patients with acromegaly. They claimed to have achieved a sensitivity of 96% and a specificity of 96%.

Boehringer et al. performed principal component analysis and linear discriminant analysis for a computer-based diagnosis among the 10 syndromes [79]. Their dataset includes 147 facial images with 10 syndromes, which means the average number of disease-specific face images for each category is approximately 15. They claimed to have achieved an accuracy of 75.7% for 10-class classification.

Shukla et al. used deep convolutional neural network to detect 6 disorders [80]. Their dataset includes 1126 facial images with 6 disorders, which means the average number of disease-specific face images for each category is approximately 188. They claimed to have achieved an accuracy of 48% for 6-class classification and an accuracy of 98.80% for binary classification.

Gurovich et al. introduced a facial image analysis framework called DeepGestalt, which employs computer vision and deep learning algorithms to quantify similarities to hundreds of syndromes [81]. Their dataset includes 17106 images with 216 different syndromes, which means the average number of disease-specific face images for each category is approximately 79. They claimed to have achieved 61.3∼68.7% top-1 accuracy and 89.4∼90.6% top-10 accuracy in identifying the correct syndrome on hundreds of images.

Jin et al. proposed using deep transfer learning from face recognition to facial diagnosis, named 'Deep Facial Diagnosis' [82]. Their dataset includes 280 images with 4 different diseases, which means the average number of disease-specific face images for each category is 70. They claimed to have achieved an overall top-1 accuracy of over 90%.

Porras et al. used deep neural networks and facial statistical shape models to screen children for genetic syndromes [83]. Their dataset includes 1,400 children images with 128 genetic conditions, which means the average number of disease-specific face images for each category is approximately 11. They claimed to have achieved an accuracy of 88% for the detection of a genetic syndrome.

Compared to two-dimensional images, three-dimensional images contain information about the spatial relationships between objects. In light of this, some researchers have started to explore facial diagnosis using three-dimensional facial images.

Hallgrímsson et al. conducted binary classification on 3D human face images using both parametric methods and machine learning techniques [84]. Their dataset includes 3327 images with 396 different syndromes, which means the average number of disease-specific face images for each category is approximately 8. They claimed to have achieved balanced accuracy was 73% and mean sensitivity 49%.

Bannister et al. performed 3D facial surface modeling using deep learning and performed 3D facial diagnosis [85]. Their dataset includes 4700 scans with 47 different syndromes, which means the average number of disease-specific face images for each category is approximately 100. They claimed to have achieved overall top-1 accuracy of 71%, and a mean sensitivity of 43% across all syndrome classes.

### 2.2.1 Transfer Learning

Pan and Yang categorize transfer learning approaches into instance-based transfer learning, feature-based transfer learning, parameter-based transfer learning, and

Table 2.1: **A summary of existing research in facial diagnosis**

| Research | Condition | No. of DSF images per category | Cls. problem | Method |
|---|---|---|---|---|
| Schneider et al. (2011) | Acromegaly | 57 | 2D Binary | Texture and geometry |
| Q. Zhao et al. (2014) | Down syndrome | 50 | 2D Binary | ICA |
| Q. Zhao et al. (2014) | Down syndrome | 50 | 2D Binary | Ensemble learning |
| Kong et al. (2018) | Acromegaly | 641 | 2D Binary | Ensemble learning |
| Boehringer et al. (2006) | 10 syndromes | 15 | 2D Multi-class | LDA |
| Shukla et al. (2017) | 6 disorders | 188 | 2D Multi-class | DCNN |
| Gurovich et al. (2019) | 216 syndromes | 81 | 2D Multi-class | DCNN |
| B. Jin et al. (2020) | 4 conditions | 70 | 2D Multi-class | Deep transfer learning |
| Porras et al. (2021) | 128 conditions | 11 | 2D Multi-class | Deep neural networks |
| Hallgrímsson et al. (2020) | 396 syndromes | 8 | 3D Multi-class | Parametric methods and ML |
| Bannister et al. (2022) | 47 syndromes | 100 | 3D Multi-class | Normalizing flows |

relation-based transfer learning [32]. Here we list some classical researches of each category.

Instance-based transfer learning is to reuse the source domain data by reweighting. Dai et al. presented TrAdaBoost to increase the instance weights that are beneficial to the target classification task and reduce the instance weights that are not conducive to the target classification task [86]. Tan et al. proposed a Selective Learning Algorithm (SLA) to solve the Distant Domain Transfer Learning (DDTL) problem with the supervised autoencoder as a base model for knowledge sharing among different domains [87].

As for feature-based transfer learning, it is to encode the knowledge to be transferred into the learned feature representation to reduce the gap between the source domain and the target domain. Pan et al. presented transfer component analysis (TCA) using Maximum Mean Discrepancy (MMD) as the measurement criterion to minimize the data distribution difference in different domains [88]. Long et al. presented Joint Adaptation Networks (JAN) to align the joint distributions-based on a joint maximum mean discrepancy (JMMD) criterion [89].

Regarding Parameter-based transfer learning is to encode the transferred knowledge into the shared parameters. It is widely used in the medical application. Razavian et al. found that CNNs trained on large-scale datasets (e.g. ImageNet) are also pretty good feature extractors [90]. Esteva et al. used Google Inception v3 CNN architecture pretrained on the ImageNet dataset (1.28 million images over 1,000 generic object classes) and fine-tuned on their own dataset of 129,450 skin lesions comprising 2,032 different diseases [91]. The high accuracy demonstrates an artificial intelligence capable of classifying skin cancer with a level of competence comparable to dermatologists. Yu et al. used a voting system-based on the output of three CNNs for medical images modality classification [92]. They fixed earlier layers of CNNs for reserving generic features of natural images, and trained high-level portion for medical image features. Shi et al. used a deep CNN based transfer learning method for pulmonary nodule detection in CT slices [93]. Raghu et al. demonstrated feature-independent benefits of transfer learning for better weight scaling and convergence speedups in medical imaging [94]. Shin et al. evaluated CNN architectures, dataset characteristics and transfer learning for thoraco-abdominal lymph node (LN) detection and interstitial lung disease (ILD) classification [95].

Besides, relation-based transfer learning is to transfer the relationship among the data in the source and target domains. Davis and Domingos utilized Markov logic to discover properties of predicates including symmetry and transitivity, and relations among predicates [96].

# Chapter 3

# Materials and Methods

In this chapter, we have introduced the materials and methods used for face recognition and facial diagnosis respectively. The materials and methods used in these two topics are both interconnected and distinct. This chapter mainly focuses on describing the methods common to both tasks, with specific differences in methods between the two tasks to be detailed separately in the next chapters.

## 3.1 Face Recognition

In this section, we propose a pseudo RGB-D face recognition framework, as illustrated in Figure 1.11. Figure 1.11 presents a modular process, where algorithms within the module lists can be selected for preprocessing, depth generation, image fusion, and feature extraction, and then combined for face recognition. For depth-generation, we specifically propose a GAN-based model, D+GAN, which performs multi-conditional image-to-image translation with facial attribute labels, transforming RGB face images into corresponding depth maps.

Generative Adversarial Network (GAN), proposed by Ian Goodfellow et al., is a model that learns a mapping from random noise vector to output images [6].

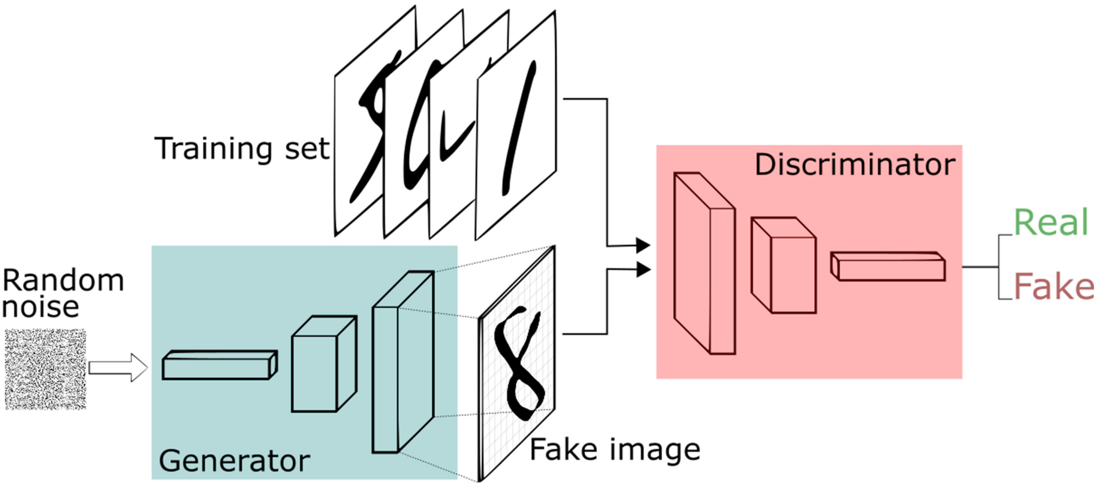


Figure 3.1: Schematic diagram of the original GAN structure from [6]

The original GAN consists of two parts which are a generator and a discriminator, which is shown as Figure 3.1. The objective of the generator is to map input Gaussian noise into a fake image, and the discriminator is to determine whether the input image comes from the generator or not, that is, to compute the probability of the input image being false. The conditional generative adversarial network (cGAN), proposed by Mehdi Mirza and Simon Osindero, is a supervised model that can generate output images with a desired condition from random noise [97]. Pix2Pix, proposed by Phillip Isola et al., could be regarded as a special case of cGAN. It takes the 2D image as the input condition of cGAN to realize the image-to-image translation [98]. ACGAN, proposed by Augustus Odena, Christopher Olah and Jonathon Shlens, is required not only to judge whether the input image is true or not, but also to classify the category of the input image in the discriminator part [99].

For adapting our task that is generating the corresponding depth from RGB face images better, we comprehensively refer to the above network structures and cooperate with some advanced skills, and propose the D+GAN. Figure 3.2 indicates the main structures of cGAN, Pix2Pix, ACGAN and D+GAN. It con-

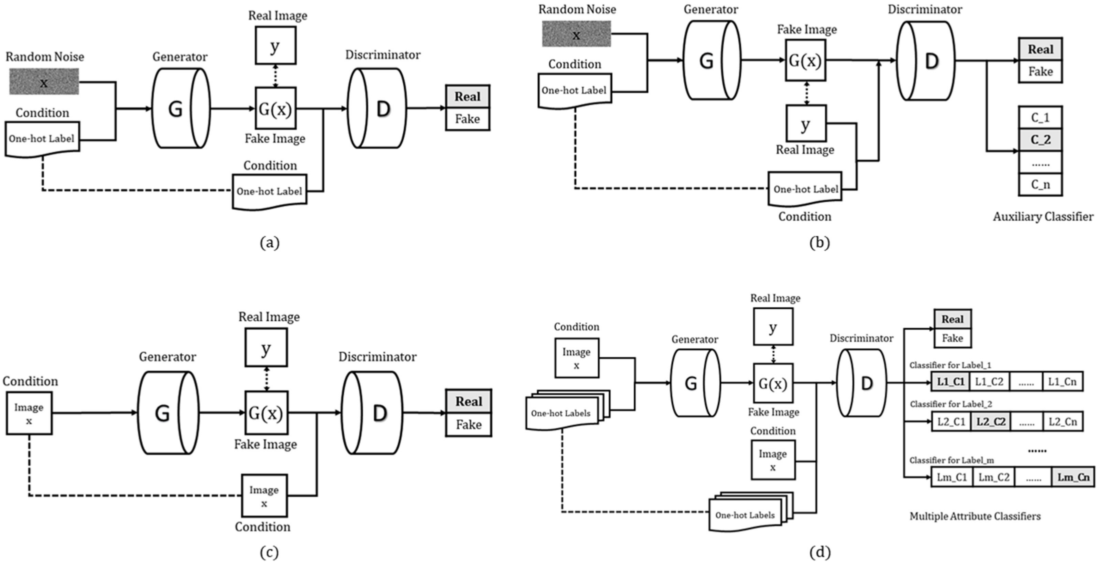


Figure 3.2: Main structures of GANs. (a) cGAN (b) ACGAN (c) Pix2Pix (d) Ours: D+GAN

cisely shows the difference between D+GAN and other GANs' main structures. They both control the generated images by introducing external conditions. For cGAN and ACGAN, the generator generates fake samples from random noise and conditions. For Pix2Pix, the generator generates fake images from images which could be regarded as conditions. Whereas, for D+GAN, the generator generates fake images from condition images and their corresponding labels. For cGAN and Pix2Pix, the discriminator determines whether the sample is the real sample that meet the condition. For ACGAN, the discriminator determines not only whether the sample is the real sample that meets the condition, but also the category of each sample. Whereas, for D+GAN, the discriminator determines not only whether the input sample is the real sample that corresponds the condition image, but also the multiple categories that each sample belongs to.

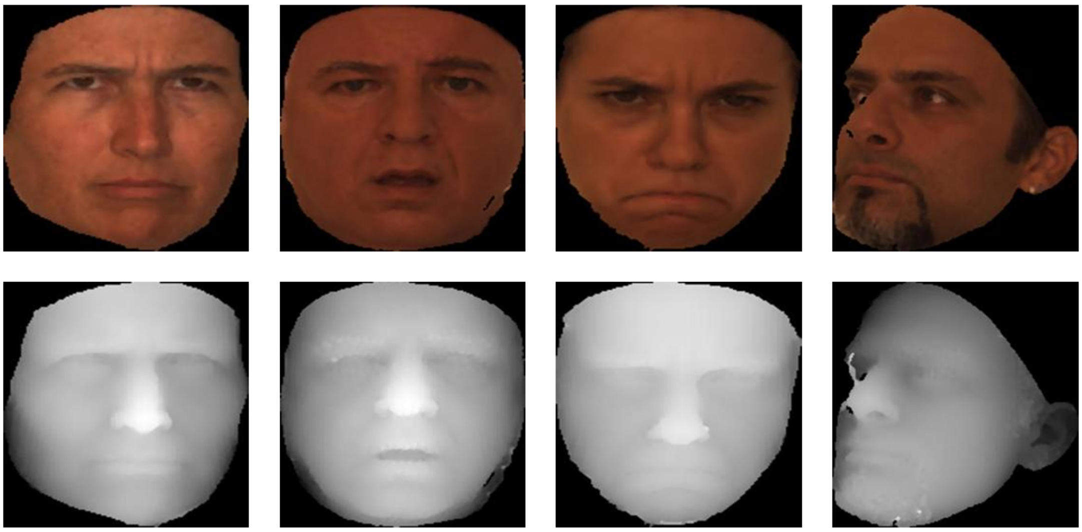


Figure 3.3: Image samples from Bosphorus 3D Face Database

### 3.1.1 Dataset

In our experiments, there are 9290 pairs of colored images and corresponding depth maps from Bosphorus 3D Face Database [57] and CASIA 3D Face Database [55] for training the GAN models. Binghamton University 3D Facial Expression (BU-3DFE) Database [56] is only for testing.

**Bosphorus 3D Face Database** Bosphorus 3D Face Database widely used for 3D face processing contains 105 subjects and 4666 faces in the database. One third of the subjects are professional actors or actresses. There are various expressions (up to 35), head poses (13 yaw and pitch rotations) and varieties of face occlusions for each subject. Facial data in the dataset is acquired by a 3D system based on the structured-light. The ground truth depth images and their corresponding color images are transformed from 3D point cloud files provided by the Bosphorus database. Some example RGB images and their corresponding depth maps from the dataset are illustrated in the Figure 3.3.

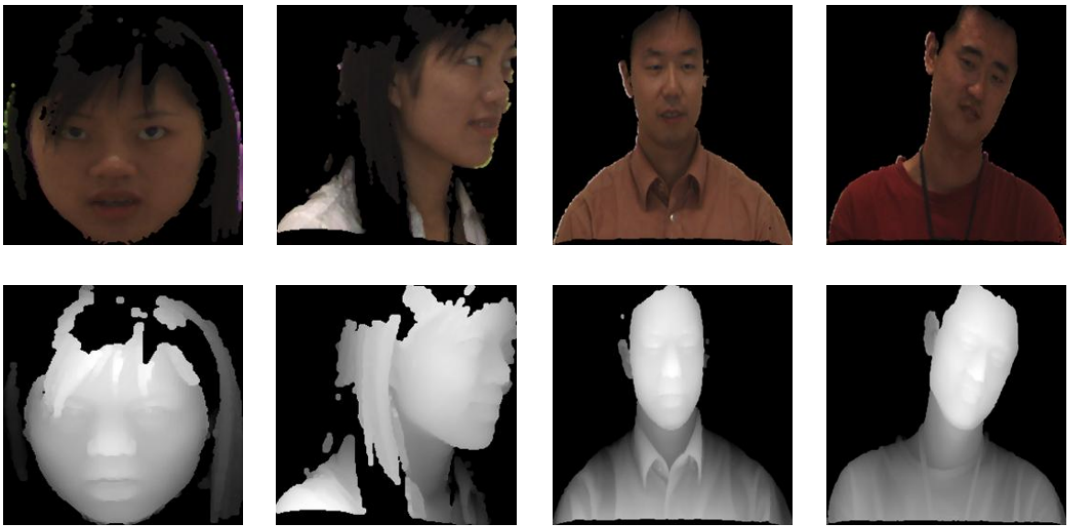


Figure 3.4: Image samples from CASIA 3D Face Database

**CASIA 3D Face Database** CASIA 3D Face Database collected by the Chinese Academy of Sciences contains 4624 scans of 123 persons. The scans are collected by the Minolta Vivid 910 which is a non-contact 3D digitizer. Each person in the database has 37 or 38 scans which include variations of poses, expressions and illuminations. Most of the persons in the database are Mongoloid. Some example RGB images and their corresponding depth maps from the dataset are illustrated in the Figure 3.4.

**Binghamton University 3D Facial Expression (BU-3DFE) Database** There are 100 subjects in the BU-3DFE Database of which 56 are male and 44 are female. The majority of subjects were undergraduates with various races. For each subject, there are 25 3D models with seven expressions which are happiness, disgust, fear, anger, surprise, sadness and neutral with different levels of intensity. Some example RGB images and their corresponding depth maps from the dataset are illustrated in the Figure 3.5.

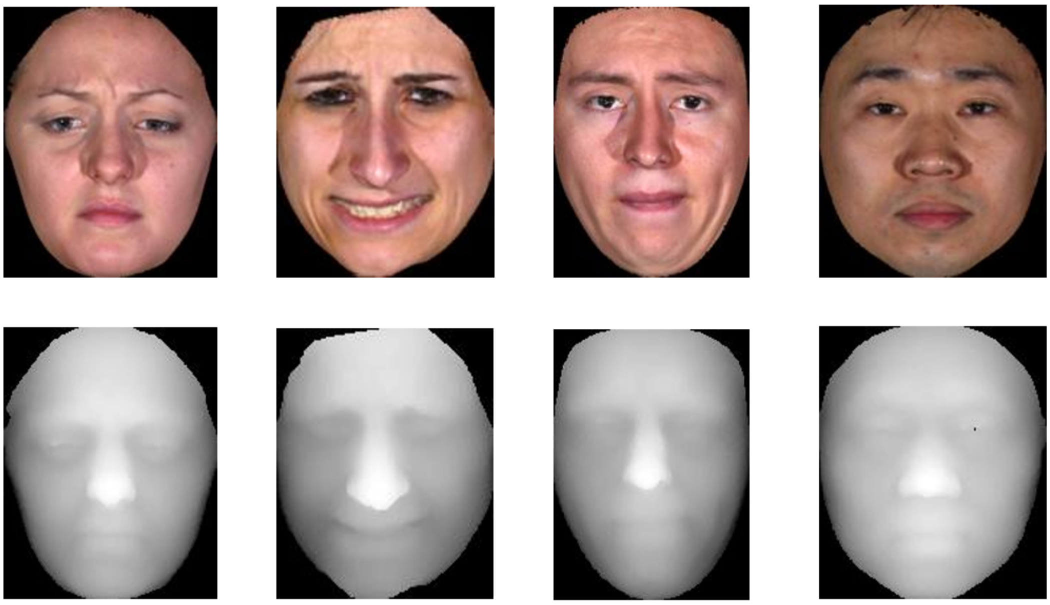


Figure 3.5: Image samples from BU-3DFE Database

### 3.1.2 Preprocessing

In practice, images always have different backgrounds which can affect the processing performance of the algorithm. Since training image pairs transformed from 3D data have black backgrounds. In this section the main purpose is to remove the image background out of the face uniformly. Firstly, the threshold is calculated by using Otsu's method [100]. Then, the image is transformed to a binary image by the threshold. Thus, 8-connected objects are labeled to locate the face based on the binary image. Next, background pixels are replaced with black pixels. Finally, an open operation which is an erosion followed by a dilation is performed to remove small objects and smooth the boundaries of larger objects of the image. The pseudo-code for removing the background is depicted as follows:

```
Func RemoveBg(Img):
Begin
```

```
    Thr = Otsu(Img)
    BinImg = Binarize(Img, Thr)
    LabImg = Label(BinImg,8-Conn)
    FaceImg = BlackBg(LabImg)
    OutImg = OpenOp(FaceImg)
    return OutImg
End
```

### 3.1.3 Depth Plus Generative Adversarial Network: D+GAN

In the task of generating face depth maps from corresponding RGB images, we propose a generative adversarial network named D+GAN for making full use of the attribute information of the human face. The generator ($G$) is composed of residual modules [5], self-attention modules [7] and convolution neural network, and its input is a $256 \times 256$ RGB image and its facial attribute labels which include the corresponding gender, age and race categories. The output is a depth map with the same size, which realizes the mapping of image to image. The discriminator ($D$) is used to identify the quality of the depth map. In our design, D+GAN not only outputs the score of the depth map, but also determines gender, age and race categories. Thus the input of the discriminator is a $256 \times 256$ depth map with its labels, and the output of the discriminator contains four scalar values which represent probabilities of true or false, age, gender and race. Figure 3.6 shows the structure of D+GAN.

#### 3.1.3.1 Generator

Specifically, the core architecture of the generator is U-shaped [101], which consists of an encoder and a decoder. The encoder is mainly used for feature extrac-

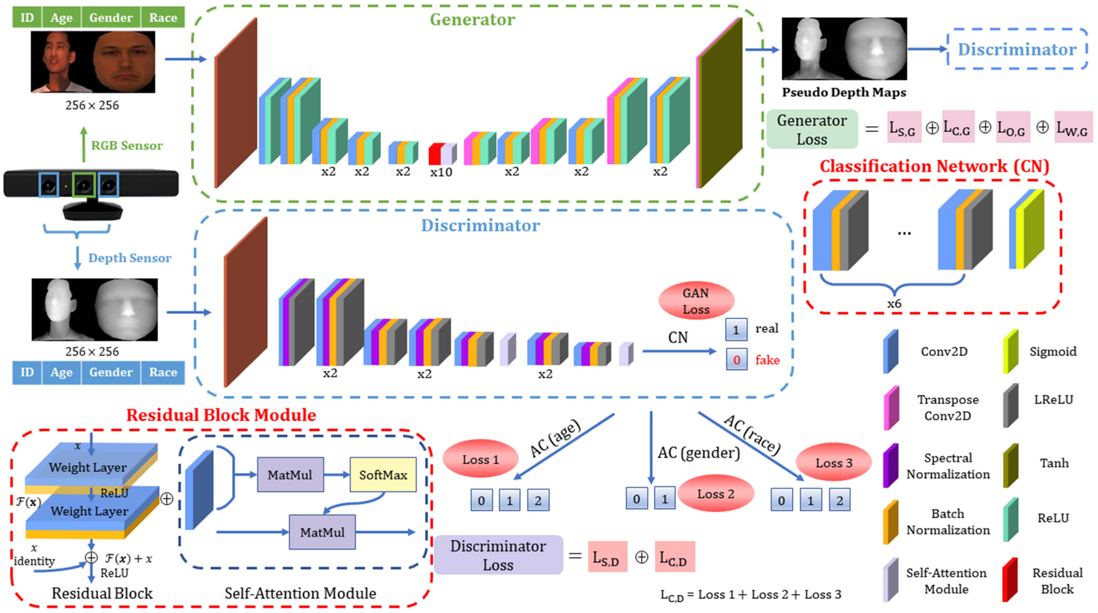


Figure 3.6: D+GAN: A GAN architecture for translating RGB images to depth maps with multiple face attributes

tion and feature compression of the image. It reduces the size of the input image and the number of feature parameters while increases the number of channels, which realizes the down-sampling process. The decoder with a symmetric and opposite structure to the encoder performs the encoding representation up-sampling successively and restores it to the same feature size as the encoder input.

The generator model also utilizes a skip connection in the convolutional layer between the encoder and decoder to build an information flow transmission approach, which can relieve the gradient disappearance problem effectively. The encoder is composed of 8 two-dimensional convolutional layers, as shown in Figure 3.6. The number of convolution kernels set is [64, 128, 256, 256, 512, 512, 512, 512] respectively, and the strides are set to [2, 1, 2, 1, 2, 1, 2, 1] sequentially. There are one Batch Normalization (BN) layer for normalizing input features to accelerate the convergence process and one layer with the *ReLU* activation function for introducing the sparsity of data to suppress the overfitting after each convolutional layer except for the first one.

The decoder is mainly composed of the convolutional neural network and deconvolutional neural network. In the decoder, the convolutional neural network is designed for feature extraction, and its calculation method is the same as that of the encoder, while deconvolutional neural network is designed for increasing the size of feature maps for up-sampling. In addition, the decoder intersperses two convolutional neural layers as shown in Figure 3.6. The number of convolution kernels set is [512, 512, 256, 256, 256, 128, 128, 128, 64, 3] respectively, and the strides are set to [2, 1, 1, 2, 1, 1, 2, 1, 1, 2] sequentially. Layer 1, 4, 7 and 10 are the deconvolutional layers. Similarly, BN layers and *ReLU* activation functions are added after each convolution layer except for the last one. Finally, the *tanh* activation function is used to normalize the output depth map at [-1, 1].

**Residual block** In order to fully extract features and increase model capacity, ten groups of residual block and self attention module combinations are used consecutively at the connection between the encoder and decoder of the generator. In our design, we use residual blocks to replace the original design of UNet. In the residual block $H(x)$, the original mapping is changed into $F(x) + x$ from $F(x)$ by using skip connections, which makes the neural network to be easier optimized. The number of convolution kernels is 256, the kernel size is $3 \times 3$, and the stride is set to 1.

**Self-attention module** Self-attention mechanism can learn from distant blocks, so it is used in both generator and discriminator in our design. The self-attention module helps to learn multi-level and long range dependencies across image regions, which is complementary to the convolution layer. In the self-attention module, the input feature $x$ with $n$ channels is transformed into query ($Q = W_Q x$),

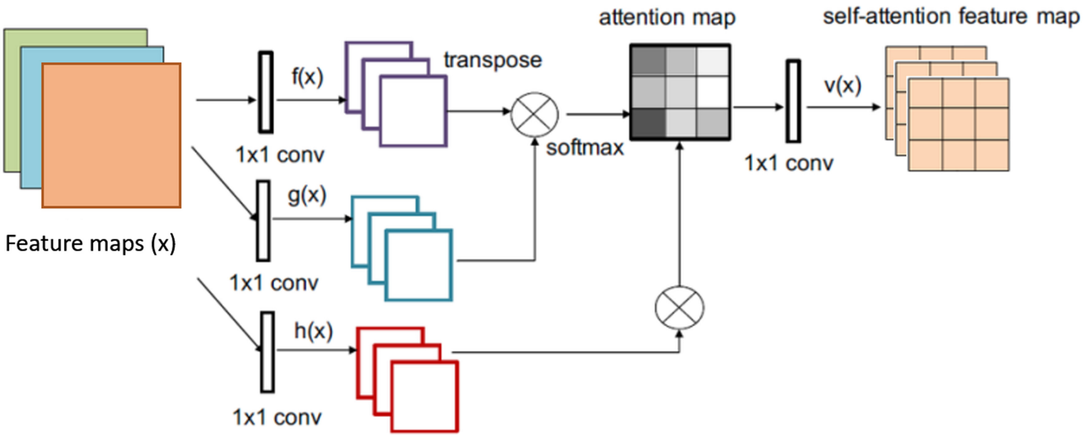


Figure 3.7: The illustration of self-attention module [7]

key ($K = W_K x$) and value ($V = W_V x$) by convolution operations. The size of $Q, K, V$ remains unchanged, but the number of channels becomes $n/8, n/8$ and $n$ respectively. Next, $Q$, $K$, and $V$ are serialized by channels so that feature map of $q_{m\times\frac{n}{8}}$, $k_{m\times\frac{n}{8}}$ and $v_{m\times n}$ are obtained respectively, where $m$ represents the feature size. The final output of attention weight distribution is computed as:

$$attention(q, k, v) = softmax(qk^T)v \tag{3.1}$$

#### 3.1.3.2 Discriminator

The discriminator of D+GAN consists of a backbone structure for distinguishing between true and false, and three branches for identifying face attributes of the image generated. In the backbone network, in order to provide more information exchange between channels and save computing resources, we insert a self-attention module after some higher convolutional layers as described above before the branch node. In detail, there are ten convolutional layers where the number of convolution kernels set is [64, 64, 64, 128, 128, 128, 256, 256, 256, 512] respectively and the strides are set to [2, 1, 1, 2, 1, 1, 2, 1, 1, 2] sequentially. The size of convolution kernels is $3 \times 3$, except the first layer is $5 \times 5$. In order to make the training process more stable, we set up spectral normalization [102] in these 10 convolutional layers to make the neural network robust to input disturbances.

**Spectral Normalization** In detail, for the weight $W_{m\times n}$ of the neural network, the spectral norm is the maximum singular value of the matrix. The maximum singular value $\sigma(W_{m\times n})$ is defined as:

$$\sigma(W_{m\times n}) = \max_{\delta} \frac{||W_{m\times n}\delta||_2}{||\delta||_2} \tag{3.2}$$

In practice, $\sigma(W_{m\times n})$ is approximately calculated by the power iteration, and then the weight $W_{m\times n}$ is updated to $W_{m\times n}/\sigma(W_{m\times n})$ in the forward direction during training, which is the process of spectral normalization.

The four branch networks get the output of the branch node as the input and perform different classification tasks. The first branch network is used to judge whether the depth map is true or false, which is essentially a binary classification task. Similarly, the second, third and fourth branch networks are used to classify age, gender and race respectively. In detail, the age label is divided into three categories which are 19-39 years old, 40-60 years old, and above 60 years old. The gender label is divided into two categories which are male and female. The race label is divided into three categories which are Caucasoid, Mongoloid and Negroid. These four branch networks have the same network structure except for the last layer, which are composed of seven two-dimensional convolutional layers, and their kernel size is $3 \times 3$. The number of convolution kernels in the first six layers is 512 with a stride of 1, and the number of kernels in the last layer is 2 or 3 with a stride of 2.

#### 3.1.3.3 Loss Function

The loss of the discriminator $L_D$ consists of two parts. The first part $L_{S,D}$, adopted from standard GAN, is used to distinguish between training samples and generated samples, which is indicated as:

$$\begin{aligned} L_{S,D} = \mathbb{E}_{Y \in P_{dat}(Y), X \in P_{dat}(X)} \left[ logD_1 \left( X, Y \right) \right] \\ + \mathbb{E}_{X \in P_{dat}(X)} \left[ log(1 - D_1(G(X), X))) \right] \end{aligned} \tag{3.3}$$

where $X$ represents the RGB face image to be translated, $Y$ represents the condition image corresponding to the real depth image, and $P_{dat}$ represents the probability distribution of the corresponding dataset. $D_1$ represents the output of the first discriminator. For the condition real image $Y$ and the generated image $G(X)$, the classifiers in the discriminator should be able to predict the classes it belongs to.

The second part $L_{C,D}$, classification loss, is the cross entropy loss of age, gender and race classification, which is indicated as:

$$\begin{aligned} L_{C,D} = \sum_{i=2}^{4} \mathbb{E}_{X \in P_{dat}(X)} [\log P(D_i = c | G(X))] \\ + \mathbb{E}_{Y \in P_{dat}(Y)} [\log P(D_i = c | Y)] \end{aligned} \tag{3.4}$$

where $D_i$ represents the $i$th discriminator, and $C_i$ represents the corresponding label. Totally, the training loss of the discriminator, $L_D$, can be expressed as:

$$L_D = \lambda_1 L_{S,D} + \lambda_2 L_{C,D} \tag{3.5}$$

For the generator, its loss function $L_G$ contains four parts. First, it is expected that the generated samples can deceive the discriminator, thus $L_{S,G}$ is defined as:

$$L_{S,G} = -\mathbb{E}_{X \in P_{dat}(X)}[\log D_1(G(X), X)] \tag{3.6}$$

In order to ensure the similarity of input and output images of the generator, L2-loss is introduced as:

$$L_{O,G} = -E_{Y \in P_{dat}(Y), X \in P_{dat}(X)} \left[ \|Y - G(X)\|_2 \right] \tag{3.7}$$

Next, the generator is expected to generate high-quality samples so that they can be correctly classified by the discriminator. Similarly, the classification loss $L_{C,G}$ is defined as:

$$L_{C,G} = \sum_{i=2}^{4} \mathbb{E}_{X \in P_{dat}(X)}[\log P(D_i = c | G(X))] \tag{3.8}$$

In addition, in order to avoid the over-fitting, the weight regularization term $L_{W,G}$ is introduced. It is expressed as:

$$L_{W,G} = \frac{1}{2}||W||^2 \tag{3.9}$$

Totally, the training loss of generator, $L_G$, can be expressed as:

$$L_G = \lambda_1 L_{S,G} + \lambda_2 L_{C,G} + \lambda_3 L_{O,G} + \lambda_4 L_{W,G} \tag{3.10}$$

The D+GAN is implemented with Python and TensorFlow. Python is a widely used, high-level, and general-purpose programming language [103]. TensorFlow is an open source software library for machine learning and artificial intelligence [104].

### 3.1.4 Face Depth Prediction by the Scene Depth

In general, 3D scene understanding dataset can be divided into outdoor scene dataset and indoor scene dataset. The representative of outdoor scene and indoor scene datasets are KITTI [105] and NYU Depth V2 [106] respectively. In this section, we have introduced deep neural network models trained with supervised or unsupervised learning using these two 3D scene datasets. The performance of these models will be compared in the experiments in the next chapter.

#### 3.1.4.1 Supervised Learning

Generally, it is required to know in advance the depth values corresponding to a large number of input pictures as training constraints, so as to back-propagate the deep neural network and train neural network for depth prediction of similar scenes. The loss function of the depth regression problem is considering the difference between the true value of the depth map and the predicted value of the depth regression network. In Densedepth [65], the loss function can be indicated as:

$$\begin{aligned} L\left(y,\tilde{y}\right) =& \frac{c}{n} * \sum_{i}^{n} \left|y_i - \tilde{y_i}\right| + \frac{1}{n} \sum_{i}^{n} \left|g_x\left(y_i,\tilde{y_i}\right) + g_y\left(y_i,\tilde{y_i}\right)\right| \\ &+ \frac{1 - SSIM\left(y,\tilde{y}\right)}{2} \end{aligned} \tag{3.11}$$

where $y$ indicates the ground truth depth map, and $\tilde{y_i}$ indicates the generated depth map. $c$ is a constant, empirically set to 0.1. $g_x$ and $g_y$ are functions of computing the differences in components $x$ and $y$ for the depth maps gradients. Structural Similarity Index (SSIM) [107] is a metric to measure the similarity between $y$ and $\tilde{y_i}$.

In this strategy, many well-known multi-layer pre-trained networks with different structures can fully utilize the advantages of deep neural networks as a function simulator.

#### 3.1.4.2 Self-supervised Learning

**Stereo training modality** In stereo vision, it is supposed that there are two cameras $L$ and $R$, and one point whose coordinates are $(x, D)$. The disparity represents the translation value required for the pixels in the left camera to form the corresponding pixels in the right camera. According to the triangle similarity law, the disparity denoted as $dis$ can be calculated as:

$$dis = x_L - x_R = \frac{f * b}{D} \tag{3.12}$$

where $f$ is the focal length of the camera, and $b$ is the distance between two cameras. Therefore, a mapping function $F$ for predicting the disparity is expected as:

$$I_L\left(dis + x_L\right) = I_L\left(F\left(x_L\right) + x_L\right) = I_R\left(x_R\right) \tag{3.13}$$

Thus, $I_L$ is used for the input, and $I_R$ is used for the reference, the model for predicting disparity can be achieved. Finally, the depth map can be obtained with disparity and camera parameters $b$ and $f$. When in the training process, the problem is formulated as the minimization of a photometric reprojection error:

$$L_p = \min \sum_{\tau} E(I_t, I_{\langle\tau\rangle}) \tag{3.14}$$

$$L_p = \sum_{\tau} \alpha \frac{1 - SSIM(I_t, I_{\langle\tau\rangle})}{2} + (1 - \alpha) \left\| I_t - I_{\langle\tau\rangle} \right\| \tag{3.15}$$

where $I_t$ represents the target image, $I_\tau$ represents the source image and $I_{\langle\tau\rangle}$ represents the sampled source image. In Monodepth2 [67], the value of $\alpha$ is fixed as 0.85 empirically, and the final loss combining per-pixel smoothness and masked photometric losses is as:

$$L = c_1 L_p + c_2 L_s \tag{3.16}$$

where

$$L_s = \left|\partial x \frac{d_t}{\overline{d_t}}\right| e^{-|\partial x I_t|} + \left|\partial y \frac{d_t}{\overline{d_t}}\right| e^{-|\partial y I_t|} \tag{3.17}$$

In the equation above, $\overline{dt}$ represents the mean depth.

**Mono training modality** Our source image $I_\tau$ could be the second view of $I_t$ in stereo training while $I_\tau$ are the temporally adjacent frames of $I_t$ in mono training, that is, $I_\tau \in \{I_{t-1}, I_{t+1}\}$. Additionally, $I_\tau$ includes both the second view and temporally adjacent frames of $I_t$ in the mix training modality.

## 3.2 Facial Diagnosis

In the facial diagnosis section, Disease-specific Faces (DSF) dataset [108] and Disease-specific Faces 2 (DSF2) dataset [109] were built in the context of thesis for training and testing models.

### 3.2.1 Dataset

**Disease-Specific Faces (DSF) Dataset** The DSF dataset [108] used includes condition-specific face images which are collected from professional medical publications, medical forums, medical websites and hospitals with definite diagnostic results. In the DSF dataset used in Section 5.1, there are totally 350 face images (JPG files) in the dataset, and there are 70 images in 4 types of condition-specific faces described in Section 1.2. The four condition-specific faces are shown in Figure 3.8. Generally the ratio of training data and testing data is from 2:1 to 4:1. In our experiments with the small dataset, the ratio is set as 4:3 for the efficient

Table 3.1: **The statistics of the races in the dataset.**

| Condition-specific face | Number of face images | | |
|---|---|---|---|
| | Caucasoid | Mongoloid | Negroid |
| Beta-thalassemia | 9 | 54 | 7 |
| Hyperthyroidism | 36 | 28 | 6 |
| Down syndrome | 48 | 16 | 6 |
| Leprosy | 12 | 37 | 21 |
| Control | 21 | 40 | 9 |
| **Total** | 126 | 175 | 49 |

evaluation. Table 3.1 shows the statistics of the races distinguished by eyes of face images in the experiments.

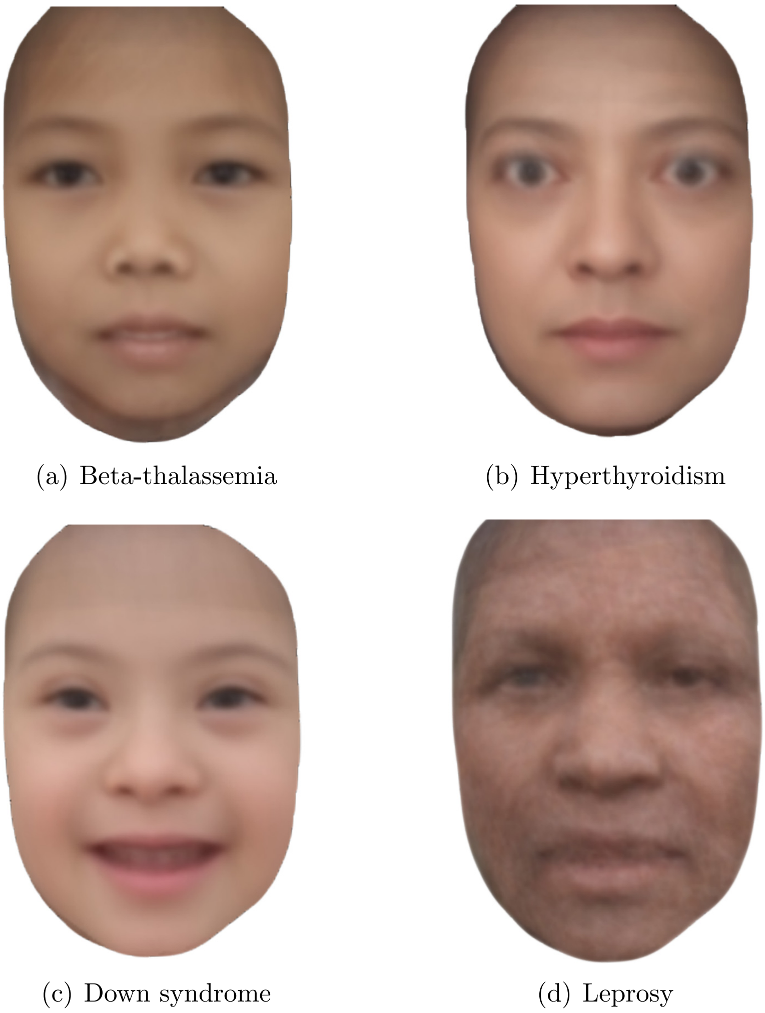


Figure 3.8: Disease-specific faces

**Disease-Specific Faces 2 (DSF2) dataset** The DSF2 dataset [109] includes 6 condition-specific faces and health controls. Six conditions are acromegaly, facial nerve paralysis, Down syndrome, leprosy, thalassemia and hyperthyroidism, which is aforementioned in Section 1.2. The DSF2 dataset utilized consists of condition-specific face images with diagnostic results from professional medical publications, medical websites, forums, and hospitals. These results were reviewed by practicing physicians to create labels for supervised learning of the

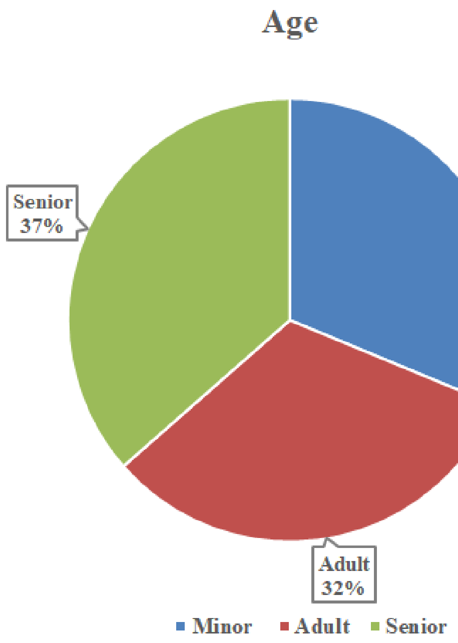


Figure 3.9: Approximate age distribution

model. And there is no evidence that the patients in the photos had two or more of the six conditions at the time the photos were taken.

To protect patient privacy, it is essential to de-identify condition-specific face image data. This entails removing all information that could potentially be used to identify an individual, such as names, birthdates, and medical record numbers. Furthermore, in order to protect patient privacy, we do not permit the direct publication of any images from the dataset in any media or publications.

The number of faces of each class is 85. There are a total of 595 images in the dataset. The proportions of age, gender, and ethnicity within the dataset are approximately represented in Figures 3.9, 3.10, and 3.11.

Acromegaly-specific face and facial nerve paralysis-specific face in the DSF2 dataset are shown in Figure 3.12 and Figure 3.13 respectively.

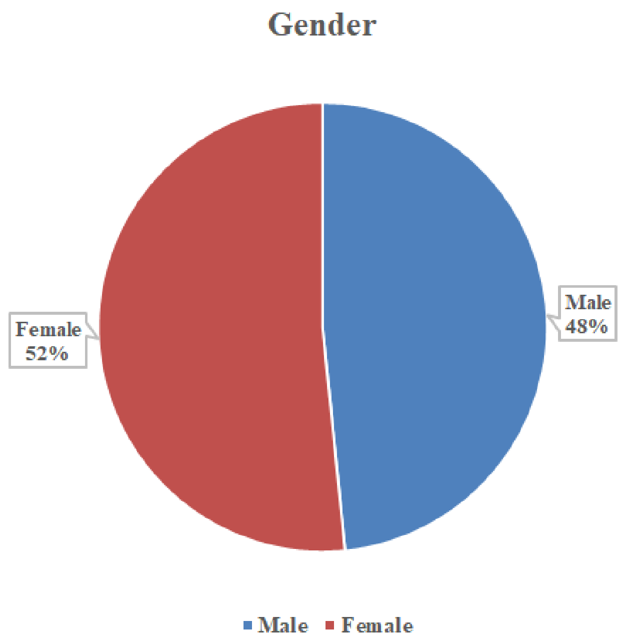


Figure 3.10: Approximate gender distribution

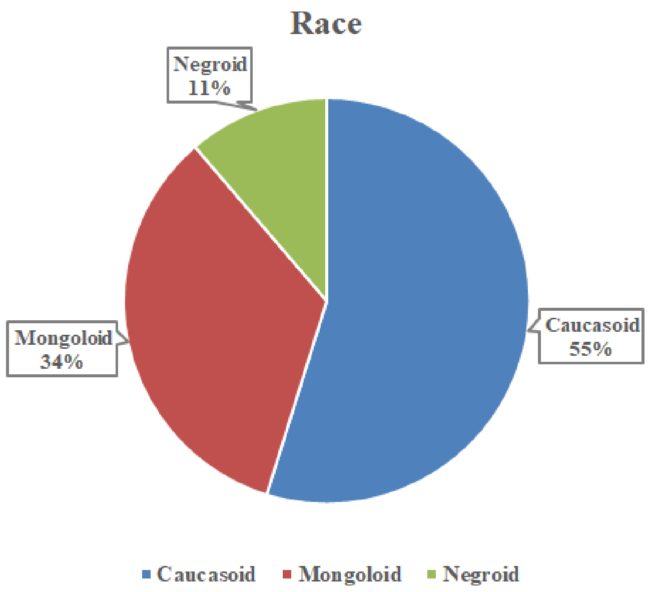


Figure 3.11: Approximate race distribution

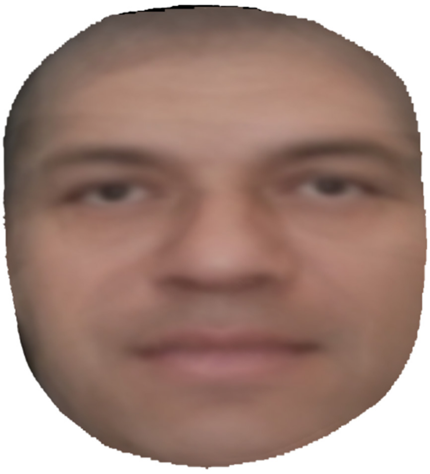


Figure 3.12: Acromegaly-specific face

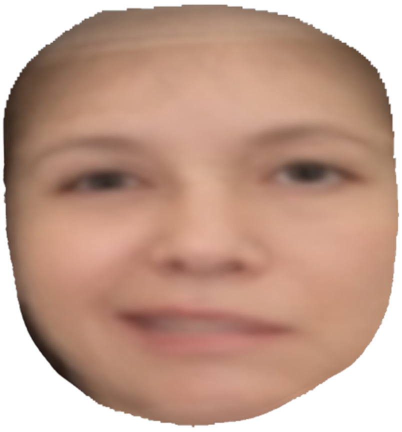


Figure 3.13: Facial nerve paralysis-specific face

### 3.2.2 Deep Transfer Learning

Training a CNN which is end to end learning from scratch will inevitably lead to over-fitting since that the training data is generally insufficient for the task of facial diagnosis. Transfer learning is applying the knowledge gained while solving one problem to a different but related problem. In the transfer learning problem [110], generally we let $\mathcal{D}_s$ indicate the source domain, $\mathcal{D}_t$ indicate the target domain and $\mathcal{X}$ be the feature space domain. $\mathcal{H}$ is assumed to be a hypothesis class on $\mathcal{X}$, and $I(h)$ is the set for characteristic function $h \in \mathcal{H}$. The definition of $\mathcal{H}$-divergence between $D_s$ and $D_t$ which is used to estimate divergence of unlabeled data is:

$$d_{\mathcal{H}}(\mathcal{D}_s, \mathcal{D}_t) = 2 \sup_{h \in \mathcal{H}} \left| \Pr_{x \in \mathcal{D}_s} [I(h)] - \Pr_{x \in \mathcal{D}_t} [I(h)] \right| \tag{3.18}$$

where $Pr$ indicates the probability distribution. sup means computing the supremum, which is the least upper bound of the set. Furthermore, the relationship between errors of target domain and source domain can be calculated as:

$$e_t\left(h\right) \leq e_s\left(h\right) + \frac{1}{2}\hat{d}_{\mathcal{H}\Delta\mathcal{H}}\left(u_s, u_t\right) + 4\sqrt{2d\log\left(2m'\right) + \frac{\log\frac{2}{\delta}}{m'}} + \lambda \tag{3.19}$$

where $u_s$ and $u_t$ are unlabeled samples from $\mathcal{D}_s$ and $\mathcal{D}_t$ respectively. Briefly, the difference in the error between source domain and task domain is bounded as:

$$|e_t - e_s| \approx \frac{1}{2}d_{\mathcal{H}\Delta\mathcal{H}}\left(\mathcal{D}_s, \mathcal{D}_t\right) \tag{3.20}$$

where $d_{\mathcal{H}\Delta\mathcal{H}}$ indicates the distance of symmetric difference hypothesis space $\mathcal{H}\Delta\mathcal{H}$. The equations above have proved that transfer learning from different domains is mathematically effective [111]. Deep transfer learning (DTL) [32], [112] is to transfer knowledge by pretrained deep neural network which originally aims to perform facial verification and recognition in this work. Thus the source task is face recognition and verification, and the target task is facial diagnosis. In this case, the feature spaces of the source domain and target domain are same while the source task and the target task are different but related. The similarity of two tasks motivates us to use deep transfer learning from face recognition to solve facial diagnosis problem with a small dataset. If divided according to transfer learning scenarios, it belongs to inductive transfer learning. If divided according to transfer learning methods, it belongs to parameter based transfer learning. In this section, two main deep transfer learning strategies [113], [114] are applied to perform comparison. In the main experiment, DCNN models pretrained by VGG-Face dataset [27] and ImageNet dataset [115] are compared with traditional machine learning methods. VGG-Face dataset contains 2.6M images over 2.6K people for face recognition and verification, and ImageNet dataset contains more than 14M images of 20K categories for visual object recognition.

The pretrained CNN is for end-to-end learning so that it can extract high-level features automatically. Since deep transfer learning is based on the fact that CNN features are more generic in early layers and more original dataset-specific in later layers, operation should be performed on the last layers of DCNN models. The diagram of facial diagnosis by deep transfer learning is shown in Figure 3.14. The implementation is based on Matlab (version: 2017b) with its CNNs toolbox for computer vision applications named MatConvNet (version: 1.0-beta25). NVIDIA CUDA toolkit (version: 9.0.176) and its library CuDNN (version: 7.4.1) are applied for GPU (model: Nvidia GeForce GTX 1060) accelerating.

#### 3.2.2.1 DTL1: Fine-tuning the Pretrained CNN Model

In this section, we replace the final fully connected layer of the pretrained CNN by initializing the weight. When fine-tuning the CNN (see PSEUDOCODE 1),

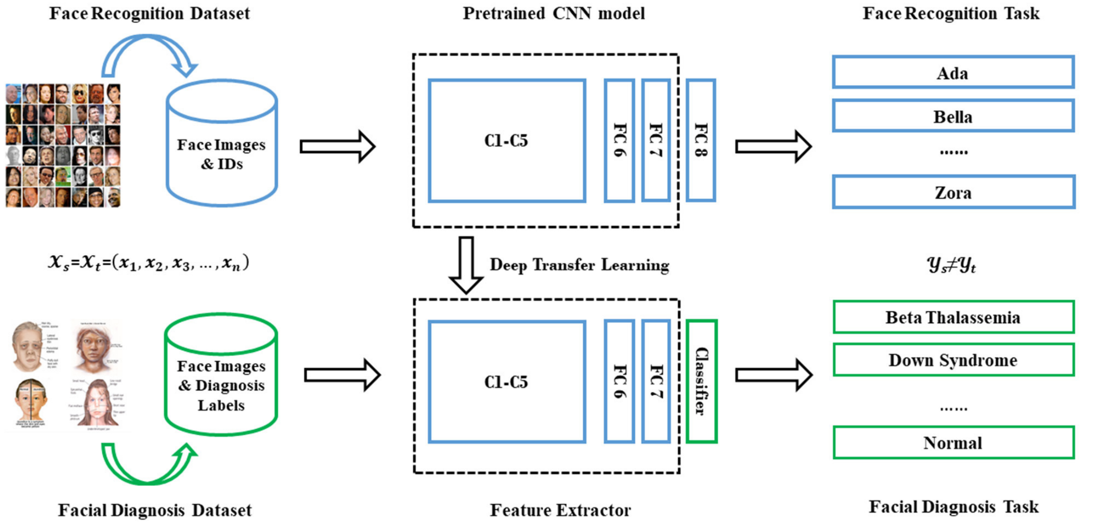


Figure 3.14: The schematic diagram of facial diagnosis by deep transfer learning

we calculate activation value through forward propagation of the convolutional layer as:

$$c_{u,v}^{l} = \sum_{i=-\infty}^{+\infty} \sum_{j=-\infty}^{+\infty} \sigma\left(i,j\right) \cdot a_{i+u,j+v}^{l-1} k_{r_{i,j}}^{\ l} + b^{l} \tag{3.21}$$

where $a$ indicates input feature map of some layer, and $k$ indicates its corresponding kernel. $\sigma$ is defined as:

$$\sigma\left(i,j\right) = \begin{cases} 1 & \text{if } 0 \leqslant i,j \leqslant 1 \\ 0 & \text{if } others \end{cases} \tag{3.22}$$

Therefore, the output value of convolution operation is calculated as $f(c_{u,v}^{l})$ in which $f$ is the activation function. When updating the weights, we calculate error term through back propagation of the convolutional layer as:

$$\begin{aligned} E_{g,h}^{l} &= \frac{\partial J(W,b;x,y)}{\partial c_{g,h}^{l}} \\ &= \sum_{i=0}^{r-1}\sum_{j=0}^{r-1}\frac{\partial J(W,b;x,y)}{\partial c_{i,j}^{(l+1)}}\cdot \\ &\frac{\partial \beta^{(l+1)}\sum_{u=ir}^{(i+1)r-1}\sum_{u=jr}^{(j+1)r-1} f\left(c_{u,v}^{l}\right)+b^{(l+1)}}{\partial c_{g,h}^{l}} \\ &= \beta^{(l+1)}E_{i+pr,j+qr}^{(l+1)} f'\left(c_{g,h}^{l}\right) \end{aligned} \tag{3.23}$$

where $f$, same with above, represents the activation function, $J$ represents the cost function, $(W, b)$ are the parameters and $(x, y)$ are the training data and label pairs. Since the pretrained model has already converged on the original training data, a small learning rate of $5 \times 10^{-5}$ is utilized. Weight Decay for avoiding overfitting to a certain extent is set as $5 \times 10^{-4}$, and momentum for accelerating convergence in mini-batch gradient descent (SGD) is set as 0.9. Here we take VGG-16 model also named VGG-Face as an example, which is the best case in the main experiment. A softmax loss layer is added for retraining by 100 epochs initially. Figure 3.15 containing three indicators Objective, Top-1 error and Top-3 error shows the process of fine-tuning the pretrained VGG-Face for the multiclass classification task. Objective is the sum loss of all samples in a batch. The loss can be calculated as:

$$L = -\sum_{i} y_i \ln p_i = -\sum_{i} y_i \ln \frac{e^{z_i}}{\sum_k e^{z_k}} \tag{3.24}$$

where $y_i$ refers to the $i$ th true classification result, $p_i$ represents the $i$ th output of the softmax function, and $z_i$ represents the $i$ th output of the convolutional neural network. The Top-1 error refers to the percentage of the time that the classifier did not correctly predict the class with the highest score. The Top-

3 error refers to the percentage of the time that the classifier did not include the correct class among its top 3 guesses. As it can be seen from Figure 3.15, all three indicators converge after retraining for about 11 epochs, which indicates fine-tuning is successful and effective. However, the validation error is higher than the training error, which is because of the limitation of the fine-tuning strategy on the small dataset. As shown in Figure 3.15, after 24 epochs the validation top-1 error rises while the training error doesn't, which indicates over-fitting may occur. So we saved the fine-tuned CNN model after retraining 24 epochs for testing. The early stopping technique is used here. The softmax layer is used for classification, which is consistent with the pretrained model.

Time complexity is the number of calculations of one model / algorithm, which can be measured with floating point operations (FLOPs). In our estimations, the Multiply-Accumulate Operation (MAC) is used as the unit of FLOPs. In CNNs, time complexity of a single convolutional layer can be estimated as:

$$O\left(M^2 \cdot K^2 \cdot C_{in} \cdot C_{out}\right) \tag{3.25}$$

where $M$ is the side length of the feature map output by each kernel, $K$ is the side length of each kernel, and $C$ represents the number of corresponding channels [116]. Thus, the overall time complexity of convolutional neural networks can be estimated as:

$$O\left(\sum_{l=1}^{D} M_l^2 \cdot K_l^2 \cdot C_{l-1} \cdot C_l\right) \tag{3.26}$$

The FLOPs of the fully connected layers can be estimated by $I \cdot O$ where $I$ indicates input neuron numbers and $O$ indicates output neuron numbers. $I$ corresponds to $C_{l-1}$ and $O$ corresponds to $C_l$ in the above formula. Because pretrained models for object and face recognition have a larger number of categories, the time complexity of adapted models by DTL1 in our task is smaller than the

PSEUDOCODE 1

```
DTL 1 Fine-tuning the pretrained CNN model
Input: training dataset $D = \{(\mathbf{x}_k, \mathbf{y}_k)\}_{k=1}^{m}$,
       pretrained CNN model,
       number of epochs $n$,
       small learning rate $\eta$
Output: adapted CNN model
  function $DTL1(D,\ CNN,\ n,\ \eta)$
    parameters ← load_model(CNN, pretrained=true)
    parameters ← initialize(parameters, layer=FC)
    repeat
      for all$(\mathbf{x}_k, \mathbf{y}_k) \in D$ do
        activation ← forward_propagation($x_k$, parameters)
        cost ← loss_function(activation, $y_k$)
        gradient ← back_propagation(activation, cost)
        parameters ← weight_update(parameters, gradient, $\eta$)
      end for
    until $n$ times
  end function
```

original corresponding pretrained model.

#### 3.2.2.2 DTL2: CNN as Fixed Feature Extractor

In this section, the CNN is used as a feature extractor directly for the smaller dataset (see PSEUDOCODE 2). During training process for facial diagnosis, we only want to utilize the partial weighted layers of the pretrained CNN model to extract features, but not to update the weights of it. As the architect Ludwig Mies van der Rohe said, ”Less is more”. We select the linear kernel for the SVM [117] model to do classification in this strategy, because the dimension of the input feature vectors is much larger than the number of samples. For the reason that CNN features are more original dataset specific in the last layers, we directly extract features of the layer which is located before the final fully connected layer of pretrained DCNN models, and then train a linear SVM classifier leveraging

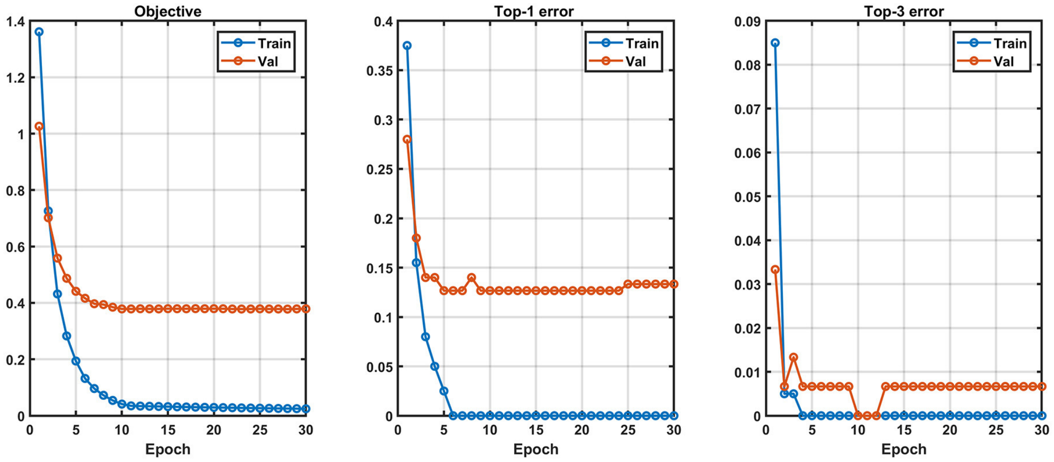


Figure 3.15: The process of fine-tuning the pretrained VGG-Face model

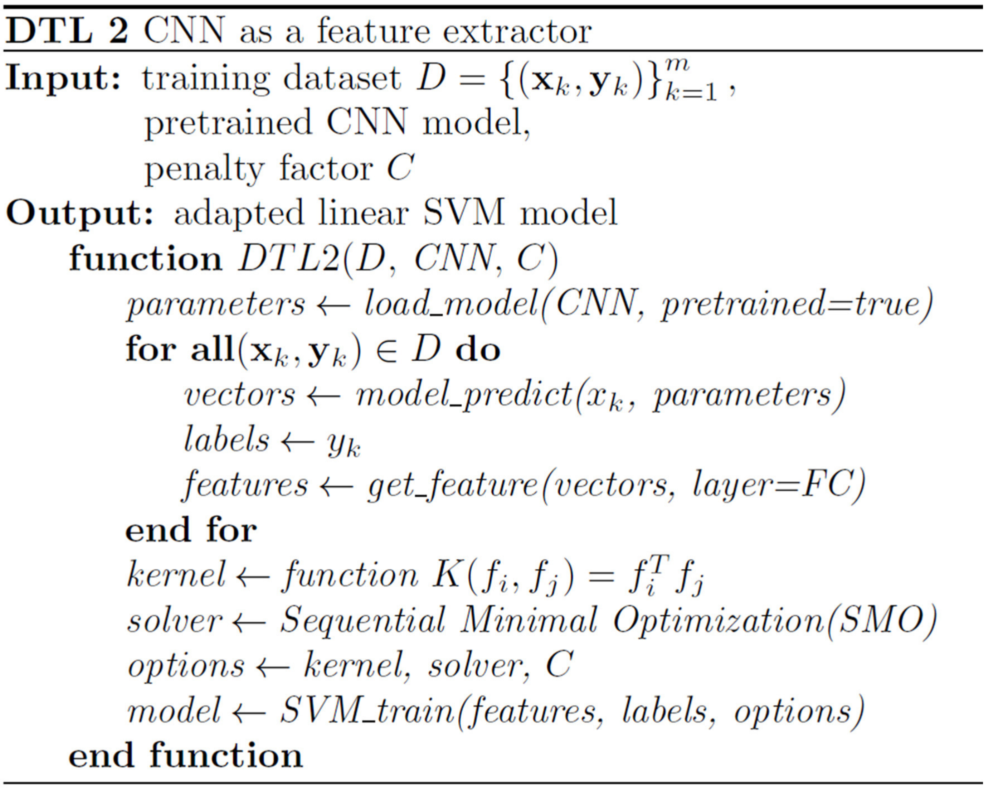
the features extracted as:

$$\min_{\mathbf{w}} \left\{ C \sum_{i} max \left(1 - y_i \mathbf{w}^T x_i, 0\right) + \frac{1}{2} \|\mathbf{w}\|^2 \right\} \tag{3.27}$$

where $C$ which is a hyper-parameter indicates a penalty factor, and $(x_i, y_i)$ represents the training data. After the training process, we could obtain the linear SVM model trained to perform testing.

During the training phase, the time complexity of SVM is different in different situations, namely whether most support vectors are at the upper bound or not, and depending on the ratio of the number of vectors and the number of training points. During the testing phase, the time complexity of SVM is $O(M \cdot N_s)$ where $M$ is the number of operations required by the corresponding kernel, and $N_s$ is the number of support vectors. For a linear SVM classifier, the algorithm complexity is $O(d_l \cdot N_s)$ where $d_l$ is the dimension of input vectors [118]. In our tasks, $N_s$ is larger than the number of output neurons of CNN final fully connected layers in DTL1, while generally smaller than it in the original corresponding pretrained models.

PSEUDOCODE 2



```
DTL 2 CNN as a feature extractor
Input: training dataset $D = \{(\mathbf{x}_k, \mathbf{y}_k)\}_{k=1}^{m}$,
       pretrained CNN model,
       penalty factor $C$
Output: adapted linear SVM model
    function $DTL2(D,\ CNN,\ C)$
        $parameters \leftarrow load\_model(CNN,\ pretrained{=}true)$
        for all$(\mathbf{x}_k, \mathbf{y}_k) \in D$ do
            $vectors \leftarrow model\_predict(x_k,\ parameters)$
            $labels \leftarrow y_k$
            $features \leftarrow get\_feature(vectors,\ layer{=}FC)$
        end for
        $kernel \leftarrow function\ K(f_i, f_j) = f_i^T f_j$
        $solver \leftarrow Sequential\ Minimal\ Optimization(SMO)$
        $options \leftarrow kernel,\ solver,\ C$
        $model \leftarrow SVM\_train(features,\ labels,\ options)$
    end function
```

# Chapter 4

# Face Recognition Experiment Results and Analysis

## 4.1 Depth Plus Generative Adversarial Network: D+GAN

In this section, we not only evaluate on the face depth map generated itself, but also validate it for the face recognition task in various datasets.

### 4.1.1 Qualitative Results and Analysis

To perform the qualitative evaluation, we calculate some indicators on the three 3D face datasets described in Section 3.2 to evaluate the quality of the obtained depth map. In this section, we present outputs of face depth maps generated by several state-of-the-art techniques for some examples. There are Monodepth2, DenseDepth (KITTI), DenseDepth (NYU-Depth V2), 3DMM, Pix2Pix, CycleGAN and D+GAN for comparison.

In this study, Monodepth2 [67] is trained on the KITTI dataset with the mono training modality. DenseDepth (KITTI) [65] is trained successively on the ImageNet and KITTI datasets, and DenseDepth (NYU-Depth V2) is trained successively on the ImageNet and NYU-Depth V2 datasets. 3D Morphable Model (3DMM) [119] is to generate a textured 3D face with parameters including vertices, triangles and attribute based on Basel Face Model (BFM). With these parameters, we render this 3D face into the depth map via a rasterization renderer. GAN models including Pix2Pix, CycleGAN and D+GAN are all trained on the Bosphorus 3D Face Database and CASIA 3D Face Database for 20 epochs, and their training curves all converge before or around 16 epochs. Adam optimizer is used for Pix2Pix and CycleGAN, while Adadelta optimizer is used for D+GAN.

The IDs of the example cases are bs016_LFAU_22_0 of Bosphorus 3D Face Database, 008-025 of CASIA 3D Face Database and F0010_FE03WH_F2D of BU-3DFE Database respectively. The ground truth depth image and its corresponding color image are transformed from 3D data provided.

#### 4.1.1.1 Case Study: bs016_LFAU_22_0 of Bosphorus 3D Face Database

Figure 4.1 presents the results for the case of bs016_LFAU_22_0 of Bosphorus 3D Face Database. Figure 4.1a shows the RGB face image which is transformed from 3D data provided, and Figure 4.1b shows the ground truth face depth map which is transformed from 3D data provided. Figure 4.1c shows the output generated by Monodepth2. The result shows vaguely the contour of the face, and the relative depth information is not accurately expressed. Figure 4.1d shows the output generated by DenseDepth (KITTI). The result can only show the outline of the face, and cannot show the depth of facial details. Figure 4.1e shows the output generated by DenseDepth (NYU-Depth V2). The result shows the depth better, but still lacks the facial detailed depth. Figure 4.1f shows the output generated by 3DMM. The result shows more face detailed depth information, however the contour of eyes, nose, mouth and the face shape showed are visually very different to the ground truth. We infer that this is because 3DMM is based on an average model. Visually, Figure 4.1g and Figure 4.1h show basically satisfactory results which are generated by Pix2Pix and CycleGAN. Figure 4.1i shows the best result, in visual inspection, which is the output generated by D+GAN. The depth values especially in eyes, nose and mouth shown by D+GAN are more precise than Pix2Pix and CycleGAN.

The autocorrelation function is usually used as the texture measure in the image. The texture coarseness of the image is proportional to the expansion of the autocorrelation function. We assume that one image is denoted as $I(x, y)$. Autocorrelation function is defined as:

$$C\left(\xi, \eta, a, b\right) = \frac{\sum_{x=a-w}^{a+w} \sum_{y=b-w}^{b+w} I(x, y) I(x-\xi, y-\eta)}{\sum_{x=a-w}^{a+w} \sum_{y=b-w}^{b+w} [I(x, y)]^2} \tag{4.1}$$

where (a, b) is the pixel in the window which size is $(2w+1) * (2w+1)$. $\xi, \eta = \pm 0, \pm 1, \pm 2 ... \pm N$. $\xi$ and $\eta$ are shifting variables on the pixels.

In the case of bs016_LFAU_22_0 of Bosphorus 3D Face Database, autocorrelation function graphs on depth maps generated by various models are shown as Figure 4.2. In the autocorrelation function graph, a larger downward trend observed as $\xi$ and $\eta$ increase indicates a greater coarseness of the corresponding image. Figure 4.2b shows the autocorrelation function graph of the ground truth depth map. Comparing with Figure 4.2a, Figure 4.2b has a smaller downward trend as $\xi$ and $\eta$ increase, which means the depth map has a lower coarseness than its corresponding grayscale image. Subjectively, the spatial details of the face should be changed regularly. Comparing with Figure 4.2b, Figure 4.2c, Figure 4.2d and Figure 4.2e has a larger downward trend as $\xi$ and $\eta$ increasing, which means the depth maps generated by Monodepth2, DenseDepth (KITTI)

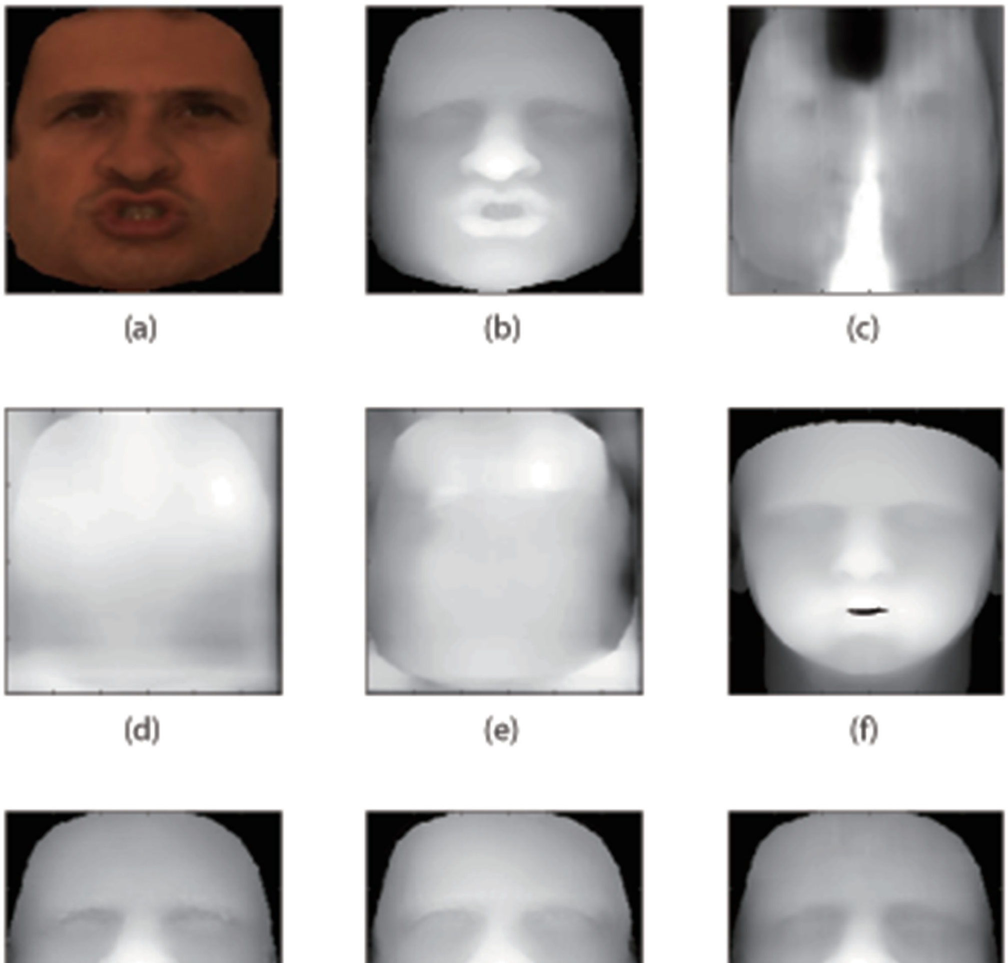


Figure 4.1: Face depth maps generated by various models in the case of bs016_LFAU_22_0. (a) Input RGB image, (b) Ground truth depth map, (c) Model: Monodepth2, (d) Model: DenseDepth (KITTI), (e) Model: DenseDepth (NYU-Depth V2), (f) Model: 3DMM, (g) Model: Pix2Pix, (h) Model: CycleGAN, (i) Proposed Model: D+GAN

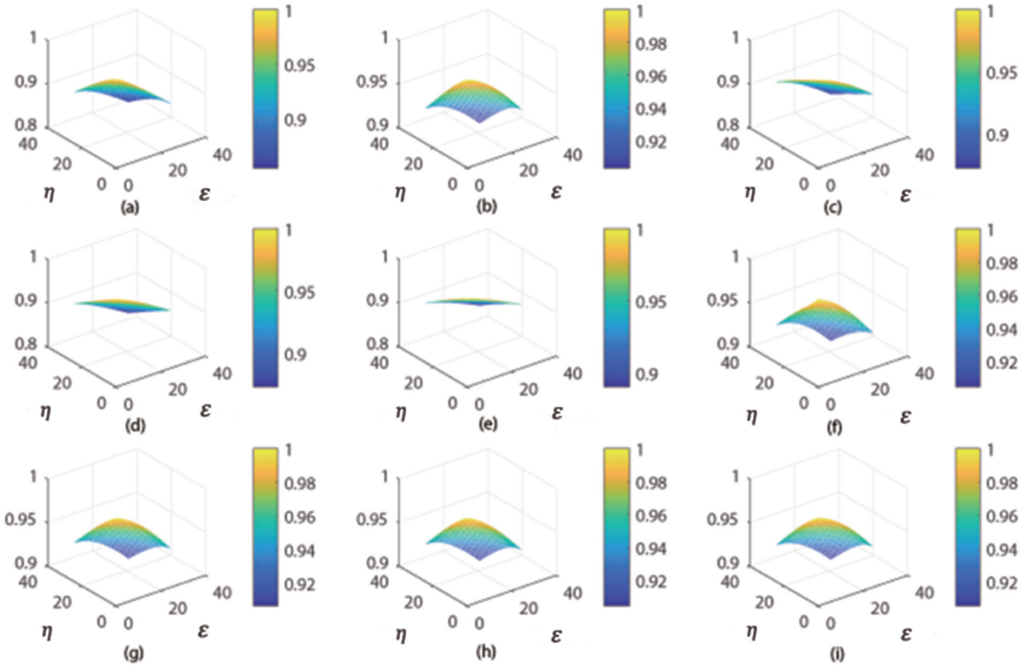


Figure 4.2: Autocorrelation function graphs of various output images: (a) Original RGB image, (b) Ground truth depth map, (c) Depth map generated by Monodepth2, (d) Depth map generated by DenseDepth (KITTI), (e) Depth map generated by DenseDepth (NYU-Depth V2), (f) Depth map generated by 3DMM, (g) Depth map generated by Pix2Pix, (h) Depth map generated by CycleGAN, (i) Depth map generated by D+GAN

and DenseDepth (NYU-Depth V2) have a higher coarseness than the ground truth depth map. Conversely, the shapes of Figure 4.2f, Figure 4.2g, Figure 4.2h and Figure 4.2i are similar with Figure 4.2b, which indicates the depth maps generated by 3DMM, Pix2Pix, CycleGAN and D+GAN have a higher quality.

In the case of bs016_LFAU_22_0 of Bosphorus 3D Face Database, local SSIM maps of the depth maps generated by various models are shown in Figure 4.3. The structural similarity index measure (SSIM) is to measure the similarity between two images. In the SSIM map, regions with smaller local SSIM values correspond to different regions from the reference image. Similarly, regions with

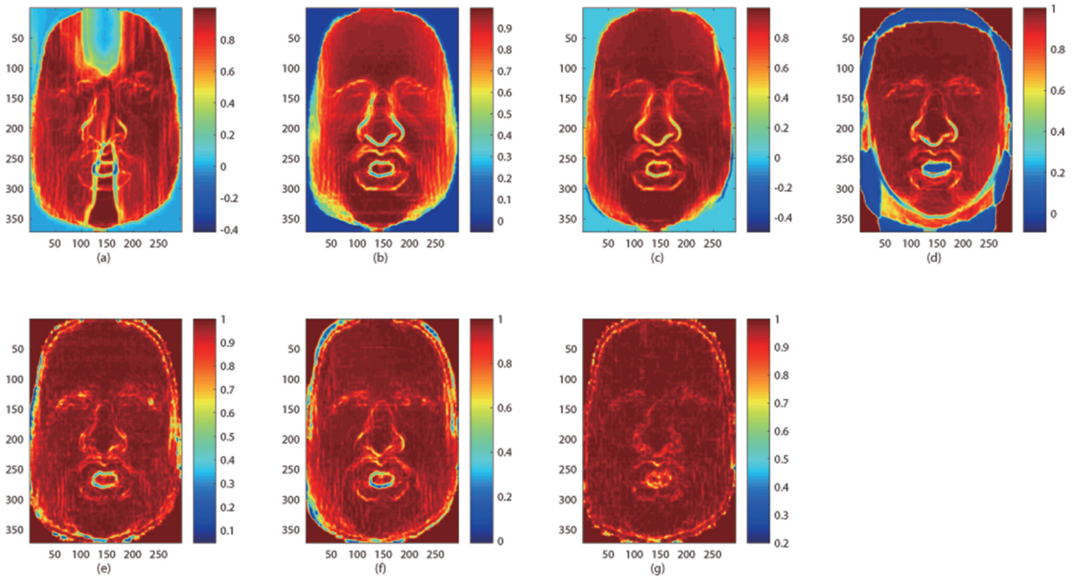


Figure 4.3: Local SSIM maps of depth maps generated by various models. (a) Model: Monodepth2, (b) Model: DenseDepth (KITTI), (c) Model: DenseDepth (NYU-Depth V2), (d) Model: 3DMM, (e) Model: Pix2Pix, (f) Model: CycleGAN, (g) Proposed Model: D+GAN

larger local SSIM values correspond to uniform regions of the reference image. The reference image here is the ground truth face depth map. From Figure 4.3, it can be observed that Figure 4.3g representing D+GAN has the most red area. Figure 4.3e representing Pix2Pix and Figure 4.3f representing CycleGAN in overall perform well except in specific areas of eyes, nose and mouth in comparison with Figure 4.3g. Figure 4.3d representing 3DMM shows a larger difference in face shape besides in eyes, nose and mouth. In addition, besides eyes, nose and mouth areas, Figure 4.3a representing Monodepth2, Figure 4.3b representing DenseDepth (KITTI) and Figure 4.3c representing DenseDepth (NYU-Depth V2) show a larger difference in four corners out of the face. Among these three, Figure 4.3c shows a less difference in the area of the human face.

#### 4.1.1.2 Case Study: 008-025 of CASIA 3D Face Database

Figure 4.4 presents the results for the case of 008-025 of CASIA 3D Face Database. Unlike the previous example, the input image in this example is a bust. In all, the performance of each model is similar to that in the above example. Figure 4.4g, Figure 4.4h and Figure 4.4i representing three GAN models show a satisfactory result. Especially for Figure 4.4i representing D+GAN, it is difficult to see the difference from the ground truth with the naked eye. It is worth mentioning that 3DMM can only be used for the human head area (see Figure 4.4f).

In the case of 008-025 of CASIA 3D Face Database, autocorrelation function graphs on depth maps generated by various models are shown as Figure 4.5. It shows the coarseness of the generated depth map. It is worth mentioning that Figure 4.5 indicates the texture coarseness of the depth map of the bust should be higher than the face (see Figure 4.2). Comparing with Figure 4.5b, Figure 4.5c, Figure 4.5d and Figure 4.5e has a smaller downward trend as $\xi$ and $\eta$ increasing, which means the depth maps generated by Monodepth2, DenseDepth (KITTI) and DenseDepth (NYU-Depth V2) have a lower coarseness than the ground truth depth map. In contrast, Figure 4.5g, Figure 4.5h and Figure 4.5i representing three GAN models have similar trends with Figure 4.5b, which implies they retain depth information well.

In the case of 008-025 of CASIA 3D Face Database, local SSIM maps of the depth maps generated by various models are shown in Figure 4.6. It shows the similarity of areas in the depth maps generated. In all, the performance of each model is similar to that in the last example. It is worth mentioning that the areas of clothes and neck in the depth map generated by CycleGAN are not as satisfactory as Pix2Pix and D+GAN (see Figure 4.6f).

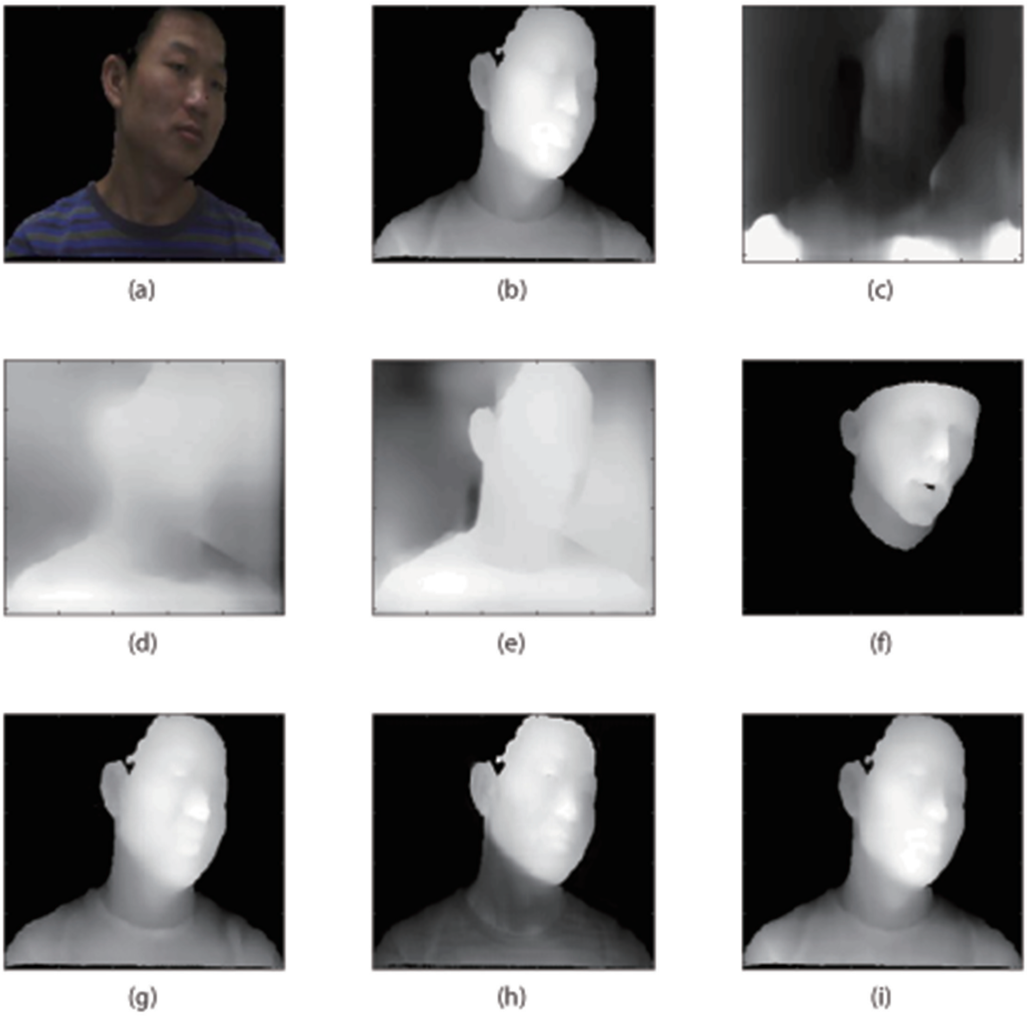


Figure 4.4: Face depth maps generated by various models in the case of 008-025. (a) Input RGB image, (b) Ground truth depth map, (c) Model: Monodepth2, (d) Model: DenseDepth (KITTI), (e) Model: Densedepth (NYU-Depth V2), (f) Model: 3DMM, (g) Model: Pix2Pix, (h) Model: CycleGAN, (i) Proposed Model: D+GAN

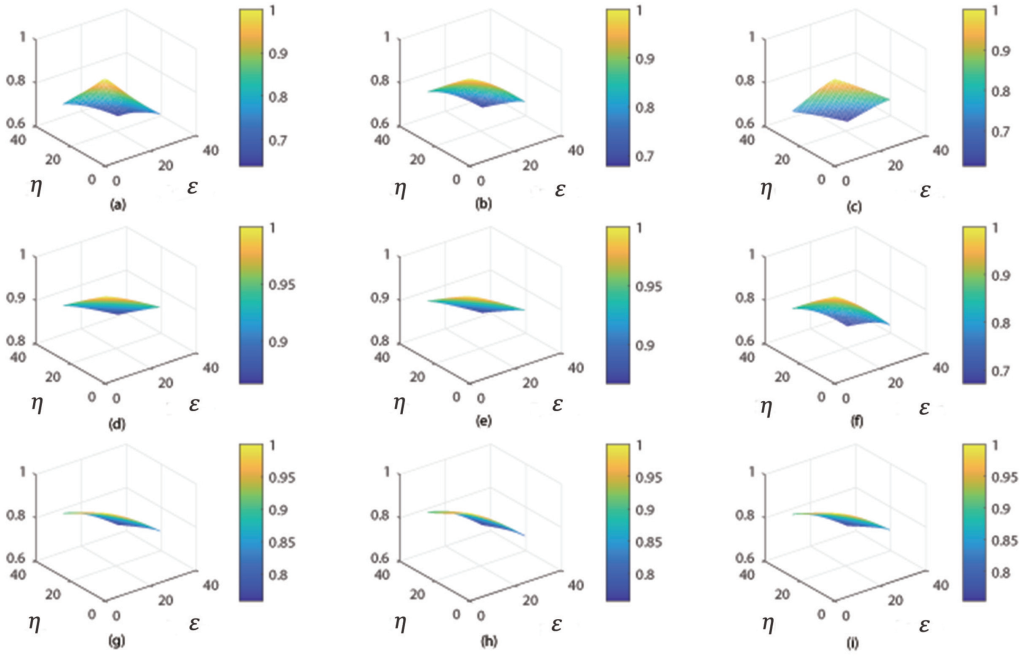


Figure 4.5: Autocorrelation function graphs of various output images: (a) Original RGB image, (b) Ground truth depth map, (c) Depth map generated by Monodepth2, (d) Depth map generated by DenseDepth (KITTI), (e) Depth map generated by DenseDepth (NYU-Depth V2), (f) Depth map generated by 3DMM, (g) Depth map generated by Pix2Pix, (h) Depth map generated by CycleGAN, (i) Depth map generated by D+GAN

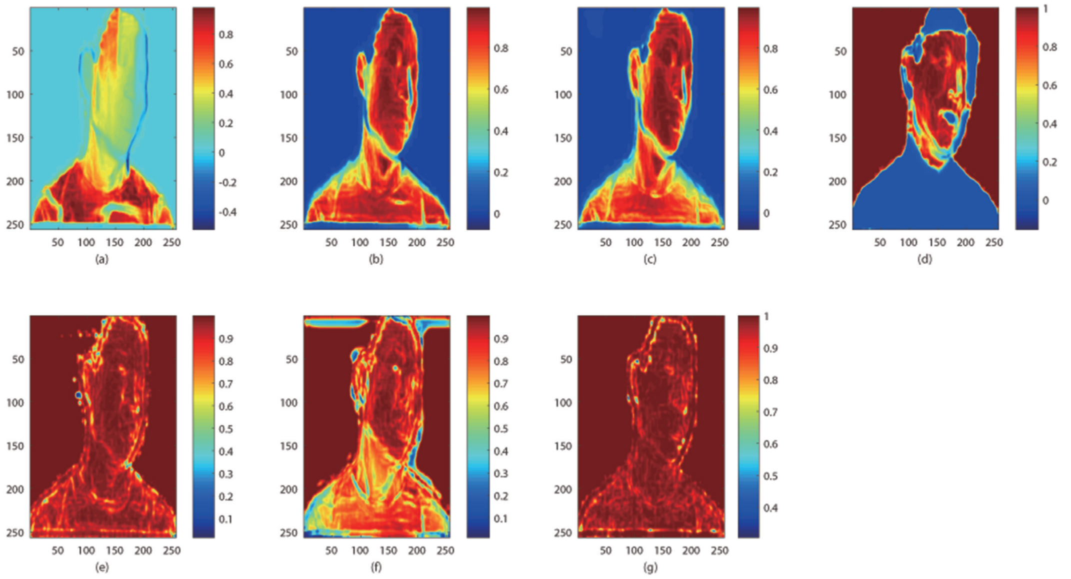


Figure 4.6: Local SSIM maps of depth maps generated by various models. (a) Model: Monodepth2, (b) Model: DenseDepth (KITTI), (c) Model: DenseDepth (NYU-Depth V2), (d) Model: 3DMM, (e) Model: Pix2Pix, (f) Model: CycleGAN, (g) Proposed Model: D+GAN

#### 4.1.1.3 Case Study: F0010_FE03WH_F2D of BU-3DFE Database

Figure 4.7 presents the results for the case of F0010_FE03WH_F2D of BU-3DFE Database. It is worth mentioning that, unlike the previous examples, GAN models are not trained by BU-3DFE Database. In all, the performance of each model is similar to that in the first example. Figure 4.7g, Figure 4.7h and Figure 4.7i representing three GAN models show a more satisfactory result than others. In detail, Figure 4.7g and Figure 4.7h representing Pix2Pix and CycleGAN respectively show a inaccurate depth in the eyes area. However, D+GAN performs well in the eyes area (see Figure 4.7i). It is worth mentioning that 3DMM generates inaccurate results in the face shape again (see Figure 4.7f).

In the case of F0010_FE03WH_F2D of BU-3DFE Database, autocorrelation function graphs on depth maps generated by various models are shown by Figure 4.8. It shows the coarseness of the depth map generated. It is worth mentioning

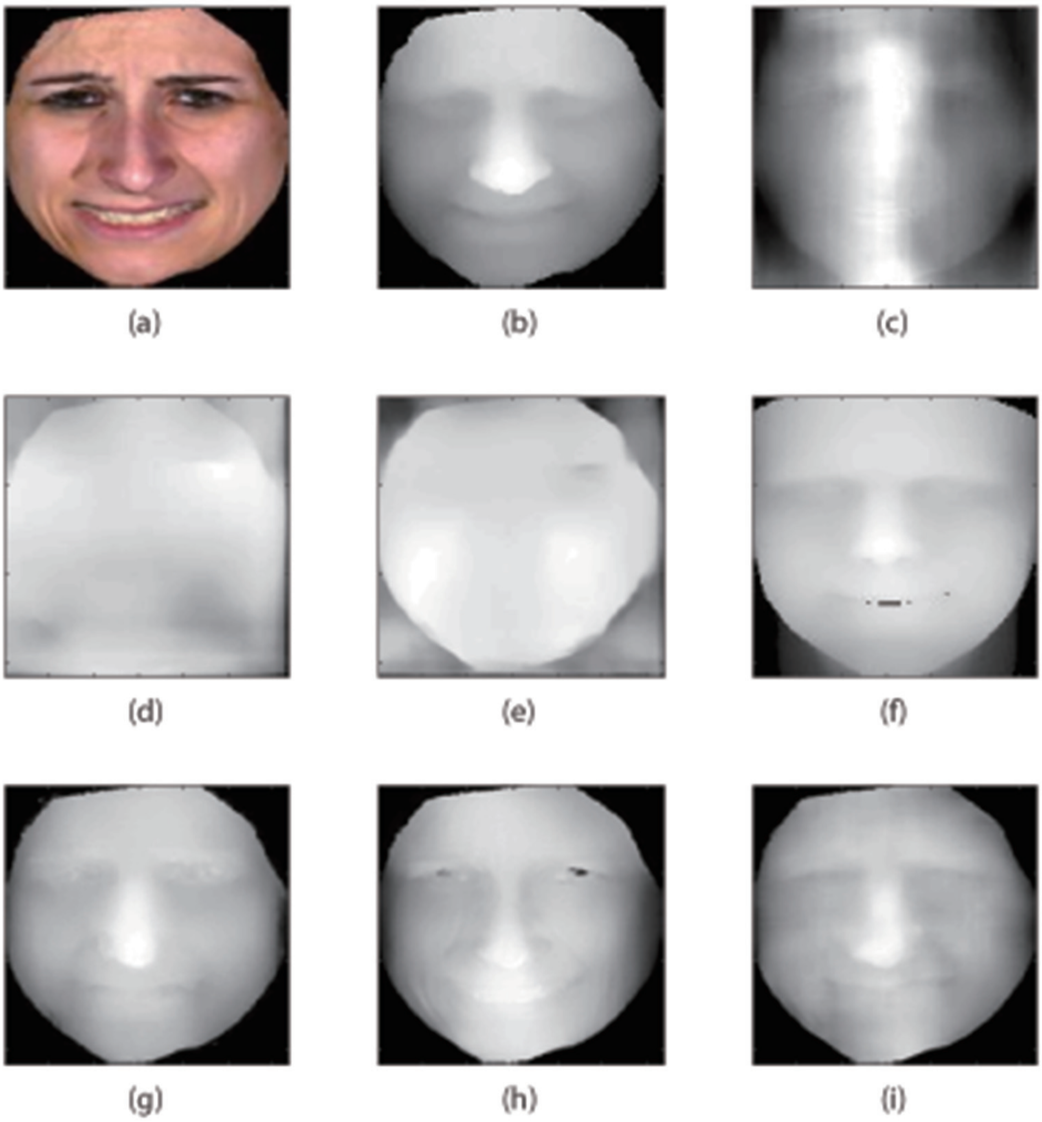


Figure 4.7: Face depth maps generated by various models in the case of F0010_FE03WH_F2D. (a) Input RGB image, (b) Ground truth depth map, (c) Model: Monodepth2, (d) Model: DenseDepth (KITTI), (e) Model: DenseDepth (NYU-Depth V2), (f) Model: 3DMM, (g) Model: Pix2Pix, (h) Model: CycleGAN, (i) Proposed Model: D+GAN

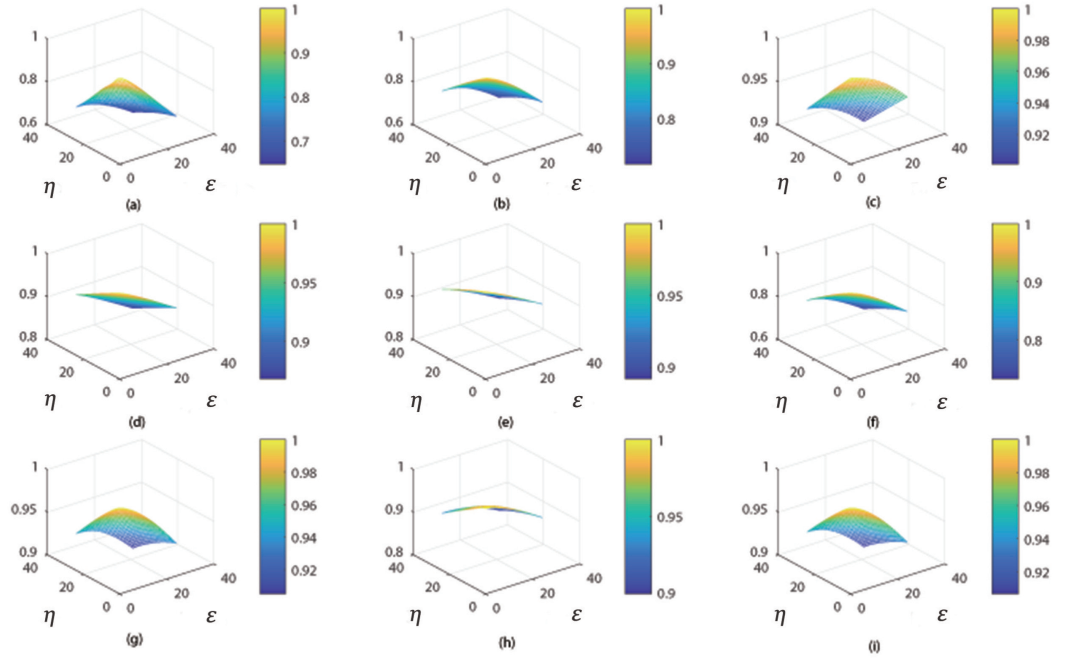


Figure 4.8: Autocorrelation function graphs of various output images: (a) Original RGB image, (b) Ground truth depth map, (c) Depth map generated by Monodepth2, (d) Depth map generated by DenseDepth (KITTI), (e) Depth map generated by DenseDepth (NYU-Depth V2), (f) Depth map generated by 3DMM, (g) Depth map generated by Pix2Pix, (h) Depth map generated by CycleGAN, (i) Depth map generated by D+GAN

that the graph shape of Figure 4.8f representing 3DMM is the most similar with Figure 4.8b representing the ground truth in this case. Figure 4.8g and Figure 4.8h has a smaller downward trend as $\xi$ and $\eta$ increasing, which means the depth maps generated for the face by Pix2Pix and CycleGAN have a lower coarseness.

In the case of F0010_FE03WH_F2D of BU-3DFE Database, local SSIM maps of the depth maps generated by various models are shown in Figure 4.9. It shows the similarity of areas in the depth maps generated. In all, the performance of each model is similar to that in the previous example. It is worth mentioning that the areas of clothes and neck in the depth map generated by CycleGAN are not as satisfactory as Pix2Pix and D+GAN (see Figure 4.9). In comparison with

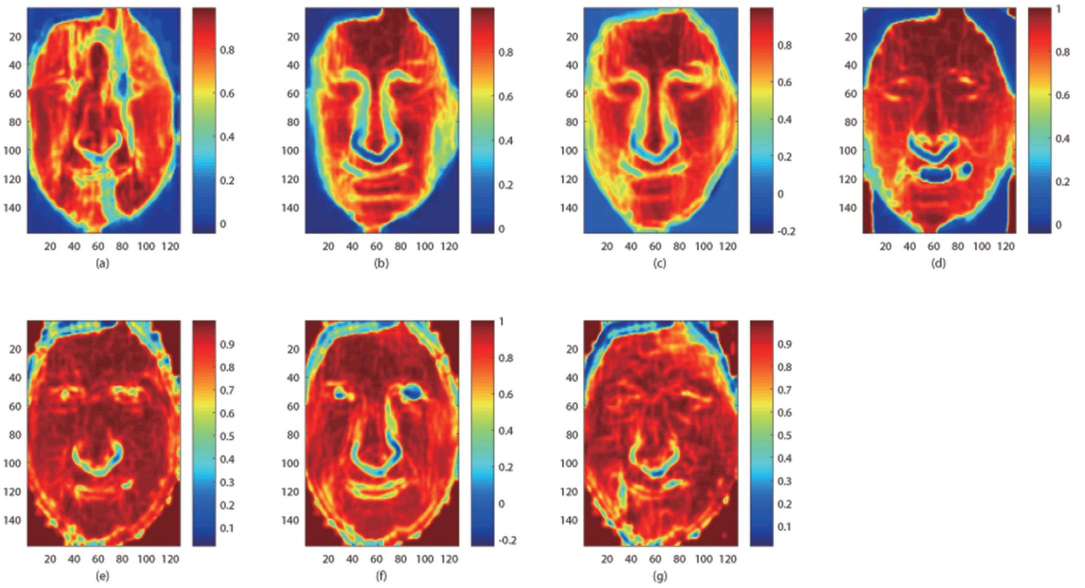


Figure 4.9: Local SSIM maps of the depth maps generated by various models. (a) Model: Monodepth2, (b) Model: DenseDepth (KITTI), (c) Model: DenseDepth (NYU-Depth V2), (d) Model: 3DMM, (e) Model: Pix2Pix, (f) Model: CycleGAN, (g) Proposed Model: D+GAN

Figure 4.9e, Figure 4.9g representing D+GAN performs better in the area of the eyes.

### 4.1.2 Quantitative Results and Analysis

In this section, quantitative analysis is carried out. The Structural Similarity Index (SSIM), Root Mean Squared Error (RMSE) and Peak Signal-to-Noise Ratio (PSNR) are selected to evaluate of the quality of the face depth map generated by several models on three datasets described before which are Bosphorus 3D Face Database, CASIA 3D Face Database and BU-3DFE Database.

The Structural Similarity Index (SSIM) [107] is the widely used standard for evaluating structural similarity in images that evaluates the quality of a processed image from a ground truth image. We calculate the SSIM for above six models as:

$$SSIM(a,b) = [l(a,b)]^{\alpha}[c(a,b)]^{\beta}[s(a,b)]^{\gamma} \tag{4.2}$$

where

$$l(a,b) = \frac{2\mu_a\mu_b + C_1}{\mu_a^2 + \mu_b^2 + C_1} \tag{4.3}$$

$$c(a,b) = \frac{2\sigma_a\sigma_b + C_2}{\sigma_a^2 + \sigma_b^2 + C_2} \tag{4.4}$$

$$s(a,b) = \frac{\sigma_{ab} + C_3}{\sigma_a\sigma_b + C_3} \tag{4.5}$$

In the above equations, there are two images denoted as a and b. $\mu_a$ and $\mu_b$ indicate the local mean values of corresponding images, $\sigma_a$ and $\sigma_b$ indicate the standard deviations and $\sigma_{ab}$ indicates the cross-covariance for images. Weights $\alpha > 0$, $\beta > 0$ and $\gamma > 0$. $C_1$, $C_2$ and $C_3$ are all constants to avoid the denominator being 0.

A lower RMSE value means a more accurate result corresponding to the reference. The RMSE between images a and b is calculated as:

$$RMSE(a,b) = \sqrt{\frac{1}{M \times N}\sum_{i=1}^{M}\sum_{i=1}^{N}(a(i,j) - b(i,j))^2} \tag{4.6}$$

where M and N are width and height of the image respectively.

The Mean Squared Erro (MSE) between images a and b is calculated as:

$$MSE(a,b) = \frac{1}{M \times N}\sum_{i=1}^{M}\sum_{i=1}^{N}(a(i,j) - b(i,j))^2 \tag{4.7}$$

PSNR, a logarithmic form using the decibel scale based on MSE, is widely used to quantify reconstruction quality for images. It is defined as:

$$PSNR = 10\log_{10}\frac{L^2}{MSE} = 20\log_{10}\frac{L}{RMSE} \tag{4.8}$$

where $L$ is the maximum possible pixel value of the image. Here, $L$ equals 255.

The histogram representations of the calculated average results of SSIM, RMSE, and PSNR for the above three data sets are shown in Figures 4.10 to 4.12, which are presented in Table 4.1. Not only qualitatively, but also quantitatively, the GAN model overall outperforms other models in these three datasets. Among them, the depth map output by D+GAN can get the best SSIM, RMSE and PSNR values.

For a 256×256 image, among the above three GAN models, Pix2Pix requires 18.6G multiply-accumulate operations (MACs) approximately, CycleGAN requires about 56.8G MACs approximately [120], and D+GAN, the embodiment showed, requires about 21.6G MACs approximately. These computations are acceptable for today's GPUs. Using the GAN model to obtain high-quality spatial information of face images will take more computation, which is a trade-off.

## 4.2 Pseudo RGB-D Face Recognition

### 4.2.1 Results and Analysis

In this section, classic machine learning and deep learning models including PCA [50], ICA [51], FaceNet [54] and InsightFace [30] are selected as face recognition methods. Five classic face recognition datasets including ORL [121], Yale [122], UMIST [123], AR [124] and FERET [125] are selected.

In order to make effective use of generated depth features in the pseudo RGB-D face recognition, image fusion algorithms are utilized. Through comparisons

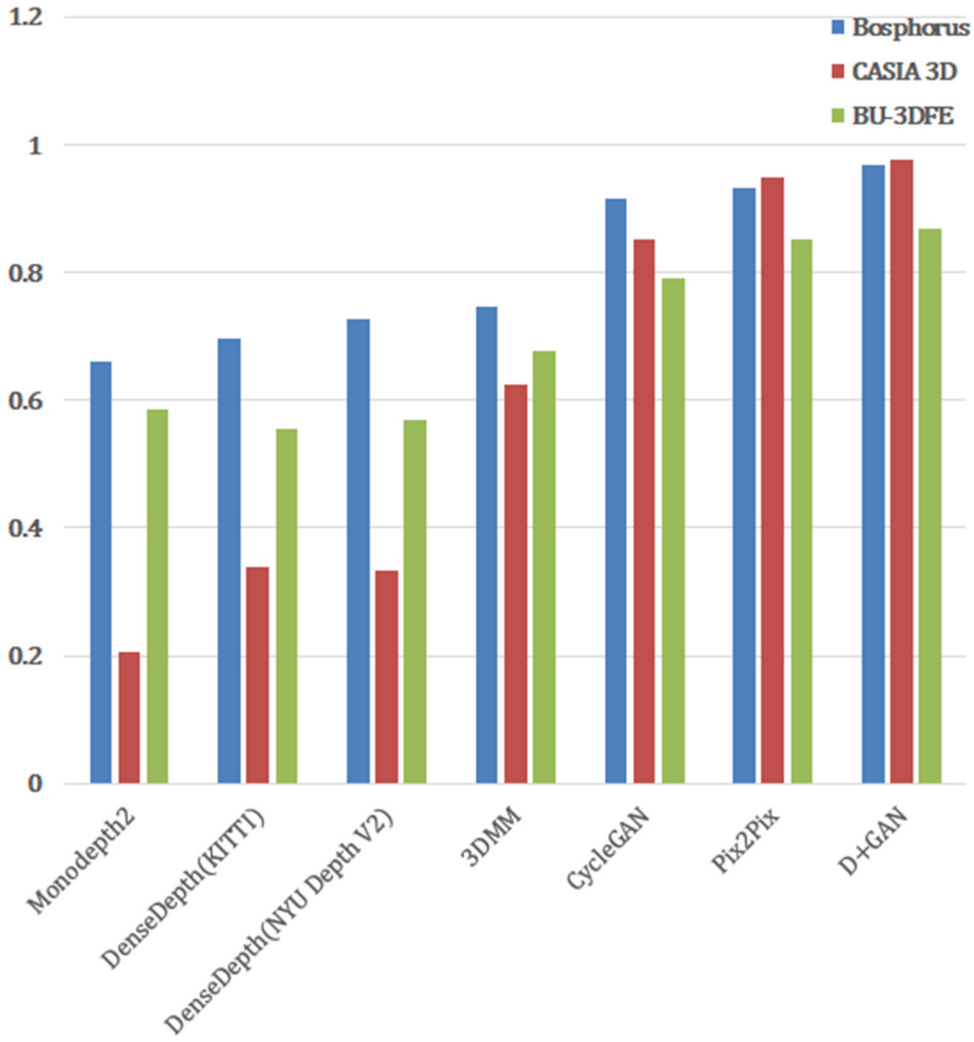


Figure 4.10: 2-D histogram representation of various models in SSIM indicators

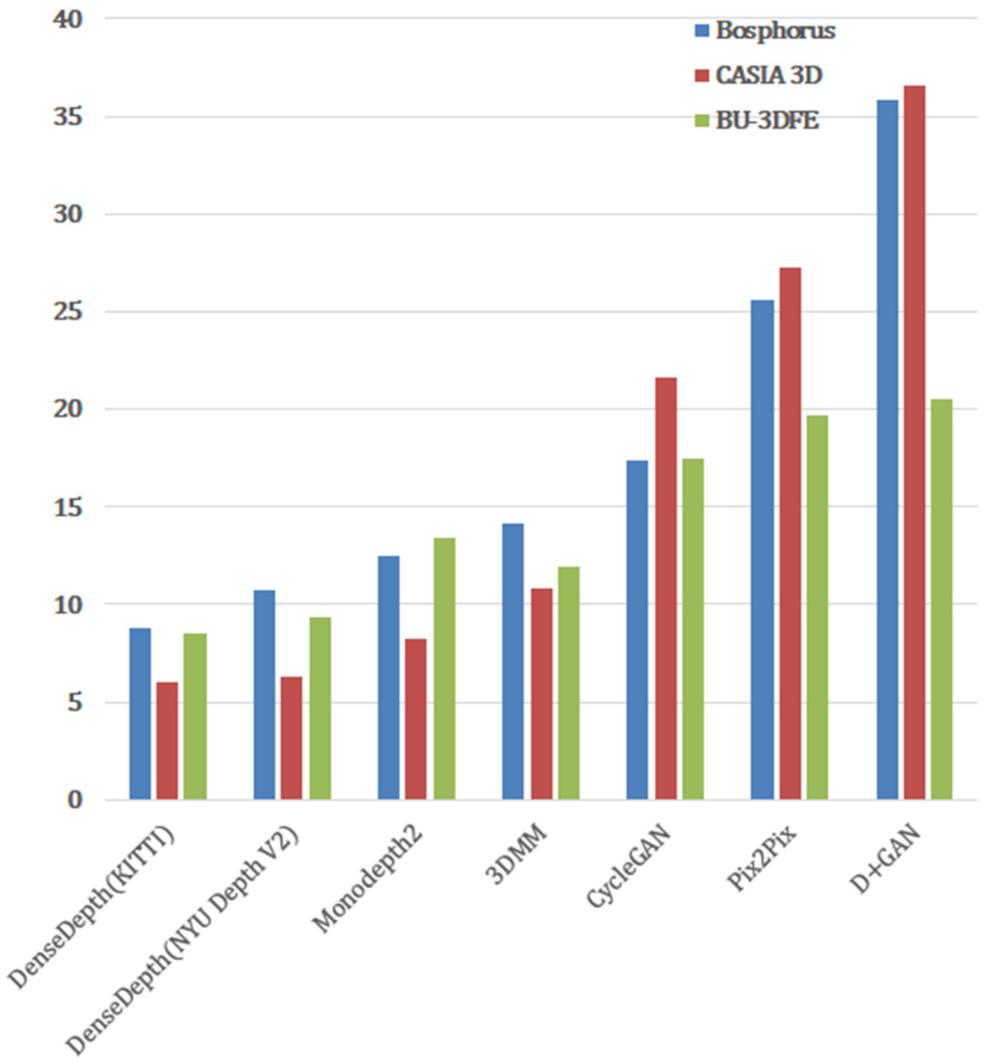


Figure 4.11: 2-D histogram representation of various models in PSNR indicators

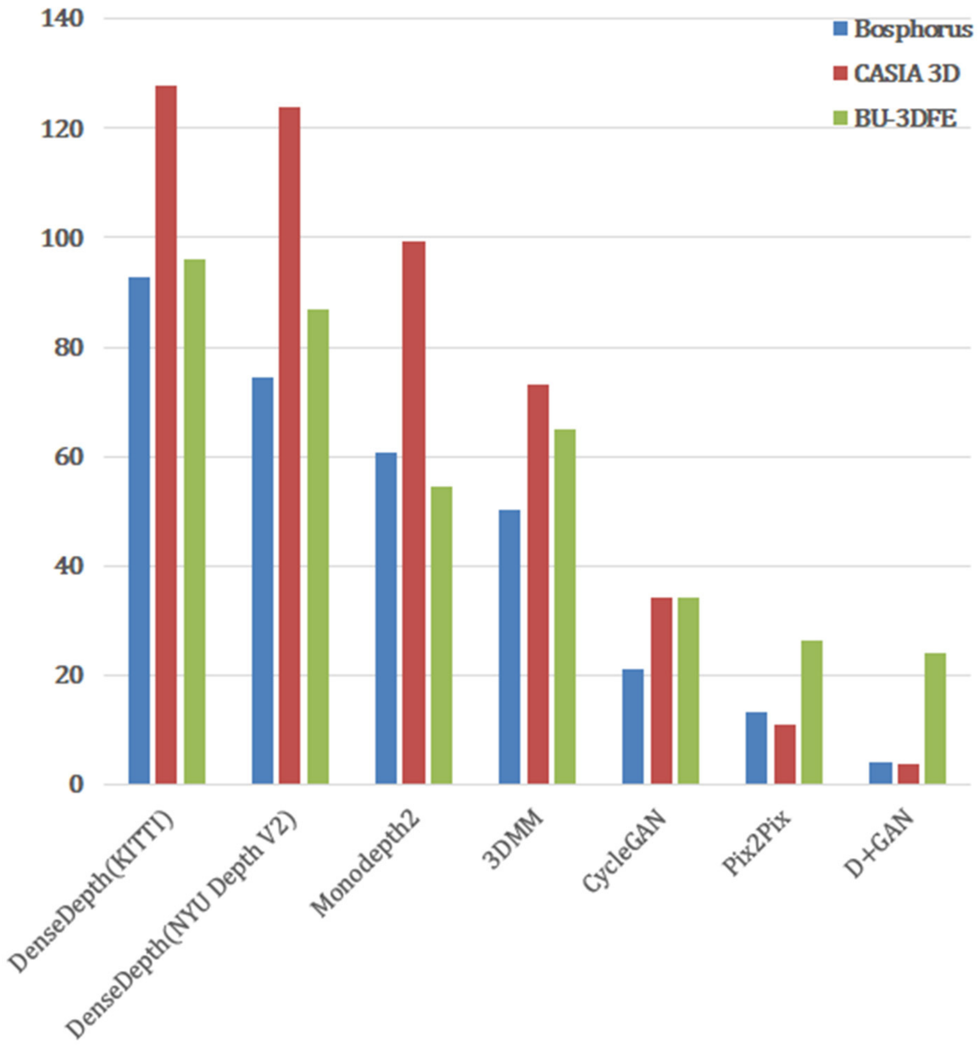


Figure 4.12: 2-D histogram representation of various models in RMSE indicators

Table 4.1: **Quantitative Index Results**

| Method | Index | Dataset | | |
|---|---|---|---|---|
| | | Bosphorus | CASIA 3D | BU-3DFE |
| Monodepth2 | SSIM | 0.660 | 0.205 | 0.585 |
| | RMSE | 60.77 | 99.15 | 54.41 |
| | PSNR | 12.46 | 8.205 | 13.42 |
| DenseDepth (KITTI) | SSIM | 0.697 | 0.339 | 0.555 |
| | RMSE | 92.70 | 127.7 | 95.91 |
| | PSNR | 8.789 | 6.007 | 8.494 |
| DenseDepth (NYU Depth V2) | SSIM | 0.728 | 0.334 | 0.570 |
| | RMSE | 74.38 | 123.7 | 86.79 |
| | PSNR | 10.70 | 6.283 | 9.361 |
| 3DMM | SSIM | 0.747 | 0.624 | 0.677 |
| | RMSE | 50.20 | 73.27 | 64.82 |
| | PSNR | 14.12 | 10.83 | 11.90 |
| Pix2Pix | SSIM | 0.933 | 0.949 | 0.852 |
| | RMSE | 13.43 | 11.11 | 26.41 |
| | PSNR | 25.56 | 27.22 | 19.70 |
| CycleGAN | SSIM | 0.916 | 0.851 | 0.792 |
| | RMSE | 21.26 | 34.36 | 34.23 |
| | PSNR | 17.41 | 21.58 | 17.44 |
| **D+GAN** | **SSIM** | **0.970** | **0.978** | **0.869** |
| | **RMSE** | **4.122** | **3.803** | **23.99** |
| | **PSNR** | **35.83** | **36.53** | **20.53** |

among Wavelet-based methods, Laplacian Pyramid and Non-subsampled Shearlet Transform (NSST) [49], NSST performs the best so as to be selected as the image fusion method for our face recognition experiments.

The shearlet system can be expressed as:

$$\begin{aligned}\Lambda_{D,S}(\Psi) = \big\{\Psi_{j,k,l}(x) = |det(D)|^{j/2}\Psi(S^l D^j x - k) :\\ j, l \in Z; k \in Z^2\}\end{aligned} \tag{4.9}$$

where $j$, $k$, and $l$ denote the scale, shift, and direction respectively. $D$, the anisotropic expanding matrix, is expressed as:

$$D = \begin{bmatrix} 4 & 0 \\ 0 & 2 \end{bmatrix} \tag{4.10}$$

and $S$, the shear matrix, is expressed as:

$$S = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \tag{4.11}$$

The NSST performs multi-scale and multi-directional decomposition on input images by Non-subsampled Pyramids (NSPs) and shearing filters in the first place. Next, according to the made fusion strategy, the high frequency and low frequency sub-band images decomposed are transformed and combined into new sub-band images. Last, the final fused image is achieved by the inverse NSST on the new sub-band images. In our embodiment, the filter set for the Laplacian Pyramid decomposition is 'maxflat'. The vector indicating decomposition directions is set to [3, 3, 4, 4]. The vector indicating the local support of the shearing filter is set to [8, 8, 16, 16]. The fusion coefficient is set to 0.5.

Besides NSST, D+GAN is selected as the preferred embodiment for generating

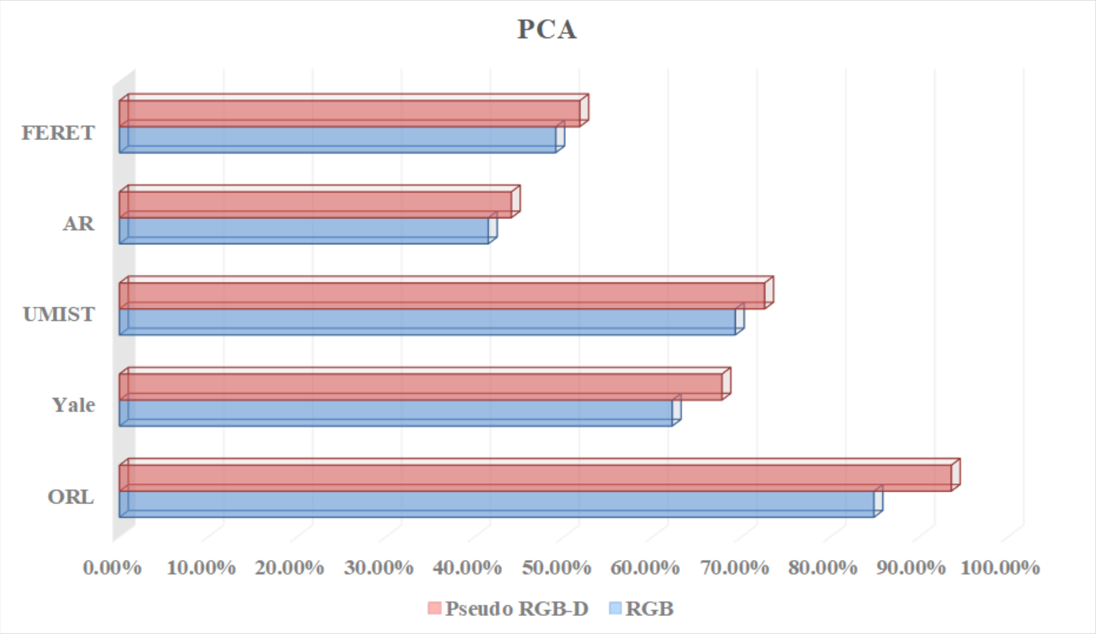


Figure 4.13: Accuracy histogram of PCA in RGB and Pseudo RGB-D modes

the pseudo face depth map in the pseudo RGB-D face recognition due to its good performance in the previous section. If the training images are sufficient, due to the great learning ability of the deep learning model, it is easy to have a 100% accuracy during testing. Therefore, in the evaluation, due to the different capabilities of ML models, we used separate experimental settings to differentiate the performance of face recognition of each model.

In experiments of testing PCA, two images of each person in the dataset are applied for testing, and the rest images of that person are for training. The number of feature face set is 30 for PCA. In this case, the mode of pseudo RGB-D face recognition improves the accuracy by 10.2%, 9.0%, 4.6%, 6.3% and 5.5% approximately on ORL, Yale, UMIST, AR and FERET dataset respectively.

In experiments of testing ICA, five images of each person in the dataset are applied for training, and rest images of that person are for testing. The number of components set is 70 for ICA. The mode of pseudo RGB-D face recognition

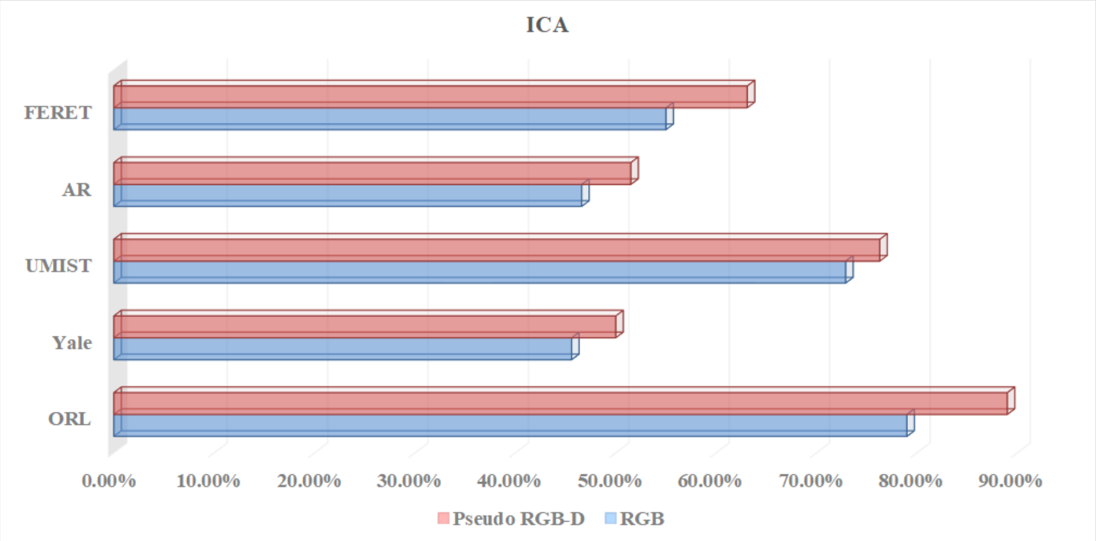


Figure 4.14: Accuracy histogram of ICA in RGB and Pseudo RGB-D modes

improves the accuracy by 12.7%, 9.6%, 3.4%, 10.6% and 14.8% approximately on ORL, Yale, UMIST, AR and FERET dataset respectively.

In experiments of testing DL models including FaceNet and InsightFace, for ORL and AR datasets, 30% of the images of each person are used for training, and 70% of the images of each person are used for testing. For Yale dataset, 20% of the images of each person are used for training, and 80% of the images of each person are used for testing. For UMIST dataset, 10% of the images of each person are used for training, and 90% of the images of each person are used for testing. For FERET dataset, 40% of the images of each person are used for training, and 60% of the images of each person are used for testing. Since the number of images of each person in the ORL and YALE datasets is relatively small, and the total number of people is also relatively small. Therefore, using the pre-trained model to directly extract features, and then training a linear SVM classifier for testing could get better results. For the datasets UMIST, AR and FERET with more images, fine-tuning the pretrained network model could be used as a conventional strategy.

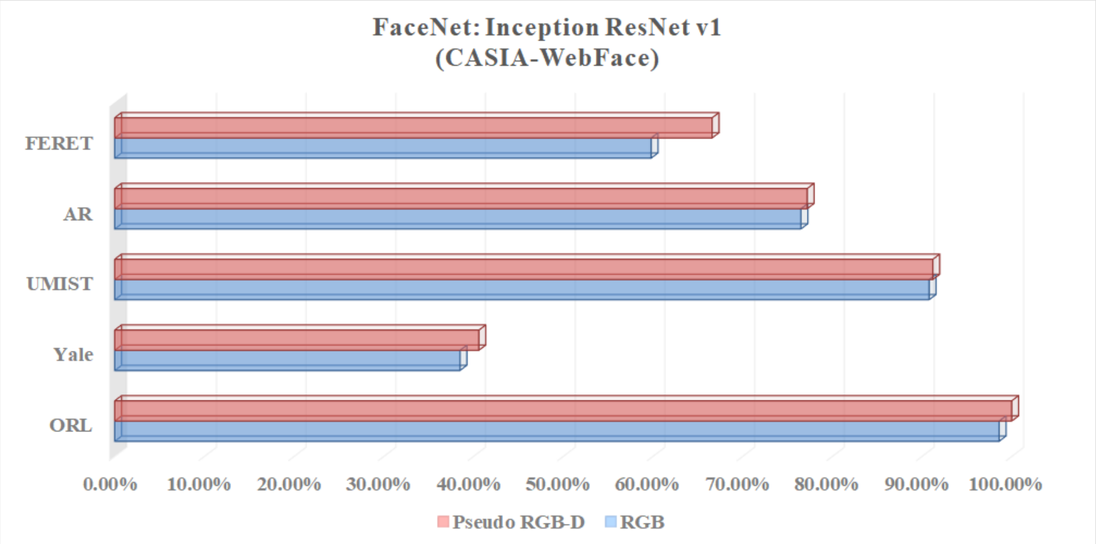


Figure 4.15: Accuracy histogram of FaceNet: Inception ResNet v1 (CASIA-WebFace) in RGB and Pseudo RGB-D modes

Table 4.2 presents the face recognition results by two modes including RGB and Pseudo RGB-D using traditional ML and advanced DL models on the five classical face recognition datasets. Table 4.3 presents the performance difference from RGB mode to Pseudo RGB-D mode.

Specifically, in experiments of testing the FaceNet: Inception ResNet v1 model pretrained by CASIA-WebFace, the mode of pseudo RGB-D face recognition improves the accuracy by 2.7%, 5.7%, 0.4%, 0.9% and 11.3% approximately on datasets ORL, Yale, UMIST, AR and FERET respectively. In experiments of testing the FaceNet: Inception ResNet v1 model pretrained by VGG-Face2, the mode of pseudo RGB-D face recognition improves the accuracy by 0%, 0%, 1.7%, 0.7% and 1.3% approximately on datasets ORL, Yale, UMIST, AR and FERET respectively. In experiments of testing the Insightface: IResNet34 model pretrained by MS1MV2, the mode of pseudo RGB-D face recognition improves the accuracy by 2.1%, 3.2%, 1.0%, 0.2% and 7.9% approximately on datasets ORL, Yale, UMIST, AR and FERET respectively. In experiments of testing the In-

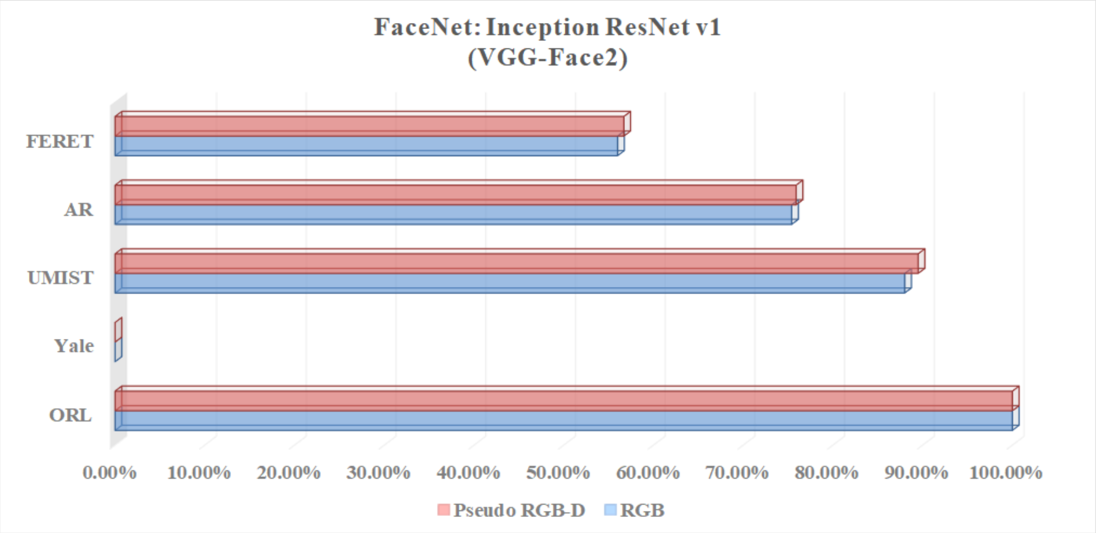


Figure 4.16: Accuracy histogram of FaceNet: Inception ResNet v1 (VGG-Face2) in RGB and Pseudo RGB-D modes

sightface: IResNet100 model pretrained by MS1MV2, the mode of pseudo RGB-D face recognition improves the accuracy by 2.7%, 5.7%, 0.3%, 2.4% and 2.5% approximately on datasets ORL, Yale, UMIST, AR and FERET respectively.

Table 4.2 shows that, in the face recognition experiments, the best performing results annotated in bold for each dataset of the five, almost all use the mode of pseudo RGB-D face recognition. It can be concluded that pseudo RGB-D face recognition proposed is able to improve the accuracy in comparison with RGB face recognition using different classic traditional ML and DL models. Especially for traditional ML models, pseudo RGB-D face recognition mode can increase the accuracy more.

## 4.3 Summary

In the field of facial recognition, many deep learning models have already achieved very high accuracy rates. To allow for comparison and to prevent all models

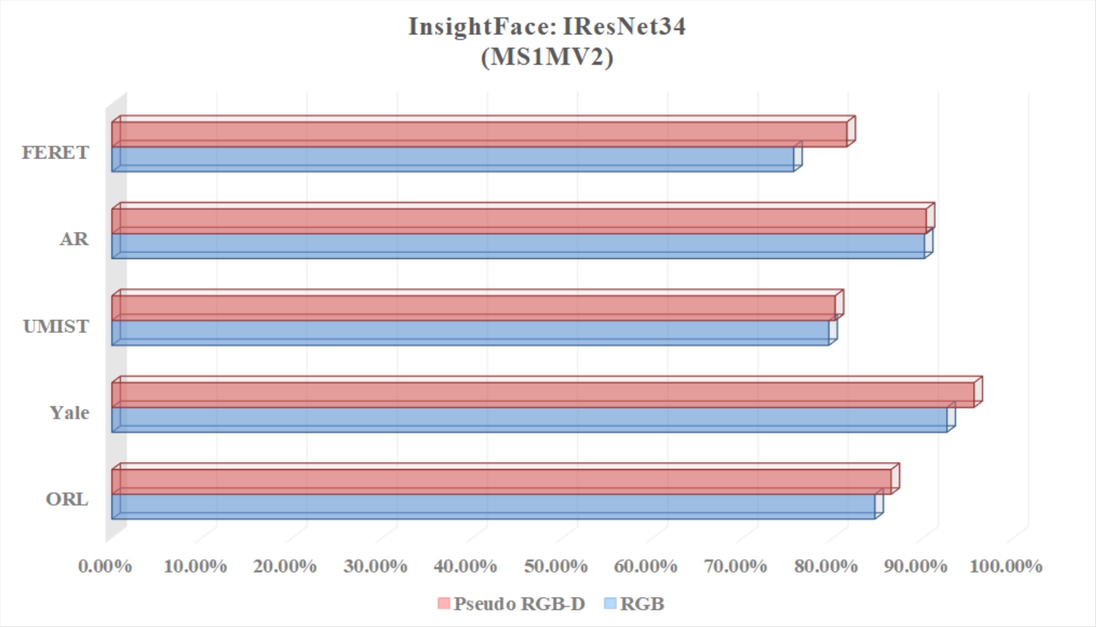


Figure 4.17: Accuracy histogram of Insightface: IResNet34 (MS1MV2) in RGB and Pseudo RGB-D modes

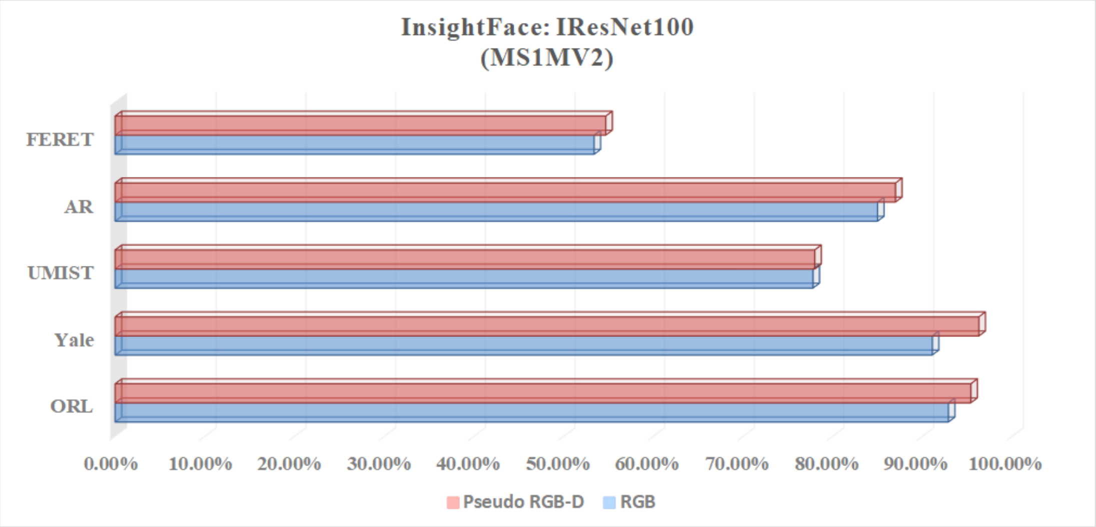


Figure 4.18: Accuracy histogram of Insightface: IResNet100 (MS1MV2) in RGB and Pseudo RGB-D modes

Table 4.2: **Experimental Results of Face Recognition**

| Mode | Method | Dataset | | | | |
|---|---|---|---|---|---|---|
| | | ORL | Yale | UMIST | AR | FERET |
| RGB<br>Face Recognition | PCA | 84.9% | 62.2% | 69.3% | 41.5% | 49.1% |
| | ICA | 79.0% | 45.6% | 72.9% | 46.6% | 55.0% |
| | FaceNet: Inception ResNet v1 (CASIA-WebFace) | 98.6% | 38.5% | 90.8% | 76.5% | 59.8% |
| | FaceNet: Inception ResNet v1 (VGG-Face2) | **100.0%** | 0.0% | 88.0% | 75.4% | 56.0% |
| | InsightFace: IResNet34 (MS1MV2) | 84.6% | 92.6% | 79.5% | 90.1% | 75.6% |
| | InsightFace: IResNet100 (MS1MV2) | 92.9% | 91.1% | 77.8% | 85.0% | 53.4% |
| **Pseudo RGB-D**<br>**Face Recognition** | PCA | 93.6% | 67.8% | 72.6% | 44.1% | 51.8% |
| | ICA | 89.0% | 50.0% | 76.3% | 51.5% | 63.1% |
| | FaceNet: Inception ResNet v1 (CASIA-WebFace) | **100.0%** | 40.6% | **91.2%** | 77.2% | 66.6% |
| | FaceNet: Inception ResNet v1 (VGG-Face2) | **100.0%** | 0.0% | 89.5% | 75.9% | 56.7% |
| | InsightFace: IResNet34 (MS1MV2) | 86.4% | 95.6% | 80.2% | **90.3%** | **81.5%** |
| | InsightFace: IResNet100 (MS1MV2) | 95.4% | **96.3%** | 78.0% | 87.0% | 54.7% |

Table 4.3: **Changes of Face Recognition Results**

| Change | Method | Dataset | | | | |
|---|---|---|---|---|---|---|
| | | ORL | Yale | UMIST | AR | FERET |
| | PCA | 8.7% | 5.6% | 3.3% | 2.6% | 2.7% |
| | ICA | 10.0% | 4.4% | 3.4% | 4.9% | 8.1% |
| | FaceNet: Inception ResNet v1 (CASIA-WebFace) | 1.4% | 2.1% | 0.4% | 0.7% | 6.8% |
| | FaceNet: Inception ResNet v1 (VGG-Face 2) | 0 | 0 | 1.5% | 0.4% | 0.7% |
| | InsightFace: IResNet34 (MS1MV2) | 1.8% | 3% | 0.7% | 0.2% | 5.9% |
| | InsightFace: IResNet100 (MS1MV2) | 2.5% | 5.2% | 0.2% | 2% | 1.3% |

from achieving 100% accuracy, we only used a small portion of the dataset for training. In our experiments, the improvement of the Pseudo RGB-D mode on the FaceNet: Inception ResNet v1 (VGG-Face 2) model was limited, which was due to the scarcity of training data provided to the model.

Inspired by the occurrence of RGB-D face recognition, we propose a pseudo RGB-D face recognition framework. In essence, the ML model is able to imitate the relative depth map from its corresponding RGB image by learning from big data to replace the depth sensors. We provide a D+GAN model for making increased use of face attribute information to generate the high quality face depth map. In cooperation with NSST, the pseudo RGB-D face recognition obtains an overall improvement in comparison with RGB face recognition. With the pseudo RGB-D face recognition framework, we could modularly adapt off-the-shelf algorithm models to promote the performance of RGB face recognition. In the facial recognition experiments we designed, the PRGB-D mode was able to further enhance the performance of advanced facial recognition models trained with RGB images. In future, we will continue to discover simple and effective models to perform the monocular face depth estimation, and efficient ways to apply them to improve the biometric recognition performance.

# Chapter 5

# Facial Diagnosis Experiment Results and Analysis

In this chapter, we will present the results of facial diagnosis experiments according to different time stages. In Section 5.1, we primarily validate the effectiveness of using pre-trained face recognition models in facial diagnosis tasks. In Section 5.2, building upon the use of pre-trained face recognition models, we initially employed pseudo-depth, as described in Pseudo RGB-D Face Recognition of Section 3.1, to enhance the performance of RGB facial diagnosis tasks. In Section 5.3, we introduced a Simulated Multimodal Framework that makes fuller use of pseudo-depth information, thereby further enhancing the performance of RGB facial diagnosis tasks.

## 5.1 Deep Facial Diagnosis

### 5.1.1 Experimental Results and Discussions

In this section, we perform the experiments on two tasks of facial diagnosis by two strategies of deep transfer learning including fine-tuning abbreviated as DTL1 and using CNN as a feature extractor abbreviated as DTL2 as presented in Section 3.2. The deep learning models pretrained for object detection and face recognition are selected for comparison. In addition, we compare the results with traditional machine learning methods using the hand-crafted feature that is Dense Scale-Invariant Feature Transform (DSIFT) [126]. DSIFT, which is often used in object recognition, performs Scale Invariant Feature Transform (SIFT) on a dense grid of locations of the image at a certain scale and orientation. The SVM algorithm for its good performance in few-shot learning, is used as the classifier for Bag of Features (BOF) models with DSIFT descriptors.

Two cases of facial diagnosis are designed in this section. One is the detection of beta-thalassemia, which is a binary classification task. The other one is the detection of four conditions which are beta-thalassemia, hyperthyroidism, Down syndrome and leprosy with the healthy control, which is a multiclass classification task and more challenging.

#### 5.1.1.1 Single Condition Detection: Beta-thalassemia

In practice, we usually need to perform detection or screening on one specific condition. In this case, we only use 140 images of the dataset which is composed of 70 beta-thalassemia-specific face images and 70 images for healthy control. 40 of each type images are for training, and 30 of each type images are for testing. It is a binary classification task. By comparing all selected machine learning methods (see TABLE 5.2), we find that the best overall top-1 accuracies can be achieved by using the strategies of deep transfer learning on the VGG-Face model (VGG-16 pretrained on the VGG-Face dataset). Furthermore, applying DTL2: CNN as a feature extractor can get a better accuracy of 95.0% than using DTL1: fine-tuning in this task, which is indicated by Figure 5.1.

Figure 5.1 shows the confusion matrices of DTL1 and DTL2 on the VGG-

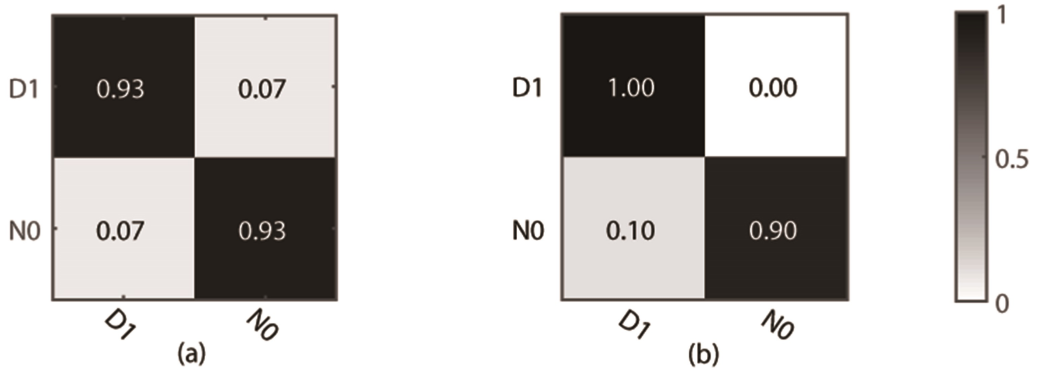


Figure 5.1: The confusion matrix for beta-thalassemia detection (a binary classification task). (a) DTL1: VGG-Face (Fine-tuning). (b) DTL2: VGG-Face (Feature Extractor) + SVM Linear. D1 represents the beta-thalassemia-specific face, N0 represents the healthy control.

Face model in this task. D1 represents the beta-thalassemia-specific face, and N0 represents the healthy control. The row in the confusion matrix indicates the predicted classes, and the column in the confusion matrix indicates the actual classes. In detail, two of thirty testing images for each type, false positives and false negatives, are misclassified by DTL1, which leads to an accuracy of 93.3%. For DTL2, thirty images belonging to the type of beta-thalassemia in actual, true positives, are all classified correctly. On the other hand, three of thirty images, false positives, are belonging to the healthy control in actual, but classified as the beta-thalassemia-specific face. Figure 5.2 shows the receiver operating characteristic (ROC) curves of the VGG-Face model by DTL1 and DTL2. The blue dotted line indicates the performance of DTL1, and the red solid line indicates the performance of DTL2. The Areas Under ROC curves (AUC) calculated are 0.969 and 0.978 correspondingly.

For comparison, deep learning models pretrained such as AlexNet, VGG16 and ResNet are used. In addition, traditional machine learning methods extracting DSIFT features on the face image and predicting with a linear or nonlinear SVM classifier [127] are selected. Five indicators that are accuracy, precision,

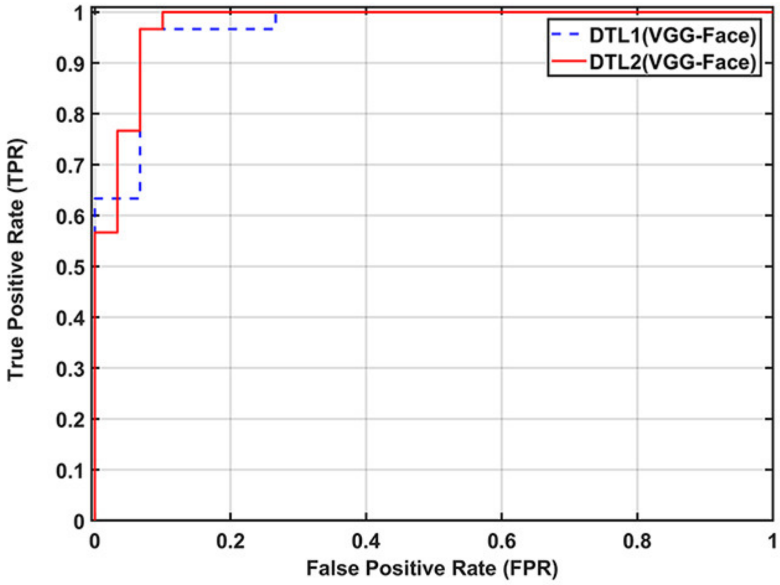


Figure 5.2: The receiver operating characteristic (ROC) curves of the VGG-Face model. The blue dotted line indicates the performance of DTL1, and the red solid line indicates the performance of DTL2.

sensitivity, specificity and F1-score which is a weighted average of the precision and sensitivity are selected to evaluate the performance of models. The indicator of FLOPs spent for forward pass is estimated to evaluate the time complexity of models. Table 5.1 lists the results of both traditional machine learning methods and fine-tuning deep learning models pretrained on the ImageNet and VGG-Face dataset in this task.

From the results, we find that the performance by traditional machine learning methods is close to the performance of fine-tuning (DTL1) deep learning models pretrained on ImageNet. However, the performance of fine-tuning (DTL1) the deep learning models pretrained on VGG-Face is overall better than ones pretrained on ImageNet, which is reasonable. Because the source domain of VGG-Face is nearer to DSF dataset than ImageNet. Table 5.2 lists the results of CNN as a feature extractor on the pretrained deep learning models (DTL2). Applying

Table 5.1: **Binary classification results on the detection of beta-thalassemia (Traditional:Row 2&3 and DTL1:Row 4-9)**

| Method<br>(Traditional and DTL1) | Pretrain | FLOPs | Accuracy | Precision | Sensitivity | Specificity | F1-Score |
|---|---|---|---|---|---|---|---|
| DSIFT<br>+ SVM Linear | N | 1.28G | 71.7% | 67.6% | 83.3% | 60.0% | 74.6% |
| DSIFT<br>+ SVM Chiˆ2 | N | 1.28G | 68.3% | 65.7% | 76.7% | 60.0% | 70.8% |
| AlexNet<br>(Fine-tuning) | ImageNet | 723M | 76.7% | 86.4% | 63.3% | 90.0% | 73.1% |
| AlexNet<br>(Fine-tuning) | VGG-Face | 723M | 80.0% | 72.5% | 96.7% | 63.3% | 82.9% |
| VGG16<br>(Fine-tuning) | ImageNet | 15.47G | 78.3% | 75.8% | 83.3% | 73.3% | 79.4% |
| **VGG16<br>(Fine-tuning)** | **VGG-Face** | **15.47G** | **93.3%** | **93.3%** | **93.3%** | **93.3%** | **93.3%** |
| ResNet50<br>(Fine-tuning) | ImageNet | 3.87G | 75.0% | 77.8% | 70.0% | 80.0% | 73.7% |
| ResNet50<br>(Fine-tuning) | VGG-Face | 3.87G | 91.7% | 100% | 83.3% | 100% | 90.9% |

DTL2: CNN as a feature extractor can get an overall better performance than traditional machine learning methods and DTL1. However, deep learning models pretrained on VGG-Face seem to behave not necessarily better than deep learning models pretrained on ImageNet in this strategy. It will be further investigated in the next experiment.

#### 5.1.1.2 Multiple Conditions Detection

In practical applications, performing the detection or screening of multiple conditions at once can greatly increase efficiency. To further evaluate the algorithm, in this case, the task dataset contains a total of 350 images, with 70 images for each facial type. During the training process, a total of 200 images (40 images of each type) are used, while during the testing process, 150 images (30 images of each type) are used. This is a multi-class classification task.

By comparing all selected machine learning methods, we find that the best overall top-1 accuracies can be achieved by using the deep transfer learning strategies based on the VGG-Face model. Furthermore, applying DTL2: VGG-Face

Table 5.2: **Binary classification results on the detection of beta-thalassemia (DTL2)**

| Method<br>(DTL2) | Pretrain | FLOPs | Accuracy | Precision | Sensitivity | Specificity | F1-Score |
|---|---|---|---|---|---|---|---|
| AlexNet<br>(Feature Extractor)<br>+ SVM Linear | ImageNet | 723M | 90.0% | 87.5% | 93.3% | 86.7% | 90.3% |
| AlexNet<br>(Feature Extractor)<br>+ SVM Linear | VGG-Face | 723M | 85.0% | 86.2% | 83.3% | 86.7% | 84.7% |
| VGG16<br>(Feature Extractor)<br>+ SVM Linear | ImageNet | 15.47G | 88.3% | 84.8% | 93.3% | 83.3% | 88.8% |
| **Best: VGG16<br>(Feature Extractor)<br>+ SVM Linear** | **VGG-Face** | **15.47G** | **95.0%** | **90.9%** | **100%** | **90.0%** | **95.2%** |
| ResNet50<br>(Feature Extractor)<br>+ SVM Linear | ImageNet | 3.87G | 91.7% | 93.1% | 90.0% | 93.2% | 91.5% |
| ResNet50<br>(Feature Extractor)<br>+ SVM Linear) | VGG-Face | 3.87G | 86.7% | 95.8% | 76.7% | 96.7% | 85.2% |

as a feature extractor can get a better accuracy of 93.3% than using DTL1: fine-tuning in this task, which is indicated by Figure 5.3. Figure 5.3 shows the confusion matrices of DTL1 and DTL2 on the VGG-Face model in this task. D1 represents the beta-thalassemia-specific face, D2 represents the hyperthyroidism-specific face, D3 represents the DS-specific face, D4 represents the leprosy-specific face and N0 represents the healthy control. The row in the confusion matrix indicates the predicted classes, and the column in the confusion matrix indicates the actual classes. From the Figure 5.3(b), four of thirty images are belonging to the hyperthyroidism-specific face in actual, but classified as other types, which indicates it is relatively difficult for the classifier to recognize hyperthyroidism from face images. For recognizing beta-thalassemia, Down syndrome and leprosy, the classifier has a very good accuracy. Figure 5.3(a) of DTL1 also shows a low accuracy on recognizing hyperthyroidism.

Table 5.3 lists the results of traditional machine learning methods and deep learning methods in the multiclass classification task as previously described.

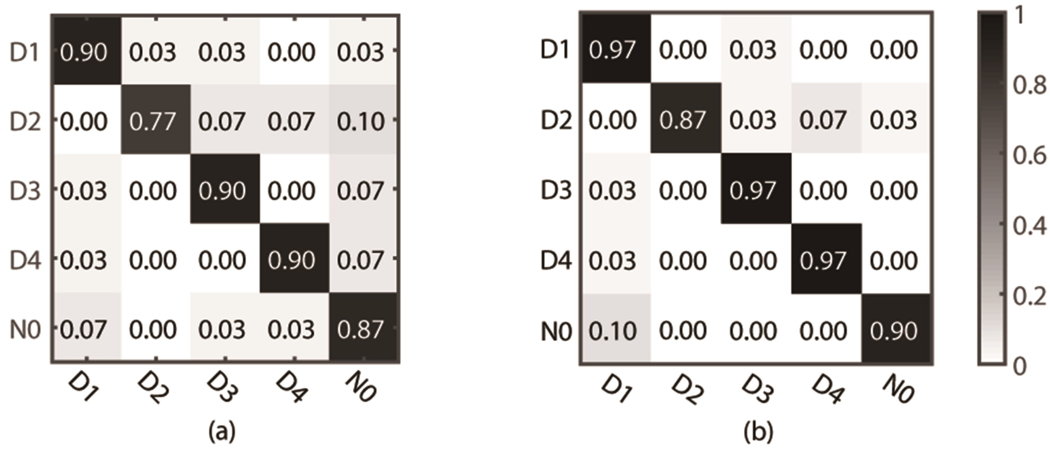


Figure 5.3: The confusion matrix for multiple conditions detection (a multiclass classification task). (a) DTL1: VGG-Face (Fine-tuning). (b) DTL2: VGG-Face (Feature Extractor) + SVM Linear. D1 represents the beta-thalassemia-specific face, D2 represents the hyperthyroidism-specific face, D3 represents the DS-specific face, D4 represents the leprosy-specific face and N0 represents the healthy control.

Since the multiclass classification task is more difficult than the binary classification task as presented before, the accuracies of machine learning models decrease generally. The results by deep transfer learning methods are much better than the results by traditional machine learning methods in this task, which is as expected. And deep learning models pretrained on VGG-Face behave generally better than deep learning models pretrained on ImageNet in both strategies. The performance of DTL2: CNN as a feature extractor is overall better than that of DTL1: Fine-tuning again, which probably is due to the relatively small dataset. Fig 5.4-5.6 provide a clear illustration of the changes in accuracy and the corresponding trends for AlexNet, VGG16, and ResNet50 under DTL1 and DTL2, respectively.

On the basis of applying DTL2, for exploring a better performance by deep transfer learning, we investigate the performance of ResNet50 and SE-ResNet50 [128] models pretrained on MS-Celeb-1M [28] and VGGFace2 [29]. MS-Celeb-1M

Table 5.3: **Multiclass classification results on the detection of four conditions**

| Method | Pretrain | FLOPs | Accuracy | Method | Pretrain | FLOPs | Accuracy |
|---|---|---|---|---|---|---|---|
| DSIFT<br>+ SVM Linear | N | 1.28G | 49.3% | DSIFT<br>+ SVM Chiˆ2 | N | 1.28G | 54.0% |
| AlexNet<br>(Fine-tuning) | ImageNet | 723M | 66.0% | AlexNet<br>(Feature Extractor)<br>+ SVM Linear | ImageNet | 724M | 78.0% |
| AlexNet<br>(Fine-tuning) | VGG-Face | 723M | 73.3% | AlexNet<br>(Feature Extractor)<br>+ SVM Linear | VGG-Face | 724M | 86.0% |
| VGG16<br>(Fine-tuning) | ImageNet | 15.47G | 72.0% | VGG16<br>(Feature Extractor)<br>+ SVM Linear | ImageNet | 15.47G | 78.0% |
| **VGG16<br>(Fine-tuning)** | **VGG-Face** | **15.47G** | **86.7%** | **Best:VGG16<br>(Feature Extractor)<br>+ SVM Linear** | **VGG-Face** | **15.47G** | **93.3%** |
| ResNet50<br>(Fine-tuning) | ImageNet | 3.87G | 77.3% | ResNet50<br>(Feature Extractor)<br>+ SVM Linear | ImageNet | 3.87G | 86.7% |
| ResNet50<br>(Fine-tuning) | VGG-Face | 3.87G | 82.0% | ResNet50<br>(Feature Extractor)<br>+ SVM Linear | VGG-Face | 3.87G | 88.7% |

is a widely used dataset of roughly 10 million photos from 100,000 individuals for face recognition. VGGFace2 is a large-scale dataset containing more than 3.3 million face images over 9K identities for face recognition. Table 5.4 lists the results of ResNet50 and SE-ResNet50 models pretrained on the different datasets. SE-ResNet50 has a more complex structure but does not get better results than ResNet50 here, which is according with the fact that "VGG-Face" model achieves the best results in our experiments. The results indicate that pretraining on more task-related datasets can improve the performance in this task. The ResNet50 pretrained on MS-Celeb-1M and finetuned on VGGFace2 improves its accuracy from 86.7% (ImageNet) to 92.7% which is closest to the best result. The testing accuracy of the specialists published [129] is about 80%. DTL2: CNN as a feature extractor still outperforms clinicians, which is promising.

Regarding the time complexity (see TABLE 5.1-5.4), as mentioned in the theoretical part, the time complexity of DTL1 and DTL2 are both smaller than that of the corresponding pretrained model, and the time complexity of DTL2 is

Table 5.4: **Multiclass classification advanced results on the detection of four conditions**

| Method(DTL2) | Pretrain | FLOPs | Accuracy |
|---|---|---|---|
| ResNet50<br>(Feature Extractor)<br>+ SVM Linear | VGGFace2 | 3.87G | 82.0% |
| SE-ResNet50<br>(Feature Extractor)<br>+ SVM Linear | VGGFace2 | 3.88G | 84.7% |
| **ResNet50<br>(Feature Extractor)<br>+ SVM Linear** | **VGGFace2<br>&<br>MS-Celeb-1M** | **3.87G** | **92.7%** |
| SE-ResNet50<br>(Feature Extractor)<br>+ SVM Linear | VGGFace2<br>&<br>MS-Celeb-1M | 3.88G | 90.0% |

a bit larger than that of DTL1. Since the FLOPs of CNN models are almost more than a few hundred millions now, the difference in FLOPs values of the adapted model and its corresponding pretrained model shown in tables is not obvious.

From these experiments, we can conclude that the performance by deep learning methods are overall better than the results by traditional machine learning methods as expected. The difference is more expressive for the multiclass classification task. In the case of the small dataset of facial diagnosis, DTL2: CNN as a feature extractor is more appropriate than DTL1: Fine-tuning. Furthermore, it is because of the similarity between the target domain and the source domain of deep learning models pretrained for face recognition that the better performance can be reached by deep transfer learning methods. Deep learning models pretrained on more datasets for face recognition can achieve a better performance on facial diagnosis by deep transfer learning.

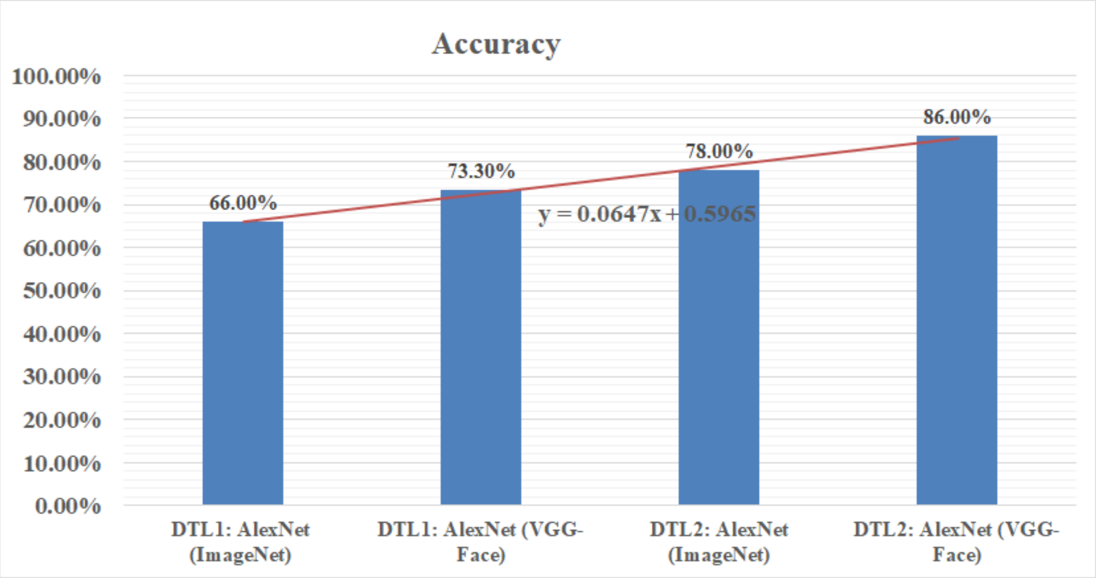


Figure 5.4: AlexNet accuracy comparison in DTL1 and DTL2

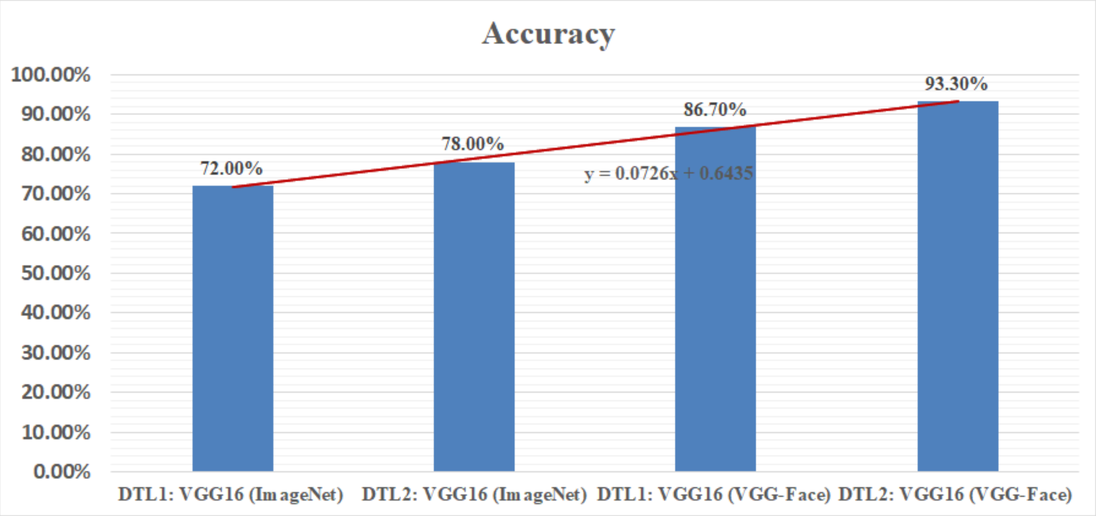


Figure 5.5: VGG16 accuracy comparison in DTL1 and DTL2

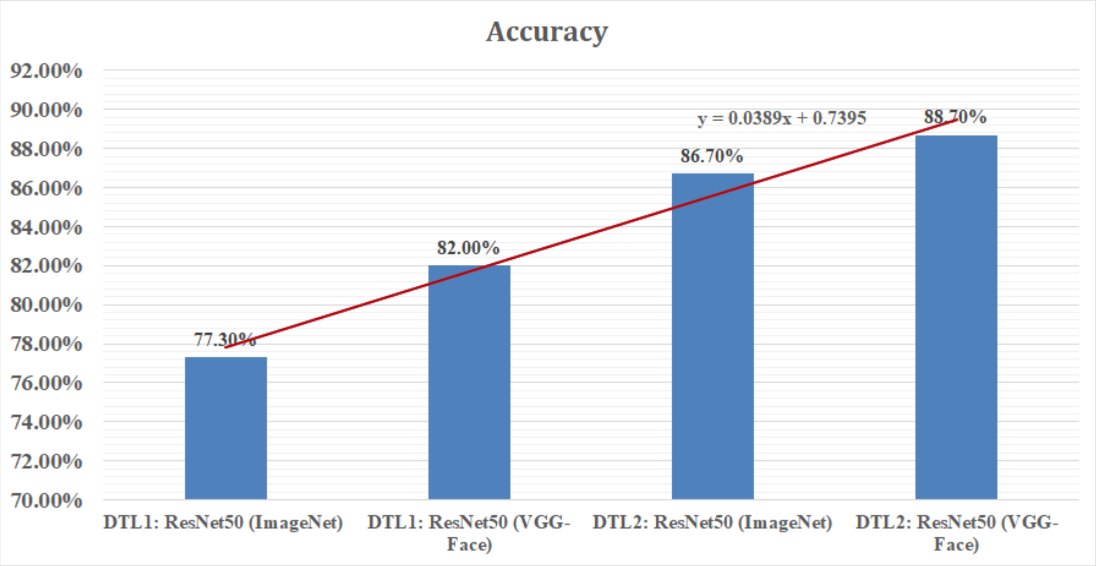


Figure 5.6: ResNet50 accuracy comparison in DTL1 and DTL2

### 5.1.2 Summary

More and more studies have shown that computer-aided facial diagnosis is a promising way for condition screening and detection. In this work, we propose deep transfer learning from face recognition methods to realize computer-aided facial diagnosis definitely and validate them on single condition and multiple conditions with the healthy control. The experimental results of above 90% accuracy have proven that CNN as a feature extractor is the most appropriate deep transfer learning method in the case of the small dataset of facial diagnosis. It can solve the general problem of insufficient data in the facial diagnosis area to a certain extent. In the future, we will continue to discover deep learning models to perform facial diagnosis effectively with the help of data augmentation methods. We hope that more and more conditions can be detected efficiently by face photographs.

## 5.2 Pseudo RGB-D Facial Diagnosis

In this section, we apply the pseudo RGB-D facial image processing framework on the facial diagnosis on 6 conditions including acromegaly, facial nerve paralysis, Down syndrome, leprosy, thalassemia and hyperthyroidism. The Disease-Specific Faces 2 (DSF2) dataset used in this section includes aforementioned six conditions and healthy controls.

### 5.2.1 Experimental Results and Analysis

Following Pseudo RGB-D Face Recognition, for pseudo-depth generation, we utilize aforementioned D+GAN. For image fusion, we propose a wavelet soft-thresholding-based method, which is robust to noise.

**Wavelet soft-thresholding image fusion** The algorithm is implemented using *Matlab*. The first step involves performing a multilevel 2-D wavelet decomposition on each image that is to be fused. The general form of the function is [C, S] = wavedec2(X, N, wname). For the inputs, $X$ represents the input matrix, $N$ denotes the level of decomposition, and *wname* specifies the wavelet used. In this embodiment, a 4-level decomposition using the Symlets 4 wavelet function [130] is performed (see Figure 5.7). For the outputs, C represents the wavelet decomposition vector, and S is the bookkeeping matrix containing the number of coefficients by level and orientation.

The second step involves obtaining the threshold value using the equation $thr = \sqrt{2 * \log(n)}$. Here, n represents the number of input images. The general form of the function is [thr, sorh, keepapp] = ddencmp(in1, in2, x). For the inputs, $x$ represents the input 2-D matrix, *in1* denotes the mode for denoising or compression, and *in2* specifies whether to use wavelets or wavelet packets. For the outputs, *thr* indicates the threshold, sorh determines whether soft or hard

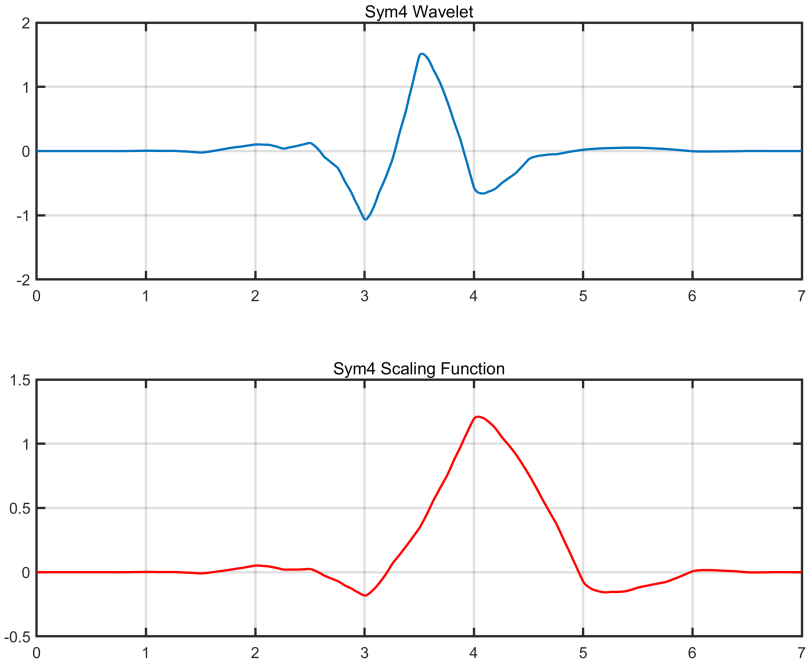


Figure 5.7: Sym4 wavelet waveform

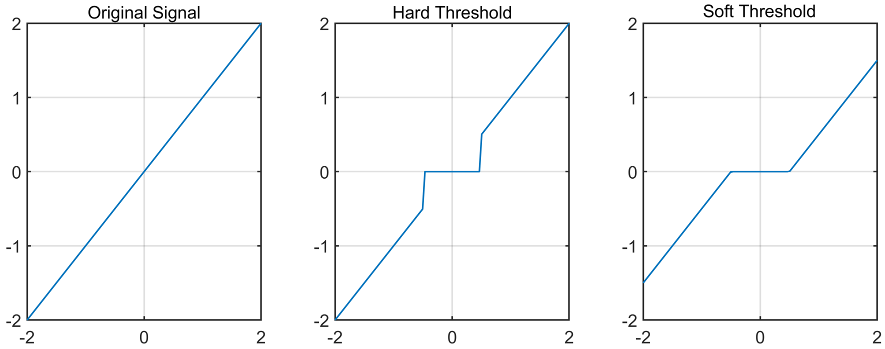


Figure 5.8: Comparison of Hard and Soft Thresholding Techniques

thresholding is used, and *keepapp* selects whether the approximation coefficients are thresholded or not for other purposes.

The third step entails performing 2-D coefficient soft thresholding [131]. The general form of the function is $NC = wthcoef2('type', C, S, N, T, SORH)$. For the inputs, *'type'* specifies the type of coefficients, $C$ represents the wavelet decomposition vector, and $S$ denotes the bookkeeping matrix containing the number of coefficients by level and orientation, which are the outputs from the first step. $N$ indicates the detail levels to be thresholded, $T$ represents the corresponding thresholds obtained in the second step, and $SORH$ determines whether soft or hard thresholding is applied (see Figure 5.8). For the outputs, $NC$ denotes the processed detail coefficients.

The fourth step involves fusing coefficients. Generally, low-frequency components represent areas in the image where brightness or grayscale values change slowly, describing the main part of the image. High-frequency components correspond to parts of the image that change drastically, describing the edges, noise, and details of the image. Two fusion strategies are adopted. The first involves averaging the corresponding coefficients of the two images. The second strategy takes the wavelet coefficients with large absolute values from the two images for high-frequency coefficients and averages the two images for low-frequency coefficients.

The final step entails performing a multilevel wavelet reconstruction of the matrix. The general form of the function is $x = waverec2(c, s, wname)$. For the inputs, $C$ represents the processed wavelet decomposition vector, and $S$ denotes the processed bookkeeping matrix containing the number of coefficients by level and orientation, which are outputs from the first step. *wname* specifies the wavelet name used. For the output, $x$ represents the reconstructed matrix.

The pseudo-code for the wavelet soft-thresholding image fusion is depicted as follows:

```
Func WST Fusion2D(Img1, Img2):
Begin
 // Perform wavelet 2D decomposition
 C1, S1 = WaveletDecomposition(Img1)
 C2, S2 = WaveletDecomposition(Img2)
 // Compute threshold
 thr = sqrt(2 * log(2))
 // Perform soft-thresholding
 C1 = SoftThreshold(C1, thr)
 C2 = SoftThreshold(C2, thr)
 // Combine coefficients
 Cf = CombineCoefficients(C1, C2)
 // Wavelet reconstruction
 FusedImage = WaveletReconstruction(Cf)
 return FusedImage
 End
```

For feature extraction and classification, we initially fine-tune the pre-trained models of FaceNet [54] and InsightFace [30], a process that mirrors the approach utilized in the study 'Pseudo RGB-D Face Recognition'. The employed InsightFace model includes two structures: InsightFace: IResNet34 and InsightFace: IResNet100, both pre-trained with the MS1MV2 dataset. Meanwhile, FaceNet models are pre-trained with the CASIA-WebFace and VGG-Face2 datasets. Fine-grained classification is applicable to classification tasks characterized by substantial intra-class differences and minor inter-class differences. Inspired by the philosophy of fine-grained classification, we introduce a bilinear operation into both InsightFace and FaceNet processing models, as illustrated in Figure 5.9. The mathematical process can be represented by following equations:

$$Bi\left(l, I, u, v\right) = u^T(l, I)v(l, I) \tag{5.1}$$

where $Bi$ represents the bilinear feature combination, $l$ denotes location, $I$ is the input image, and $u$ and $v$ are two feature functions.

$$\psi\left(I\right) = \sum_{l} Bi(l, I, u, v) \tag{5.2}$$

where $\psi$ represents the feature map for the entire image.

$$x = vec(\psi(I)) \tag{5.3}$$

$$y = sign(x)\sqrt{|x|} \tag{5.4}$$

$$z = \frac{y}{\|y\|_2} \tag{5.5}$$

where $z$ represents the final fused feature utilized for classification.

For a more comprehensive assessment, three different cases are performed.

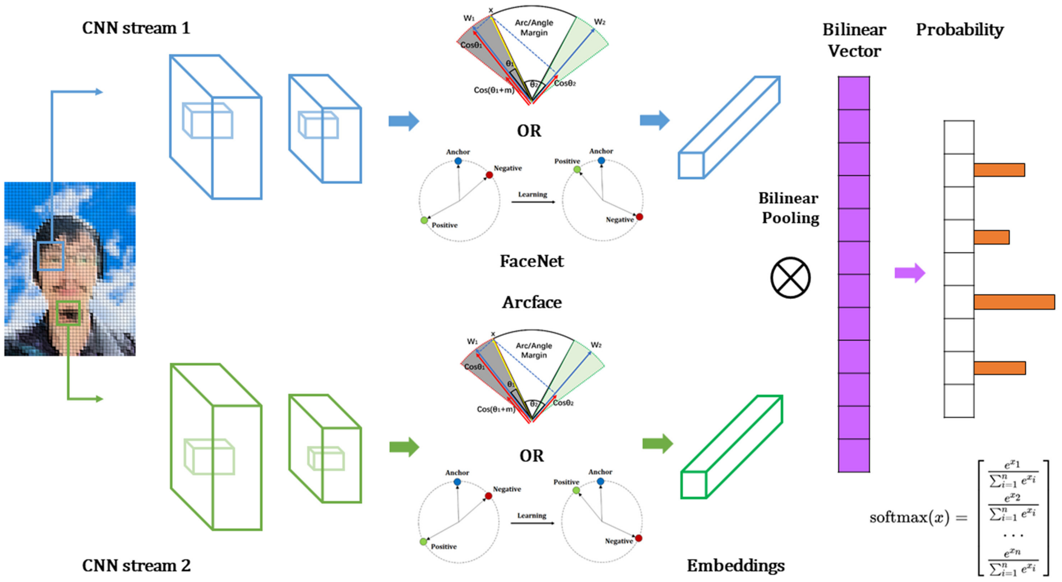


Figure 5.9: Bilinear model for fine-grained facial diagnosis

Case 1 is that 45 images of each category for training, and 40 images of each category for testing. Case 2 is that 50 images of each category for training, and 35 images of each category for testing. Case 3 is that 55 images of each category for training, and 30 images of each category for testing.

For evaluation, in addition to accuracy being of significance for facial diagnosis, Matthews Correlation Coefficient is selected as an alternative indicator.

The Matthews Correlation Coefficient (MCC) [132] is a widely used metric for evaluating the performance of classification models, including multi-class classification tasks. It takes into account the confusion matrix to provide a comprehensive assessment of classification accuracy. The MCC ranges from -1 to 1, with -1 indicating a completely incorrect classification, 1 indicating a perfect classification, and 0 signifying a random classification. For binary classification, the MCC is calculated using the formula:

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{5.6}$$

where TP means True Positives, TN means True Negatives, FP means False Positives, and FN means False Negatives. For multiclass classification problems, the MCC can be calculated by treating each class as binary (i.e., class i versus the rest) and averaging the MCCs for each binary problem.

Compared to the RGB mode of the InsightFace34 model, the Pseudo RGB-D mode of the Bilinear-InsightFace model has different degrees of improvement in the aforementioned three cases, which is as shown in Table 5.5, Table 5.6 and Table 5.7.

Compared to the RGB mode of the InsightFace100 model, the Pseudo RGB-D mode of the Bilinear-InsightFace model does not have an improvement effect in all cases, which is as shown in Table 5.7.

Compared to the RGB mode of the FaceNet (CASIA-WebFace) model, the Pseudo RGB-D mode of the Bilinear-FaceNet model has different degrees of improvement in the aforementioned three cases, which is as shown in Table 5.8, Table 5.9 and Table 5.10.

Compared to the RGB mode of the FaceNet (VGG-Face2) model, the Pseudo RGB-D mode of the Bilinear-InsightFace model does not have an improvement effect in all cases, which is as shown in Table 5.9 and Table 5.10.

It is worth noting that sometimes using only the bilinear model can achieve better results.

Pseudo RGB-D facial image processing framework presents a modular process. Algorithms within the module lists can be selected for preprocessing, depth-generating, image fusion, and feature extraction & recognition. There is no need to reinvent the wheel. Nowadays, there exist many models that have been trained with a lot of computing resources. Especially for facial diagnosis, pretrained DL models for face recognition should be made full use of. Pseudo RGB-D facial image processing framework can promote facial diagnosis performance with small

Table 5.5: InsightFace Performance @Training: 45; Testing: 40

| Mode | Method | Accuracy Score (ACC) | Matthews Correlation Coefficient (MCC) |
|---|---|---|---|
| RGB | InsightFace: IResNet34 (MS1MV2) | 0.557 | 0.497 |
| | Bilinear-InsightFace: IResNet34 (MS1MV2) | 0.596 | 0.532 |
| | InsightFace: IResNet100 (MS1MV2) | 0.557 | 0.492 |
| | Bilinear-InsightFace: IResNet100 (MS1MV2) | **0.711** | **0.667** |
| Pseudo RGB-D | InsightFace: IResNet34 (MS1MV2) | 0.529 | 0.459 |
| | Bilinear-InsightFace: IResNet34 (MS1MV2) | **0.636** | **0.575** |
| | InsightFace: IResNet100 (MS1MV2) | 0.571 | 0.509 |
| | Bilinear-InsightFace: IResNet100 (MS1MV2) | 0.682 | 0.632 |

Table 5.6: InsightFace Performance @Training: 50; Testing: 35

| Mode | Method | Accuracy Score (ACC) | Matthews Correlation Coefficient (MCC) |
|---|---|---|---|
| RGB | InsightFace: IResNet34 (MS1MV2) | 0.506 | 0.437 |
| | Bilinear-InsightFace: IResNet34 (MS1MV2) | 0.612 | 0.550 |
| | InsightFace: IResNet100 (MS1MV2) | 0.629 | 0.578 |
| | Bilinear-InsightFace: IResNet100 (MS1MV2) | **0.718** | **0.675** |
| Pseudo RGB-D | InsightFace: IResNet34 (MS1MV2) | 0.514 | 0.451 |
| | Bilinear-InsightFace: IResNet34 (MS1MV2) | **0.620** | **0.563** |
| | InsightFace: IResNet100 (MS1MV2) | 0.559 | 0.498 |
| | Bilinear-InsightFace: IResNet100 (MS1MV2) | 0.678 | 0.625 |

Table 5.7: InsightFace Performance @Training: 55; Testing: 30

| Mode | Method | Accuracy Score (ACC) | Matthews Correlation Coefficient (MCC) |
|---|---|---|---|
| RGB | InsightFace: IResNet34 (MS1MV2) | 0.581 | 0.517 |
| | Bilinear-InsightFace: IResNet34 (MS1MV2) | 0.610 | 0.546 |
| | InsightFace: IResNet100 (MS1MV2) | 0.719 | 0.675 |
| | Bilinear-InsightFace: IResNet100 (MS1MV2) | **0.742** | **0.702** |
| Pseudo RGB-D | InsightFace: IResNet34 (MS1MV2) | 0.562 | 0.495 |
| | Bilinear-InsightFace: IResNet34 (MS1MV2) | **0.610** | **0.547** |
| | InsightFace: IResNet100 (MS1MV2) | 0.657 | 0.604 |
| | Bilinear-InsightFace: IResNet100 (MS1MV2) | 0.657 | 0.601 |

Table 5.8: FaceNet Performance @Training: 45; Testing: 40

| Mode | Method | Accuracy (ACC) | Matthews correlation coefficient (MCC) |
|---|---|---|---|
| RGB | FaceNet: Inception ResNet v1 (CASIA-WebFace) | 0.411 | **0.352** |
| | Bilinear-FaceNet: Inception ResNet v1 (CASIA-WebFace) | 0.407 | 0.323 |
| | FaceNet: Inception ResNet v1 (VGG-Face2) | 0.382 | 0.322 |
| | Bilinear-FaceNet: Inception ResNet v1 (VGG-Face2) | **0.454** | **0.372** |
| Pseudo RGB-D | FaceNet: Inception ResNet v1 (CASIA-WebFace) | 0.357 | 0.284 |
| | Bilinear-FaceNet: Inception ResNet v1 (CASIA-WebFace) | **0.421** | 0.342 |
| | FaceNet: Inception ResNet v1 (VGG-Face2) | 0.375 | 0.297 |
| | Bilinear-FaceNet: Inception ResNet v1 (VGG-Face2) | 0.404 | 0.340 |

Table 5.9: FaceNet Performance @Training: 50; Testing: 35

| Mode | Method | Accuracy (ACC) | Matthews correlation coefficient (MCC) |
|---|---|---|---|
| RGB | FaceNet: Inception ResNet v1 (CASIA-WebFace) | 0.400 | 0.309 |
| | Bilinear-FaceNet: Inception ResNet v1 (CASIA-WebFace) | 0.420 | 0.332 |
| | FaceNet: Inception ResNet v1 (VGG-Face2) | **0.408** | **0.331** |
| | Bilinear-FaceNet: Inception ResNet v1 (VGG-Face2) | 0.314 | 0.218 |
| Pseudo RGB-D | FaceNet: Inception ResNet v1 (CASIA-WebFace) | 0.404 | 0.311 |
| | Bilinear-FaceNet: Inception ResNet v1 (CASIA-WebFace) | **0.469** | **0.390** |
| | FaceNet: Inception ResNet v1 (VGG-Face2) | 0.404 | 0.322 |
| | Bilinear-FaceNet: Inception ResNet v1 (VGG-Face2) | 0.298 | 0.238 |

Table 5.10: FaceNet Performance @Training: 55; Testing: 30

| Mode | Method | Accuracy (ACC) | Matthews correlation coefficient (MCC) |
|---|---|---|---|
| RGB | FaceNet: Inception ResNet v1 (CASIA-WebFace) | 0.376 | 0.298 |
| | Bilinear-FaceNet: Inception ResNet v1 (CASIA-WebFace) | 0.519 | 0.459 |
| | FaceNet: Inception ResNet v1 (VGG-Face2) | 0.386 | 0.292 |
| | Bilinear-FaceNet: Inception ResNet v1 (VGG-Face2) | **0.419** | **0.331** |
| Pseudo RGB-D | FaceNet: Inception ResNet v1 (CASIA-WebFace) | 0.381 | 0.320 |
| | Bilinear-FaceNet: Inception ResNet v1 (CASIA-WebFace) | **0.552** | **0.483** |
| | FaceNet: Inception ResNet v1 (VGG-Face2) | 0.390 | 0.324 |
| | Bilinear-FaceNet: Inception ResNet v1 (VGG-Face2) | 0.371 | 0.306 |

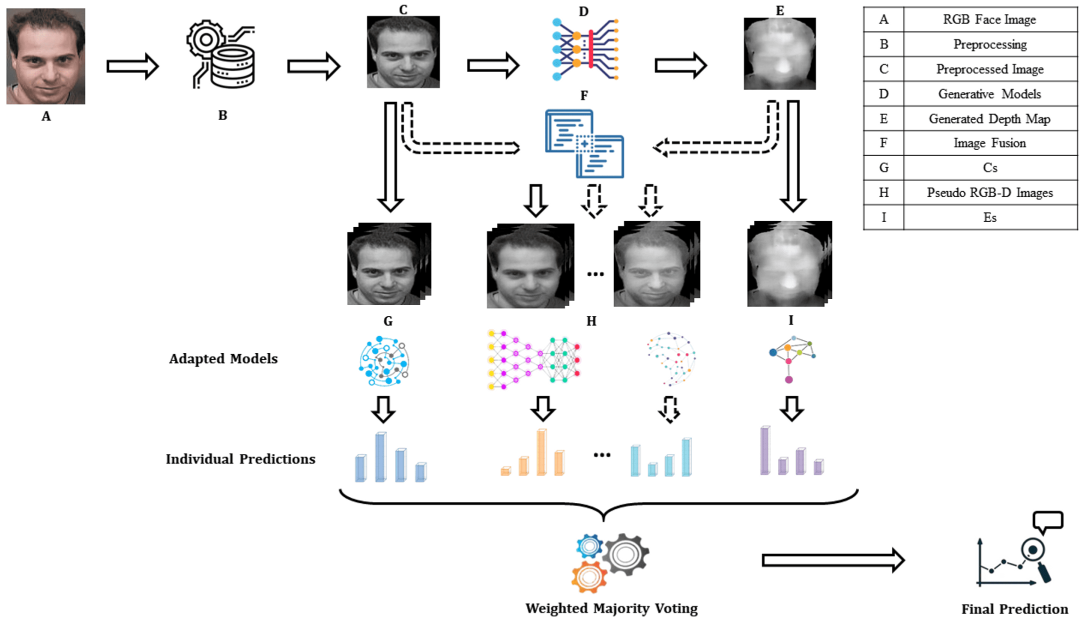


Figure 5.10: Simulated multimodal facial image processing framework for enhanced performance

training data.

## 5.3 Simulated Multimodal Deep Facial Diagnosis

In the last chapter, we used the pseudo RGB-D face recognition framework in the face diagnosis task. But the results show that the pseudo RGB-D face recognition framework does not improve performance for every selected model. In order to better utilize the pseudo-depth generated, we propose a simulated multimodal facial image processing framework, which is shown as Figure 5.10.

It makes RGB images to generate multiple simulated multi-modal images through the generative model and image fusion strategies, and trains and predicts the simulated multi-modal images respectively. It obtains the final prediction result through weighted majority voting of individual results. Behind this approach lies the thought of self-evolution.

In our practice, this method has a significant improvement over the RGB face diagnosis. In our embodiment, we use D+GAN to generate the pseudo-depth map, and perform the wavelet soft-thresholding image fusion aforementioned before on the pseudo-depth map with 2 different strategies.

Strategy 1 is to use the mean value for both low-frequency and high-frequency parts of the two images. By Strategy 1, we get Pseudo RGB-D 1 images.

Strategy 2 is that in the part of high-frequency coefficients, the wavelet coefficients with large absolute values in the two pictures are applied; in the part of low-frequency coefficients, the mean wavelet coefficients of the two pictures are applied. Through Strategy 2, we get Pseudo RGB-D 2 images.

In our embodiment, we have four modes of images which are RGB images, Pseudo RGB-D 1 images, Pseudo RGB-D 2 images and pseudo-depth images to perform training and predicting respectively. Four modes of image examples are displayed in Figure 5.11. For comparison, all models were trained for 150 epochs with a low learning rate, and their loss functions all converged. The final prediction results are obtained by weighted majority voting of the predictions from each model. The prediction weights assigned are positively correlated with the accuracy of the models on the training set.

### 5.3.1 Experimental Results and Analysis

For a more comprehensive assessment, three different cases are performed. Case 1 is that 45 images of each category for training, and 40 images of each category for testing. Case 2 is that 50 images of each category for training, and 35 images of each category for testing. Case 3 is that 55 images of each category for training, and 30 images of each category for testing. For fair comparison, the training

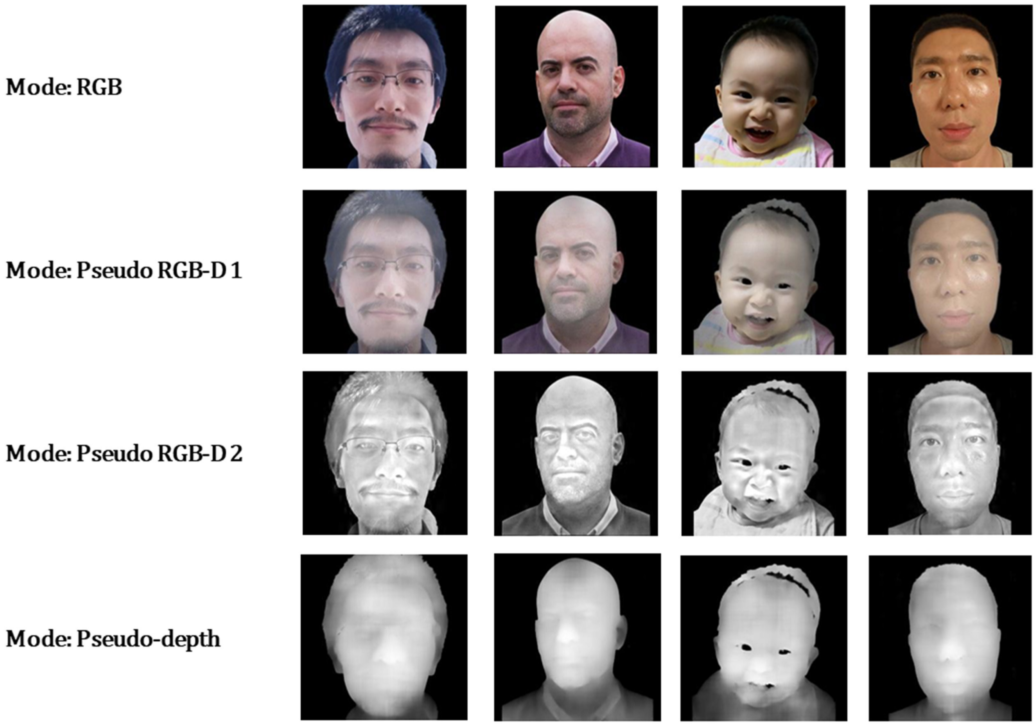


Figure 5.11: Simulated multimodal image samples in DSF2 dataset

period of all models in this section is set to 200 epochs.

Figures 5.12-5.17 visually illustrate the performance changes of various models under different modes.

In an experiment using the FaceNet: Inception ResNet v1 (CASIA-Webface) model, with each category containing 45 training images and 40 testing images, the simulated multimodal framework improved the ACC by 0.73% and the MCC by 0.85% in comparison to the RGB mode, as demonstrated in Figures 5.12 and 5.13.

In an experiment using the bilinear FaceNet: Inception ResNet v1 (CASIA-Webface) model, with each category containing 45 training images and 40 testing images, the simulated multimodal framework improved the ACC by 14.00% and the MCC by 18.58% in comparison to the RGB mode, as demonstrated in Figures 5.12 and 5.13.

In an experiment using the FaceNet: Inception ResNet v1 (VGG-Face2) model, with each category containing 45 training images and 40 testing images, the simulated multimodal framework improved the ACC by 3.66% and the MCC by 3.11% in comparison to the RGB mode, as demonstrated in Figures 5.12 and 5.13.

In an experiment using the bilinear FaceNet: Inception ResNet v1 (VGG-Face2) model, with each category containing 50 training images and 35 testing images, the simulated multimodal framework improved the ACC by 2.86% and the MCC by 5.09% in comparison to the RGB mode, as demonstrated in Figures 5.14 and 5.15.

In an experiment using the InsightFace: IResNet34 (MS1MV2) model, with each category containing 45 training images and 40 testing images, the simulated multimodal framework improved the ACC by 0.72% and the MCC by 0.60% in comparison to the RGB mode, as demonstrated in Figures 5.12 and 5.13.

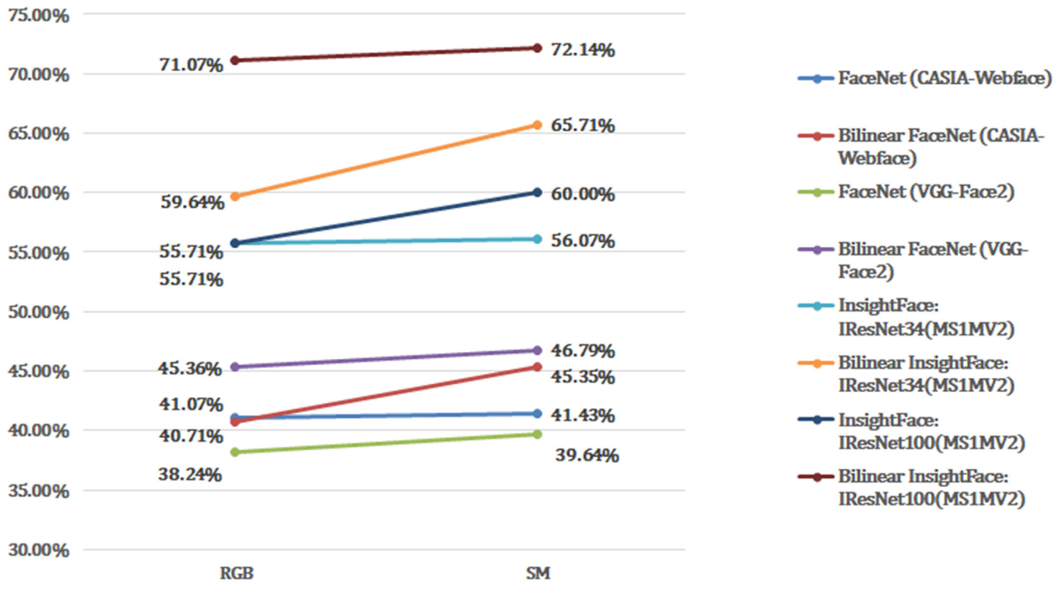


Figure 5.12: Accuracy of Simulated Multimodal Framework for Facial Diagnosis in Case 1

In an experiment using the bilinear InsightFace: IResNet34 (MS1MV2) model, with each category containing 45 training images and 40 testing images, the simulated multimodal framework improved the ACC by 10.24% and the MCC by 13.16% in comparison to the RGB mode, as demonstrated in Figures 5.12 and 5.13.

In an experiment using the InsightFace: IResNet100 (MS1MV2) model, with each category containing 45 training images and 40 testing images, the simulated multimodal framework reduced the ACC by 7.71% and the MCC by 9.76% in comparison to the RGB mode, as demonstrated in Figures 5.12 and 5.13.

In an experiment using the bilinear InsightFace: IResNet100 (MS1MV2) model, with each category containing 45 training images and 40 testing images, the simulated multimodal framework improved the ACC by 1.41% and the MCC by 1.65% in comparison to the RGB mode, as demonstrated in Figures 5.12 and 5.13.

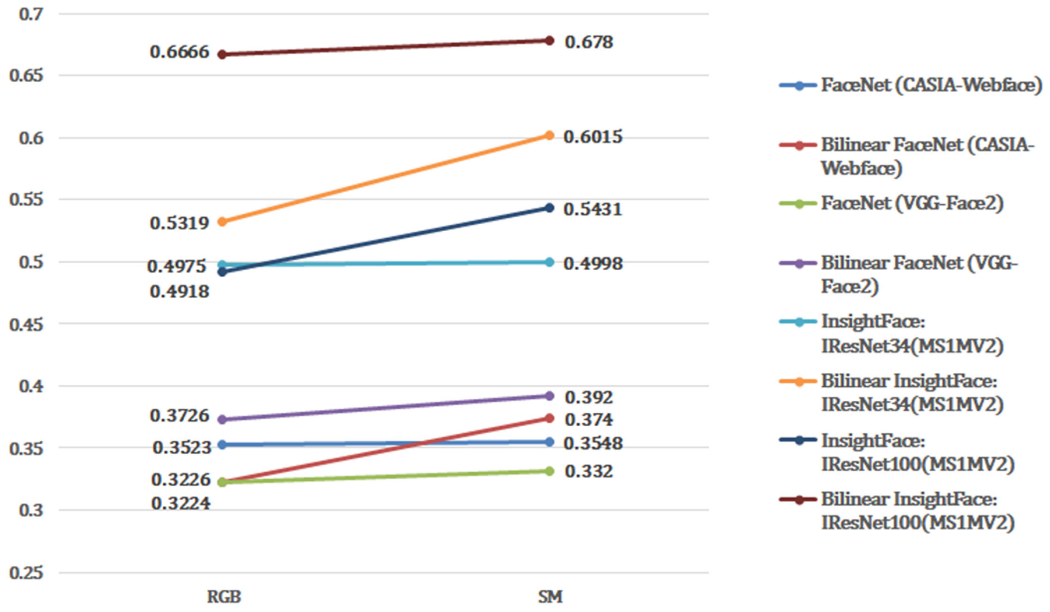


Figure 5.13: MCC of Simulated Multimodal Framework for Facial Diagnosis in Case 1

In an experiment using the FaceNet: Inception ResNet v1 (CASIA-Webface) model, with each category containing 50 training images and 35 testing images, the simulated multimodal framework improved the ACC by 15.25% and the MCC by 22.65% in comparison to the RGB mode, as demonstrated in Figures 5.14 and 5.15.

In an experiment using the bilinear FaceNet: Inception ResNet v1 (CASIA-Webface) model, with each category containing 50 training images and 35 testing images, the simulated multimodal framework improved the ACC by 2.34% and the MCC by 2.79% in comparison to the RGB mode, as demonstrated in Figures 5.14 and 5.15.

In an experiment using the FaceNet: Inception ResNet v1 (VGG-Face2) model, with each category containing 50 training images and 35 testing images, the simulated multimodal framework improved the ACC by 12.01% and the MCC by 16.92% in comparison to the RGB mode, as demonstrated in Figures 5.14 and 5.15.

In an experiment using the bilinear FaceNet: Inception ResNet v1 (VGG-Face2) model, with each category containing 50 training images and 35 testing images, the simulated multimodal framework improved the ACC by 6.69% and the MCC by 10.55% in comparison to the RGB mode, as demonstrated in Figures 5.14 and 5.15.

In an experiment using the InsightFace: IResNet34 (MS1MV2) model, with each category containing 55 training images and 30 testing images, the simulated multimodal framework improved the ACC by 3.17% and the MCC by 5.25% in comparison to the RGB mode, as demonstrated in Figures 5.14 and 5.15.

In an experiment using the bilinear InsightFace: IResNet34 (MS1MV2) model, with each category containing 50 training images and 35 testing images, the simulated multimodal framework improved the ACC by 10.67% and the MCC by 14.00% in comparison to the RGB mode, as demonstrated in Figures 5.14 and 5.15.

In an experiment using the InsightFace: IResNet100 (MS1MV2) model, with each category containing 50 training images and 35 testing images, the simulated multimodal framework reduced the ACC by 1.27% and the MCC by 1.56% in comparison to the RGB mode, as demonstrated in Figures 5.14 and 5.15.

In an experiment using the bilinear InsightFace: IResNet100 (MS1MV2) model, with each category containing 50 training images and 35 testing images, the simulated multimodal framework improved the ACC by 3.48% and the MCC by 3.39% in comparison to the RGB mode, as demonstrated in Figures 5.14 and 5.15.

In an experiment using the FaceNet: Inception ResNet v1 (CASIA-Webface) model, with each category containing 55 training images and 30 testing images,

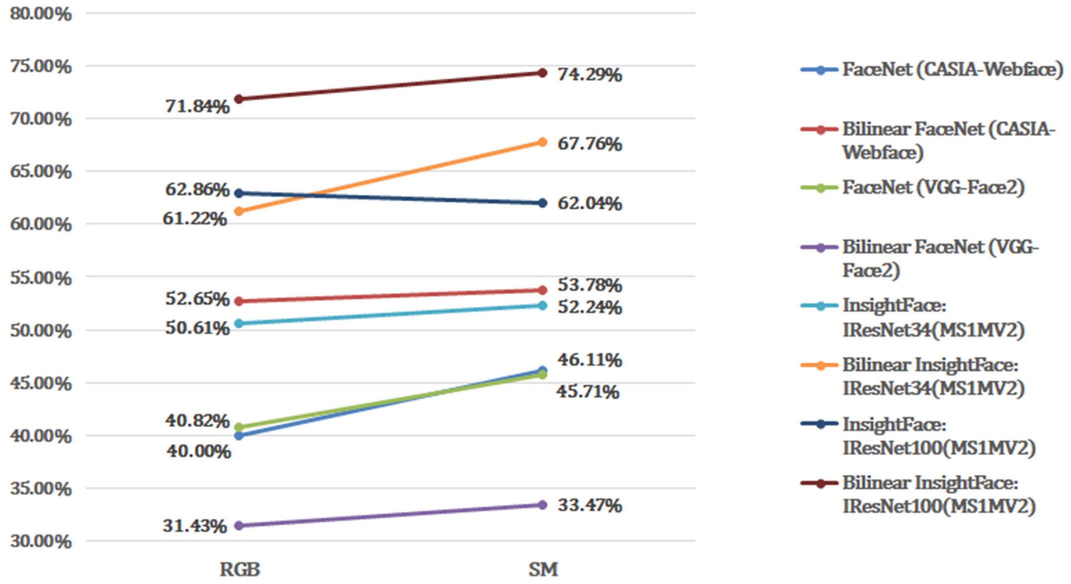


Figure 5.14: Accuracy of Simulated Multimodal Framework for Facial Diagnosis in Case 2

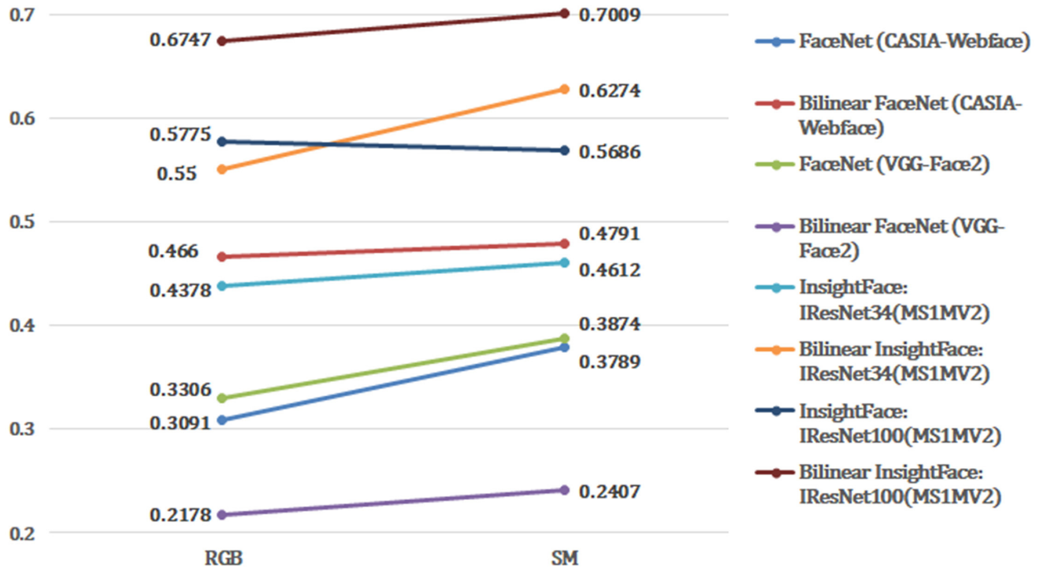


Figure 5.15: MCC of Simulated Multimodal Framework for Facial Diagnosis in Case 2

the simulated multimodal framework improved the ACC by 16.18% and the MCC by 24.83% in comparison to the RGB mode, as demonstrated in Figures 5.16 and 5.17.

In an experiment using the bilinear FaceNet: Inception ResNet v1 (CASIA-Webface) model, with each category containing 55 training images and 30 testing images, the simulated multimodal framework improved the ACC by 12.91% and the MCC by 14.60% in comparison to the RGB mode, as demonstrated in Figures 5.16 and 5.17.

In an experiment using the FaceNet: Inception ResNet v1 (VGG-Face2) model, with each category containing 55 training images and 30 testing images, the simulated multimodal framework improved the ACC by 8.55% and the MCC by 14.38% in comparison to the RGB mode, as demonstrated in Figures 5.16 and 5.17.

In an experiment using the bilinear FaceNet: Inception ResNet v1 (VGG-Face2) model, with each category containing 55 training images and 30 testing images, the simulated multimodal framework improved the ACC by 10.26% and the MCC by 16.31% in comparison to the RGB mode, as demonstrated in Figures 5.16 and 5.17.

In an experiment using the InsightFace: IResNet34 (MS1MV2) model, with each category containing 55 training images and 30 testing images, the simulated multimodal framework improved the ACC by 2.41% and the MCC by 3.10% in comparison to the RGB mode, as demonstrated in Figures 5.16 and 5.17.

In an experiment using the bilinear InsightFace: IResNet34 (MS1MV2) model, with each category containing 55 training images and 30 testing images, the simulated multimodal framework improved the ACC by 6.25% and the MCC by 8.24% in comparison to the RGB mode, as demonstrated in Figures 5.16 and 5.17.

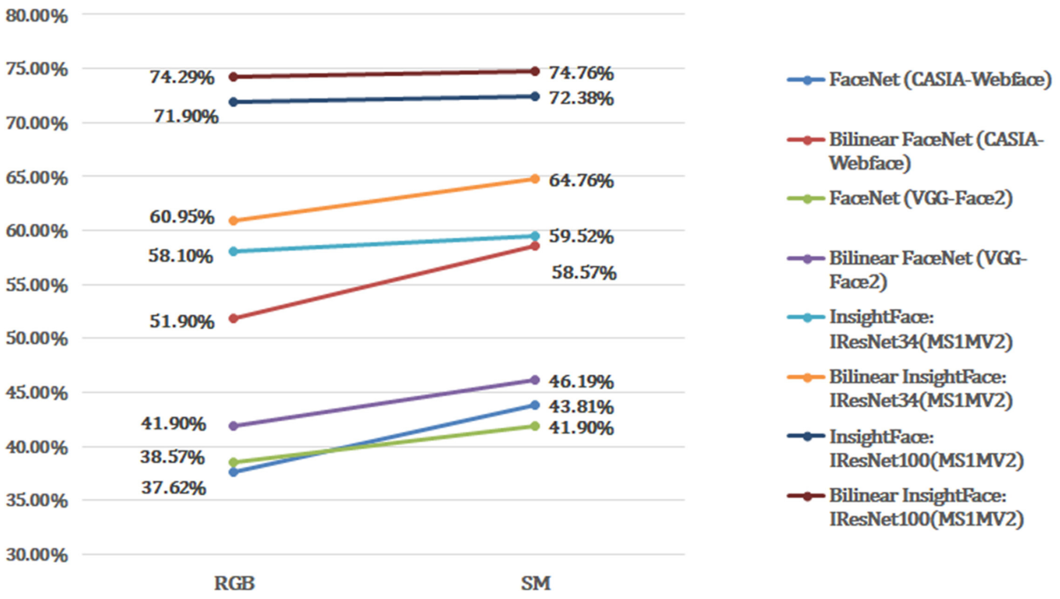


Figure 5.16: Accuracy of Simulated Multimodal Framework for Facial Diagnosis in Case 3

In an experiment using the InsightFace: IResNet100 (MS1MV2) model, with each category containing 55 training images and 30 testing images, the simulated multimodal framework improved the ACC by 0.70% and the MCC by 0.88% in comparison to the RGB mode, as demonstrated in Figures 5.16 and 5.17.

In an experiment using the bilinear InsightFace: IResNet100 (MS1MV2) model, with each category containing 55 training images and 30 testing images, the simulated multimodal framework improved the ACC by 0.63% and the MCC by 0.60% in comparison to the RGB mode, as demonstrated in Figures 5.16 and 5.17.

The experimental results for the three cases are listed in Tables 5.11-5.13. From the tables, it is observed that for the SM mode enhancing the RGB mode, out of the 24 experiments conducted, only one case did not show any improvement, which resulted in an effectiveness rate of 95.83%. In these 24 experiments, the ACC improved by an average of approximately 6.22%, while the MCC im-

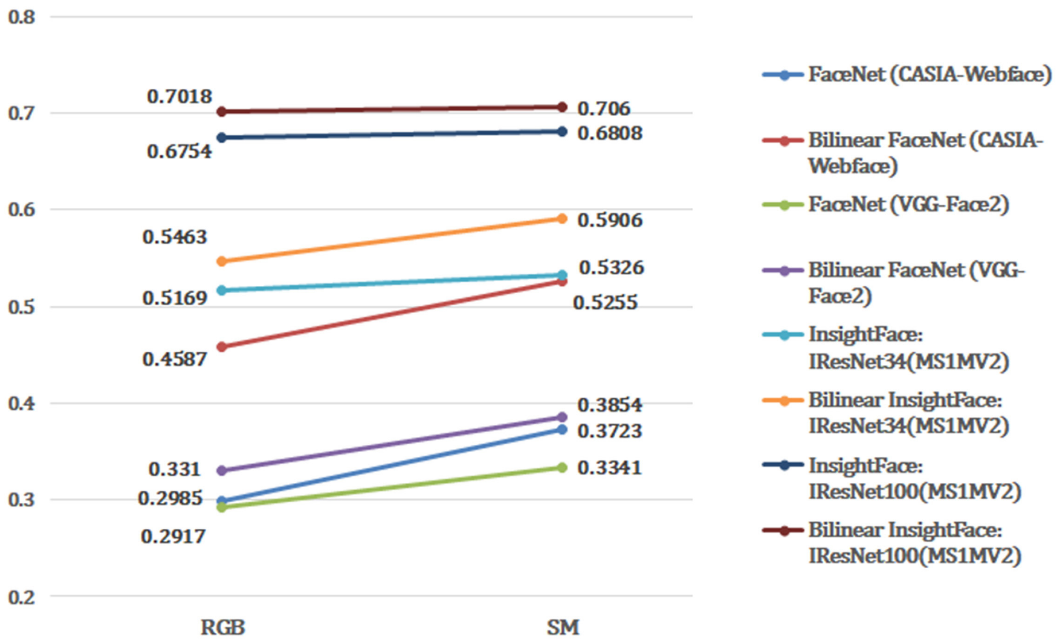


Figure 5.17: MCC of Simulated Multimodal Framework for Facial Diagnosis in Case 3

proved by an average of about 8.67%. From the tables, it is observed that in terms of the improvement effect of the SM mode and the bilinear structure model on the RGB mode and non-bilinear structure models, only one out of 12 experiments showed no improvement, leading to an effectiveness rate of 91.67%. Moreover, in these 12 experiments, the ACC improved by an average of approximately 19.97%, and the MCC improved by an average of about 25.50%.

Table 5.11: Comparison results of models in Case 1

| Train-test split ratio | Model | Evaluation metrics | Mode | |
|---|---|---|---|---|
| | | | RGB | SM |
| 45:40 | FaceNet (CASIA-Webface) | ACC | 41.07% | 41.43% |
| | | MCC | 0.3523 | 0.3548 |
| | Bilinear FaceNet (CASIA-Webface) | ACC | 40.71% | 45.35% |
| | | MCC | 0.3226 | 0.3740 |
| | FaceNet (VGG-Face2) | ACC | 38.24% | 39.64% |
| | | MCC | 0.3224 | 0.3320 |
| | Bilinear FaceNet (VGG-Face2) | ACC | 45.36% | 46.79% |
| | | MCC | 0.3726 | 0.3920 |
| | InsightFace: IResNet34 (MS1MV2) | ACC | 55.71% | 56.07% |
| | | MCC | 0.4975 | 0.4998 |
| | Bilinear InsightFace: IResNet34 (MS1MV2) | ACC | 59.64% | 65.71% |
| | | MCC | 0.5319 | 0.6015 |
| | InsightFace: IResNet100 (MS1MV2) | ACC | 55.71% | 60.00% |
| | | MCC | 0.4918 | 0.5431 |
| | Bilinear InsightFace: IResNet100 (MS1MV2) | ACC | 71.07% | **72.14%** |
| | | MCC | 0.6666 | **0.6780** |

Table 5.12: Comparison results of models in Case 2

| Train-test split ratio | Model | Evaluation metrics | Mode | |
|---|---|---|---|---|
| | | | RGB | SM |
| 50:35 | FaceNet (CASIA-Webface) | ACC | 40.00% | 46.11% |
| | | MCC | 0.3091 | 0.3789 |
| | Bilinear FaceNet (CASIA-Webface) | ACC | 52.65% | 53.78% |
| | | MCC | 0.4660 | 0.4791 |
| | FaceNet (VGG-Face2) | ACC | 40.82% | 45.71% |
| | | MCC | 0.3306 | 0.3874 |
| | Bilinear FaceNet (VGG-Face2) | ACC | 31.43% | 33.47% |
| | | MCC | 0.2178 | 0.2407 |
| | InsightFace: IResNet34 (MS1MV2) | ACC | 50.61% | 52.24% |
| | | MCC | 0.4378 | 0.4612 |
| | Bilinear InsightFace: IResNet34 (MS1MV2) | ACC | 61.22% | 67.76% |
| | | MCC | 0.5500 | 0.6274 |
| | InsightFace: IResNet100 (MS1MV2) | ACC | 62.86% | 62.04% |
| | | MCC | 0.5775 | 0.5686 |
| | Bilinear InsightFace: IResNet100 (MS1MV2) | ACC | 71.84% | **74.29%** |
| | | MCC | 0.6747 | **0.7009** |

Table 5.13: Comparison results of models in Case 3

| Train-test split ratio | Model | Evaluation metrics | Mode | |
|---|---|---|---|---|
| | | | RGB | SM |
| 55:30 | FaceNet (CASIA-Webface) | ACC | 37.62% | 43.81% |
| | | MCC | 0.2985 | 0.3723 |
| | Bilinear FaceNet (CASIA-Webface) | ACC | 51.90% | 58.57% |
| | | MCC | 0.4587 | 0.5255 |
| | FaceNet (VGG-Face2) | ACC | 38.57% | 41.90% |
| | | MCC | 0.2917 | 0.3341 |
| | Bilinear FaceNet (VGG-Face2) | ACC | 41.90% | 46.19% |
| | | MCC | 0.3310 | 0.3854 |
| | InsightFace: IResNet34 (MS1MV2) | ACC | 58.10% | 59.52% |
| | | MCC | 0.5169 | 0.5326 |
| | Bilinear InsightFace: IResNet34 (MS1MV2) | ACC | 60.95% | 64.76% |
| | | MCC | 0.5463 | 0.5906 |
| | InsightFace: IResNet100 (MS1MV2) | ACC | 71.90% | 72.38% |
| | | MCC | 0.6754 | 0.6808 |
| | Bilinear InsightFace: IResNet100 (MS1MV2) | ACC | 74.29% | **74.76%** |
| | | MCC | 0.7018 | **0.7060** |

# Chapter 6

# Discussion

”Disease”, ”disorder”, ”syndrome”, and ”condition” are terms often interchangeably used in daily use. However, in medical terminology, they bear slightly different meanings. These definitions are not strict, and their usage can overlap and vary among medical providers. A disease typically refers to an abnormality in body function or structure caused by specific etiological factors such as bacteria, viruses, environmental influences, etc. Diseases usually present with clear symptoms and signs that can be definitively diagnosed through medical examination, such as pneumonia and diabetes. Nevertheless, it’s also important to note that in some contexts, the term ”disease” may be used more broadly to encompass various long-term or short-term physical or mental health issues. The term ”condition” is a very general term and can be used to describe any situation that affects an individual’s state of health, regardless of its severity. Diseases, disorders, and syndromes can all be classified as conditions. Therefore, in this thesis, we tend to use of the term ”condition”.

In the field of facial diagnosis, the amount of data used in various studies varies greatly. In most cases, no more than 100 facial images are available for each condition category. Many studies do not clearly specify the number of images used for training and testing. Furthermore, the majority of the datasets are private and not publicly accessible. For binary classification tasks, the models reported in the literature generally perform well. However, for multi-classification tasks, there is a substantial discrepancy in the publicly reported recognition results, with accuracy rates ranging from 48% to 93%. Due to these factors, there are doubts surrounding the findings of many research studies, yet it is not possible to verify them.

In the field of facial diagnosis, due to the scarcity of training data, we first proposed and applied transfer learning from facial recognition tasks and achieved good results. Facial recognition is a relatively mature research field, and many models have reached recognition accuracies of over 99.5% in various datasets, leaving limited room for improvement. Inspired by depth estimation, we utilized pseudo-depth to enhance facial recognition performance with a limited number of training images. In the expanded facial diagnosis task dataset, the recognition task is more difficult, and using only pseudo-depth does not guarantee improved results in every experiment. Therefore, we introduced the concept of fine-grained classification and employed a bilinear model structure. In combination with pseudo-depth, facial diagnosis performance is improved in most cases. However, the improvement still has a certain degree of probability. Based on this, we proposed a simulated multimodal structure, using the same processing model structure for recognition comparison, to increase the likelihood of improvement. The research results are reproducible.

## 6.1 Ethical Discussion

In recent years, computer-aided facial diagnosis has emerged as a promising tool in the field of healthcare, enabling the identification of various diseases and conditions based on facial features. While this technology offers significant benefits in terms of diagnostic efficiency and accuracy, it also raises a number of ethical considerations that must be addressed before widespread adoption. These considerations include patient privacy and data security, potential bias in algorithms, informed consent, overreliance on the technology, transparency, interdisciplinary collaboration, potential misuse, and accessibility.

To harness the potential of computer-aided facial diagnosis while upholding the highest standards of patient care and privacy, it is crucial to engage in ongoing ethical discussions and assessments. This includes monitoring the technology's impact on patient care, ensuring diverse and representative datasets, obtaining consent from patients, fostering transparency in algorithm development, promoting interdisciplinary collaboration among experts, and addressing accessibility concerns in resource-limited settings. By maintaining an ongoing dialogue and implementing appropriate safeguards, it is possible to leverage the benefits of computer-aided facial diagnosis while mitigating potential risks and upholding the highest ethical standards.

# Chapter 7

# Conclusion

In the Deep Facial Diagnosis section, we propose the use of deep transfer learning from face recognition to perform facial diagnosis, achieving better results than those obtained using traditional methods. In the Pseudo RGB-D Face Recognition section, we propose a pseudo RGB-D facial image processing framework, with core components including pseudo-depth generation, RGB-D image fusion, and feature extraction. To generate more accurate depth maps, we also propose a generative adversarial network, D+GAN, for multi-conditional image-to-image translation using face attributes. In the Pseudo RGB-D Face Recognition section, the combination of D+GAN and NSST results in an overall improvement compared to RGB face recognition. In the Pseudo RGB-D Facial Diagnosis section, we propose applying the Pseudo RGB-D Face Image Processing to facial diagnosis. Specifically, we use wavelet-based soft thresholding image fusion in the image fusion part, and introduce the idea of fine-grained classification in the feature extraction part. We employ bilinear InsightFace and FaceNet models for training and testing to further improve the accuracy of facial diagnosis for the six conditions under consideration.

In order to extract 3D spatial features from 2D RGB images and utilize these features, we propose a pseudo RGB-D face image processing framework. The advantages of this framework include the modular selection of algorithms for core components, as well as the ability to fully leverage pre-trained models. This is particularly important in the facial diagnosis task, where training data is scarce. It is essential to make the most of pre-trained models, originally designed for face recognition, to obtain a new inference model for facial diagnosis.

In order to more effectively utilize pseudo-depth features, we propose the Simulated Multimodal Framework, which is an improved pseudo RGB-D facial image processing framework designed to enhance performance. We introduce early fusion and late fusion strategies into the Pseudo RGB-D facial image processing framework for individual training and prediction of RGB, pseudo-depth, and pseudo RGB-D images, followed by weighted majority voting. Experimental results show that this approach significantly improves the performance of RGB face diagnosis with high probability.

In future work, we plan to collect more real-world data for training and testing facial diagnosis models and prepare the necessary software and hardware for practical applications in society. Furthermore, the Simulated Multimodal Framework proposed is not limited to the field of computer-aided facial diagnosis; we believe it has significant potential in other domains, such as autonomous driving and target tracking, for classification and detection tasks. We will collaborate with various research departments to apply and evaluate the Simulated Multimodal Framework in these contexts.

# Biography

**Jin Bo** (IEEE Member) was born in Nanjing, China. He received the B.Sc. and M.Sc. degrees both from the Department of Electrical and Computer Engineering, University of Macau, Macau SAR, China. He was doing the Ph.D. research with the Visual Information Security Team, Institute of Systems and Robotics (ISR), Department of Electrical and Computer Engineering (DEEC), University of Coimbra, Coimbra, Portugal.

He has a wide range of research interests, especially in computers, robotics and genes. He mainly published the research results related with Deep Facial Diagnosis, which was awarded the national invention patent, the People's Republic of China (PRC).

### Publications during the Ph.D. study

- Journal Paper

  1. B. Jin, L. Cruz, and N. Gonçalves, "Pseudo RGB-D Face Recognition", ***IEEE Sensors Journal***, vol. 22, no. 22, pp. 21780–21794, 2022, DOI: 10.1109/JSEN.2022.3197235.
  2. B. Jin, L. Cruz, and N. Gonçalves, "Deep facial diagnosis: deep transfer learning from face recognition to facial diagnosis", ***IEEE Access***, vol. 8, pp. 123649–123661, 2020. DOI: 10.1109/ACCESS.2020.3005687.

3. Z. Jiang, B. Jin and Y. Song, ”A Novel Pet Trajectory Prediction Method for Intelligent Plant Cultivation Robot”, ***IEEE Sensors Letters***, vol. 7, no. 2, pp. 1-4, Feb. 2023, Art no. 6000904, DOI: 10.1109/LSENS.2023.3238468.
4. H. Lin, Y. Han, W. Cai, and B. Jin, ”Traffic Signal Optimization Based on Fuzzy Control and Differential Evolution Algorithm”, ***IEEE Transactions on Intelligent Transportation Systems***, 2022. DOI: 10.1109/TITS.2022.3195221.
5. F. Zhang, W. Liu, L. Deng, Z. Li, Y. Wang and B. Jin, ”The Relationship Between Chinese College Student Offspring’s Physical Activity and Father Physical Activity During COVID-19 Pandemic”, ***Frontiers in Public Health***, vol. 10, 2022.
DOI: 10.3389/fpubh.2022.896087.
6. G. Wu, F. He, Y. Zhou, Y. Jing, X. Ning, C. Wang and B. Jin, ”AC-GAN: Age-Compensated Makeup Transfer Based on Homologous Continuity Generative Adversarial Network Model”, ***IET Computer Vision***, 2022. DOI: 10.1049/cvi2.12138.
7. Q. Li, L. Ma, Z. Jiang, M. Li and B. Jin, ”TECMH: Transformer-Based Cross-Modal Hashing For Fine-Grained Image-Text Retrieval”, ***Computers, Materials & Continua***, vol. 75, no. 2, pp. 3713–3728, 2023. DOI: 10.32604/cmc.2023.037463.

- Conference Paper

1. B. Jin, L. Cruz, and N. Gonçalves, ”Face Depth Prediction by the Scene Depth”, ***IEEE/ACIS 19th International Conference on Computer and Information Science (ICIS)***, pp. 42–48, 2021. DOI: 10.1109/ICIS51600.2021.9516598

# References

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, pp. 1097–1105, 2012. ix, 19, 20

[2] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European conference on computer vision*, pp. 818–833, Springer, 2014. ix, 19, 21

[3] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014. ix, 20, 22

[4] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9, 2015. ix, 20, 22

[5] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016. ix, 20, 22, 37

[6] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014. ix, 31, 32

[7] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, “Self-attention generative adversarial networks,” in *International conference on machine learning*, pp. 7354–7363, PMLR, 2019. x, 37, 40

[8] B. Jin, “Deep learning facial diagnosis system, ZL201711255031.1.” Patent, 2022. Publication of CN108806792B. 1

[9] B. Jin, L. Cruz, and N. Gonçalves, “Deep facial diagnosis: Deep transfer learning from face recognition to facial diagnosis,” *IEEE Access*, vol. 8, pp. 123649–123661, 2020. 1

[10] Y. Wang and M. Kosinski, “Deep neural networks are more accurate than humans at detecting sexual orientation from facial images.,” *Journal of personality and social psychology*, vol. 114, no. 2, p. 246, 2018. 1

[11] G. Wu, F. He, Y. Zhou, Y. Jing, X. Ning, C. Wang, and B. Jin, “Acgan: Age-compensated makeup transfer based on homologous continuity generative adversarial network model,” *IET Computer Vision*, 2022. 2

[12] Q. Li, L. Ma, Z. Jiang, M. Li, and B. Jin, “Tecmh: Transformer-based cross-modal hashing for fine-grained image-text retrieval.,” *Computers, Materials & Continua*, vol. 75, no. 2, 2023. 2

[13] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, *et al.*, “Imagenet large scale visual recognition challenge,” *International journal of computer vision*, vol. 115, no. 3, pp. 211–252, 2015. 2, 19

[14] Z. Chen, F. Silvestri, J. Wang, H. Zhu, H. Ahn, and G. Tolomei, “Relax: Reinforcement learning agent explainer for arbitrary predictive models,” in *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pp. 252–261, 2022. 2

[15] Z. Chen, F. Silvestri, G. Tolomei, J. Wang, H. Zhu, and H. Ahn, "Explain the explainer: Interpreting model-agnostic counterfactual explanations of a deep reinforcement learning agent," *IEEE Transactions on Artificial Intelligence*, pp. 1–15, 2022. 2

[16] C. Darwin, *On the origin of species, 1859.* Routledge, 2004. 3

[17] A. Downey, *Think complexity: complexity science and computational modeling.* " O'Reilly Media, Inc.", 2018. 3

[18] Z. Zhang, "Microsoft kinect sensor and its effect," *IEEE multimedia*, vol. 19, no. 2, pp. 4–10, 2012. 3, 25

[19] M. Carfagni, R. Furferi, L. Governi, M. Servi, F. Uccheddu, and Y. Volpe, "On the performance of the intel sr300 depth camera: metrological and critical characterization," *IEEE Sensors Journal*, vol. 17, no. 14, pp. 4508–4519, 2017. 3

[20] G. Goswami, S. Bharadwaj, M. Vatsa, and R. Singh, "On rgb-d face recognition using kinect," in *2013 IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, pp. 1–6, IEEE, 2013. 3

[21] Y.-C. Lee, J. Chen, C. W. Tseng, and S.-H. Lai, "Accurate and robust face recognition from rgb-d images with a deep learning approach.," in *BMVC*, vol. 1, p. 3, 2016. 3, 23

[22] X. Xiong, X. Wen, and C. Huang, "Improving rgb-d face recognition via transfer learning from a pretrained 2d network," in *Benchmarking, Measuring, and Optimizing: Second BenchCouncil International Symposium, Bench 2019, Denver, CO, USA, November 14–16, 2019, Revised Selected Papers*, pp. 141–148, Springer, 2020. 3

[23] P. U. Unschuld, “Huang di nei jing su wen,” in *Huang Di Nei Jing Su Wen*, University of California Press, 2003. 5

[24] J. Fanghänel, T. Gedrange, and P. Proff, “The face-physiognomic expressiveness and human identity,” *Annals of Anatomy-Anatomischer Anzeiger*, vol. 188, no. 3, pp. 261–266, 2006. 5

[25] B. Zhang, X. Wang, F. Karray, Z. Yang, and D. Zhang, “Computerized facial diagnosis using both color and texture features,” *Information Sciences*, vol. 221, pp. 49–59, 2013. 5

[26] H. J. Schneider, R. P. Kosilek, M. Günther, J. Roemmler, G. K. Stalla, C. Sievers, M. Reincke, J. Schopohl, and R. P. Würtz, “A novel approach to the detection of acromegaly: accuracy of diagnosis by automatic face classification,” *The Journal of Clinical Endocrinology & Metabolism*, vol. 96, no. 7, pp. 2074–2080, 2011. 5

[27] O. M. Parkhi, A. Vedaldi, and A. Zisserman, “Deep face recognition,” 2015. 6, 20, 53

[28] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, “Ms-celeb-1m: A dataset and benchmark for large-scale face recognition,” in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*, pp. 87–102, Springer, 2016. 6, 93

[29] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, “Vggface2: A dataset for recognising faces across pose and age,” in *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pp. 67–74, IEEE, 2018. 6, 93

[30] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, “Arcface: Additive angular margin loss for deep face recognition,” in *Proceedings of the IEEE/CVF*

*Conference on Computer Vision and Pattern Recognition*, pp. 4690–4699, 2019. 6, 23, 74, 101

[31] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Learning face representation from scratch," *arXiv preprint arXiv:1411.7923*, 2014. 6

[32] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, pp. 1345–1359, 2010. 6, 29, 52

[33] A. Lavrentaki, A. Paluzzi, J. A. Wass, and N. Karavitaki, "Epidemiology of acromegaly: review of population studies," *Pituitary*, vol. 20, pp. 4–9, 2017. 8

[34] K. Ho, *Growth hormone related diseases and therapy: a molecular and physiological perspective for the clinician.* Springer Science & Business Media, 2011. 8

[35] H.-S. Kim, J. Jung, S. H. Dong, S. H. Kim, S. Y. Jung, and S. G. Yeo, "Association between high neutrophil to lymphocyte ratio and delayed recovery from bell's palsy," *Clin Exp Otorhinolaryngol*, vol. 12, no. 3, pp. 261–6, 2019. 9

[36] J. D. Tiemstra and N. Khatkhate, "Bell's palsy: diagnosis and management," *American family physician*, vol. 76, no. 7, pp. 997–1002, 2007. 9

[37] S. E. Coulson, N. J. O'Dwyer, R. D. Adams, and G. R. Croxson, "Expression of emotion and quality of life after facial nerve paralysis," *Otology & neurotology*, vol. 25, no. 6, pp. 1014–1019, 2004. 9

[38] C. for Disease Control, P. (CDC, *et al.*, "Improved national prevalence estimates for 18 selected major birth defects–united states, 1999-2001,"

*MMWR. Morbidity and mortality weekly report*, vol. 54, no. 51, pp. 1301–1305, 2006. 10

[39] S. B. Freeman, E. G. Allen, C. L. Oxford-Wright, S. W. Tinker, C. Druschel, C. A. Hobbs, L. A. O'Leary, P. A. Romitti, M. H. Royle, C. P. Torfs, *et al.*, "The national down syndrome project: design and implementation," *Public Health Reports*, vol. 122, no. 1, pp. 62–72, 2007. 10

[40] M. Rodrigues, J. Nunes, S. Figueiredo, A. Martins de Campos, and A. F. Geraldo, "Neuroimaging assessment in down syndrome: a pictorial review," *Insights into imaging*, vol. 10, pp. 1–13, 2019. 10

[41] O. mondiale de la Santé, W. H. Organization, *et al.*, "Global leprosy (hansen disease) update, 2019: time to step-up prevention initiatives–situation de la lèpre (maladie de hansen) dans le monde, 2019: le moment est venu d'intensifier les initiatives de prévention," *Weekly Epidemiological Record= Relevé épidémiologique hebdomadaire*, vol. 95, no. 36, pp. 417–438, 2020. 11

[42] W. H. Organization *et al.*, "Global leprosy update, 2013; reducing disease burden," *Weekly Epidemiological Record= Relevé épidémiologique hebdomadaire*, vol. 89, no. 36, pp. 389–400, 2014. 11

[43] J.-S. Lee, T.-M. Rhee, K. Jeon, Y. Cho, S.-W. Lee, K.-D. Han, M.-W. Seong, S.-S. Park, and Y. K. Lee, "Epidemiologic trends of thalassemia, 2006–2018: A nationwide population-based study," *Journal of Clinical Medicine*, vol. 11, no. 9, p. 2289, 2022. 11

[44] B. Modell and M. Darlison, "Global epidemiology of haemoglobin disorders and derived service indicators," *Bulletin of the World Health Organization*, vol. 86, no. 6, pp. 480–487, 2008. 11

[45] “Chapter 40 - thalassemia syndromes,” in *Hematology (Seventh Edition)* (R. Hoffman, E. J. Benz, L. E. Silberstein, H. E. Heslop, J. I. Weitz, J. Anastasi, M. E. Salama, and S. A. Abutalib, eds.), pp. 546–570.e10, Elsevier, seventh edition ed., 2018. 11

[46] E. S. A. Alhaija, F. N. Hattab, and M. A. Al-Omari, “Cephalometric measurements and facial deformities in subjects with $\beta$-thalassaemia major,” *The European Journal of Orthodontics*, vol. 24, no. 1, pp. 9–19, 2002. 11

[47] J. Muñoz-Ortiz, M. C. Sierra-Cote, E. Zapata-Bravo, L. Valenzuela-Vallejo, M. A. Marin-Noriega, P. Uribe-Reina, J. P. Terreros-Dorado, M. Gómez-Suarez, K. Arteaga-Rivera, and A. De-La-Torre, “Prevalence of hyperthyroidism, hypothyroidism, and euthyroidism in thyroid eye disease: a systematic review of the literature,” *Systematic reviews*, vol. 9, no. 1, pp. 1–12, 2020. 12

[48] A. Manifold, “Hyperthyroidism, thyroid storm, and graves’ disease,” *E-medicine*, vol. 4, pp. 1–18, 2005. 12

[49] G. Easley, D. Labate, and W.-Q. Lim, “Sparse directional image representations using the discrete shearlet transform,” *Applied and Computational Harmonic Analysis*, vol. 25, no. 1, pp. 25–46, 2008. 15, 79

[50] M. Turk and A. Pentland, “Eigenfaces for recognition,” *Journal of cognitive neuroscience*, vol. 3, no. 1, pp. 71–86, 1991. 18, 74

[51] M. S. Bartlett, J. R. Movellan, and T. J. Sejnowski, “Face recognition by independent component analysis,” *IEEE Transactions on neural networks*, vol. 13, no. 6, pp. 1450–1464, 2002. 18, 74

[52] P. Phillips, “Support vector machines applied to face recognition,” *Advances in Neural Information Processing Systems*, vol. 11, pp. 803–809, 1998. 18

[53] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, “Deepface: Closing the gap to human-level performance in face verification,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1701–1708, 2014. 19

[54] F. Schroff, D. Kalenichenko, and J. Philbin, “Facenet: A unified embedding for face recognition and clustering,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 815–823, 2015. 20, 74, 101

[55] CASIA, “Casia-3d face v1.” Website, 2004. `http://biometrics.idealtest.org/`. 23, 34

[56] L. Yin, X. Wei, Y. Sun, J. Wang, and M. J. Rosato, “A 3d facial expression database for facial behavior research,” in *7th international conference on automatic face and gesture recognition (FGR06)*, pp. 211–216, IEEE, 2006. 23, 34

[57] A. Savran, N. Alyüz, H. Dibeklioğlu, O. Çeliktutan, B. Gökberk, B. Sankur, and L. Akarun, “Bosphorus database for 3d face analysis,” in *European workshop on biometrics and identity management*, pp. 47–56, Springer, 2008. 23, 34

[58] H. B. Abebe and C.-L. Hwang, “Rgb-d face recognition using lbp with suitable feature dimension of depth image,” *IET Cyber-Physical Systems: Theory & Applications*, vol. 4, no. 3, pp. 189–197, 2019. 23

[59] D. Kim, M. Hernandez, J. Choi, and G. Medioni, “Deep 3d face identification,” in *2017 IEEE international joint conference on biometrics (IJCB)*, pp. 133–142, IEEE, 2017. 23

[60] H. Zhang, H. Han, J. Cui, S. Shan, and X. Chen, “Rgb-d face recognition via deep complementary and common feature learning,” in *2018 13th*

*IEEE International Conference on Automatic Face  Gesture Recognition (FG 2018)*, pp. 8–15, 2018. 23

[61] L. Jiang, J. Zhang, and B. Deng, "Robust rgb-d face recognition using attribute-aware loss," *IEEE transactions on pattern analysis and machine intelligence*, vol. 42, no. 10, pp. 2552–2566, 2019. 23

[62] Z. Jiang, B. Jin, and Y. Song, "A novel pet trajectory prediction method for intelligent plant cultivation robot," *IEEE Sensors Letters*, vol. 7, no. 2, pp. 1–4, 2023. 24

[63] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," *Advances in neural information processing systems*, vol. 27, pp. 2366–2374, 2014. 24

[64] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab, "Deeper depth prediction with fully convolutional residual networks," in *2016 Fourth international conference on 3D vision (3DV)*, pp. 239–248, IEEE, 2016. 24

[65] I. Alhashim and P. Wonka, "High quality monocular depth estimation via transfer learning," *arXiv preprint arXiv:1812.11941*, 2018. 24, 44, 61

[66] C. Godard, O. Mac Aodha, and G. J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 270–279, 2017. 25

[67] C. Godard, O. Mac Aodha, M. Firman, and G. J. Brostow, "Digging into self-supervised monocular depth estimation," in *Proceedings of the IEEE international conference on computer vision*, pp. 3828–3838, 2019. 25, 45, 60

[68] S.-H. Lai, C.-W. Fu, and S. Chang, "A generalized depth estimation algorithm with a single image," *IEEE Computer Architecture Letters*, vol. 14, no. 04, pp. 405–411, 1992. 25

[69] Z.-L. Sun and K.-M. Lam, "Depth estimation of face images based on the constrained ica model," *IEEE transactions on information forensics and security*, vol. 6, no. 2, pp. 360–370, 2011. 25

[70] Z.-L. Sun, K.-M. Lam, and Q.-W. Gao, "Depth estimation of face images using the nonlinear least-squares model," *IEEE transactions on image processing*, vol. 22, no. 1, pp. 17–30, 2012. 25

[71] J. Cui, H. Zhang, H. Han, S. Shan, and X. Chen, "Improving 2d face recognition via discriminative face depth estimation," in *2018 International Conference on Biometrics (ICB)*, pp. 140–147, IEEE, 2018. 26

[72] S. Pini, F. Grazioli, G. Borghi, R. Vezzani, and R. Cucchiara, "Learning to generate facial depth maps," in *2018 International Conference on 3D Vision (3DV)*, pp. 634–642, IEEE, 2018. 26

[73] A. T. Arslan and E. Seke, "Face depth estimation with conditional generative adversarial networks," *IEEE Access*, vol. 7, pp. 23222–23231, 2019. 26

[74] B. Jin, L. Cruz, and N. Gonçalves, "Face depth prediction by the scene depth," in *2021 IEEE/ACIS 19th International Conference on Computer and Information Science (ICIS)*, pp. 42–48, IEEE, 2021. 26

[75] H. J. Schneider, R. P. Kosilek, M. Günther, J. Roemmler, G. K. Stalla, C. Sievers, M. Reincke, J. Schopohl, and R. P. Würtz, "A novel approach

to the detection of acromegaly: accuracy of diagnosis by automatic face classification," *The Journal of Clinical Endocrinology & Metabolism*, vol. 96, no. 7, pp. 2074–2080, 2011. 26

[76] Q. Zhao, K. Okada, K. Rosenbaum, L. Kehoe, D. J. Zand, R. Sze, M. Summar, and M. G. Linguraru, "Digital facial dysmorphology for genetic screening: Hierarchical constrained local model using ica," *Medical image analysis*, vol. 18, no. 5, pp. 699–710, 2014. 26

[77] Q. Zhao, N. Werghi, K. Okada, K. Rosenbaum, M. Summar, and M. G. Linguraru, "Ensemble learning for the detection of facial dysmorphology," in *2014 36th annual international conference of the IEEE engineering in medicine and biology society*, pp. 754–757, IEEE, 2014. 26

[78] X. Kong, S. Gong, L. Su, N. Howard, and Y. Kong, "Automatic detection of acromegaly from facial photographs using machine learning methods," *EBioMedicine*, vol. 27, pp. 94–102, 2018. 27

[79] S. Boehringer, T. Vollmar, C. Tasse, R. P. Wurtz, G. Gillessen-Kaesbach, B. Horsthemke, and D. Wieczorek, "Syndrome identification based on 2d analysis software," *European Journal of Human Genetics*, vol. 14, no. 10, pp. 1082–1089, 2006. 27

[80] P. Shukla, T. Gupta, A. Saini, P. Singh, and R. Balasubramanian, "A deep learning frame-work for recognizing developmental disorders," in *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 705–714, 2017. 27

[81] Y. Gurovich, Y. Hanani, O. Bar, G. Nadav, N. Fleischer, D. Gelbman, L. Basel-Salmon, P. M. Krawitz, S. B. Kamphausen, M. Zenker, *et al.*,

"Identifying facial phenotypes of genetic disorders using deep learning," *Nature medicine*, vol. 25, no. 1, pp. 60–64, 2019. 27

[82] B. Jin, L. Cruz, and N. Gonçalves, "Deep facial diagnosis: deep transfer learning from face recognition to facial diagnosis," *IEEE Access*, vol. 8, pp. 123649–123661, 2020. 27

[83] A. R. Porras, K. Rosenbaum, C. Tor-Diez, M. Summar, and M. G. Linguraru, "Development and evaluation of a machine learning-based point-of-care screening tool for genetic syndromes in children: a multinational retrospective study," *The Lancet Digital Health*, vol. 3, no. 10, pp. e635–e643, 2021. 28

[84] B. Hallgrímsson, J. D. Aponte, D. C. Katz, J. J. Bannister, S. L. Riccardi, N. Mahasuwan, B. L. McInnes, T. M. Ferrara, D. M. Lipman, A. B. Neves, *et al.*, "Automated syndrome diagnosis by three-dimensional facial imaging," *Genetics in medicine*, vol. 22, no. 10, pp. 1682–1693, 2020. 28

[85] J. J. Bannister, M. Wilms, J. D. Aponte, D. C. Katz, O. D. Klein, F. P. J. Bernier, R. A. Spritz, B. Hallgrímsson, and N. D. Forkert, "A deep invertible 3-d facial shape model for interpretable genetic syndrome diagnosis," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 7, pp. 3229–3239, 2022. 28

[86] W. Dai, Q. Yang, G.-R. Xue, and Y. Yu, "Boosting for transfer learning," in *Proceedings of the 24th International Conference on Machine Learning*, ICML '07, (New York, NY, USA), p. 193–200, Association for Computing Machinery, 2007. 29

[87] B. Tan, Y. Zhang, S. Pan, and Q. Yang, "Distant domain transfer learning,"

in *Proceedings of the AAAI conference on artificial intelligence*, vol. 31, 2017. 29

[88] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang, “Domain adaptation via transfer component analysis,” *IEEE transactions on neural networks*, vol. 22, no. 2, pp. 199–210, 2010. 30

[89] M. Long, H. Zhu, J. Wang, and M. I. Jordan, “Deep transfer learning with joint adaptation networks,” in *International conference on machine learning*, pp. 2208–2217, PMLR, 2017. 30

[90] A. Sharif Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, “Cnn features off-the-shelf: an astounding baseline for recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 806–813, 2014. 30

[91] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, “Dermatologist-level classification of skin cancer with deep neural networks,” *nature*, vol. 542, no. 7639, pp. 115–118, 2017. 30

[92] Y. Yu, H. Lin, J. Meng, X. Wei, H. Guo, and Z. Zhao, “Deep transfer learning for modality classification of medical images,” *Information*, vol. 8, no. 3, p. 91, 2017. 30

[93] Z. Shi, H. Hao, M. Zhao, Y. Feng, L. He, Y. Wang, and K. Suzuki, “A deep cnn based transfer learning method for false positive reduction,” *Multimedia Tools and Applications*, vol. 78, pp. 1017–1033, 2019. 30

[94] M. Raghu, C. Zhang, J. Kleinberg, and S. Bengio, “Transfusion: Understanding transfer learning for medical imaging,” *Advances in neural information processing systems*, vol. 32, 2019. 30

[95] H.-C. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, and R. M. Summers, “Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning,” *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1285–1298, 2016. 30

[96] J. Davis and P. Domingos, “Deep transfer via second-order markov logic,” in *Proceedings of the 26th annual international conference on machine learning*, pp. 217–224, 2009. 30

[97] M. Mirza and S. Osindero, “Conditional generative adversarial nets,” *arXiv preprint arXiv:1411.1784*, 2014. 32

[98] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, “Image-to-image translation with conditional adversarial networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1125–1134, 2017. 32

[99] A. Odena, C. Olah, and J. Shlens, “Conditional image synthesis with auxiliary classifier gans,” in *International conference on machine learning*, pp. 2642–2651, PMLR, 2017. 32

[100] N. Otsu, “A threshold selection method from gray-level histograms,” *IEEE transactions on systems, man, and cybernetics*, vol. 9, no. 1, pp. 62–66, 1979. 36

[101] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241, Springer, 2015. 37

[102] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral normalization for generative adversarial networks," *arXiv preprint arXiv:1802.05957*, 2018. 41

[103] G. vanRossum, "Python reference manual," *Department of Computer Science [CS]*, no. R 9525, 1995. 43

[104] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, *et al.*, "{TensorFlow}: a system for {Large-Scale} machine learning," in *12th USENIX symposium on operating systems design and implementation (OSDI 16)*, pp. 265–283, 2016. 43

[105] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1231–1237, 2013. 44

[106] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from rgbd images.," *ECCV (5)*, vol. 7576, pp. 746–760, 2012. 44

[107] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004. 44, 72

[108] B. Jin, "Disease-specific faces." IEEE Dataport. `https://dx.doi.org/10.21227/rk2v-ka85`. 46

[109] B. Jin, "Disease-specific faces 2." IEEE Dataport. `https://dx.doi.org/10.21227/zqra-nh98`. 46, 48

[110] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W.

Vaughan, “A theory of learning from different domains,” *Machine learning*, vol. 79, pp. 151–175, 2010. 52

[111] S. Ben-David, J. Blitzer, K. Crammer, and F. Pereira, “Analysis of representations for domain adaptation,” *Advances in neural information processing systems*, vol. 19, 2006. 52

[112] L. Shao, F. Zhu, and X. Li, “Transfer learning for visual categorization: A survey,” *IEEE transactions on neural networks and learning systems*, vol. 26, no. 5, pp. 1019–1034, 2014. 52

[113] D. Sarkar, “A comprehensive hands-on guide to transfer learning with real-world applications in deep learning,” *Towards Data Science*, vol. 20, p. 2020, 2018. 53

[114] S. Ruder *et al.*, “Transfer learning-machine learning’s next frontier,” *Accessed: April*, 2017. 53

[115] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255, Ieee, 2009. 53

[116] P. Molchanov, S. Tyree, T. Karras, T. Aila, and J. Kautz, “Pruning convolutional neural networks for resource efficient inference,” *arXiv preprint arXiv:1611.06440*, 2016. 56

[117] J. A. Suykens and J. Vandewalle, “Least squares support vector machine classifiers,” *Neural processing letters*, vol. 9, pp. 293–300, 1999. 57

[118] C. J. Burges, “A tutorial on support vector machines for pattern recognition,” *Data mining and knowledge discovery*, vol. 2, no. 2, pp. 121–167, 1998. 58

[119] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter, "A 3d face model for pose and illumination invariant face recognition," in *2009 sixth IEEE international conference on advanced video and signal based surveillance*, pp. 296–301, Ieee, 2009. 61

[120] S. Li, M. Lin, Y. Wang, C. Fei, L. Shao, and R. Ji, "Learning efficient gans for image translation via differentiable masks and co-attention distillation," *IEEE Transactions on Multimedia*, 2022. 74

[121] F. S. Samaria and A. C. Harter, "Parameterisation of a stochastic model for human face identification," in *Proceedings of 1994 IEEE workshop on applications of computer vision*, pp. 138–142, IEEE, 1994. 74

[122] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection," in *European conference on computer vision*, pp. 43–58, Springer, 1996. 74

[123] D. B. Graham and N. M. Allinson, "Characterising virtual eigensignatures for general purpose face recognition," in *Face Recognition*, pp. 446–456, Springer, 1998. 74

[124] A. Martinez and R. Benavente, "The ar face database: Cvc technical report, 24," 1998. 74

[125] P. J. Phillips, H. Wechsler, J. Huang, and P. J. Rauss, "The feret database and evaluation procedure for face-recognition algorithms," *Image and vision computing*, vol. 16, no. 5, pp. 295–306, 1998. 74

[126] J.-G. Wang, J. Li, C. Y. Lee, and W.-Y. Yau, "Dense sift and gabor descriptors-based face representation with applications to gender recognition," in *2010 11th International Conference on Control Automation Robotics & Vision*, pp. 1860–1864, IEEE, 2010. 88

[127] C. Shan, S. Gong, and P. W. McOwan, "Robust facial expression recognition using local binary patterns," in *IEEE International Conference on Image Processing 2005*, vol. 2, pp. II–370, IEEE, 2005. 89

[128] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7132–7141, 2018. 93

[129] P. Kruszka, Y. A. Addissie, D. E. McGinn, A. R. Porras, E. Biggs, M. Share, T. B. Crowley, B. H. Chung, G. T. Mok, C. C. Mak, *et al.*, "22q11. 2 deletion syndrome in diverse populations," *American Journal of Medical Genetics Part A*, vol. 173, no. 4, pp. 879–888, 2017. 94

[130] C. Torrence and G. P. Compo, "A practical guide to wavelet analysis," *Bulletin of the American Meteorological society*, vol. 79, no. 1, pp. 61–78, 1998. 98

[131] D. L. Donoho, "De-noising by soft-thresholding," *IEEE transactions on information theory*, vol. 41, no. 3, pp. 613–627, 1995. 100

[132] B. W. Matthews, "Comparison of the predicted and observed secondary structure of t4 phage lysozyme," *Biochimica et Biophysica Acta (BBA)-Protein Structure*, vol. 405, no. 2, pp. 442–451, 1975. 103

[133] E. P. Vichinsky, "Changing patterns of thalassemia worldwide," *Annals of the New York Academy of Sciences*, vol. 1054, no. 1, pp. 18–24, 2005.

[134] S. Crisafulli, N. Luxi, J. Sultana, A. Fontana, F. Spagnolo, G. Giuffrida, F. Ferraù, D. Gianfrilli, A. Cozzolino, M. Cristina De Martino, *et al.*, "Global epidemiology of acromegaly: a systematic review and meta-analysis," *European Journal of Endocrinology*, vol. 185, no. 2, pp. 251–263, 2021.

[135] M. Gheorghiu, "News in acromegaly," *Acta Endocrinologica (Bucharest)*, vol. 13, no. 1, p. 129, 2017.

[136] C. I. Cha, C. K. Hong, M. S. Park, and S. G. Yeo, "Comparison of facial nerve paralysis in adults and children," *Yonsei medical journal*, vol. 49, no. 5, pp. 725–734, 2008.

[137] A. P. Presson, G. Partyka, K. M. Jensen, O. J. Devine, S. A. Rasmussen, L. L. McCabe, and E. R. McCabe, "Current estimate of down syndrome population prevalence in the united states," *The Journal of pediatrics*, vol. 163, no. 4, pp. 1163–1168, 2013.

[138] D. L. Donoho, "De-noising by soft-thresholding," *IEEE transactions on information theory*, vol. 41, no. 3, pp. 613–627, 1995.

[139] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning*, pp. 448–456, PMLR, 2015.

[140] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *International conference on machine learning*, pp. 214–223, PMLR, 2017.

[141] E. Turkof, B. Richard, O. Assadian, B. Khatri, E. Knolle, and S. Lucas, "Leprosy affects facial nerves in a scattered distribution from the main trunk to all peripheral branches and neurolysis improves muscle function of the face," *The American journal of tropical medicine and hygiene*, vol. 68, no. 1, pp. 81–88, 2003.

[142] A. F. Abate, P. Barra, S. Barra, C. Molinari, M. Nappi, and F. Narducci, "Clustering facial attributes: Narrowing the path from soft to hard biometrics," *IEEE Access*, vol. 8, pp. 9037–9045, 2019.

[143] S. E. Antonarakis, B. G. Skotko, M. S. Rafii, A. Strydom, S. E. Pape, D. W. Bianchi, S. L. Sherman, and R. H. Reeves, "Down syndrome," *Nature Reviews Disease Primers*, vol. 6, no. 1, p. 9, 2020.

[144] P. N. Taylor, D. Albrecht, A. Scholz, G. Gutierrez-Buey, J. H. Lazarus, C. M. Dayan, and O. E. Okosieme, "Global epidemiology of hyperthyroidism and hypothyroidism," *Nature Reviews Endocrinology*, vol. 14, no. 5, pp. 301–316, 2018.

[145] Q. Zhao, K. Rosenbaum, R. Sze, D. Zand, M. Summar, and M. G. Linguraru, "Down syndrome detection from facial photographs using machine learning techniques," in *Medical Imaging 2013: Computer-Aided Diagnosis*, vol. 8670, pp. 9–15, SPIE, 2013.

[146] Q. Zhao, K. Okada, K. Rosenbaum, D. J. Zand, R. Sze, M. Summar, and M. G. Linguraru, "Hierarchical constrained local model using ica and its application to down syndrome detection," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2013: 16th International Conference, Nagoya, Japan, September 22-26, 2013, Proceedings, Part II 16*, pp. 222–229, Springer, 2013.

[147] Q. Zhao, K. Okada, K. Rosenbaum, L. Kehoe, D. J. Zand, R. Sze, M. Summar, and M. G. Linguraru, "Digital facial dysmorphology for genetic screening: Hierarchical constrained local model using ica," *Medical image analysis*, vol. 18, no. 5, pp. 699–710, 2014.

[148] T. Shu, B. Zhang, and Y. Y. Tang, "An extensive analysis of various texture feature extractors to detect diabetes mellitus using facial specific regions," *Computers in biology and medicine*, vol. 83, pp. 69–83, 2017.

[149] S. Hadj-Rabia, H. Schneider, E. Navarro, O. Klein, N. Kirby, K. Huttner, L. Wolf, M. Orin, S. Wohlfart, C. Bodemer, *et al.*, "Automatic recognition of the xlhed phenotype from facial images," *American Journal of Medical Genetics Part A*, vol. 173, no. 9, pp. 2408–2414, 2017.

[150] P. Kruszka, Y. A. Addissie, D. E. McGinn, A. R. Porras, E. Biggs, M. Share, T. B. Crowley, B. H. Chung, G. T. Mok, C. C. Mak, *et al.*, "22q11. 2 deletion syndrome in diverse populations," *American Journal of Medical Genetics Part A*, vol. 173, no. 4, pp. 879–888, 2017.

[151] S. Boehringer, T. Vollmar, C. Tasse, R. P. Wurtz, G. Gillessen-Kaesbach, B. Horsthemke, and D. Wieczorek, "Syndrome identification based on 2d analysis software," *European Journal of Human Genetics*, vol. 14, no. 10, pp. 1082–1089, 2006.

[152] Y. Gurovich, Y. Hanani, O. Bar, G. Nadav, N. Fleischer, D. Gelbman, L. Basel-Salmon, P. M. Krawitz, S. B. Kamphausen, M. Zenker, *et al.*, "Identifying facial phenotypes of genetic disorders using deep learning," *Nature medicine*, vol. 25, no. 1, pp. 60–64, 2019.

[153] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "Vggface2: A dataset for recognising faces across pose and age," in *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pp. 67–74, IEEE, 2018.

[154] B. Jin, L. Cruz, and N. Gonçalves, "Pseudo rgb-d face recognition," *IEEE Sensors Journal*, vol. 22, no. 22, pp. 21780–21794, 2022.

[155] P. U. Unschuld, *Huang Di Nei Jing Su Wen: Nature, knowledge, imagery in an ancient Chinese medical text: With an appendix: The doctrine of the*

*five periods and six Qi in the Huang Di Nei Jing Su Wen*. Univ of California Press, 2003.

[156] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *nature*, vol. 521, no. 7553, pp. 436–444, 2015.