Cyclic Style Generative Adversarial Network for Near Infrared and Visible Light Face Recognition

by

Fangzheng Huang

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Applied Science
in
Electrical and Computer Engineering

Waterloo, Ontario, Canada, 2022

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Face recognition in the visible light (VIS) spectrum has been widely utilized in many practical applications. With the development of the deep learning method, the recognition accuracy and speed have already reached an excellent level, where face recognition can be applied in various circumstances. However, in some extreme situations, there are still problems that face recognition cannot guarantee performance. One of the most significant cases is under poor illumination. Lacking light sources, images cannot show the true identities of detected people. To address such a problem, the near infrared (NIR) spectrum offers an alternative solution to face recognition in which face images can be captured clearly. Studies have been made in recent years, and current near infrared and visible light (NIR-VIS) face recognition methods have achieved great performance.

In this thesis, I review current NIR-VIS face recognition methods and public NIR-VIS face datasets. I first list public NIR-VIS face datasets that are used in most research. For each dataset, I represent their characteristics, including the number of subjects, collection environment, resolution of images, and whether paired or not. Also, I conclude evaluation protocols for each dataset, helping with further analyzing of performances. Then, I classify current NIR-VIS face recognition methods into three categories, image synthesis-based methods, subspace learning-based methods, and invariant feature-based methods. The contribution of each method is concisely explained. Additionally, I make comparisons between current NIR-VIS face recognition methods and propose my own opinion on the advantages and disadvantages of these methods.

To improve the shortcomings of current methods, this thesis proposes a new model, Cyclic Style Generative Adversarial Network (CS-GAN), which is a combination of image synthesis-based method and subspace learning-based method. The proposed CS-GAN improves the visualization results of image synthesis between the NIR domain and VIS domain as well as recognition accuracy. The CS-GAN is based on the Style-GAN 3 network which was proposed in 2021. In the proposed model, there are two generators from pre-trained Style-GAN 3 which generate images in the NIR domain and VIS domain, respectively. The generators consist of a mapping network and synthesis network, where the mapping network disentangles the latent code for reducing correlation between features, and the synthesis network synthesizes face images through progressive growing training. The generators have different final layers, a to-RGB layer for the VIS domain and a to-grayscale layer for the NIR domain. Generators are embedded in a cyclic structure, in which latent codes are sent into the synthesis network in the other generator for recreated images, and recreated images are compared with real images which in the same domain to ensure domain consistency. Besides, I apply the proposed cyclic subspace learning. The cyclic subspace learning is composed of two parts. The first part introduces the proposed latent loss which is to have better controls over the learning of latent subspace. The latent codes influence both details and locations of features through continuously inputting into the synthesis network. The control over latent subspace can strengthen the feature consistency between synthesized images. And the second part improves the style-transferring process by controlling high-level features with perceptual loss in each domain. In the perceptual loss, there is a pre-trained VGG-16 network to extract high-level features which can be regarded as the style of the images. Therefore, style loss can control the style of images in both domains as well as ensure style consistency between synthesized images and real images. The visualization results show that the proposed CS-GAN model can synthesize better VIS images that are detailed, corrected colorized, and with clear edges. More importantly, the experimental results show that the Rank-1 accuracy on CASISA NIR-VIS 2.0 database reaches 99.60% which improves state-of-the-art methods by 0.2%.

Acknowledgements

I would like to thank those who supported and helped me during my postgraduate program at the University of Waterloo. First, I would like to gratefully thank my supervisor, professor Dayan Ban, who offered me the chance to study here and gave me lots of important advice on my research. Besides, I would also like to thank all lab mates of Prof. Dayan's group for sharing new ideas with me. Also, I would like to thank all group members of Prof. Guangqiang Yin's group who proposed a lot of useful suggestions about my experiment work. Finally, I would like to acknowledge my parents, who supported me with their encouragement and selfless care.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

2D two dimensional 19, 20

3D three dimensional 19, 20

3DMM 3D Morphable Model 20

CCA Canonical-Correlation Analysis 27, 28, 40

CdR Cross-domain compact Representation 22

CFDAL cross-domain feature distribution alignment Learning 37, 38

CNN Convolutional Neural Networks 2, 17, 42

CS-GAN Cyclic Style Generative Adversarial Network 10, 56, 58

CV Computer vision 1, 6, 13

D discriminator 9

DoG Difference-of-Gaussian 28

DRA Deep Representation Alignment 39

FAR False Accept Rate 25, 35, 54

FFA Feature-level Face Alignment 22

FFHQ-U unaligned Flickr-Faces-HQ Dataset 49

G generator 9

GAN Generative Adversarial Networks 9, 10, 18, 19, 22, 26

HFR Heterogeneous Face Recognition 3, 6

HOG Histogram of Oriented Gradients 28, 30, 36, 37

HPR Heterogeneous face recognition across Pose and Resolution 15

IFA Image-level Face Alignment 22

InGaAs Indium Gallium Arsenide 4, 5

InP indium phosphide 5

JFDAL Joint Feature Distribution Alignment Learning 37

k-SVD K-Singular Value Decomposition 17

kNN k Nearest Neighbors 16

LBP Linear Binary Pattern 16, 28

LCHMHE Linear Cross-modal Hash Encoding 30

LDA Linear Discriminant Analysis 27, 28

LDML Logistic Discriminant based Metric Learning 31

LED light emitting diode 4, 5, 42, 54, 57

LPDM logarithmic pixel difference matrix 30

LPDV Logarithmic Pixel Difference Vector 30

LSNA Local Structure of Normalized Appearance 28

LTP Long-Term Potentiation 30

MB-LBP Multi-Block Local Binary Pattern 28

MFM Max-Feature-Map 35

MI Modality-Invariant 38

MMD Maximum Mean Discrepancy 29, 30, 35

MSDA Multi-view Smooth Discriminant Analysis 30

MTCNN Multi-Task Cascaded Convolutional Network 24

NIR near infrared 2–6, 8–10, 13–18, 20–22, 24–29, 31–34, 36–39, 42–45, 48–50, 52, 54, 56–58

NIR-VIS Near infrared and visible light 3, 4, 6, 8–11, 13–16, 18–22, 24, 25, 27, 33–37, 39, 41, 44, 48, 51, 54, 56–59

O-Net output network 24

OMD orthogonal modality disentanglement 38, 39

OMDRA Orthogonal Modality Disentanglement and Representation Alignment 38

P-Net proposed network 24

PACH Pose Aligned Cross-spectral Hallucination 21

PCA Principle Component Analysis 2, 27–29, 37

PSNR peak signal-to-noise ratio 17

R-Net refining network 24

RBM Restricted Boltzmnn Machine 28, 29

ROIC readout integrated circuit 5

SFDAL source-domain feature distribution alignment learning 37, 38

SIFT Scale-Invariant Feature Transform 30

SST semi-Siamese Training network 33

StEn-CNNs Student-Encoder CNNs 22

SWIR shortwave infrared 5

TeEn-CNNs Teacher-Encoder CNNs 22

TPS Texture Prior Synthesis 21

UFA Unsupervised Face Alignment 21

VIS visible 2–6, 8–10, 13–22, 24–29, 31–34, 36–39, 42, 44, 45, 48–50, 52, 54, 56–59

VIS-VIS visible light to visible light 15

YCbCr brightness, blue minus luma and red minus luma 17

Chapter 1

Introduction

1.1 Background

1.1.1 Computer Vision

Computer vision (CV) is a scientific field in which researchers develop various methods to help machines, especially computers, to fully understand digital images and videos. Concretely, the objective of CV is to infer something about the world through images and videos[60]. The wide-spread use and application of photodetectors and the Internet over recent decades was accompanied by the development of numerous types of devices (e.g., mobile phones, Bayonet cameras, surveillance cameras) that are capable of capturing images and videos, most of which were uploaded to the Internet. Computers are no longer constrained in location and can connect to other devices across the world. However, computers cannot correctly extract all useful information from images and videos. Different from the human vision system, computers lack the capability to process images abstractly and are adversely impacted by many aspects of objects(e.g. orientation, occlusion, lighting). Despite recent research advancements, the performance of CV is still far from human vision. Thus, significant room for research and development in CV remains.

1.1.2 Face Recognition

Face recognition is a sub-task in CV which aims to detect face and determine the identity of faces in images and videos. The process involves matching the face in question with those

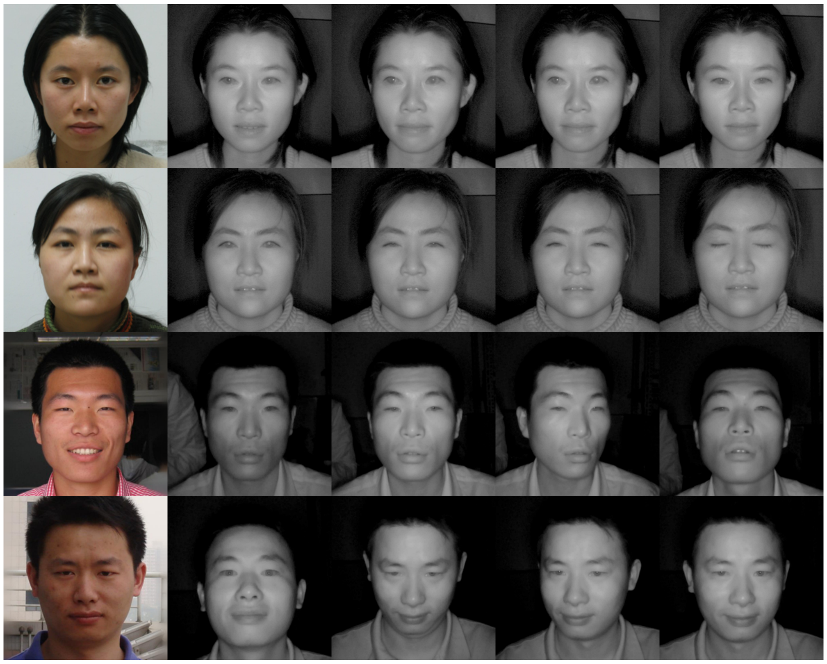


Figure 1.1: Examples of NIR and VIS images.

available in dataset galleries. Face recognition has been used to date in the civilian and law enforcement sectors(e.g. ID verification and criminal tracking). Principle Component Analysis (PCA) was integrated into face recognition research in the 1990s[21]. More recent advances in And in 2010s, face recognition techniques made during the 2010s focused on the application of deep learning approaches, primarily based on Convolutional Neural Networks (CNN)[39]; these approaches enabled more rapid and accurate execution of face recognition tasks asks in research settings.However, the related research predominantly uses idealized datasets comprised of clear, unobscured faces in high resolution; in contrast, face recognition techniques used in real-world applications must be capable of resolving numerous factors which can detract from image quality and/or clarity, such as illumination, pose variation and occlusion. Future research is needed to advance face recognition techniques for real applications.

1.1.3 NIR-VIS Face Recognition

To address the limitations associated with current approaches, the Near infrared (NIR) spectrum offers an alternative solution to face recognition on top of the conventional Visible

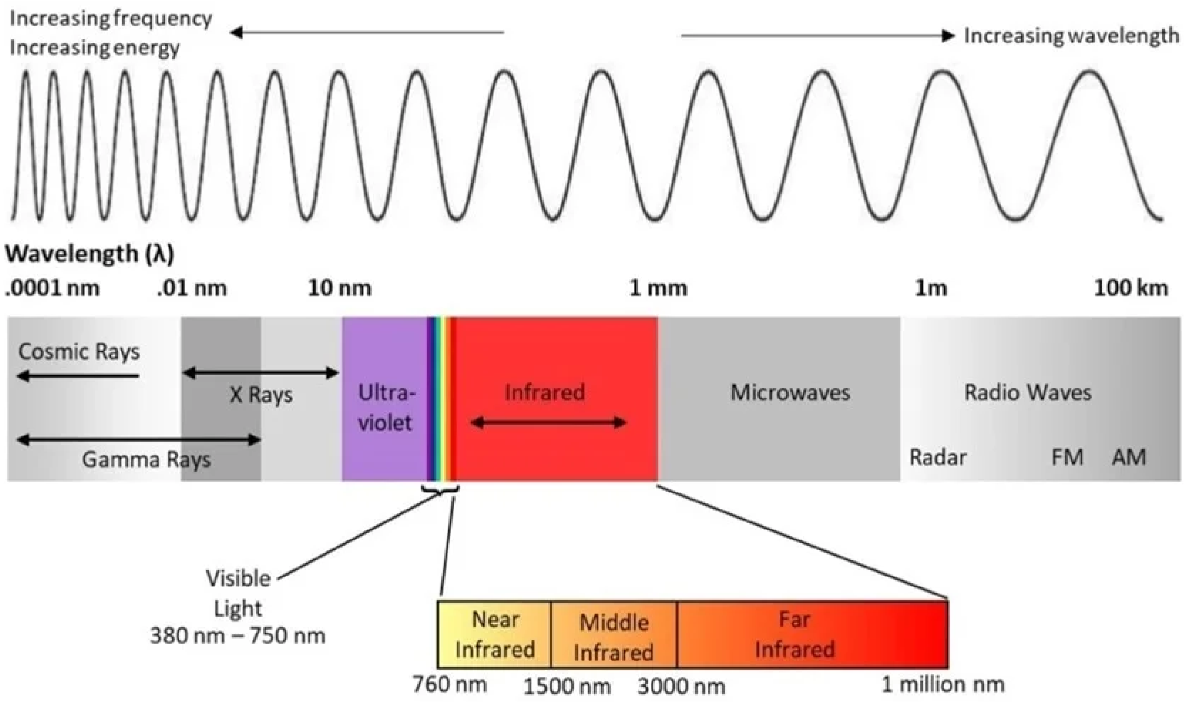


Figure 1.2: The electromagnetic spectrum, showing the details of infrared range.

(VIS) spectrum. Heterogeneous Face Recognition (HFR)) techniques attempt to match probe face images in one modality to gallery face images in another modality. Near infrared and visible light (NIR-VIS) Face recognition is one such example of a HFR technique, which identifies face images by matching NIR probe face images to VIS gallery face images. Fig.1.1 shows a series of NIR images and the corresponding VIS images.

Property of NIR-VIS Face Recognition

Fig.1.2 shows the spectrum with increasing wavelength and increasing frequency. The wavelength of NIR images is between 0.76μm and 1.5μm and VIS images is between 0.38μm and 0.75μm. This difference provides NIR-VIS face recognition with several advantages; first and foremost, NIR-VIS face recognition is a more robust method for face recognition under unconstrained illumination, and is particularly robust relative to other contemporary approaches under poor illumination conditions. Captured under low-light environment, NIR images can still preserve the main features of targeting objects. The modality gap between NIR images and VIS images increases with the increasing wavelength difference. Secondly,

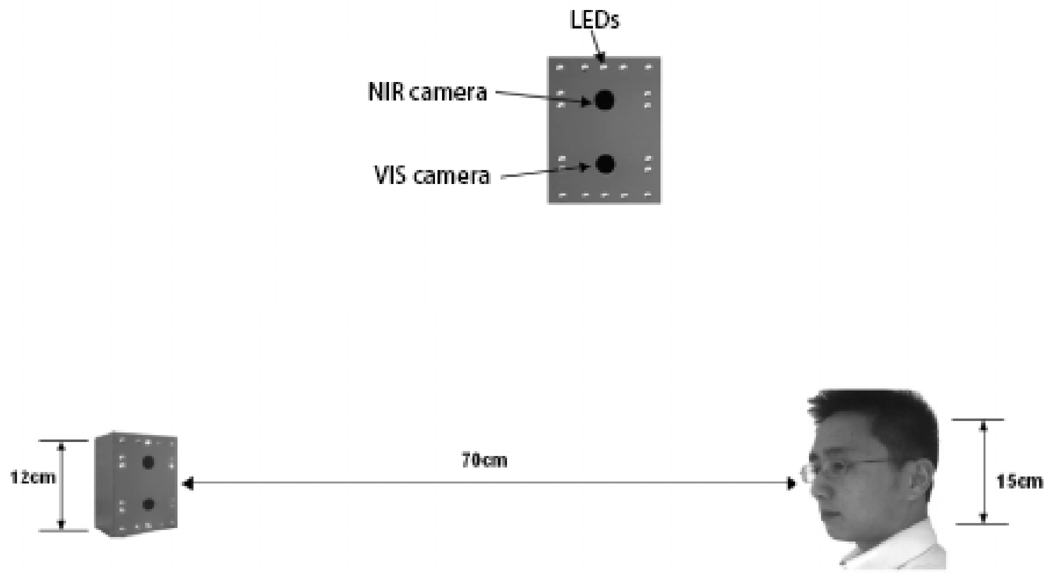


Figure 1.3: Example of NIR imaging system. It consists of NIR camera, VIS camera and 18 NIR LEDs.

NIR imaging is a lower-cost solution relative to imagers in other wavelength ranges, such as mid-wave infrared($3 - 5\mu m$) and long-wave infrared($8 - 15\mu m$). Although NIR imagers(e.g. short-wave infrared, $1.4 - 3\mu m$) are more expansive than VIS imagers, they are still much more economical relative to mid-wave infrared and long-wave infrared imagers. Thirdly, NIR is invisible to human being eyes, as a result, NIR reflection imaging, even under strong illumination from NIR Light emitting diode (LED) illuminators, which is safe to naked human eyes. With such superiority, NIR-VIS face recognition could be used in many face identification and authorization circumstances, such as night-time surveillance and E-passports.

Device for NIR-VIS Face Recognition

In NIR-VIS face recognition, the critical part is a special-designed NIR imaging system. Fig.1.3 shows one example of special-designed NIR imaging systems. In the system, there is a NIR camera. Currently, cameras for NIR imaging are Indium Gallium Arsenide (InGaAs) cameras, which consist of hybrid components (InGaAs plus CMOS). InGaAs cameras are used for applications that require high sensitivity over the $0.9 - 1.7\mu m$ wavelength range,

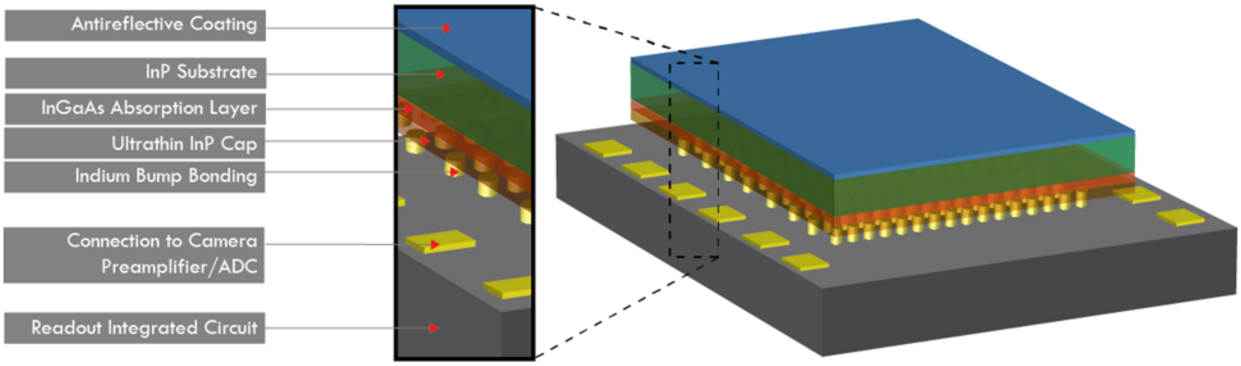


Figure 1.4: Diagram showing the different layers of a typical InGaAs sensor.

referred to as Shortwave infrared (SWIR). A InGaAs focal plane array in sensors is made of a two-dimensional photodiode array. This array consists of an Indium phosphide (InP) substrate, an InGaAs absorption layer, and an ultrathin InP cap that has been indium bump bonded to a Readout integrated circuit (ROIC), which are shown in Fig.1.4. The InGaAs two-dimensional array detects SWIR incident light, by collecting the photon-generated charge. The ROIC clocks and converts the collected charge into voltage, transferring the signal to off-chip electronics where it is used to create an image. These NIR cameras have the advantages of ease of use, compact camera dimensions, cost-efficient manufacturing, and no requirement for cooling systems.

Also, for taking NIR face images, there are active NIR illuminators, LEDs mounted around the cameras, which illuminate the face from near front direction and then capture front-lighted NIR face images. The process is like using a camera flash for VIS imagers, instead, these illuminators work in the invisible NIR spectrum. The general setting of NIR imaging system is important for capturing images with suitable pixel intensities. Firstly, the active NIR light should provide strong frontal lighting to override environmental light. Secondly, the camera exposure should be set to a low level in order to produce clear frontal-lighted face images in such a dark environment. Thirdly, there should be a long-pass optical filter that cuts off visible light of wavelength shorter than $750nm$. For outdoor scenes, the hardware should be further optimized to reduce the negative influence of sunlight. The sunlight contains a much stronger NIR component. Thus, in order to maintain the illumination from active NIR illuminators, the systems require strong active NIR pulse illuminators and NIR cameras that synchronize together.

1.1.4 Current methods of NIR-VIS Face Recognition

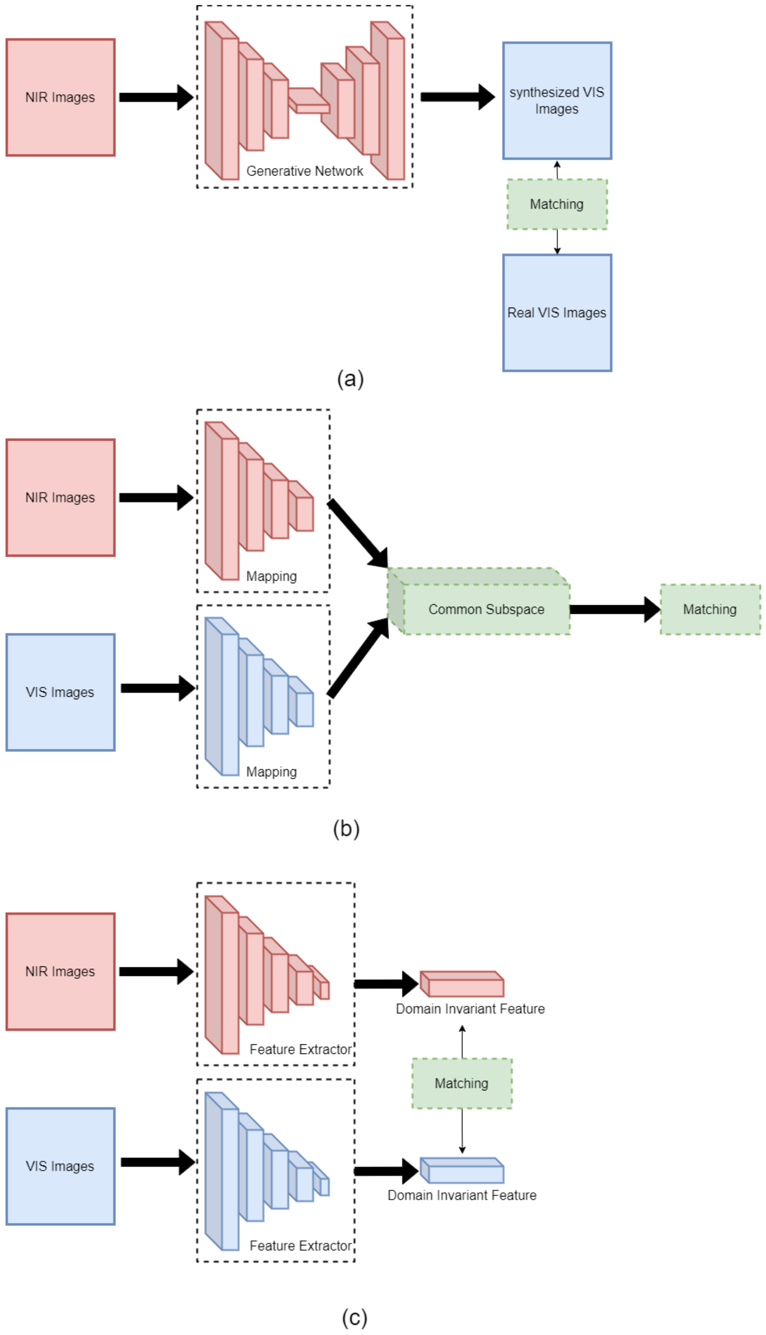
As one of the earliest HFR, NIR-VIS face recognition was originally developed in 2007[74]. Successful application of this technique requires minimization of the sensing gap between the NIR images and VIS images of targeting faces and preservation of identifiable information for comparison. Based on improvements in Pattern Recognition and CV field, various methodologies have been proposed for use in NIR-VIS face recognition field; these methodologies can be broadly categorized into one of three types: 1) image synthesis-based methods, which synthesize face images from one modality into the other and then conduct face matching. The related literature primarily involves the synthesis of VIS images from NIR images, thereby enabling the application of a high-accuracy face recognition system to the synthesized VIS image; 2) subspace learning-based methods, which project both NIR and VIS facial images into a common space so as to minimize the modality gap; 3) invariant feature-based methods, which extract modality-invariant face features from both NIR and VIS images for NIR-VIS face recognition. With recent progress in deep learning, these methods can all deliver excellent performance. The general flowcharts of three kinds of methods are shown in Fig.1.5.

1.2 Current Limitation

Though current researches can achieve state-of-the-art performance, challenges remain in NIR-VIS face recognition.

1.2.1 Dataset

One of the most intractable problems is that current NIR-VIS face datasets are not as available nor readily abundant as traditional VIS datasets. Through years of development, VIS face can be easily gathered through multiple resources, and researchers can build up large-size VIS face datasets. For example, in the CelebFaces Attributes Dataset (CelebA), Liu et al. [48] collected face images of 10,177 celebrities. Images of celebrities are now everywhere on the Internet which makes it easy to collect various images with the same identity. What is more, these images are captured in various scenes, which brings multi-appearance of the same person with high resolution. However, for NIR face dataset, it is not easy to get NIR face images as the demand of special NIR imaging system to capture. All existing NIR-VIS face datasets are collected manually, and thus it is too difficult to build up a dataset with the same size as VIS face datasets. Also, equipment for NIR



(a)

(b)

(c)

Figure 1.5: The general flowchart of (a) image synthesis-based methods, (b) subspace learning-based methods and (c) invariant feature-based methods.

imaging is much more expensive than VIS imagers. Researchers will have to spend extra money for NIR imaging system, which includes a light source and NIR imagers. For the existing NIR-VIS face dataset, the unpaired property is another obstacle. Captured with different imagers, NIR images and VIS images are necessarily unpaired. In some research, face matching between paired datasets can bring lots of benefits to network learning. But now, researchers have to find a way to deal with challenges like a transaction. And also, current NIR-VIS face datasets are all close-set. The close-set means that query images definitely have their gallery images in the dataset. However, such close-set properties violate the practical scenes of the face recognition task. In practical applications, it is not guaranteed that every detected person has their corresponding record in the system. Therefore, an open-set NIR-VIS face dataset is demanded. In sum, traditional VIS face datasets consist of thousands of images with various appearances which makes the matching progress more challenging. In contrast, NIR-VIS face recognition datasets typically only comprise the order of hundreds of identities, which can introduce challenges of over-fitting. Current state-of-the-art methods have already reached an extremely high accuracy, but for practical usage, there is a need for a larger and more robust NIR-VIS face dataset.

1.2.2 Sensing Gap

In addition, sensing gap between NIR modality and VIS modality, as well as pose variation, bluriness and occlusion, can negatively impact or even inhibit recognition. The goal of NIR-VIS face recognition is to recognize the face identities without the influence from the sensing gap. Those mentioned three types of methods, image synthesis-based methods, subspace learning-based methods and invariant feature-based methods are designed from different starting points, but all with difficulties: 1) image synthesis-based methods try to eliminate sensing gap by synthesizing VIS images based on NIR images. In VIS images, face images have different appearance with NIR images, i.e. deepen wrinkles and colorization. The generative networks form image synthesis-based methods learn to synthesize these extra details. However, the synthesis process cannot get the same results as VIS imaging where contour blurring, distortion and wrong colorization may occur. Such failed synthesis will influence the matching between synthesis images and real VIS images; 2) subspace learning-based methods deal with sensing gap by mapping both images into one common subspace for matching. The common subspace shall have the property that images of same identity are closer to each other than images of other identities. Ideally, subspace learning-based methods have the least requirement of computing resources than other methods. Nonetheless, it is troublesome to find such a common subspace. Up to now, researchers make lots of efforts in this way, but their common subspace is abstract and hard to analysis, which have not reached the expected consequent; 3) invariant feature-based methods solve the sensing gap problem by decoupling domain-variant and domain-invariant features. NIR images and VIS images can be regarded as the combination of these two kinds of features. It is essential to decouple with feature extractors. As entangling as high-level features, current methods are not able to separate them perfectly. In which case, domain-invariant features will be influenced by domain-variant feature, which further influence the recognition. On the method level, sensing gap still remains the biggest challenge in NIR-VIS face recognition.

These challenges necessitate further research to determine optimal methods for NIR-VIS face recognition.

1.3 Significance and Contribution

1.3.1 Basic Network

Current image thesis-based methods are modifications of Generative Adversarial Networks (GAN). GANs are firstly introduced in 2014[17]. The main structure of GAN includes two networks, a Generator (G) and a Discriminator (D). Two networks (generator and discriminator) are trained at the same time and compete in a minimax algorithm. This adversarial method avoids some difficulties in the practical application of some traditional generative models and cleverly approximates some unsolvable loss functions through adversarial learning. Numerous GANs are proposed in recent years with a different practical purpose, e.g. Cycle-GANs[78] are designed for unpaired image-to-image translation. Typically, the structured processing of a typical GAN generator is that coarse, low-resolution features are hierarchically refined through upsampling layers, locally blended through convolutions, and non-linearly introduced to introduce new details. Such architecture may essentially restore the surface features of the image, but it does not naturally synthesize a more realistic image, that is, the rough features ensure the presence of image details, but do not control their precise position, details are fixed in image coordinates. Style-GAN 3[33] was proposed by NVIDIA in 2021, which is the 3rd generation of Style-GAN[34][35]. Differently, Style-GAN 3 interprets all signals in the network as continuous and makes slight adjustments to the architecture to ensure that unwanted information does not leak into the layered synthesis process. At the same time, its internal representation is significantly improved, in which absolute translation and rotation can be achieved even at the sub-pixel scale. These advantages of Style-GAN 3 are extremely suitable for NIR-VIS face recognition task.

1.3.2 Proposed Method

In this thesis, I propose a brand-new network, call Cyclic Style Generative Adversarial Network (CS-GAN). The network introduces Style-GAN architecture into NIR-VIS face recognition field for the first time. The CS-GAN is composed of two pre-trained Style-GANs, one for VIS image synthesis and another for NIR image translation. Both GANs are composed of a mapping network, a synthesis network, and a discriminator network. The mapping network takes latent vector z as input which is sent into a mapping network with 2 fully-connected layers. Through the mapping network, a new latent code w is generated in the latent space W. The latent code w is not used as the feature map of the generated image, but to control the feature map of the following synthesis network, thereby indirectly controlling the features of the output image. Because of generating images in different domains, the two synthesis networks have different structures. The synthesis network for VIS image synthesis has the FTheier transform layer, 14 synthesis layers, and an extra to-RGB layer, which is as same as the original Style-GAN; the synthesis network for NIR image synthesis has only the Fourier transform layer and 14 synthesis layers. The discriminator networks in both GANs have the same structure, which is applied from Style-GAN 2. The discriminator network first converts the image into a feature matrix with a mapping network, then reduces the feature dimension through the down-sampling of n blocks, finds the standard deviation in a mini-batch, and finally passes a custom convolution and a self-defined fully connected layer to output the final image classification result. The result of the discriminator is measured by logistic loss. Because of the unpaired property of NIR-VIS face datasets, I adapt the same general structure of Cycle-GAN, using the cycle loss to constrain the process of synthesis. More specifically, I use the same latent code w to synthesize images in both domains and calculated the cycle loss between images in the same domain. Image synthesis between different domains has the same purpose as style transferring. Based on such property, I proposed the cyclic subspace learning method. I control the learning of the network in two-stage, latent subspace, and final synthesized style. For the first stage, I use the proposed latent loss to control the learning of latent subspace W which has further control over features of synthesized images. In the second stage, I apply a pre-trained VGG-16 [46] networks to calculate the style loss which consists of two perceptual loss[29], to have further control of style (domain property). The model is trained and tested on CASIA NIR-VIS 2.0 database[42] with 99.60% accuracy and 98.22% TAR@FAR = 0.1%, which is better than state-of-the-art methods.

1.4 Objectives

The objectives of this thesis are as follows:

- To evaluate properties of current NIR-VIS face datasets and find out their advantages and disadvantages.
- To evaluate state-of-the-art NIR-VIS face recognition methods and analysis their contribution.
- To apply Style-GAN network to NIR-VIS face recognition task for the first time and combine it with Cycle-GAN structure, in which I take NIR-VIS recognition task as style transferring and image translation task.
- To propose cyclic subspace learning for having better control over latent subspace and style transferring simultaneously.
- To improve the accuracy of NIR-VIS face recognition on the CASIA NIR-VIS 2.0 database.

1.5 Thesis Overview

The structure of the thesis is shown as follows:

Chapter 1 introduces the research background and the practical target of the selected topic, demonstrating the property and advantages of NIR-VIS face recognition. Secondly, it discusses the currently existing problem in NIR-VIS face recognition field and the potential improvements. Thirdly, the objective of the thesis is illustrated.

Chapter 2 reviews the current research status of NIR-VIS face recognition field. First and foremost, it introduces current public NIR-VIS face datasets with their characteristics and protocols. Besides, it discusses state-of-the-art relevant research from three categories, image synthesis-based methods, subspace learning-based methods, and invariant feature-based methods, and the potential improvements.

Chapter 3 introduces the proposed Cyclic-Style Generative Adversarial Networks. Based on the foundation of Style-GAN 3 networks, I introduce a cyclic derivative network structure, which is fine-tuned for cross-domain synthesis. Also, it proposes cyclic subspace learning in which the learning of features is controlled by the proposed latent loss function, and the learning of image style is controlled by the style loss function.

Chapter 4 represents the experimental settings and results of the proposed method. By testing on the mainstream dataset, the proposed method achieves the state-of-the-art result. Additionally, it displays the visualization of results for understanding the improvement in image level.

Chapter 5 summarizes the overall work and provides an overview of future direction based on existing progress.

Chapter 2

Literature Review

2.1 NIR-VIS face Dataset

Datasets are vital for helping computers to learn CV tasks, and a critical gap in the advancement of NIR-VIS face recognition techniques is the development of robust datasets. Capturing face images in NIR is much more challenging than in VIS; for instance, traditional VIS face images can be collected on the Internet. However, NIR images are much more uncommon on these publicly available platforms. As such, the cost associated with developing a dataset comprised of NIR face images would be prohibitively expensive when considering the high costs of NIR photography equipment and special capturing conditions (i.e., low light). The largest NIR-VIS face dataset that is presently available is the CASIS NIR-VIS 2.0 Face Database, however, this dataset is still far smaller than typical VIS face datasets, such as CASIA-WebFace Dataset[76] or Labeled Faces in the Wild[26]. This section will introduce four popular datasets used in NIR-VIS face recognition research. Table 2.1 summarizes currently available NIR-VIS face datasets with their properties, such as the number of subjects, capture environment, resolution of images, and whether or not the datasets contain paired images.

Dataset	# of subjects	Environment	Resolution	Paired
The CASIA NIR-VIS 2.0 Face Database	725	Indoor	640 × 480	No
Oulu-CASIA NIR&VIS Facial Expression Database	80	Indoor	320 × 240	No
The BUAA-VisNir Face Database	150	Indoor	640 × 480	Yes
Heterogeneous Face Recognition across Pose and Resolution	200	Indoor	-	No

Table 2.1: Summary of the available datasets for NIR-VIS face recognition.

2.1.1 The CASIA NIR-VIS 2.0 Face Database

The CASIA NIR-VIS 2.0 Face Database is the largest NIR-VIS face dataset presently available and contains 17,580 images of 725 subjects (each of which has a different number of corresponding images). VIS and NIR face images of each subject range in quantity from 1 to 22 and 5 to 50, respectively. Each image has a raw resolution of 640 × 480 pixels and is cropped to 128×128 pixels by eye coordinates. The dataset features two protocols, algorithm development, and performance reporting. In the algorithm development, the dataset is divided into a training set and a testing set; the testing set consists of VIS gallery images and NIR probe images. In algorithm development, parameters can be tuned and fixed via training and testing. In the performance reporting, the dataset is divided into ten sub-experiments, in which the Rank-1 recognition rate and verification rate can be calculated.

2.1.2 Oulu-CASIA NIR & VIS Facial Expression Database

The Oulu-CASIA NIR & VIS Facial Expression Database[68] was not exclusively designed for NIR-VIS face recognition task, and instead was developed for research of facial expression recognition. This dataset consists of 80 subjects with 2,880 video sequences in the resolution of 320×240 pixels. Image frames can be extracted from the video sequences for use in the NIR-VIS face recognition research. The dataset includes facial images spanning several races and traits of human beings. One challenge with this dataset is that video sequences were mainly collected under three different conditions, including normal light, weak light, and nearly dark; thus, the number of NIR images in nearly dark is quite limited. In addition, because the dataset was not intended specifically for NIR-VIS face recognition tasks, an external protocol must be used for performance evaluation.

2.1.3 The BUAA-VisNir Face Database

The BUAA-VisNir Face Database[25] was developed in 2012 and includes 150 subjects with 40 images per subject. For each subject, there are 9 NIR-VIS image pairs at a resolution of 640×480 pixels. Paired images were captured simultaneously using a multi-spectral imaging device. Among these subjects, 50 subjects are used for training and 100 subjects are intended for testing. Similarly to the Oulu-CASIA NIR&VIS Facial Expression Database, the BUAA-VisNir Face Database was not intended specifically for NIR-VIS face recognition, and as a result, faces in images are not aligned to coordinates in this dataset, which are presented in different views. Therefore, pre-processing is needed to facilitate NIR-VIR face recognition. The BUAA-VisNir Face Database does not provide performance evaluation protocols, thereby necessitating the development of external protocols to measure results.

2.1.4 Heterogeneous face recognition across Pose and Resolution

The Heterogeneous face recognition across Pose and Resolution (HPR) dataset[56] is based on the purpose that poses variation and imaging distance will affect the recognition performance. For NIR images, pose variation will cause blurring or noise, and distance variations will affect the resolution of the face image; this dataset addresses the challenges associated with pose variation and distance variation. The dataset includes 200 subjects, 50 of which are used for training and 150 for testing. The dataset includes human facial images with different indoor scenes as backgrounds. There are no aligned human faces in the dataset. The dataset contains three protocols to address the effect of the resolution, pose variation, and distance variation, respectively. For evaluation, the Rank-1 recognition rate is calculated.

2.2 Methods

In addition to datasets, the methodology used for NIR-VIS face recognition is also critically important. Current methods are mostly based on previous research on traditional VIS face recognition, in some cases, combine with other means to adapt to the particularities of NIR-VIS face recognition. For instance, protocols are typically utilized to alleviate domain gaps, low resolution, and over-fitting. Through years of research, there are now diverse methods that can be roughly classified into three categories according to their motivation, (A) Image Synthesis-based methods; (B) Subspace Learning-based methods; (C) Invariant Feature-based methods. This section will review each of these three state-of-the-art methodologies.

2.2.1 Image Synthesis Based Methods

Image synthesis-based methods compose VIS face images from NIR face images, and perform recognition tasks between generated VIS face images and gallery VIS face images; this process thus changes NIR-VIS face recognition to Visible light to visible light (VIS-VIS) face recognition. Currently, VIS-VIS face recognition research developed robust systems with extraordinary performance. There are two main ways with which to generate VIS images from NIR images; the traditional way involves mapping of NIR image patches with their corresponding VIS image patches, then combining these target VIS image patches for a full-scale VIS face. Researchers tend to utilize encoder-decoder networks to synthesize VIS images, such that NIR images are encoded to features, and the synthesized VIS images are outputted with transformed features. Image synthesis-based methods provide the means to artificially generate new images from existing datasets. The synthetic images can be different from their source while crossing different modalities. Because the NIR and VIS images are collected in different spectra, the images develop discrepancies in contours, textures, and color. Therefore, the core of NIR-VIS face recognition is to solve the randomness of modality variation between the NIR and VIS fields. Based on current results, image synthesis-based methods can effectively and directly transform NIR images into VIS images. Image generated through image synthesis-based methods can be used in alternative ways. For example, synthetic images can be created and used as training data, thereby addressing some existing limitations mentioned in Section II. This subsection will focus on advanced image synthesis-based methods and summarizes their main contributions.

Traditional Method

Traditional methods in NIR-VIS image synthesis use the implicit local linear mapping between the NIR and VIS spectra. The overall network structure consists of two parts. In the first part, researchers apply different extractors to extract features – primarily, structural information – from both spectral mains. In the second part, the researcher then attempts to conduct reliable linear mapping between features, using constraints such as the relative position of facial features. Once trained, networks synthesize VIS images by mapping NIR images to their corresponding VIS images.

One representative work is [7], in which the authors used manifold learning to accomplish the mapping function. Specifically, they divided images into patches, extracted patch features through Linear Binary Pattern (LBP)[57], and learned the corresponding relationship among VIS patches and NIR patches through K Nearest Neighbors (kNN)[63]. While learning with KNN, the authors used a constant constraint between the distance of pairs to ensure that the distance between patches would not change significantly after mapping to preserve local geometry. Through the learning of these patches, the authors obtained a NIR manifold and a VIS manifold which was approximately isometric, and VIS images could be combined like the pieces of a puzzle by finding counterparts in the NIR manifold and the VIS manifold. In this work, manifold learning alleviated the problem of non-linearity between the NIR and VIS images. The simultaneous learning of two different manifolds reduced image dimensions and provided a mechanism through which linear mapping between two different spectra could be accomplished. The authors of [31] used a similar idea of turning non-linearity into linear mapping but with dictionary learning. They built up a joint framework to learn dictionaries in both domains while constraining sparse representation of pair-wise images. The sparse representation of K-Singular Value Decomposition (k-SVD)[1] helped their algorithm learn dictionaries which needed only very sparse information when expressing and accurately reconstructing original images after mapping.To accurately reconstruct VIS images from NIR images, sparse matrix X was shared between dictionaries, as

$$\begin{gathered} \underset{D_N, D_V, X}{\arg\min} \left\| \begin{pmatrix} Y_V \\ Y_N \end{pmatrix} - \begin{pmatrix} D_V \\ D_N \end{pmatrix} X \right\|_F^2 \\ \text{subject to } \forall i, \|x_i\|_0 < K, \end{gathered} \tag{2.1}$$

where Ys are data, Ds are the dictionaries and K is the optimal sparsity level determined by Peak signal-to-noise ratio (PSNR). The authors then generated VIS images y_V following linear mapping,

$$\begin{aligned} & x = \underset{x}{\arg\min} \|y_N - D_N x\|_F^2, \\ & \text{subject to } \forall i, \|x_i\|_0 < K \end{aligned} \tag{2.2}$$

$$y_V = D_V x. \tag{2.3}$$

The NIR-VIS CASIA database used in [31] was not strictly pair-wise, which was consistent with assumptions made by the authors. Thus, distortion appeared in their visual results.

CNN-based Method

Generative networks evolved into the CNNs, the functionality of which surpassed traditional methods. Due to the small number of training samples, image synthesis-based methods that utilize CNNs were pre-trained on VIS dataset to extract features in the NIR domain and then synthesize VIS images through extracted features. In [40], the authors improved a classical pre-trained CNN, adding a cross-spectral hallucination part. Their basic architecture was a CNN which was pre-trained on a VIS dataset. To achieve the full potential of the network, the authors developed a cross-spectral hallucination CNN to transform NIR images into VIS images. The hallucination network was composed of three hourglass structured sub-networks working for Brightness, blue minus luma and red minus luma (YCbCr) color space. The hourglass structure mimicked the encoder-decoder scheme,

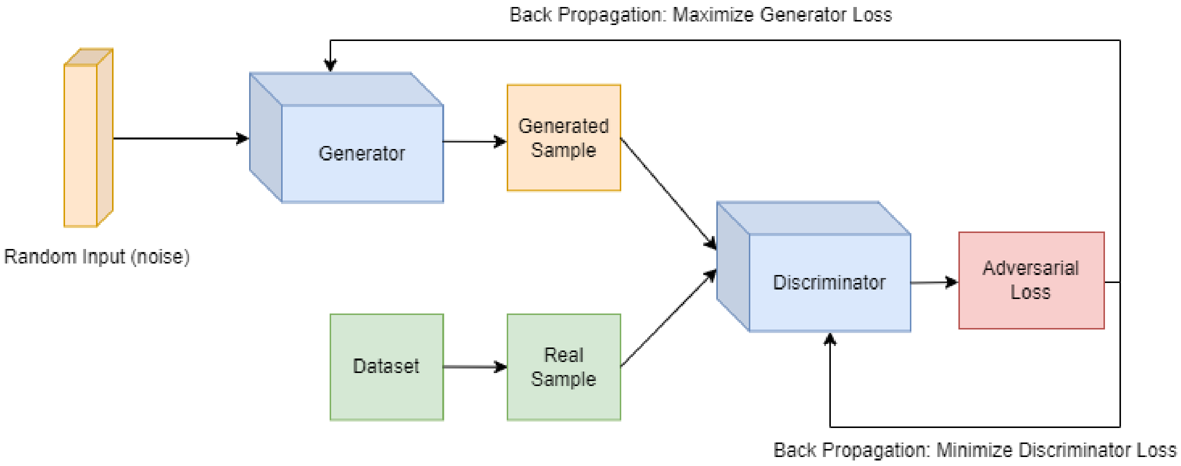


Figure 2.1: General structure of Generative Adversarial Networks.

where in middle layers had a narrower depth than the first and last layers. According to their designation, the luminance Y channel had the largest sub-network, because luminance difference was the major problem in the case of NIR-VIS face recognition. Through this hallucination network, mined NIR patch inputs were transformed into the corresponding VIS spectrum. In addition to image generation, the original NIR images were blended in their generated VIS images, on the luminance Y channel, as a safeguard mechanism to preserve information. Although improvements were made, the approach used by the author in [40] had a drawback. For instance, generation in each channel will cause misalignment and blend with original NIR images.

GAN-based Method

More recently, adversarial learning is a heated topic in the field of machine learning, especially since Goodfellow et al. proposed GANs in 2014. As a generative network, GAN can synthesize images in an adversarial way to yield state-of-the-art results. GANs are composed of two different networks – generator network and discriminator network. The goal of the generator is to generate passable results with which to fool the discriminator without being caught. In contrast, the goal of the discriminator is to discriminate whether the results from the generator are fake. The overall structure of GANs is shown in Fig.2.1.

And the loss function of GANs is called adversarial loss (L_{adv}),

$$\begin{aligned} L_{adv} =& E_{x \sim p_{data}(x)}[\log D(x)] \\ &+ E_{z \sim p_z x}[\log(1 - D(G(z)))], \end{aligned} \tag{2.4}$$

where D represents discriminator network, G represents generator network and z is the random input vector.

Song et al. first applied GAN in their work[66] and made adjustments for better performance when applied to NIR-VIS face recognition. In addition to generating the global features of faces, the authors added an extra GAN to serve as a local path for periocular regions, which are indispensable in face recognition. Using this approach, the generated VIS face images were vivid enough for VIS face recognition with images in the gallery. With pristine VIS images and generated VIS images in discriminative feature learning space, the authors used Adversarial Loss (L_{adv} in Equation 2.4 a Class-wise Variance Discrepancy (L_{CVD} in Equation 2.5) and Cross-Entropy Loss (L_{cls} in Equation 2.6) for identity recognition to guide the network to learn domain-invariant face representation. Loss functions were organized as follows,

$$L_{CVD} = \sum_{c=1}^{C} E(\|\sigma(F_c^V) - \sigma(F_c^N)\|), \tag{2.5}$$

$$\sigma(F) = E((F - E(F))^2), \tag{2.6}$$

$$L_{cls} = \frac{1}{|N| + |V|} \sum_{i \in \{N,V\}} \mathcal{L}(WF_i, y_i), \tag{2.7}$$

in which $\sigma(*)$ is the variance function; F_c^V and F_c^N are c_{th} class's features; W is the Softmax normalization; $\mathcal{L}$ is the cross-entropy loss. GAN is a network structure with a very high upper limit. To fully explore it, researchers make an extra analysis of the property of the target field. He et al.[19] concentrated on the difference between pose and texture difference. Based on these differences, they established an adversarial learning framework, consisting of a pose correction network, texture-inpainting network, and fusion-warping network. For a given input image, the pose correction network estimated normalized shape information whereas the texture-inpainting network produced pose-invariant facial texture representation. The pose correction network worked with an estimation of a dense UV correspondence field (two-dimensional texture coordinates that correspond with the vertex information for geometry). The UV field was a combination of UV facial texture space and RGB image space, where UV facial texture space was a contiguous Two dimensional (2D) atlas that contained a manifold of Three dimensional (3D) face. The UV field is mostly used for 3D analysis. The application of the UV field in NIR-VIS face recognition can be beneficial. On the one hand, this approach allows the preservation of complete shape information as the UV field specifies the pixel-wise relationship between facial texture and 2D map. On the other hand, representations in the UV field are flattened surfaces of 3D faces, which approaches 3D-aware. To train the pose correction network, the authors first obtained mean ground truth UV field $\overline{UV}$ of faces by 3D Morphable Model (3DMM)[58] and extracted estimated 3D shape information. Next, the authors used a cylindrical unwrapping method to map shape information to the UV space[5]. Then, while training the generative network, pose correction network G_p was guided within the mean ground truth $\overline{UV}$, as UV loss (L_{uv}) is defined below.

$$L_{uv} = \|G_p(X) - \overline{UV}\|_1. \tag{2.8}$$

NIR images capture less texture information than VIS images. To alleviate this limitation, the texture inpainting network G_t improved performance by encoding face texture into identity representations and decoding into the VIS domain. Further, to synthesize images that are more realistic and to eliminate intra-class variation, the authors adjusted the loss function of G_p discriminator L_{D_t} from the original adversarial loss, as follows,

$$L_{D_t} = E_{X \sim p_{data}}[-\log(1 - D_t(G_t(X))) - \log(D_t(X))]. \tag{2.9}$$

In this formulation, D_t integrated both the synthesized image and the pristine images as input. Therefore, G_t should have better performance to deceive D_t. In circumstances in which G_p and G_t worked simultaneously, the authors proposed a fusion-warping network to combine the output of two generative networks. The fusion-warping network was comprised of several convolution layers which were fed with the output of G_p, the output of D_t, and the output of the second last layer of G_t (i.e., the facial texture feature map). To supervise the fusion warping net, authors employed a multi-scale discriminator D_r which could achieve high-resolution face completion. More specifically, Haar wavelet decomposition was applied to input data. To supervise the fusion-warping network, the authors employed a multi-scale discriminator D_r which could achieve high-resolution face completion. More specifically, Haar wavelet decomposition was applied to input data. The discriminator supervised Haar wavelets[53] in two different frequencies, including low frequency (D_{rl}) and high frequency (D_{rh}). Thus, the generators can create globally and locally consistent results. The loss of the fusion-warping network was

$$L_{G_F} = E[-\log(D_{rl}(\phi_{rl}(F(X)))) - \lambda \log(D_{rh}(\phi_{rh}(F(X))))], \tag{2.10}$$

where $F(x)$ is the warped fusion from G_t and G_p; $\phi()$ is the decomposed wavelet coefficients. In their experiment, λ was set to 10, for emphasizing high-frequency information. The final output of their framework was matched with ground truth VIS images with perceptual loss and pixel-wise L1 loss. In [13], the authors attempted to address misalignment in the NIR-VIS field via the development of a novel framework called Pose Aligned Cross-spectral Hallucination (PACH). PACH used two-stage procedure, Unsupervised Face Alignment (UFA) and Texture Prior Synthesis (TPS). UFA aligned facial shape of NIR images to corresponding VIS images in an unsupervised manner. The network of UFA consisted of shape encoder Enc_s, identity encoder Enc_i, AdaIN residual blocks $AdaRes$[27], and decoder Dec. Enc_s was the extractor for facial information which used the UV map (M_N) of NIR images as input. And Enc_i extracted identity information that was irrelevant to facial information. The AdaIN residual blocks worked as

$$AdaIN(z, \gamma, \beta) = \gamma(\frac{z - u(z)}{\sigma(z)}) + \beta, \tag{2.11}$$

where γ and β are identity information extracted by Enc_i; z is the means of the facial information, and $\mathbf{u}(z)$ and $\sigma(z)$ were the channel-wise mean and standard deviation of the facial information. Through $AdaRes$, identity information and facial information were disentangled and decoded into image space. In general, UFA worked as a generator that aligned the facial shape of NIR images to the corresponding VIS images by changing the UV map. The authors adopted pixel-wise L1 loss to constrain the output with input and applied another Identity Preserving loss (L_{ip}) to ensure that identity information was preserved:

$$L_{ip} = E_{I'_N, I_N}[\|D_{ip}(I'_N) - D_{ip}(I_N)\|_2], \tag{2.12}$$

wherein D_{ip} is pre-trained LightCNN to extract identity features from I_N and I'_N. Like other GANs, there was an adversarial loss to improve the visual quality of I'_N. UFA aligned images in UV map of NIR domain, then TPS replaced UV map with VIS domain with a texture prior T which provided specific guidance related to texture information. In the TPS stage, texture prior T was concatenated with the aligned NIR images and was fed into a generator to synthesize VIS images. Through this process, the pixel-wise translation was supervised by pixel loss. The authors used a Total Variation Regularization to reduce artifacts:

$$\begin{aligned} L_{tv} = &\sum_{c=1}^{C} \sum_{w,h=1}^{W,H} |G(I''_N, T)_{w+1,h,c} - G(I''_N, T)_{w,h,c}| \\ &+ |G(I''_N, T)_{w,h+1,c} - G(I''_N, T)_{w,h,c}|, \end{aligned} \tag{2.13}$$

where W and H are size of images. Besides these two loss functions, there were still adversarial loss and cross-entropy loss. Hu et al.[23] focused their efforts on misaligned images.

In this case, the authors created VIS neutral faces, containing almost no face variations, and built up their Dual Face Alignment Learning based on these images. The VIS neutral faces were face images that had residual-unrelated discriminative VIS face features. To generate and make full use of VIS neutral faces, their approach was composed of three parts, including Feature-level Face Alignment (FFA), Image-level Face Alignment (IFA), and Cross-domain compact Representation (CdR). The baseline of their network was a pre-trained Teacher-Encoder CNNs (TeEn-CNNs) and a Student-Encoder CNNs (StEn-CNNs) with a decoder that learned knowledge from TeEn-CNNs and gradually gained the ability to extract neutral features from non-neutral images in both domains. FFA was presented for guiding the learning of VIS neutral facial representations. This FFA could efficiently help TeEn-CNNs guide StEn-CNNs to learn how to encode domain-invariant and residual-independent representations, thus reducing the intra-class variations. Once StEn-CNNs learned encoding neutral features correctly, there should be another image-level alignment between reconstructed face images and VIS neutral images with which IFA worked. IFA introduced intensity constraints in both NIR and VIS reconstructed images to enforce content consistency. These intensity constraints could maintain low-frequency information but can result in over-smoothing. To preserve more detailed information, there were other constraints in IFA, called texture consistency constraints, which converted images into gray-scale and used a Prewitt filter as a prominent part extractor. The final CdR part focused on identity information in images, in which image perspective, intra- and inter-modality negative pairs, inter-modality positive pairs, and inter-semantic relationships were considered. The negative pairs should be stretched to increase semantic variation, through cosine similarity, and the positive pairs should be compressed for a compact representation. In addition, the inter-semantic relationship was constrained to be consistent. These GAN based methods make full use of the functionality of GANs in adversarial learning and compete to produce more realistic images. In GANs, the discriminator uses real VIS images to determine whether the generated images are real or not; therefore, the pair-wise problem remains. Researchers gradually developed methods with GANs and their variations in NIR-VIS face recognition field. Some studies generated pair-wise images for further matching, and others generated VIS images and then exploited a well-developed VIS face recognition system. Cycle-GAN was proposed by Zhu et al. in 2017 and is capable of image-to-image translation in the absence of paired-wise images. This approach is suitable for NIR-VIS face recognition, and is optimized by two loss functions, adversarial loss (L_{adv} in Equation 2.4) and cycle-consistency loss (L_{cyc}) which is defined below:

$$L_{cyc}(G,F) = E_{x\ p_{data}(x)}[\|F(G(x)) - x\|_1] + E_{y\ p_{data}(y)}[\|G(F(y)) - y\|_1], \tag{2.14}$$

, where G and F were two generators. The general structure is shown in Fig.2.2.

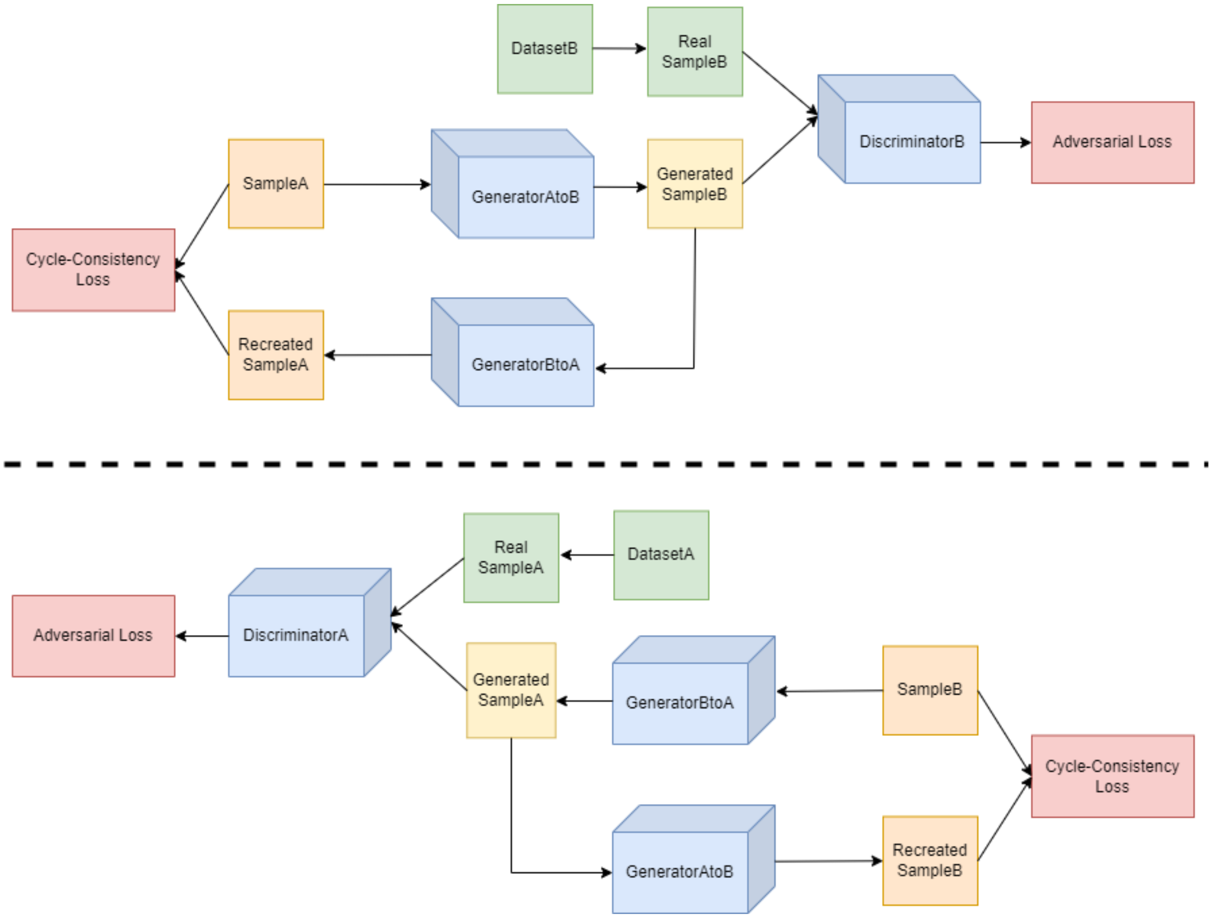


Figure 2.2: General structure of Cycle-GANs. The A and B represent samples in two domains, respectively. The GeneratorAtoB converts images from domain A to domain B, while GeneratorBtoA vice versa.

Modified Cycle-GANs are now becoming mainstream research focus in Image Synthesis-based methods. Among the literature, recent efforts focused on both applying Cycle-GAN to image recognition tasks and furthering the development and robustness of frameworks to maximize the potential of this approach. In [72], the authors built up a robust system for NIR-VIS face recognition, from face detection and alignment, to NIR-VIS image translation than to the final translated VIS face image recognition. In the first stage, the detection and alignment were done by Multi-Task Cascaded Convolutional Network (MTCNN). The MTCNN consisted of three networks, Proposed network (P-Net), Refining network (R-Net), and Output network (O-Net). The P-Net generated a list of candidate windows, whereas R-Net rejected the wrong candidates and O-Net outputted five facial landmarks. In the second stage, aligned NIR face images were translated to VIS images with the same identity using the Cycle-GAN framework. In addition to the loss function in the Cycle-GAN framework, the authors added a similarity preservation function, constrastive loss (L_{con}), to constrain the learning of the mapping function,

$$L_{con}(l, i_1, i_2) = (1 - l)\{\max(0, m - d)\}^2 + ld^2, \tag{2.15}$$

where is are input vectors selected unsupervised; d is the cosine distance; and l is the binary label calculated from input images(i.e. l equaled one if they were positive pairs, otherwise l equaled zero. With further study, Dou et al.[11] denoted NIR-VIS translation task as an asymmetric translation task where translation between domains was uneven complexity. Based on this denotation, the authors designed Asymmetric Cycle-GAN to deal with this asymmetric translation. Unlike the original Cycle-GAN, two generators G_1 and G_2 in Asymmetric Cycle-GAN were different, a simple U-Net for VIS-to-NIR translation (complex to simple) and a complex U-Net for NIR-to-VIS translation (simple to complex). VIS-to-NIR translation was a dimension-reduced image translation process and NIR-to-VIS translation was dimension-ascending image translation respectively. Given that the numbers of down-sampling operations in U-Net were modifiable, G_1 had 5 down-sampling convolution layers for NIR-to-VIS translation and G_2 had only 3 layers whereas 3 down-sampling convolution layers were able to extract most shallow information. In addition to the use of different generative networks, the authors added another pre-trained U-net as an edge detection network E_d which extracted edge for an additional edge loss. This edge loss (L_{Edge}) was used for retaining necessary edge details in generated VIS images. Wang et al.[70] also adopted Cycle-GAN and added an extra pixel consistency loss between the generated images and the pristine images to constrain generated images. Pixel consistency loss (L_{pc}) was formed as,

$$L_{pc}(G, F) = E_{i_V \ P(i_V)}[\|G(i_V) - i_N\|_1] + E_{i_N \ P(i_N)}[\|F(i_N) - i_V\|_1]. \tag{2.16}$$

Apart from improving image translation, in [3], the authors used a two-step framework that combined image translation and feature learning. The image translation part used Cycle-GAN as their baseline. To resolve structural variations between the two domains, a Siamese network was added to preserve the contents of images. In the training stage, the Siamese network was inputted with generated images and their positive and negative pairs, and calculated contrastive loss was. Generated images were more realistic when the loss between them and their corresponding positive pairs was smaller, and between negative pairs, larger. Therefore, the loss function of the first part was that the original loss function of Cycle-GAN was integrated with contrastive loss as

$$L_{con}(x_1, x_2.x_3) = L^n_{con}(x_1, x_2) + L^p_{con}(x_1, x_3), \tag{2.17}$$

$$L^n_{con}(x_1, x_2) = \max(0, m - \|x_1 - x_2\|^2_2), \tag{2.18}$$

$$L^p_{con}(x_1, x_3) = \|x_1 - x_3\|^2_2. \tag{2.19}$$

where x_2 and x_3 were positive sample and negative sample respectively, and x_1 is the query sample. For the feature learning part, the authors adopted pre-trained ResNet-101[43] as their backbone network. Besides the original network, an additional angle margin loss L_{angle} was added,

$$L_{angle} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{e^{s(\cos(\theta_{y_i}+m))}}{e^{s(\cos(\theta_{y_i}+m))} + \sum_{j=1,j\neq y_i}^{n} e^{s\cos\theta_j}}, \tag{2.20}$$

where θ_{y_i} is the ground truth angle; m is the angular margin penalty; and s is the feature scale. These researchers demonstrated that Cycle-GAN is an effective framework for the NIR-VIS facial recognition.

Performance Evaluation

Table2.2.1 summarizes state-of-the-art image synthesis-based methods. Current methods can be used to obtain excellent performance on publicly available datasets. Although these datasets are relatively small, high rank-1 accuracy shows the high-standard ability of these approaches in facial recognition. However, their verification rates at 0.1% False Accept Rate (FAR) are not as good as their recognition rates, thereby indicating that these approaches can lead to incorrect facial recognition [70].

The central idea of the traditional way is to develop a highly accurate linear mapping between the NIR and VIS domain. For both domains, though different in manifold

Methods	Dataset	Rank-1	0.1%FAR	Characteristic
[7]	-	97.3%	-	Manifold learning; local geometry preservation.
[31]	CASIA	78.46%	85.8%	Dictionary learning; not strictly paired.
[40]	CASIA NIR-VIS 2.0	96.41%	-	YCbCr color space; misalignment in appearance.
[66]	CASIA NIR-VIS 2.0	98.15%	97.18%	Extra local periocular regions.
	BUAA-Visnir	95.2%	88%	
	Oulu-CASIA	95.5%	60.7%	
[19]	CASIA NIR-VIS 2.0	99.5%	97.5%	3D-aware of face; emphasizing high-frequency information.
	BUAA-Visnir	99.7%	97.8%	
	Oulu-CASIA	99.9%	90.7%	
[13]	CASIA NIR-VIS 2.0	98.9%	98.3%	Align facial shape between NIR and VIS; Pixel-wise translation.
	BUAA-Visnir	98.6%	93.5%	
	Oulu-CASIA	100%	88.2%	
[23]	CASIA NIR-VIS 2.0	98.9%	93.8%	VIS neutral faces with no face variations; pre-trained generators guidance.
	BUAA-Visnir	100%	94.0%;	
	Oulu-CASIA	100%	93.8%	
[72]	ONVF	99.8%	-	Face detection and Alignment.
[11]	-	-	-	Asymmetric translation; edge correction.
[70]	WHU VIS-NIR	99.3%	64.0%	Pixel-wise translation.
	Oulu-CASIA	96.5%	61.3%	
[3]	CASIA NIR-VIS 2.0	99.40%	98.74%	Siamese structure for more contents of image; positive- and negative-pair learning.

Table 2.2: Performance of Image Synthesis Based Methods. [7] used their dataset which was not named, thus there is no name on it. [11] is an image-image translation task, which has no recognition evaluation.

and dictionary, they all tried to deal with the non-linear mapping between NIR and VIS images. However, a comparison with recent works illustrates that the images generated using the traditional approach is relatively low quality and are often affected by distortion. Such a result is reasonable given that similarity between local patches cannot guarantee the similarity between global faces. CNN-based methods can be regarded as a transition approach, in that they offer better performance than traditional approaches, but there is no extra supervision of image synthesis procedures, and their network structures are auto-encoders composed of CNN. GAN-based methods further improved image synthesis by adding discriminator networks to ensure that synthesized images are similar to real images. GAN-based methods are capable of yielding impressive results, however, the related literature is dominated by studies that assume that images in the VIS and NIR domains are paired. This assumption violates the unpaired nature of most publicly available datasets. GAN-based methods are anticipated to yield strong performance in applications that utilize reliable paired datasets. More generally, the final process of recognition associated with image synthesis-based methods may be characterized by some amount of redundancy because features will be extracted from synthesized images and real images. This redundancy increases the complexity and time-intensive nature of these methods. Despite these limitations, image synthesis-based methods are worthwhile approaches for continued development improvements to the robustness of generative networks will further drive performance improvements.

2.2.2 Subspace Learning Based Methods

Subspace learning aims to map from a high-dimensional space to a low-dimensional subspace while preserving as much useful information as possible. It is critical to map simply and efficiently to minimize calculations and improve the learning of valuable features. Several algorithms have been developed and well-used, such as Linear Discriminant Analysis (LDA)[16], PCA and Canonical-Correlation Analysis (CCA)[38]. Subspace learning-based methods were the mainstream approach for use in NIR-VIS face recognition before the revolution of CNNs. Because NIR and VIS face images lie in different modalities, simple projections between two spaces result in low-quality images that are affected by distortion, representing a key heterogeneous face recognition problem. To map between the two spaces, researchers derived novel solutions involving the projection of images into a low-dimensional common subspace, where matching between NIR images and VIS images can be done more straightforwardly. This section will provide an introduction to recent subspace learning-based methods in NIR-VIS face recognition.

Traditional Method

Among subspace learning-based methods, there are ways of projecting into a common subspace through dimension reduction, in which data reduction and interpretation are done simultaneously. Traditionally, researchers preferentially extracted features first and utilized subspace learning methods to project features in different domains into one common subspace. Thereafter, the recognition problem would typically be solved as a general eigenvalue problem.

$$S_{inter}V = \lambda S_{intra}V. \tag{2.21}$$

CCA which explored the relationships between two vectors from the same identity provides one mechanism through which to solve this challenge. In [74], Yi et al. reformulated the comparison problem between NIR images and VIS images as correlational regression and therefore applied LDA for reduction, and CCA for interpretation. The authors made use of LDA to transform NIR images and VIS images into feature spaces of lower dimension, which were based on intra-class and extra-class scatter matrices of images in each spectrum. Then in the second step, CCA was used to find two linear projection matrices by maximizing the correlation,

$$\begin{aligned}\rho(w_X, w_Y) &= \frac{E[xy^T]}{\sqrt{E[\|x\|^2]E[\|y\|^2]}} \\ &= \frac{w_X^T C_{XY} w_Y}{\sqrt{w_X^T C_{XX} w_X w_Y^T C_{YY} w_Y}},\end{aligned} \tag{2.22}$$

where w_X and w_Y are two linear projection matrices; x and y are NIR and VIS images in CCA subspace; C_XY, C_XX and C_YY are correlation matrices.Finally, the matching score was computed as the correlation between NIR and VIS images in the CCA subspace. Another alternative for the pre-processing image was identified using a Difference-of-Gaussian (DoG) filter[69]. In [44], the authors used a Lambertian model and adopted DoG filtering to normalize the appearance of input face images from both NIR and VIS spectra; this approach projected images from both spectra into a common space. The DoG filter reduced illumination variation in the low-frequency domain, and image noise and aliasing in the high-frequency domain, which is computed as,

$$D\left(x, y|\sigma_0, \sigma_1\right) = \left(G\left(x, y, \sigma_0\right) - G\left(x, y, \sigma_1\right)\right) * I\left(x, y\right), \tag{2.23}$$

$$G\left(x, y, \sigma\right) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x^2+y^2)/2\sigma^2}. \tag{2.24}$$

To treat all appearance normalization in the same space, the authors decided to apply Multi-Block Local Binary Pattern (MB-LBP) operator[77], computing average values of block sub-regions, to encode local image structures called Local Structure of Normalized Appearance (LSNA), while using a histogram of MB-LBP to represent final feature set. MB-LBP generated an over-complete representation, so the Gentle AdaBoost algorithm[15] was required to remove redundancy and to build effective classifiers. At the final stage, R-LDA was proposed to construct a universal subspace for identifying different individuals. PCA is also a classic and effective method. Klare et al.[37] adopted PCA and built up an ensemble classifier for NIR and VIS images on a random subspace. Their random subspace method used NIR and VIS features extracted through a Histogram of Oriented Gradients (HOG)[49] and uniform LBP[14]. For each iteration, they randomly sampled α feature vectors and computed the mean class vector for each subject using both NIR and VIS images. The intra-class and inter-class scatter matrices were then constructed. Through these scatter, the matrix of eigenvectors $V^{(k)}$ was computed as Equation (19). Then, the final discriminative projection matrix was generated with a PCA projection matrix. Through PCA, mapping was improved with the projection matrix while concentrating on the identity information. In [75], Yi et al. used Restricted Boltzmnn Machine (RBM)[54], a generative stochastic neural network, to learn nonlinear relationships between VIS and NIR images in the space of Gabor wavelet. Gabor wavelet extracted local features from all face points by aligning a face with facial points, then used RBM and PCA to transform heterogeneous data into a common subspace from their extracted features. Their proposed RBM was multi-modal, with an energy function defined by

$$E(\hat{v}_1, \hat{v}_2, h; \Theta) = \frac{1}{2}(\hat{v}_1 - a)^T(\hat{v}_1 - a) + r\frac{1}{2}(\hat{v}_2 - b)^T(\hat{v}_2 - b) - c^T h - \hat{v}_1^T W_1 h - \hat{v}_2^T W_2 h, \tag{2.25}$$

where $\hat{v}_1$ and $\hat{v}_2$ are face images in different domains; W_1 and W_2 are weight matrix for domains; a and b are biases of visible and hidden units; and h is the stochastic hidden units. According to this energy function, the model learned the shared representations. The dimensions of these representations were reduced by PCA and projected into a common subspace where similarity was matched by Cosine distance. Dimension reduction is a straightforward way with which to find a common subspace for the projection. However, extracted features are relatively more relevant, and cannot be strictly constrained. These features may lack some identity information for matching.

Current publicly available datasets include probe images which are found in the gallery sets, and are referred to as “closed-set datasets”. Analyses by some researchers focused on closed-set datasets. Zhu et al. used an assumption that, for each probe image, there existed gallery images with unknown labels, and built up their Transductive Heterogeneous Face Matching(THFM) through two studies. In the first study[80], the authors found a common feature space wherein they could minimize intra-class variation and Maximum Mean Discrepancy (MMD)[6] while maximizing inter-class variation. The MMD was formulated as

$$MMD(X^G, X^P) = \| \frac{1}{N_G}\sum_{p,i} f(x_{p,i}^{Gallery}) - \frac{1}{N_P}\sum_{p,j} f(x_{p,j}^{Probe}) \|, \tag{2.26}$$

where f is the linear function of mapping. Due to such a proposal, their final objective function could be represented as a general eigenvalue problem,

$$S_{inter} w = \lambda(S_{intra} + M + \eta I)w, \tag{2.27}$$

where $M = XLX^T$, $X = [X^G, X^P]$, η is a constant for Tikhonov regularization and $L = [L_{ij}]$ with $L_{ij} = \frac{1}{N_G^2}$ if both x from gallery, $L_{ij} = \frac{1}{N_P^2}$ if both x from probe or $L_{ij} = -\frac{1}{N_G N_P}$ otherwise.The solution w was the identity vector for the probe set. In the second study by Zhu et al.[79], the authors added their THFM to alleviate the domain variance in their transductive subspace. The first part of their method used an approach that was the same as their previous work and focused on minimizing intra-class variance, maximizing inter-class variance, and adding MMD penalization. In the second part, the unified kernel learning was adopted in THFM learned a low dimensional feature space. According to the empirical kernel map K, their object function could be formulated as,

$$K\tilde{S}_{inter}K\omega = \lambda(K(\tilde{S}_{intra} + \beta M + \gamma L)K + \alpha I)\omega, \tag{2.28}$$

where K can be decomposed $K = (KK^{(-1/2)})(K^{(-1/2)}K)$. In [41], the authors introduced the Multi-view Smooth Discriminant Analysis (MSDA), finding projection matrices to a common space that could be seen as a linear transformation. In the feature extraction stage, the authors combined features extracted from multiple methods, HOG, Long-Term Potentiation (LTP)[51] and Scale-Invariant Feature Transform (SIFT))[24]. In the next stage, MSDA sent combined features through Laplacian smoothing[30], specifically, the Discretized Laplacian smoothing method which smoothed the basis vectors of face data from different views. After calculating intra-class and between-class scatter matrices, using the same approach was done for most subspace learning-based methods, the objective function of MSDA could be reformulated as generalized eigenvalue decomposition. Bhattacharya et al.[4] adopted a hash-encoding-based descriptor, Linear Cross-modal Hash Encoding (LCHMHE), to deal with the domain gap. The first part is Logarithmic Pixel Difference Vector (LPDV), in which they compared the central pixel to neighboring pixels without thresholding. LPDV eliminated the luminance part but left the reflected difference, as follows,

$$\tilde{I}_i(x,y) - \tilde{I}_c(x,y) \approx \log_2[G_i(x,y) - \log_2[G_c(x,y)]], \tag{2.29}$$

where i and c represent neighboring and central pixels. After the window slid through the image, there was a Logarithmic pixel difference matrix (LPDM) for each image in both domains. The authors also applied an Intra-similarity Preservation method to preserve the neighboring relationship after mapping into a common subspace; using this method, they used k-means clustering to generate 256 centroids for approximating 256 most representative data points in the LPDV, in which they used k-means clustering method to generate 256 centroids for approximating 256 most representative data points in LPDV. Then, in the third stage, Inter-similarity Preservation, the authors projected LPDM into a common Hamming space where similar identities should display the same binary codes. The transformation was formulated as,

$$\begin{aligned} &\min_{W^{(1)},W^{(2)}} \|Z^{(1)}W^{(1)} - Z^{(2)}W^{(2)}\|_F^2 \\ &\text{subject to} W^{(1)^T}W^{(1)} = I, W^{(2)^T}W^{(2)} = I, \end{aligned} \tag{2.30}$$

where Zs are LPDM and Ws are transformation matrix, hash functions. The such formulation was essentially an eigenvalue problem. Then, the mapped matrix Y was calculated through

$$Y^{(i)} = tr(Z^{(i)}W^{(i)}). \tag{2.31}$$

Through Y and the mean vector $u^{(i)}$, Y was encoded into binary codes as

$$\begin{cases} b_{jl}^{(i)} = 1 \quad \text{if} y_{jl}^{(i)} \geq u_l^{(i)} \\ b_{jl}^{(i)} = 0 \quad \text{if} y_{jl}^{(i)} \leq u_l^{(i)}. \end{cases} \tag{2.32}$$

In this way, all images could be projected into a common space in the form of an 8-bit string. Ultimately, the authors performed the matching function using chi-sq distance and achieved excellent results. The success of this study demonstrated that the transduction assumption is tenable in the close-set datasets. Despite the practicality of this approach, the transduction assumption may cause negative effects; namely, for a public system, there will always be individuals whose identities are unknown. These methods could match the incorrect identities of these individuals

CNN-based Method

Subspace learning-based methods are becoming less popular than they once were, but recent initiatives improve these methods by combining them with CNNs. CNNs work as powerful feature extractors, thus allowing the extracted features to be projected into a common subspace at a low cost. CNNs are typically used in research environments to facilitate the extraction of high-level or low-level features and to project these features into a common subspace. Saxena et al.[65] used metric learning to align domains, such that they employed a CNN model to separate projection matrices to project NIR and VIS images into a common subspace. Their work was founded on the assumption that domain variance could be treated as one of the nuisance factors in heterogeneous face recognition. Thus, they utilized Logistic Discriminant based Metric Learning (LDML) to learn Mahalanobis matrices from pairwise supervision. Their pre-trained CNN was fine-tuned on inter-intra domain pairs to enable learning of the shared projection matrices. In instances in which case matrices were obtained, images in both domains were projected into a common subspace where the domain difference was reduced as much as possible. In [20], the authors explored a low-dimensional subspace while using Wasserstein distance to measure the distance between NIR and VIS distributions. To remove spectrum information, they developed three orthogonal mapping matrices,

$$\begin{gathered} f_i = \begin{bmatrix} f_{shared} \\ f_{unique} \end{bmatrix} = \begin{bmatrix} WX_i \\ P_iX_i \end{bmatrix} (i \in \{N, V\}) \\ P_i^T W = 0 (i \in \{N, V\}) \end{gathered} \tag{2.33}$$

, where WX_i is the unique feature that contained mostly spectrum information and P_iX_i is the shared identity feature. Assuming that NIR and VIS images followed Gaussian distribution in the mapped space, Wasserstein distance[2] measured distance between NIR and VIS distribution:

$$W_2(X.Y)^2 = \frac{1}{2}[\|m_N - m_V\|_2^2 + \|\sigma_N - \sigma_V\|_2^2], \tag{2.34}$$

$$\sigma_N = \sqrt{\frac{1}{n}\sum_{i=0}^{n} x_i^2 - m_N^2}, \tag{2.35}$$

$$\sigma_N = \sqrt{\frac{1}{n}\sum_{i=0}^{n} y_i^2 - m_V^2}, \tag{2.36}$$

where X and Y follow Gaussian distribution; m_N and m_V are means of X and Y.The Siamese structure is an advanced methodology to map the input to the new space, forming a representation of the input in the new space, during feature extraction through CNNs. In [62][61], Reale et al. trained Googlenet[67] on VIS dataset and optimized the trained network to extract coupled features from VIS and NIR images. Training on a large VIS dataset could help CNN extract facial features while dealing with both VIS and NIR images in the latter stage. For heterogeneous face recognition, the network was adjusted to fit data from different domains. Firstly, the authors reduced the number of parameters in the network to alleviate issues related to over-fitting while training NIR images. They removed a fully- connected soft-max classifier(FC layer). Secondly, they coupled two networks by creating a Siamese network, shown in Fig.2.3. The coupled networks were trained on a NIR-VIS dataset simultaneously, but without sharing weights. The authors used two contrastive losses as their loss function:

$$\begin{aligned} L_{l1}(x,y) &= \begin{cases} \|x-y\|_1 & \text{if} l_x = l_y \\ \max(0, (p - \|x-y\|_1)) & \text{otherwise,} \end{cases} \\ L_{l2}(x,y) &= \begin{cases} \|x-y\|_2^2 & \text{if} l_x = l_y \\ \max(0, (p - \|x-y\|_2^2)) & \text{otherwise.} \end{cases} \end{aligned} \tag{2.37}$$

where x and y are different feature vectors.CNNs in Siamese structures can be different; in structures where two networks are not sharing parameters, the structure is called semi-Siamese. Du et al.[12] utilized a Semi-Siamese Training network (SST) for NIR-VIS face recognition and included an additional constraint that face images were synthesized with masks on their faces. They adopted PR-Net to accomplish the mask synthesis approach.

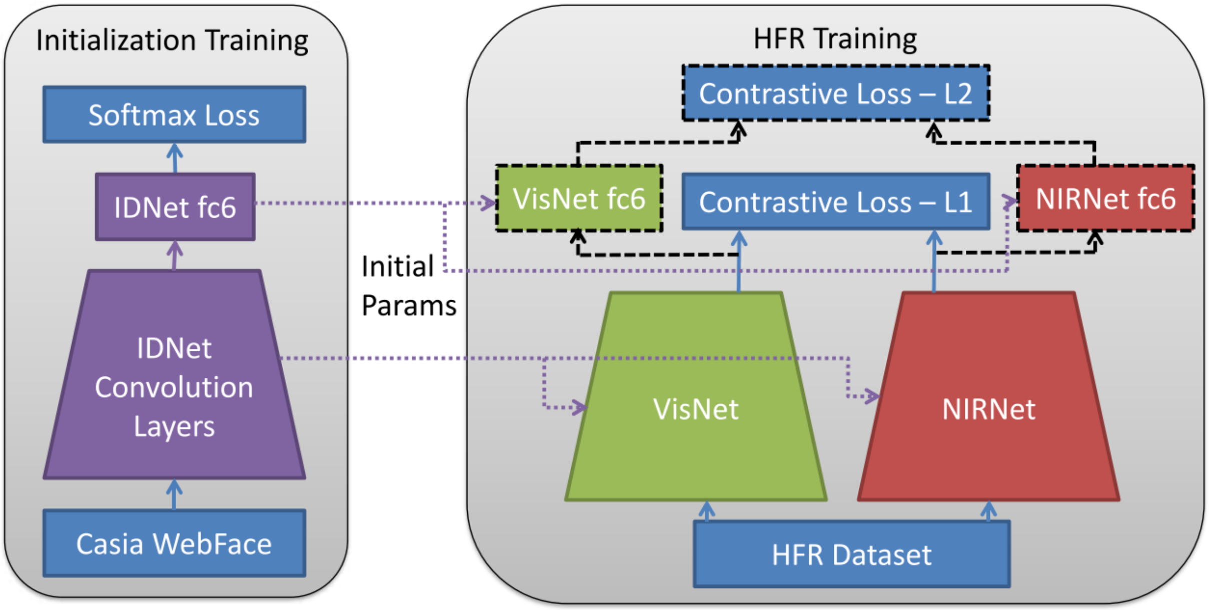


Figure 2.3: Siamese network structure of [61]: VisNet and NIRNet shared parameters which initialized from the pre-trained network.

They first segmented facial masks from mask VIS images and the UV texture map T_M. Second, they combined the mask template, T_M, into the UV texture map, T_I, of the non-masked face image, $\hat{T}_I$, in which corresponding regions of the face masks were removed. Finally, face images were recovered with T_{MI} and UV position map P_I that were extracted from the original non-masked face image. Once the authors obtained masked face images, they trained their semi-siamese network with the input of positive pairs of heterogeneous faces, masked NIR faces and the original VIS faces. This semi-siamese network consisted of two sub-network, a probe network, and a gallery network, which was both pre-trained on VIS dataset. Probe-net ϕ_p embedded features of probe images and Gallery-net ϕ_g updated their proposed prototype queues. There were two prototype queues, including one for NIR faces and VIS faces. NIR prototype queue was used for computing training loss with probe network's output features of NIR images, whereas VIS prototype queue computed training loss of VIS features from a gallery network. The training loss was a reformulated softmax loss,

$$L(I_N, I_V) = -\log \frac{e^{s\phi_p(I_N)\phi_g(I_V)}}{e^{s\phi_p(I_N)\phi_g(I_V)} + \sum_{j=1}^{n} e^{s\phi_p(I_N)f_j^V}}, \tag{2.38}$$

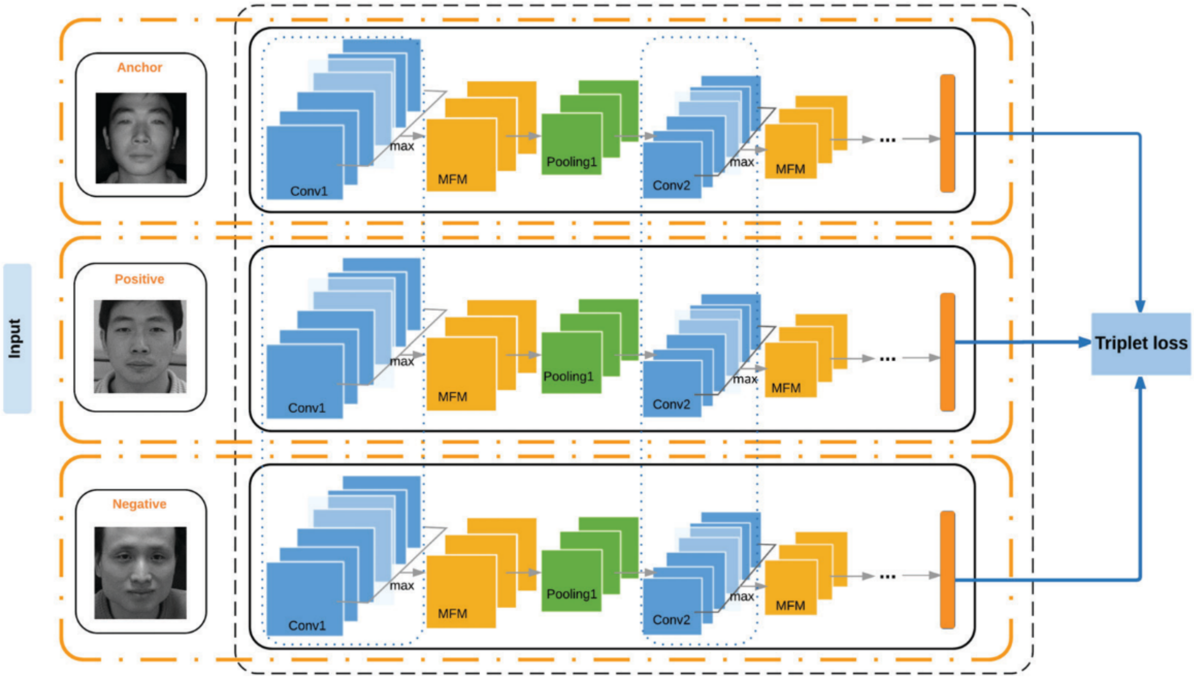


Figure 2.4: Trivet network structure of [47]: three inputs were sent into the same feature space; through learning, positive samples were closer to each other and negative sample got further.

where f_j^V is the jth feature of VIS prototype queue. While minimizing the loss, face representation was able to be spread in feature space where the same identity from both domains could be closer. In [47], the authors not only utilized pre-trained CNNs but also took coupled three networks into a trivet architecture to map three inputs into a single feature space for better learning; this network structure is shown in Fig.2.4. Their idea was based on the triplet loss,

$$L_{trip} = \sum_{i}^{N} [\|f(x_i^a) - f(x_i^p)\|_2^2 - \|f(x_i^a) - f(x_i^n)\|_2^2 + \alpha]_+, \tag{2.39}$$

where x^a is the anchor; x^p is the positive example; and x^n is the negative example. The loss function prompted the network to focus on individual distinction where domain-variant features were eliminated as much as possible. Their triplet loss was along with an example selection strategy called Hard NIR-VIS Triplets Selection. For a NIR image (anchor), hard positive samples were VIS images with the same identity but a lower score; hard negative

Methods	Dataset	Rank-1	0.1%FAR	Characteristic
[74]	-	-	93.1%	LDA reduction; CCA subspace; linear correlation maximization.
[44]	-	-	67.5%	Reduction of illumination variation; MB-LBP subspace.
[37]	-	-	93.45%	PCA subspace.
[75]	CASIA NIR-VIS 2.0	86.18%	81.29%	Subspace of Gabor wavelet
[80]	HFB	90.0%	-	Transductive subspace; MMD penalization.
[79]	HFB	99.28%	98.42%	Transductive subspace; kernel learning.
[41]	HFB	77.5%	-	MSDA subspace.
[4]	CASIA NIR-VIS 2.0	98.5%	99.7%	Reflected difference; Hamming subspace.
[65]	CASIA NIR-VIS	85.9%	-	Metric learning.
[20]	CASIA NIR-VIS 2.0	98.7%	98.4%	Subspace of Wasserstein distance.
	Oulu-CASIA	98.0%	54.6%	
	BUAA-Visnir	97.4%	91.9%	
[62]	CASIA NIR-VIS 2.0	87.1%	74.5%	Siamese structure.
[61]	CASIA NIR-VIS 2.0	92.6%	-	Siamese structure.
[12]	CASIA NIR-VIS 2.0	98.6%	98.58%	Semi-Siamese structure; 3D aware.
	Oulu-CASIA	91.3%	83.0%	
	BUAA-Visnir	98.4%	70.6%	
[47]	CASIA NIR-VIS 2.0	95.74%	91.03%	Triplet structure.

Table 2.3: Performance of Subspace Learning Based Methods. [74], [44], [37] used their dataset which was not public, thus there is no name on it. [80][79], [41] used the HFB dataset which is one of the former versions of the CASIA NIR-VIS 2.0 dataset with much fewer images.

samples were VIS images with different identities but higher scores. Such a strategy was insurance for networks that paid more attention to hard-distinguished features. Further, the authors abandoned the commonly-used sigmoid or ReLU activation function and instead, chose Max-Feature-Map (MFM)[73], an ordinal activation function, which extracted the maximum of candidate nodes in two corresponding feature maps. The advantage of such replacement included: 1) MFM lightened the whole network but also selected compact and remarkable features; 2)MFM could reduce the number of parameters which was useful in small-scaled NIR-VIS training datasets; 3) MFM improved the running speed of the network. CNN-based methods are a combination of subspace learning-based methods and invariant featured-based methods and were demonstrated to provide state-of-the-art results. CNNs can extract semantic information from images, which work as invariant feature-based methods and, when used in conjunction with other modifications like MFM and MMD to further discriminate the feature subspaces, performance can be improved.

Performance Evaluation

Table2.2.2 provides a performance summary of the previously mentioned subspace learning-based methods. CASIA NIR-VIS 2.0 is the current largest dataset. Methods used over previous decades, such as that employed in[75], have recognition rates higher than 80%, which represent the best results in that period. And subspace learning-based methods now can perform extremely impressive results. The best results come from methods that combine invariant feature-based methods and subspace learning-based methods. However, these results cannot reach the same level as state-of-the-art image synthesis-based methods, both in terms of recognition rate and verification rate.

Subspace learning-based methods have provided reasonable performance before the 2010s. Using subspace learning methods, matching and mapping between identities can be performed at a much more reasonable cost relative to other methods. However, the extracted features are generally low-level and contain mostly structural information. In NIR-VIS face recognition, faces in different domains are characterized by exhibiting different contours and textures; structure information is thus not definitive, therefore indicating the key limitation which prevents the more robust performance of these methods. After applying CNN, subspace learning-based methods are combinations of subspace learning and invariant feature extraction, such that the CNNs extract semantic information from images while subspace learning-based methods learn the relationships among these high-level features. Up to now, these new methods yield state-of-the-art performance.

2.2.3 Invariant Feature-Based Methods

Some studies used an assumption that NIR and VIS images should have shared common features which can be regarded as identity information. Based on this assumption, different types of methods were proposed for extracting these shared features, referred to as modality-invariant features. Therefore, invariant feature-based methods are used to alleviate the sensing gap problem in NIR-VIS face recognition by extracting modality-invariant features for direct matching. Users must be cautious in the use of these approaches to mitigate the loss of too much information.

Traditional Method

early application of invariant feature-based methods to NIR-VIS face recognition, invariant feature-based methods could rarely compete with subspace learning-based methods, because of a lack of powerful extractors. Some methods have structures that are quite similar to subspace learning-based methods but without a common subspace. In [10], the authors proposed that HOG was an ideal way to match VIS and NIR face images; HOG measured the edge orientation information of images and, in the context of NIR-VIS face recognition, the edge orientation changed to a very minor extent between NIR and VIS images. In their approach, the authors fixed $m \times m$ key points on an image by specifying step-size to cell-size ratio along both height and width, then extracted the magnitude and gradient from these key points. For every key point, extracted HOG was stacked and the final feature descriptor of the image was generated at a size of $m \times m \times d$. Through this fixed key points-based approach, the facial shape as an invariant feature was extracted and subjected to PCA for dimension reduction, because the dimension $m \times m \times d$ could be extremely large. Finally, matching between probe images and gallery images was done by using cosine distance. Without a common subspace, the extracted features were mostly structural features which still vary considerably in different spectra. Therefore, such methods can yield the best results.

CNN-based Method

These approaches were demonstrated to function as ideal feature extractors and can extract very high-level semantic features, yielding phenomenal results for facial recognition. Modifications to CNNs have delivered an impressive performance on VIS face recognition and enabled fine-tuning in NIR-VIS face recognition. Salim et al.[64] modified ResNet for feature extraction, first by using HOG for preprocessing images. The authors then modified ResNet-34, such that in each convolution layer, the dimensions halved and yielded final average pooling with dimensions of the fully connected layer of 128 rather than 1000. Such modifications were suitable for NIR-VIS dataset while making the whole network less complicated and less prone to over-fitting. Because of the 128-dimensional features, the authors chose to use Support Vector Machine[8] with Radial Basis Function as a kernel. The authors used the revised method to apply NIR-VIS face recognition in practical case studies. Even in NIR images capturing, illumination can affect output; thus, the authors divided images in both domains into three sets based on strong illumination, weak illumination, and dark illumination. NIR images in the testing stage were barely affected by illumination, but the illumination nation of VIS images in the training stage could dramatically influence performance; the lighting condition of VIS images was improved under strong illumination rather than weak or dark. Miyamoto et al.[55] represented their Joint Feature Distribution Alignment Learning (JFDAL), consisting of Cross-domain feature distribution alignment Learning (CFDAL) and Source-domain feature distribution alignment learning (SFDAL). CFDAL was used to reduce the distance between feature distributions in different domains. In CFDAL, the authors utilized LResNet50E[9] as their baseline and fine-tuned on the NIR-VIS dataset. There were two loss functions in CFDAL, including a softmax loss (which was used as a face classification loss function) and their proposed L_{dom} which measured distance between distributions, as below

$$L_{dom} = \frac{1}{M} \sum_{i} \|\mu_N^i - \mu_V^i\|, \tag{2.40}$$

$$\mu_d^i = \frac{1}{|B_d|} \sum_{I_d^i \in B_d} F(I_d^i, \Theta), \tag{2.41}$$

where B_d is the NIR or VIS domain subset. And SFDAL was proposed for keeping VIS distributions from their original points. They used pre-trained LResNet50E with fixed parameters. This pre-trained network worked as guidance of the training model, therefore the distribution of VIS features would not be changed too much during training. To reduce domain variance and retain VIS feature distribution simultaneously, the authors jointly applied CFDAL and SFDAL by keeping the total loss function of their network. In [22], the authors proposed their Orthogonal Modality Disentanglement and Representation Alignment (OMDRA) network. In this work, they used their proposed Modality-Invariant (MI) loss to control both intra-class cross-domain constraint and inter-class cross-domain constraint, which formed as

$$\begin{aligned} l_{MI} =& \kappa_1 \frac{1}{c} \sum_{i=1}^{c} \|m_i^V - m_i^N\|_2^2 + \kappa_2 \frac{1}{c} \sum_{i=1}^{c} \|\zeta_i^V - \zeta_i^N\|_2^2 \\ &+ \kappa_3 \frac{1}{c} \sum_{i=1}^{c} \sum_{j=i+1}^{c} [\alpha - \|m_i^V - m_i^N\|_2^2]_+ + \kappa_4 \frac{1}{c} \sum_{i=1}^{c} \sum_{j=i+1}^{c} [\alpha - \|m_i^V - m_i^V\|_2^2]_+ \\ &+ \kappa_5 \frac{1}{c} \sum_{i=1}^{c} \sum_{j=i+1}^{c} [\alpha - \|m_i^N - m_i^N\|_2^2]_+ \end{aligned} \tag{2.42}$$

where m is the mean vector in each domain; ζ represents variance of each class in each domain; and c is the number of classes. MI loss helped the network learn domain-independent and identity-discriminative representations. The authors presented their Orthogonal modality disentanglement (OMD) to separate modality-invariant features in their network. The high-level hybrid facial feature layer consisted of two parts, the identity-related layer, and the modality-related layer. These two layers decomposed features through two orthogonal matrices and extracted identity features $y^I$and modality features y^M. The overall loss function in OMD could be expressed as:

$$\begin{aligned} l_{OMD}(y_i^M;\Theta,W^M) =& \frac{1}{n}\sum_{p}^{\{N,V\}}\sum_{i=1}^{n^p} I(l_i^M = p)\|y_i^M - m^p\|_2^2 + \frac{1}{2}[\eta - \|m^V - m^N\|_2^2]_+ \\ &+ \sum_{i=1}^{d_1}\sum_{j=i+1}^{d_1}\mu_{ij}^M \|\frac{W_i^{M^T} W_j^M}{\|W_i^M\|\|W_j^M\|}\|_F^2 + \sum_{i=1}^{d_1}\sum_{j=1}^{d_1}\mu_{ij}\|\frac{W_i^{I^T} W_j^M}{\|W_i^I\|\|W_j^M\|}\|_F^2 \end{aligned} \tag{2.43}$$

where μ is Lagrange multiplier; W is the mapping matrix.; and m is the mean vector. They applied their Deep Representation Alignment (DRA) to eliminate residual variation among images, which was high-level representation alignment, reduced within-class variation, and increase between-class variation. These CNN-based means yielded significant improvement to NIR-VIS face recognition, such that different loss functions could be used to improve network performance for NIR-VIS face recognition In addition, pre-training and fine-tuning were demonstrated to mitigate the problem of over-fitting from small NIR-VIS face datasets while learning comprehensive information.

Besides fine-tuning pre-trained networks and the formulation of different loss functions, other post-processing methods can be applied to improve matching accuracy. Peng et al.[59] adopted the Re-ranking methodology[45] in their NIR-VIS face recognition network, which further improved matching accuracy. The feature extraction part of the network comprised two relatively simple patch-level CNNs, including one for NIR images and the other for VIS images. The authors' main contribution derived from their local linear re-ranking algorithm, which consisted of a KNN selection, locally linear Jaccard distance[71], and top neighbors enhancement. The re-ranking was added behind the initial face recognition such that the re-ranking was only performed over a small proportion of fulsome galleries. The authors first selected K Nearest Neighbors based on the assumption that probe images had more neighbors in the ranking list with true targets than false ones. Once nearest neighbors were determined, the authors measured neighborhood similarity through locally linear Jaccard distance:

$$d_j(y_n, x_v) = 1 - \frac{\sum_{m=1}^{M} \min(w_{n,.m}, w_{v,m})}{\max(w_{n,.m}, w_{v,m})}, \tag{2.44}$$

$$w_{n,k} = \frac{\sum_{m=1}^{K} R(k,m)}{\sum_{i=1}^{K}\sum_{j=1}^{K} R(i,j)}, \tag{2.45}$$

where R is the Euclidean distance matrix between y_n and its KNN, and $w_{n,k}$ is the dimensional weight vector. The Jaccard distance was used to identify the top T neighbors, which were subjected to an enhancement strategy by adding weight to yield generalized averages of these new neighbors. The weight would penalize all false positive samples in T new neighbors. In [56], the authors not only applied re-ranking but also utilized dictionary learning in their work, where they extracted invariant features from images through two orthogonal dictionaries and tested them with their proposed re-ranking approach. In their network, they extracted domain invariant features by learning domain-specific orthogonal dictionaries separately. The optimization function of their dictionaries was

$$\begin{gathered}\min_{D_x,\Lambda} \|X - [A_x, D_x]\Lambda\|_2^2 + \alpha\|\Lambda\|_0^2 \\ s.t. D_x^T D_x = I_m, A_x^T D_x = 0.\end{gathered} \tag{2.46}$$

Dictionary $\bar{D_x}$ has two sub-dictionary D_x and A_x, $\bar{D_x} = [A_x, D_x]$, where D_x is the learned atoms from input with size m and A_x controls the number of atoms. And Λ in this function represented the sparse vector. Because of training separately, the authors adopted cluster CCA to learn a common space and Bipartite Graph Matching to learn the correspondence of atoms between dictionaries. The objective function of bipartite graph matching was calculated as

$$H(\phi) = \sum_i C(d_y^i, d_x^{\phi(i)}), \tag{2.47}$$

where ϕ is to permute atoms for one-to-one correspondence.Two permuted dictionaries, D_x^c and D_y^cwere generated by applying permutations to dictionaries; these dictionaries had one-to-one correspondence between columns. To further reduce domain shift, the authors formulated another mapping function on D_x^c

$$\hat{T} = \arg\min_T \|D_x^c T - D_y^c\|_2^2, \tag{2.48}$$

where T is the mapping function. Finally, the optimum value $\hat{T}$ was derived as $\hat{T} = D_x^{cT} D_y^c$, and the dictionary in x-domain was aligned to $D_x^{c,a} = D_x^c D_x^{cT} D_y^c$.The two aligned dictionaries could be used to create rank lists for face recognition. The authors proposed a re-ranking algorithm while dealing with rank lists from their dictionary algorithm and CBFD[50] which was another face recognition network. They first divided the gallery data into two rank lists into three sets – strongly similar, strongly neutral, and strongly dissimilar - based on the appearance of the gallery. If the gallery was in the top k elements of both rank lists, it was denoted as strongly similar, and if it was in the last k of both rank lists, it was denoted as strongly dissimilar. The remaining intersection elements would be divided

Methods	Dataset	Rank-1	0.1%FAR	Characteristic
[10]	CASIA NIR-VIS 2.0	73.3%	-	Dimension reduction.
[64]	Oulu-CASIA	95.56%	-	Dimension reduction of average pooling.
[55]	Oulu-CASIA	-	98.94%	Joint learning.
[22]	CASIA NIR-VIS 2.0	99.4%	97.8%	Orthogonal disentanglement; high-level representation alignment.
	Oulu-CASIA	98.5%	81.7%	
	BUAA-Visnir	99.6%	99.3%	
[59]	CASIA NIR-VIS 2.0	98.7%	96.5%	Re-ranking methodology; top neighbors enhancement.
	Oulu-CASIA	98.9%	61.7%	
[56]	CASIA NIR-VIS 2.0	68.3%	-	Re-ranking methodology; dictionary learning.
[65]	CASIA NIR-VIS	85.9%	-	Metric learning.
[20]	CASIA NIR-VIS 2.0	98.7%	98.4%	Subspace of Wasserstein distance.
	Oulu-CASIA	98.0%	54.6%	
	BUAA-Visnir	97.4%	91.9%	
[62]	CASIA NIR-VIS 2.0	87.1%	74.5%	Siamese structure.
[61]	CASIA NIR-VIS 2.0	92.6%	-	Siamese structure.
[12]	CASIA NIR-VIS 2.0	98.6%	98.58%	Semi-Siamese structure; 3D aware.
	Oulu-CASIA	91.3%	83.0%	
	BUAA-Visnir	98.4%	70.6%	
[47]	CASIA NIR-VIS 2.0	95.74%	91.03%	Triplet structure.

Table 2.4: Performance of Invariant Feature-Based Methods. In this table, [65],[20], [61], [12] and [47] were methods that combine invariant feature-based methods and subspace learning-based methods.

into a strongly neutral set. After the gallery data were classified, the authors refined rank lists from CBFD by backward re-query with galleries in the strongly similar set, whereas galleries in the strongly dissimilar and strongly neutral sets worked as penalization to push these elements away. In the final step, the refined rank list from CBFD was combined with rank lists from the proposed dictionary algorithm to determine the final distance score. These re-ranking-based methods improved the verification rates of their algorithms and showed that NIR-VIS face recognition has great potential when used in conjunction with other state-of-the-art methodologies.

Performance Evaluation

Comparison with state-of-the-art performances in image synthesis-based methods, such as those used in[59], [65],[20],[62], [61] and [12], achieved similar rank-1 recognition rates and even higher verification rates. Further, these methods are also capable of running faster than image synthesis-based methods. In future studies, more state-of-the-art networks and methods can be applied in invariant feature-based methods to build more robust and accurate systems.

The performance of invariant feature-based methods improved greatly when used in conjunction with CNNs, and the resulting networks were capable of extracting more semantic features with key identity information. In comparison with state-of-the-art image synthesis-based methods, these invariant feature-based methods are more efficient. Another benefit of invariant feature-based methods is that they accommodate the integration of other post-processing methods, such as re-ranking, providing a more robust system overall.

2.2.4 Summary

Currently, state-of-the-art methods have already achieved high accuracy, where the highest is 99.40%. Such accuracy reaches the same level as traditional face recognition systems. However, with such small-size datasets, these performances cannot ensure practical results. To be a matter of fact, there are still shortcomings in current methods.

Most image synthesis-based methods paid lots of attention to facial features, like edges[11], identity[13]. They proposed different network architectures and different loss functions, for getting better results. However, rare researchers tried to optimize the colorization part of synthesis images, not even mention optimizing both colorization and facial appearance. One of the most significant differences between NIR and VIS images is their different color spectrum. Generators can learn how to color the NIR images, but it is not sufficient. While generating face images, networks should learn to synthesize face images with true identities and correct color. What is more, in [40], the authors considered the unique imaging NIR system. They blend their synthesized images with original VIS images. Nonetheless, such a blend caused the misalignment in facial appearance, which resulted from the un-paired NIR and VIS images. It is important to take restoring NIR imaging scenes into consideration. In NIR images, LEDs put light on faces, therefore faces in NIR domain are brighter than normal VIS images.

For current subspace learning-based methods and invariant feature-based methods, their solutions were that reduce the domain variance in some domain-invariant subspaces. Subspace learning-based methods map both images into a common subspace which minimizes the domain invariance, while invariance feature-based methods extract domain-invariant features which can be regarded as a domain-invariant subspace. State-of-the-art methods are the combination of both two methods, in which CNNs try to extract domain invariant features and have subspace learning approaches in the domain-invariant subspace[65][20] [61][12][47]. Such methods can work well as long as the extractors extract domain invariant features strictly. However, in images, domain features and domain-invariant features are entangled with each other. For example, in [22], the authors used orthogonal dictionaries to disentangle the coupled features and achieved great results. But domain features are not strictly orthogonal (i.e., some wrinkles are coupled with NIR domain). Hence, it is essential to have a better subspace where features can be disentangled well.

Chapter 3

Methodology and Results

In recent years, various generative models have been proposed, which can now generate realistic images. Based on different practical scenes, generative models developed different categories, i.e., style transferring [78], [28]. Researchers paid much attention to the structure of these generative models to improve their performance. Especially, the Cycle-GAN model and the Style-GAN series models have been designed and optimized in different ways but both work well.

As mentioned, the current NIR-VIS face dataset inevitably has the unpaired property because of different domains and imaging devices. Thanks to the Cycle-GAN, it is easy to solve such unpaired image-to-image translation problems. In the model, I adapt the general structure of the Cycle-GAN. Specifically, there are two GANs in the model, one for generating images in NIR modality (G_N) and another for generating images in VIS modality (G_V). Unfortunately, the original Cycle-GAN model does not have the ideal performance for generating detailed face images. It is because the Cycle-GAN is not specially designed for face images and thus loses some detailed information about faces during the feature learning process. Therefore, I replace the generator network and discriminator network with Style-GAN 3 but keep the general cyclic structure of Cycle-GAN. The Style-GAN series models show state-of-the-art results in image synthesis tasks. As the 3rd generation, Style-GAN 3 has the alias-free-translation property which is also suitable for NIR-VIS face dataset. Additionally, I develop a style loss which consists of two perceptual losses to improve the style transferring part and a net loss (latent loss) to ensure the identity of generated images in both domains. In the following two subsections, I first introduced the proposed network architecture, and the second subsection is about the loss function which includes the new proposed latent loss.

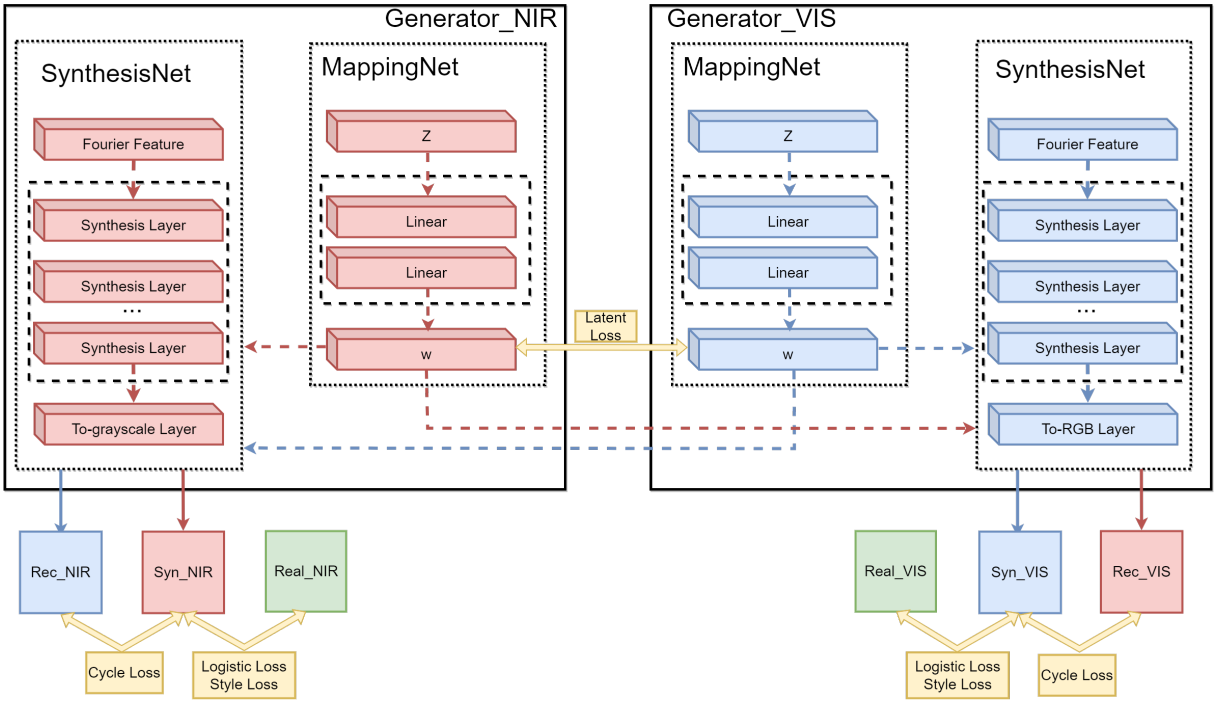


Figure 3.1: Overall network architecture of proposed CS-GAN.

3.1 Network Structure

The overall flowchart of the network is shown in Fig. 3.1. The detailed structure of the generator network is represented in Fig.3.2 and Fig.3.3. In the Cycle-GAN, the generator network consists of an encoder, translate module, and decoder, which is a traditional generator structure. However, in the network, I replaced the Cycle-GAN generator with a mapping network and synthesis network from Style-GAN. In Fig.3.1, the network G_N synthesis face images in NIR field and compares with NIR query images, while networking G_V working in VIS field.

Shown in Fig.3.1, the proposed network consists of two generators (G_N) and (G_V), each of which has a different synthesis task. G_N has learned to synthesizes images in NIR domain and G_N in VIS domain. Both generators consist of two parts, a mapping network, and a synthesis network. Generators first synthesize images in each domain, which results in Syn_NIR and Syn_VIS. Then, to further strengthen the consistency between synthesized images, I design the cyclic structure following Cycle GAN, in which latent code w is sent into the synthesis network of the other generator to synthesize the re-created images(i.e.,

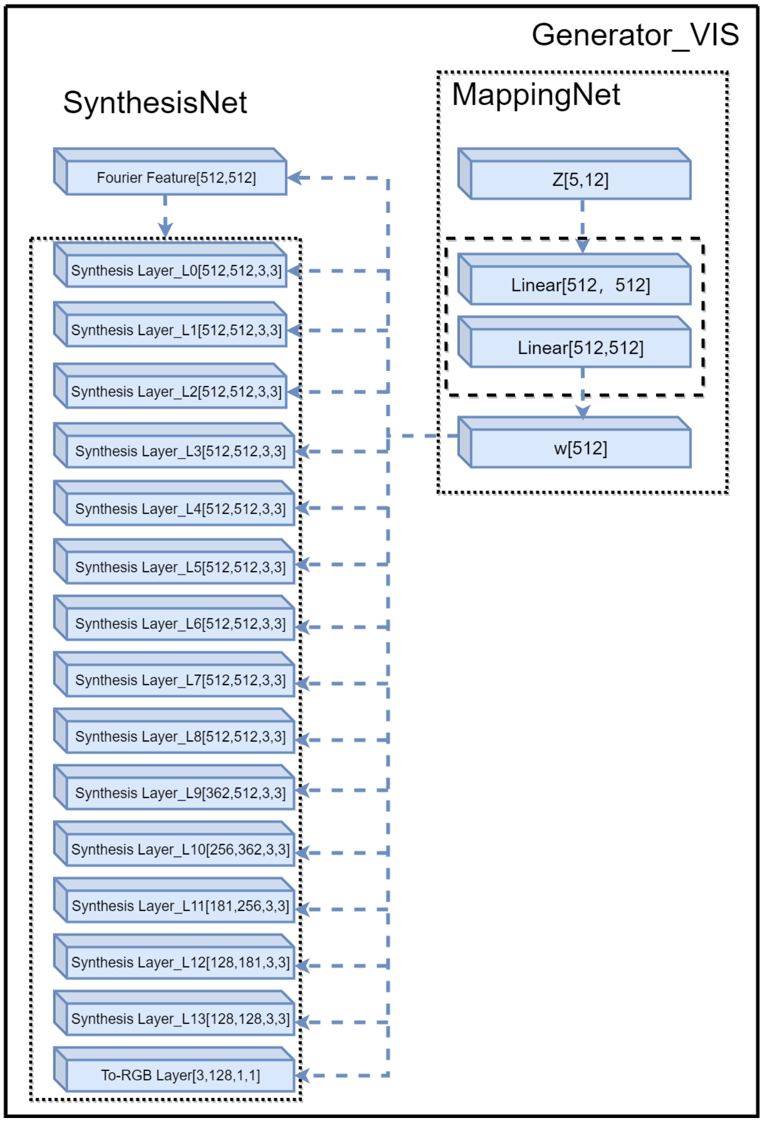

Generator_VIS
SynthesisNet
MappingNet
Fourier Feature[512,512]
Z[5,12]
Synthesis Layer_L0[512,512,3,3]
Linear[512， 512]
Synthesis Layer_L1[512,512,3,3]
Linear[512,512]
Synthesis Layer_L2[512,512,3,3]
Synthesis Layer_L3[512,512,3,3]
w[512]
Synthesis Layer_L4[512,512,3,3]
Synthesis Layer_L5[512,512,3,3]
Synthesis Layer_L6[512,512,3,3]
Synthesis Layer_L7[512,512,3,3]
Synthesis Layer_L8[512,512,3,3]
Synthesis Layer_L9[362,512,3,3]
Synthesis Layer_L10[256,362,3,3]
Synthesis Layer_L11[181,256,3,3]
Synthesis Layer_L12[128,181,3,3]
Synthesis Layer_L13[128,128,3,3]
To-RGB Layer[3,128,1,1]

Figure 3.2: An illustration of $Generator_{VIS}$. It shows the detailed weight size in each layer. The latent code w controls every layer in the synthesis layer.

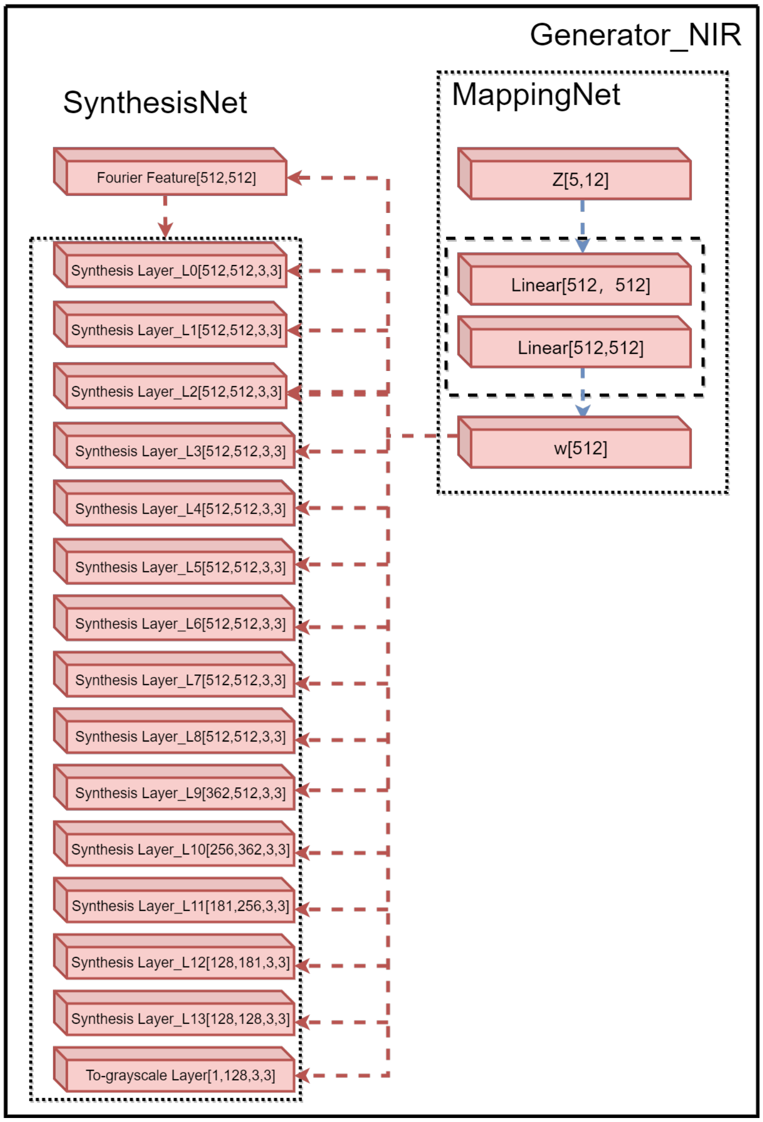

Generator_NIR
SynthesisNet
MappingNet
Fourier Feature[512,512]
Z[5,12]
Linear[512，512]
Linear[512,512]
w[512]
Synthesis Layer_L0[512,512,3,3]
Synthesis Layer_L1[512,512,3,3]
Synthesis Layer_L2[512,512,3,3]
Synthesis Layer_L3[512,512,3,3]
Synthesis Layer_L4[512,512,3,3]
Synthesis Layer_L5[512,512,3,3]
Synthesis Layer_L6[512,512,3,3]
Synthesis Layer_L7[512,512,3,3]
Synthesis Layer_L8[512,512,3,3]
Synthesis Layer_L9[362,512,3,3]
Synthesis Layer_L10[256,362,3,3]
Synthesis Layer_L11[181,256,3,3]
Synthesis Layer_L12[128,181,3,3]
Synthesis Layer_L13[128,128,3,3]
To-grayscale Layer[1,128,3,3]

Figure 3.3: An illustration of $Generator_{NIR}$. It shows the detailed weight size in each layer. The latent code w controls every layer in the synthesis layer

latent code w from G_N will be sent into synthesis network in G_V for Rec_VIS). In this way, I can make sure that correct features are embedded in latent code w.

As illustrated in Fig.3.3 and Fig.3.2, the two generators have mostly the same architectures with a difference in the final layers. The input of generators is different from traditional GANs. In Cycle-GAN, the generator encodes images into feature vectors. Then, through the translate module, extracted features are translated from NIR to VIS domain. The synthesized VIS images are decoded from these translated feature vectors. In the network, I do not simply apply NIR images as input. Instead, the input of the network is random latent vectors z with uniform or Gaussian distribution. It is as same as Style-GAN. Because of the uniform or Gaussian distribution of latent codes z, the coupling between latent codes z is relatively large. For example, while generating face images, hair length, and masculinity, according to the distribution of latent codes z, then there will be a close relationship between these two characteristics. If the hair is short, the masculinity will decrease or increase, but in reality, both short-haired men and long-haired men can have a strong masculinity. Also, latent codes z have a limited capacity to control visual features, because latent codes z must follow the probability density of the training data. Therefore, there is a mapping network from Style-GAN which consists of two fully-connected layers. The mapping network provides a learning channel for feature decoupling of latent code z. Through this mapping network, latent codes z are mapped into a latent subspace W as latent code w. Because of reducing the correlation between features, latent codes w do not have to follow the distribution of the training data and features can be changed separately. In the synthesis network, the latent codes w are firstly affined into Fourier features, which are possible to sample mapping in infinite space. Using Fourier transform in low-dimensional space can make the model better understand the information in high-dimensional space. Thus, such Fourier transform will result in better performance in translation between images. In NIR-VIS face datasets, face images are unaligned at different angles. I can regard such an un-alignment as a translation between images. And Fourier transform can resolve this problem. The Fourier features are then sent into 14 synthesis layers. For each layer, input is convoluted through a convolution layer and blend with its affined latent code w. The affine process of latent code w is called Weight Demodulation [35], in which the low-level style feature and the high-level content feature can be significantly decoupled. The combination of 14 synthesis layers follows the idea of progressive growing training [32]. Progressive growing training means that the network first trains a small-resolution image, and then gradually transitions to higher-resolution images after training step by step.

I applied this Style-GAN 3 as the baseline network. But in the network G_N, I remove the to-RGB layers and replace them with a to-grayscale layer, thus the final output of G_N

remains in grayscale. The Style-GAN 3 network was trained on Unaligned Flickr-Faces-HQ Dataset (FFHQ-U) [34], an unaligned face dataset, in their work. This pre-trained network can precisely synthesize face features in VIS domain. To further improve the performance in NIR domain, I apply the pre-trained Style-GAN 3 network and fine-tuned it on the CASIA NIR-VIS 2.0 dataset.

3.2 Cyclic Subspace Learning

The goal of the proposed method is to synthesize realistic VIS images with the same identity as NIR images. Specifically, it can be regarded as two parts, one for style and one for identity. The style means the overall visual domain of synthesis images, which in this case is the VIS domain. Identity is the most essential part of face images, which contain visual details of faces. Feature reconstruction at high layers (high-dimension) tends to preserve image content and structure, while feature reconstruction at low layers (low dimension) preserves color, texture, detail shape, etc. Thus, in the proposed model, I build up my own cyclic subspace learning method for supervising feature learning in multiple dimensions. As mentioned, the latent subspace W controls the style and content of synthesized images. Therefore, in each cycle, the latent codes w from both mapping networks are sent into the other synthesis network for recreated images, i.e. latent code w from G_N's mapping network will be the input of both generators for synthesis image and recreated image. In each cycle, Through this process, by constraining recreated images and real images, the synthesis image in the other field will have the same identity features. To have better control over style, I adapt the same way as [29]. In their original paper, they demonstrated that the summation of output difference of multiple relu layers could retain some common semantic information representing the whole image, and this commonality happens to be the artistic style of the image. I follow their idea that for the synthesis image and real image, a pre-trained VGG-16 network extracts semantic features from both images and calculate the perceptual loss for style control in each domain. By doing so, the artistic style in each domain can be well-constrained.

3.3 Loss Function

The loss function of the proposed model consists of four different types of loss functions, logistic loss, cycle-consistency loss, perceptual loss, and the proposed latent loss.

3.3.1 Logistic Loss

The logistic losses for the generators (G_N and G_V) and discriminators (D_N and D_V) are formulated as follows:

$$\begin{aligned} L_{logistic} = {} & E_{i_N \sim P(i_N)}(\log(\exp(D_N(G_N)) + 1) \\ & + \log(\exp(-D_N(i_N)) + 1)) \\ & + E_{i_V \sim P(i_V)}(\log(\exp(D_V(G_V)) + 1) \\ & + \log(\exp(-D_N(i_V)) + 1)), \end{aligned} \tag{3.1}$$

in which i_N and i_V are query and gallery images from NIR domain I_N and VIS domain I_N, respectively. In this process, generators (G_N and G_V) try to minimize the objective, and discriminators (D_N and D_V) try to maximize it.

3.3.2 Cycle-Consistency Loss

The cycle-consistency loss is formulated as follows:

$$\begin{aligned} L_{cyc} = {} & E_{i_N \sim P(i_N)}(\|G_N(w_2) - i_N\|) \\ & + E_{i_V \sim P(i_V)}(\|G_V(w_1) - i_V\|) \end{aligned} \tag{3.2}$$

where w_1 and w_2 are latent codes w from generators (G_N and G_V) respectively. The objective is to make sure that generated images and real images are as same as possible.

3.3.3 Style Loss

Following [29], I apply perceptual loss in the model to have further control of generated styles. In the perceptual loss, there is a comparison between the feature obtained by convolution of the real image (pre-trained vgg-16) and the feature obtained by convolution of the synthesized image, making the high-level information close. The style loss is consisted of perceptual loss in each domain and is formulated as follows:

$$\begin{aligned} L_{style} = {} & E_{i_N \sim P(i_N)}(L_{perceptual}(G_N, i_N)) \\ & + E_{i_V \sim P(i_V)}(L_{perceptual}(G_V, i_V)), \end{aligned} \tag{3.3}$$

$$\begin{aligned} L_{perceptual}(I_1, I_2) = {} & MSE_{relu_{1\_1}}(i_1, i_2) + MSE_{relu_{1\_2}}(i_1, i_2) \\ & + MSE_{relu_{3\_2}}(i_1, i_2) + MSE_{relu_{4\_2}}(i_1, i_2), \end{aligned} \tag{3.4}$$

$$MSE(X,Y) = (x-y)^2, \tag{3.5}$$

where MSE means mean squared error; $relu_{1\_1}$, $relu_{1\_2}$, $relu_{3\_2}$ and $relu_{4\_2}$ represent output features of each ReLU layers in the pre-trained VGG-16, respectively.

3.3.4 Latent Loss

As shown in Fig.1, I apply a latent loss in the latent subspace W to have further control of the synthesis process. The latent loss is formulated as follows:

$$L_{latent} = E_w(\| \frac{w_1}{\|w_1\|_2} - \| \frac{w_2}{\|w_2\|_2} \|_1), \tag{3.6}$$

where $\| \cdot \|_2$s the L2 norm. In the latent loss, the L2 normalization of latent codes w ensures that such a learning process in the subspace will not violate the alias-free property of generators.

3.3.5 Total Loss

Therefore, the full objective is formulated as follows:

$$L = \lambda_1 * L_{cyc} + \lambda_2 * L_{latent} + L_{style} + L_{logistic}, \tag{3.7}$$

where the parameter λ_1 and λ_2 controls the relative importance of different term.

3.4 Experiment and Analysis

3.4.1 Dataset

For pre-trained Style-GAN 3, I used FFHQ-U (unaligned FFHQ), a high-quality human face dataset for GAN research. I crop the size of images to 256x256 resolution. There are about 70k PNG images with variations in terms of age, ethnicity, and image background. For NIR-VIS face recognition research, I train the proposed model on CASIA NIR-VIS 2.0 Database, the largest public NIR-VIS face dataset. CASIA NIR-VIS 2.0 Database was collected in four sessions from 2007 to 2010. There are 725 subjects (each of which has a different number of corresponding images) in this dataset with wide variations of lighting, expression, pose, and distance, which contains the most practical scenes. For each subject, there are 1-22 VIS and 5-50 NIR images, a total of 17,580 images. The dataset features two protocols, algorithm development, and performance reporting. In the algorithm development, the dataset is divided into a training set and a testing set; the testing set consists of VIS gallery images and NIR probe images. In algorithm development, parameters can be tuned and fixed via training and testing. In the performance reporting, the dataset is divided into ten sub-experiments, in which TAR@FAR=0.1%, and Rank-1 identification rate can be calculated.

3.4.2 Experiment Settings

I choose the Style-GAN 3 generator as the baseline generator, which is pre-trained on FFHQ-U (unaligned FFHQ). The generator G_V has the same architecture as Style-GAN 3, and the output dimension of the generator is [3, 256, 256]. As for the generator G_N, the final to-RGB layer is changed into a to-grayscale layer, which results in gray-scale synthesized images with output dimension [1, 256, 256]. During training, I use Adam [36] as the optimizer with a learning rate of 0.00025. The discriminators in both GANs have the same architecture, which is an encoder with a residual net structure. To prepare face image samples, I crop the images in both VIS and NIR domain to 256x256 resolution with shape predictor 68 face landmarks[52]. The hyper-parameterλ_1 and λ_2 in Eq. 3.7 are set to 10 and 5 during the training phase. For the face recognition part in VIS domain, I adopt a pre-trained VGG-16 for face recognition.

3.4.3 Results and Analysis

Visualization

Compared with state-of-the-art methods, the proposed model has achieved better results, where the synthesized images are more realistic and detailed in facial expressions. The results and comparison with related methods are shown in Fig.3.4. For other generators in Cycle-GAN architecture, there is morphing in images, especially around edges. In the proposed models, such a problem has been well alleviated, where the face will not easily blend with the background. Additionally, the facial features of the model are very close to reality. In related works, the difference between synthesized images and real images is implicit, in which synthesized images have blurred facial appearances like wrinkles, lighting, etc. The cause of the previous phenomenon lies in the existence of the proposed

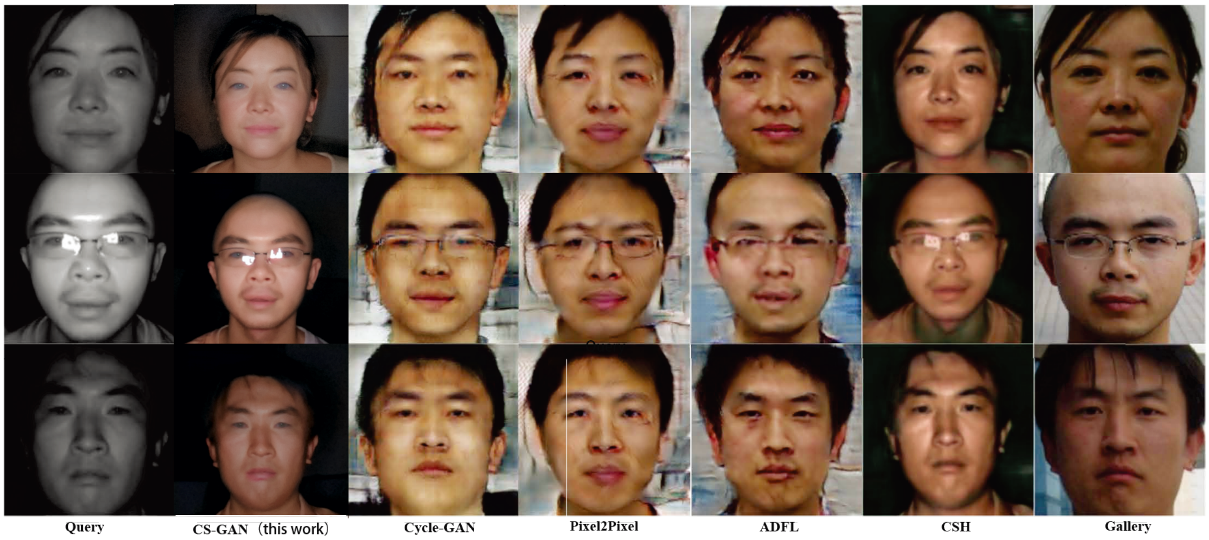


Figure 3.4: Visualization results of different methods. From left to right, the first low is query images in the NIR domain, and the last row is the gallery images in the VIS domain. The second row is the results in which VIS images are synthesized from NIR images. From the third to sixth rows, there are results from state-of-the-art methods. There are results from Cycle-GAN[78], Pixel2Pixel[28], ADFL[66] and CSH[40]

cyclic subspace learning. Through cyclic subspace learning, features in latent space can get their cross-domain connections through learning from both synthesized images in the NIR domain and VIS domain. What is more, the color of synthesized images remains a huge challenge. Color is the best problem in NIR-VIS image translation field. With simply colorizing images, the color will mostly be close to the situation in dusk hours, dark and leaning towards orange. The color of images is gradually improved through the development of research methods, but still not realistic enough. What is more, according to the method of capturing NIR images, there are LEDs in front of faces, which result in brightness on faces. The results show such property exactly, in which facial appearance is a little brighter than a normal case but shows correct skin color and detailed textures. Such close-to-reality color proves the successful control over style transferring.

Accuracy Rate and Verification Rate

Table 3.4.3 lists recognition results on the CASIA NIR-VIS 2.0 Database, which shows the Rank-1 accuracy and verification rate of 0.1% FAR. I compare the NIR-VIS face recognition results with state-of-the-art image synthesis-based methods and other deep learning-based methods which have great performance in traditional VIS face recognition tasks. In the top part, there are 6 image synthesis-based methods. Current image synthesis-based methods have already achieved great results. Among compared methods, CSH method[40] shows the lowest performance in the Rank-1 accuracy and ADFL method with the lowest TAR@FAR = 0.1%, whereas NVVT[3] has the highest accuracy and CFC has the highest verification rate. The proposed method shows the highest performance compared with the state-of-the-art method. Compared with NVVT, the Rank-1 accuracy is 0.20% higher. As for CFC[19], the verification rate is only 0.59% lower. What is more, the comparison between the proposed method and traditional deep learning methods shows a great improvement in both the accuracy rate and verification rate. Such a result indicates that coupling features are well disentangled during the synthesis process.

Ablation Study

To further analyze the effect of my proposed structure, I analyze the influence of each component on the network by superimposing each objective function (i.e., Cyclic structure (L_{cycle}), L_{style}, and L_{latent}) into the model. According to the results, the following conclusions can be drawn.

Firstly, the baseline network (Style-GAN 3) has rather a great accuracy, which demonstrates that the baseline network can generate rather good NIR images. However, the

Methods	Rank-1	0.1%FAR
CSH[40]	96.41%	-
ADFL[66]	98.15%	97.18%
CFC[19]	99.21%	**98.81%**
PACH[13]	98.90%	98.30%
IFA[23]	98.90%	98.70%
NVVT[3]	99.40%	98.74%
VGG-16[46]	55.30%	37.70%
VGG-19[46]	58.50%	41.35%
ResNet-50[18]	64.10%	55.60%
ResNet-101[18]	65.80%	62.10%
CS-GAN(this work)	**99.60%**	98.22%

Table 3.1: The comparison of Rank-1 accuracy (%) and verification rate (%) on the CASIA NIR-VIS 2.0 database.

Method	Rank-1	0.1%FAR
Baseline (Style-GAN 3)	88.33%	78.20%
Baseline+Cyclic structure (L_{cycle})	90.30%	89.44%
Baseline+Cyclic structure (L_{cycle})+L_{style}	97.88%	96.93%
Baseline+Cyclic structure (L_{cycle})+L_{style}+L_{latent}	**99.60%**	**98.22%**

Table 3.2: Ablation study on the CASIA NIR-VIS 2.0 database.

verification rate is not ideal at only 78.20%. Such a shortcoming suggests that the baseline network is still not suitable for NIR-VIS translation task. The situation is well improved through embedding the baseline network into a cyclic structure and fine-tuning it on the CASIA NIR-VIS 2.0 database. The verification rate is significantly improved while slightly improving the accuracy. This phenomenon indicates that the cyclic structure can help the baseline network reduce the sensing gap between NIR and VIS images.

Secondly, the performance is further improved while adding L_{style}. The style loss consists of two perceptual loss, which guides the network to learn abstract style feature separately. The style loss in the proposed network indeed improves the accuracy rate and verification rate to 97.88% and 96.93% which are close to the state-of-the-art methods.

Finally, adding the L_{latent} helps the model surpass all other methods on the accuracy rate, reaching 99.60%. Such improvement indicates that the proposed latent loss has great potential for learning identity features even underlying in different domains. Latent loss can enhance identity-discriminative representations by mining between-class information, between-domain information as well as inter-semantic relationship.

Therefore, each component in the CS-GAN can improve the performance of NIR-VIS face recognition task. And jointly applying them can effectively eliminate the modality-related and identity-related discrepancies.

Chapter 4

Conclusion and Future Work

4.1 Conclusion

With years of development, NIR-VIS face recognition field has already developed methods with great performance, especially image synthesis-based methods. However, these NIR-VIS face recognition methods have not yet been widely applied in practical scenes. Current image synthesis-based images can perform well on the accuracy of current public NIR-VIS face datasets. Regarding the size of datasets, such accuracy cannot guarantee practical performance. Besides, through visualization results, synthesized images still have severe problems which will affect the recognition results, and the problems and reasons can be attributed to the following points: (1) Distortion of synthesized images. NIR images and VIS images are taken from different angles, which makes the images un-paired. Traditional image translation methods cannot deal with such misalignment well, which results in the distortion of synthesized images. Specifically, the edges of synthesized faces are a blur and easily blended with the background environment; (2) Colorization of synthesized images. One of the most conspicuous characteristics of NIR images is that they are all in grayscale. Hence, to synthesize realistic face images, it is important to colorize images. Current methods have poor performance in the colorization part. Additionally, NIR images are collected under illumination from NIR LEDs. Most researchers have not considered such property; (3) Details of Synthesized images. Facial appearance consists of numerous facial details. However, in NIR images, some of these details will lose. Current synthesized images tend to lose these details, which results in a flat facial appearance. To address the above problems, this thesis proposes novel methods to improve the performance of the image synthesis-based method in NIR-VIS face recognition field.

To begin with, I review current state-of-the-art methods in NIR-VIS face recognition field. I firstly make a detailed analysis of current public NIR-VIS face datasets. Then, I evaluate three different types of methods in this field, image synthesis-based methods, subspace learning-based methods, and invariant feature-based methods, with their novelty and performance. Also, I include the perspective of these methods.

In addition, I propose the CS-GAN. I first adapt the general structure from Cycle-GAN which uses a cyclic architecture to maintain the consistency between synthesized images and ground truth images in the same domain. In the general cyclic architecture, there are two different generators, G_N, and G_V, for NIR domain and VIS domain, respectively. Secondly, in consideration of details, the generator from Style-GAN 3 is utilized as the generator. The Style-GAN 3 is one of the best generative networks in style transferring tasks. The generators are pre-trained on the unaligned VIS face datasets. Then, I fine-tune them in different domains. Noticeably, the structure of G_N is modified, replacing the last to-RGB layer with the to-grayscale layer, to fit the characteristic of NIR images. Last but not least, I propose latent subspace learning, in which the style and features of synthesized images are further controlled. The generators used the latent space W to control the details of synthesized images. Therefore, I apply further control over the latent space. For the same identity, they shall have a similar latent code w. Besides, perceptual loss is adopted for style consistency between synthesized images and real images.

Finally, there are the experiment settings and results. The model has been trained on CASIA NIR-VIS 2.0 dataset. I also compare the results with state-of-the-art image synthesis-based methods and methods in related tasks. In the visualization part, I compare the VIS synthesized images with other image synthesis-based methods and image-to-image translation tasks, in which the model has great improvement over the above-mentioned aspects. For accuracy, I list several other image synthesis-based methods and deep learning methods. The Rank-1 accuracy has achieved 99.60% which is the highest among all NIR-VIS face recognition methods. The verification rate is also impressive with 98.22% TAR@0.1%FAR.

4.2 Future Work

While the proposed method proves to solve some existing problems of the NIR-VIS face recognition field and improves the accuracy and visualization results, it still has some limitations which can be improved. The following points illustrate my future research:

- Construct a sufficient public dataset. The NIR-VIS face recognition task is datadriven, and existing methods have already achieved excellent results. But public NIR-VIS face datasets all have the problem that the size is too small, especially compared to VIS face datasets. Current accuracy cannot show the value of the practical application. Additionally, current datasets lack diversity. The appearance of people is monotonous, in which scenes are not enough and the color of skins is mostly yellow. Therefore, if I could construct large and more diverse datasets, the model can be more suitable for practical applications.

- Pre-train the baseline Style-GAN 3 network on a larger VIS dataset. The visualization of the model shows great improvement, though there are still flaws in synthesized images. The iris areas are not close to reality. As shown in the visualization, some of the iris areas are blue, which does not match with ground truth VIS images. Besides, the background information is a lack in the visualization. Training the baseline network on a larger VIS dataset can help the model to have a better understanding of the iris area, and more importantly, can further depict facial details in synthesized images.

- Build up an end-to-end face recognition system. In the methods, the image synthesis part and recognition part is separated, which will low down the speed and cost of extra computing resource. In deep learning, when using multi-steps and multi-models to solve a complex task, an obvious disadvantage is that the training objectives of each module are inconsistent. The objective function of a certain module may deviate from the macro-objective of the system. In this way, it is difficult for the trained system to reach the optimal level. performance; another problem is the accumulation of errors, the deviation produced by the previous module may affect the latter module. Thus, it is important to build up an end-to-end NIR-VIS face recognition system.

References

[1] M. Aharon, M. Elad, and A. Bruckstein. K-svd: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Transactions on Signal Processing*, pages 4311–4322, 2006.

[2] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, pages 214–223, 2017.

[3] Han Byeol Bae, Taejae Jeon, Yongju Lee, and etc. Non-visual to visual translation for cross-domain face recognition. *IEEE Access*, pages 50452–50464, 2020.

[4] Shubhobrata Bhattacharya. Linear cross-modal hash encoding (lcmhe) for visual and near-infrared face recognition. *IEEE Sensors Letters*, pages 1–4, 2021.

[5] James Booth and Stefanos Zafeiriou. Optimal uv spaces for facial morphable model construction. In *2014 IEEE International Conference on Image Processing (ICIP)*, pages 4672–4676, 2014.

[6] Karsten M. Borgwardt, Arthur Gretton, Malte J. Rasch, and etc. Integrating structured biological data by Kernel Maximum Mean Discrepancy. *Bioinformatics*, pages e49–e57, 2006.

[7] Jie Chen, Dong Yi, Jimei Yang, and etc. Learning mappings for face synthesis from near infrared to visual light images. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 156–163, 2009.

[8] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, pages 273–297, 1995.

[9] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[10] Tejas Indulal Dhamecha, Praneet Sharma, Richa Singh, and Mayank Vatsa. On effectiveness of histogram of oriented gradient features for visible to near infrared face matching. In *2014 22nd International Conference on Pattern Recognition*, pages 1788–1793, 2014.

[11] Hao Dou, Chen Chen, Xiyuan Hu, and Silong Peng. Asymmetric cyclegan for unpaired nir-to-rgb face image translation. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1757–1761, 2019.

[12] Hang Du, Hailin Shi, Yinglu Liu, Dan Zeng, and Tao Mei. Towards nir-vis masked face recognition. *IEEE Signal Processing Letters*, pages 768–772, 2021.

[13] Boyan Duan, Chaoyou Fu, Yi Li, Xingguang Song, and Ran He. Cross-spectral face hallucination via disentangling independent factors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[14] Yuchun Fang, Jie Luo, and Chengsheng Lou. Fusion of multi-directional rotation invariant uniform lbp features for face recognition. In *2009 Third International Symposium on Intelligent Information Technology Application*, pages 332–335, 2009.

[15] Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, pages 119–139, 1997.

[16] Keinosuke Fukunaga. *Introduction to Statistical Pattern Recognition (2nd Ed.)*. Academic Press Professional, Inc., 1990.

[17] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, and etc. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

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

[19] Ran He, Jie Cao, Lingxiao Song, and etc. Adversarial cross-spectral face completion for nir-vis face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1025–1037, 2020.

[20] Ran He, Xiang Wu, Zhenan Sun, and Tieniu Tan. Wasserstein cnn: Learning invariant features for nir-vis face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1761–1773, 2019.

[21] H. Hotelling. *Analysis of a Complex of Statistical Variables Into Principal Components*. Warwick & York, 1933.

[22] Weipeng Hu and Haifeng Hu. Orthogonal modality disentanglement and representation alignment network for nir-vis face recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, pages 3630–3643, 2022.

[23] Weipeng Hu, Wenjun Yan, and Haifeng Hu. Dual face alignment learning network for nir-vis face recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, pages 2411–2424, 2022.

[24] Yuxiao Hu, Zhihong Zeng, Lijun Yin, and etc. Multi-view facial expression recognition. In *2008 8th IEEE International Conference on Automatic Face Gesture Recognition*, pages 1–6, 2008.

[25] D. Huang, J. Sun, and Y. Wang. The buaa-visnir face database instructions. 2012.

[26] Gary B. Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In *Workshop on Faces in 'Real-Life' Images: Detection, Alignment, and Recognition*. Erik Learned-Miller and Andras Ferencz and Frédéric Jurie, 2008.

[27] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[28] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[29] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision – ECCV 2016*, pages 694–711, 2016.

[30] Jürgen Jost. *Riemannian Geometry and Geometric Analysis*. Universitext. 2008.

[31] Felix Juefei-Xu, Dipan K. Pal, and Marios Savvides. Nir-vis heterogeneous face recognition via cross-spectral joint dictionary learning and reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2015.

[32] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation, 2017.

[33] Tero Karras, Miika Aittala, Samuli Laine, and etc. Alias-free generative adversarial networks. In *Advances in Neural Information Processing Systems*, pages 852–863. Curran Associates, Inc., 2021.

[34] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks, 2018.

[35] Tero Karras, Samuli Laine, Miika Aittala, and etc. Analyzing and Improving the Image Quality of StyleGAN. *arXiv e-prints*, 2019.

[36] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2014.

[37] Brendan Klare and Anil K. Jain. Heterogeneous face recognition: Matching nir to visible light images. In *2010 20th International Conference on Pattern Recognition*, pages 1513–1516, 2010.

[38] Thomas R. Knapp. Canonical correlation analysis: A general parametric significance-testing system. *Psychological Bulletin*, pages 410–416, 1978.

[39] A. Krizhevsky, I. Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. *Commun. ACM*, page 84–90, 2017.

[40] Jose Lezama, Qiang Qiu, and Guillermo Sapiro. Not afraid of the dark: Nir-vis face recognition via cross-spectral hallucination and low-rank embedding, 2016.

[41] Jie Li, Yi Jin, and Qiuqi Ruan. Matching nir face to vis face using multi-feature based msda. In *2014 12th International Conference on Signal Processing (ICSP)*, pages 1443–1447, 2014.

[42] Stan Z. Li, Dong Yi, Zhen Lei, and Shengcai Liao. The casia nir-vis 2.0 face database. In *2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 348–353, 2013.

[43] Xuelei Li, Liangkui Ding, Li Wang, and Fang Cao. Fpga accelerates deep residual learning for image recognition. In *2017 IEEE 2nd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)*, pages 837–840, 2017.

[44] Shengcai Liao, Dong Yi, Zhen Lei, and etc. Heterogeneous face recognition from local structures of normalized appearance. In Massimo Tistarelli and Mark S. Nixon, editors, *Advances in Biometrics*, pages 209–218. Springer Berlin Heidelberg, 2009.

[45] W.-H. Lin, R. Jin, and A. Hauptmann. Web image retrieval re-ranking with relevance model. In *Proceedings IEEE/WIC International Conference on Web Intelligence (WI 2003)*, pages 242–248, 2003.

[46] Shuying Liu and Weihong Deng. Very deep convolutional neural network based image classification using small training sample size. In *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*, pages 730–734, 2015.

[47] Xiaoxiang Liu, Lingxiao Song, Xiang Wu, and Tieniu Tan. Transferring deep representation for nir-vis heterogeneous face recognition. In *2016 International Conference on Biometrics (ICB)*, pages 1–8, 2016.

[48] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2015.

[49] David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, pages 91–110, 2004.

[50] Jiwen Lu, Venice Erin Liong, Xiuzhuang Zhou, and Jie Zhou. Learning compact binary face descriptor for face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 2041–2056, 2015.

[51] MA Lynch. Long-term potentiation and memory. *Physiological reviews*, page 87—136, 2004.

[52] Palaniappan M, Sowmia K R, and Aravindkumar S. Real time fatigue detection using shape predictor 68 face landmarks algorithm. In *2022 International Conference on Innovative Trends in Information Technology (ICITIIT)*, pages 1–5, 2022.

[53] S.G. Mallat. A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 674–693, 1989.

[54] J.L. Mcclelland, D.E. Rumelhart, and P.D.P.R. Group. *Parallel Distributed Processing, Volume 2: Explorations in the Microstructure of Cognition: Psychological and Biological Models*. MIT Press, 1987.

[55] Takaya Miyamoto, Hiroshi Hashimoto, Akihiro Hayasaka, Akinori F. Ebihara, and Hitoshi Imaoka. Joint feature distribution alignment learning for nir-vis and vis-vis face recognition. In *2021 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–8, 2021.

[56] Sivaram Prasad Mudunuri, Shashanka Venkataramanan, and Soma Biswas. Dictionary alignment with re-ranking for low-resolution nir-vis face recognition. *IEEE Transactions on Information Forensics and Security*, pages 886–896, 2019.

[57] T. Ojala, M. Pietikainen, and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 971–987, 2002.

[58] Pascal Paysan, Reinhard Knothe, Brian Amberg, and etc. A 3d face model for pose and illumination invariant face recognition. In *2009 Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance*, pages 296–301, 2009.

[59] Chunlei Peng, Nannan Wang, Jie Li, and Xinbo Gao. Re-ranking high-dimensional deep local representation for nir-vis face recognition. *IEEE Transactions on Image Processing*, pages 4553–4565, 2019.

[60] S.J.D. Prince. *Computer Vision: Models, Learning, and Inference*. Cambridge University Press, 2012.

[61] Christopher Reale, Hyungtae Lee, and Heesung Kwon. Deep heterogeneous face recognition networks based on cross-modal distillation and an equitable distance metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, July 2017.

[62] Christopher Reale, Nasser M. Nasrabadi, Heesung Kwon, and Rama Chellappa. Seeing the forest from the trees: A holistic approach to near-infrared heterogeneous face recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 320–328, 2016.

[63] Sam T. Roweis and Lawrence K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, pages 2323–2326, 2000.

[64] Nilu R. Salim, Umarani Jayaraman, and V Srinath. Face recognition in the dark: A unified approach for nir- vis and vis- nir face matching. In *2020 IEEE 4th Conference on Information Communication Technology (CICT)*, pages 1–12, 2020.

[65] Shreyas Saxena and Jakob Verbeek. Heterogeneous face recognition with cnns. In Gang Hua and Hervé Jégou, editors, *Computer Vision – ECCV 2016 Workshops*, pages 483–491. Springer International Publishing, 2016.

[66] Lingxiao Song, Man Zhang, Xiang Wu, and Ran He. Adversarial discriminative heterogeneous face recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018.

[67] Christian Szegedy, Wei Liu, Yangqing Jia, and etc. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[68] Matti Taini, Guoying Zhao, Stan Z. Li, and Matti Pietikainen. Facial expression recognition from near-infrared video sequences. In *2008 19th International Conference on Pattern Recognition*, pages 1–4, 2008.

[69] Xiaoyang Tan and Bill Triggs. Enhanced local texture feature sets for face recognition under difficult lighting conditions. *IEEE Transactions on Image Processing*, pages 1635–1650, 2010.

[70] Huijiao Wang, Haijian Zhang, Lei Yu, Li Wang, and Xulei Yang. Facial feature embedded cyclegan for vis-nir translation. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1903–1907, 2020.

[71] Kathleen B. Watson. *Categorical Data Analysis*, pages 601–604. Springer Netherlands, 2014.

[72] Fangyu Wu, Weihang You, Jeremy S. Smith, and etc. Image-image translation to enhance near infrared face recognition. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 3442–3446, 2019.

[73] Xiang Wu, Ran He, Zhenan Sun, and Tieniu Tan. A light cnn for deep face representation with noisy labels. *IEEE Transactions on Information Forensics and Security*, pages 2884–2896, 2018.

[74] D. Y, R Liu, R. Chu, and etc. Face matching between near infrared and visible light images. pages 523–530, Berlin, Heidelberg, 2007. Springer Berlin Heidelberg.

[75] Dong Yi, Zhen Lei, and Stan Z. Li. Shared representation learning for heterogenous face recognition. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–7, 2015.

[76] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z. Li. Learning Face Representation from Scratch. *arXiv e-prints*, 2014.

[77] Lun Zhang, Rufeng Chu, Shiming Xiang, and etc. Face detection based on multi-block lbp representation. In Seong-Whan Lee and Stan Z. Li, editors, *Advances in Biometrics*, pages 11–18. Springer Berlin Heidelberg, 2007.

[78] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2017.

[79] Jun-Yong Zhu, Wei-Shi Zheng, Jian-Huang Lai, and Stan Z. Li. Matching nir face to vis face using transduction. *IEEE Transactions on Information Forensics and Security*, pages 501–514, 2014.

[80] Jun-Yong Zhu, Wei-Shi Zheng, and Jianhuang Lai. Transductive vis-nir face matching. In *2012 19th IEEE International Conference on Image Processing*, pages 1437–1440, 2012.