On Heterogeneous Face Recognition

by

Soumyadeep Ghosh

Roll no: PhD 14006

Advisors
Prof. Mayank Vatsa
Prof. Richa Singh

A thesis submitted in partial fulfillment for the
degree of Doctor of Philosophy

to the

Department of Computer Science and Engineering
IIIT Delhi

INDRAPRASTHA INSTITUTE *of*
INFORMATION TECHNOLOGY DELHI

March 2023

Certificate

This is to certify that the thesis titled "**On Heterogeneous Face Recognition**" being submitted by **Soumyadeep Ghosh** to the Indraprastha Institute of Information TechnologyDelhi, for the award of the degree of Doctor of Philosophy, is an original research work carriedout by him under my supervision. In my opinion, the thesis has reached the standards fulfilling therequirements of the regulations relating to the degree.

The results contained in this thesis have not been submitted in part or full to any other universityor institute for the award of any degree/diploma.

March 2023
Dr. Richa Singh
Professor

RSingh

March 2023
Dr. Mayank Vatsa
Professor

Indraprastha Institute of Information Technology Delhi
New Delhi

Acknowledgements

It gives me immense pleasure to be at this juncture where I can make an attempt to offer my sincere thanks and acknowledge all those wonderful people who played a role in the culmination of this dissertation. During the course of this work, several people played a key role in supporting me, not only for research but also stretched out a helping hand when I was in need.

I would like to begin by acknowledging my advisors **Prof. Mayank Vatsa** and **Prof. Richa Singh**. From the day I joined their research group, they were a constant source of motivation and guidance for my research and growth. One of the most striking qualities that they posses is leading by example and it was very difficult to keep up with their level of hard work, and this always encouraged me to push my boundaries. Sometimes their way of thinking out of the box led to the development of several research directions, later on which paved the way for some of my research achievements. I would also like to express my sincere gratitude to **Prof. Afzel Noore** who supervised me on some of my research projects, when I visited his lab at West Virginia University for about a year. I would also like to thank **Prof. Nalini Ratha** for supporting me in several of my research problems. I would also like to thank **Tata Consultancy Services** and **Ministry of Electronics and IT, Government of India** for funding my research.

The next set of individuals that I would like to thank are my colleagues at the IAB lab in IIIT Delhi. **Rohit Keshari** my batchmate at our lab was always available for a discussion, and some of my research outcomes have not been possible without him. My other labmates **Akshay, Saheb, Puspita, Maneet, Shruti** have supported me through thick and thin. This section would be heavily incomplete if I do not express my sincere thanks for my labmate **Aakarsh Malhotra** who never stopped supporting me during the entire duration of my research. I am not sure if there can be better friend and companion than him, and he supported me in ways which words cannot express. I would like to thank **Dr. Tejas Dhamecha**, who mentored me for my first research endeavour during this journey. His passion and dedication for research, and eye for minute details have played a key role in the way I approach a research problem. I would also like to thank my other lab senior **Dr. Anush Sankaran**, who happily accommodated me in the lab and made sure that I was comfortable, patiently answered all my questions whenever I used to bug him. I would never forget the day I came in the lab, the warmth and kindness he exhibited to make me feel at home.

Further, I would also like to express my gratitude to **IIIT Delhi**, the institute which changed my life. The professionalism and positive attitude of the people of this institute have been a pillar of support throughout my PhD journey. I would like to specially thank **Dr. Amita Puri**, without her kindness and emotional support, this dissertation would not have seen the light of the day.

Finally, I would like to express my regards and thanks to the most important people in my life, my **parents**, who never stopped believing in me. I would not have this far if they would not have supported me for higher studies. Coming from a humble economic background, having a career in research would never have been possible without their support and motivation. I have not been able to spend a lot of time with them during my research tenure, something for which they never complained and made sure that I was motivated for carrying out this work.

Soumyadeep Ghosh
New Delhi
March 2023

Abstract

Face recognition under controlled and constrained scenarios have reached a significant level of maturity with respect to performance and reliability. However, under unconstrained and uncontrolled settings, current state-of-the-art face recognition systems fail to yield a consistent level of performance. In recent years, several countries have experienced a high number of terrorist attacks, events of public unrest and cross border intrusions. As a preventive and investigative measure, governments around the world have installed surveillance cameras in public places such as railway and bus stations, airports, shopping malls, and so on. Images acquired from these cameras (probes) are captured in an unconstrained and non-cooperative environment, hence their quality in terms of resolution, illumination, pose, spectrum and so on may vary heavily. Images captured by these cameras are matched with a background database which contain images collected from government records such as passport, driving licenses and so on. Such images (gallery) have much better and consistent quality. The matching of poor quality probes with good quality gallery images is a challenging problem, which involves utilizing auxiliary information (such as depth maps), improving the quality of the captured images, learning of heterogeneity aware models and matching to optimize the top-k identification accuracy. This dissertation attempts to develop effective algorithms for face recognition in unconstrained and non-cooperative scenarios where images captured are either in low resolution and/or in NIR (Near-Infrared) with low quality and inherent noise due to the in-the-wild image capture setup commonly encountered in surveillance settings.

The first contribution is primarily aimed at utilizing auxiliary sources of information for training a shared representation for face recognition in unconstrained environments. Low cost depth sensors have opened new avenues for their usage in video surveillance scenarios. The depth information has been utilized in most RGB-D face recognition methods by fusing it with RGB information which results in enhanced recognition performance. However, in real world surveillance scenarios, cameras are placed at a distance too large for low cost depth sensors to capture good quality depth information. Such poor quality depth information may not contribute significantly to face recognition. The first contribution is on learning a shared representation of RGB and depth information using a reconstruction based deep neural network. The proposed network, once trained in offline mode, can generate the shared representation of RGB and depth using only the RGB image. This feature rich representation is then utilized for face identification. This allows the framework to be used in scenarios where low quality or no depth image is

captured. Experiments on two real-world RGB-D datasets, namely Kasparov and IIITD RGB-D, show the efficacy of the proposed method.

The second contribution proposes a Generative Adversarial Network (GAN) based approach to learn an image to image transformation model for enhancing the resolution of a face image. Unsupervised GAN based transformation methods in their native formulation might alter useful discriminative information in the transformed face images. This affects the performance of face recognition algorithms when applied on the transformed images. We propose a Supervised Resolution Enhancement and Recognition Network (SUPREAR-NET), which does not corrupt the useful class-specific information of the face image and transforms a low resolution probe image into a high resolution one, followed by effective matching with the gallery using a trained discriminative model. We show the results for cross-resolution face recognition on three datasets including the FaceSurv face dataset, containing poor quality low resolution videos captured at a standoff distance up to 10 meters from the camera.

The next three contributions propose novel deep metric learning algorithms, that have been utilized to learn discriminative and generalizable models which are effective for classifying unseen classes. The third contribution addresses one of the most challenging scenarios of face recognition that is matching images in presence of multiple covariates such as cross-spectral and cross-resolution. Law enforcement agencies across the world face this arduous task for which the existing face recognition algorithms do not yield the desired level of performance. In this work, we propose a Subclass Heterogeneity Aware Loss (SHEAL) to train a deep convolutional neural network model such that it produces embeddings suitable for heterogeneous face recognition. The performance of the proposed SHEAL function is evaluated on databases in terms of the recognition performance as well as convergence in time and epochs.

In the fourth contribution, a novel noise tolerant deep metric learning algorithm is proposed. The proposed method, termed as Density Aware Metric Learning, enforces the model to learn embeddings that are pulled towards the most dense region of the clusters for each class. It is achieved by iteratively shifting the estimate of the center towards the dense region of the cluster thereby leading to faster convergence and higher generalizability. In addition to this, the approach is robust to noisy samples in the training data, often present as outliers. The fifth contribution presents an elegant solution for enhancing the top-k recognition performance under the purview of deep metric learning. The algorithm first uses a clustering algorithm to identify *superclusters*, which are made of classes that are similar and are mapped close to each other in the embedding space. The compactness of these *superclusters* are then enhanced while protecting the discriminative properties of the individual classes, which results in improved top-k matching performance during testing. Results on five datasets, including the challenging SCface database, show that these solutions outperform other traditional and recent deep metric learning algorithms.

Contents

List of Tables

Abbreviations

CMC	Cumulative Match Curve
CCTV	Close Circuit Television
CRFR	Cross Resolution Face Recognition
CSFR	Cross Spectral Face Recognition
CSCR	Cross Spectral Cross Resolution
CSCR-FR	Cross Spectral Cross Resolution Face Recognition
EER	Equal Error Rate
FR	Face Recognition
FAR	False Accept Rate
GAN	Generative Adversarial Network
GAR	Genuine Accept Rate
HR	High Resolution
LR	Low Resolution
NIR	Near Infrared (Spectrum)
ROC	Receiver Operator Characteristic
VIS	Visible (Spectrum)

Dedicated to my grandfather, who could not be alive to see this day...

Chapter 1

Introduction

Face recognition is the science of matching and correctly identifying a person. Face has been used as one of the foremost identifying traits of an individual from time immemorial. Humans posseses extraordinary capability to remember and identify each other by faces. Although humans do it with an ease, computer scientists, since the last five decades, have been trying to imbibe this capability in automated systems. In recent years there has been a massive increase in interest in the field of face recognition, which led to a large number of face recognition methods appearing in the literature. There has been sufficient progress in face recognition [7] in constrained environments. However, most of the real world images are not taken in constrained environments, thus the focus of the research community in this area moved on to semi-constrained and unconstrained scenarios (Figure 1.1). In semi constrained scenarios [8] (Figure 1.1, 2^{nd} row), the pose, illumination, expression of the face may vary to some extent. Such images can be found in several applications on social media, human-computer interaction and so on. On the other hand, recent applications require to perform under highly unconstrained and uncooperative situations (Figure 1.1, 3^{rd} row). In these scenarios, illumination, pose, expression, distance of the subject from the camera, disguise and other environmental factors will not be in our control. Thus, modern face recognition methods are mostly focused on constrained and semi-constrained environments which has yielded several methods which are pose [8], [9], rotation [10], illumination [11], resolution, distance [12], spectrum, occlusion [13], age, disguise [14], and expression invariant.

One of the most challenging scenarios for unconstrained face recognition is faced by security agencies worldwide as surveillance cameras are widely used for detecting suspects and preventing intrusion. A study [16] conducted by the British Security Industry Authority (BSIA) estimated that UK has has 5.9 million CCTV cameras, thus there is one camera for every 11 people. Due to the distance of the surveillance camera from the subjects, the effective resolution of the

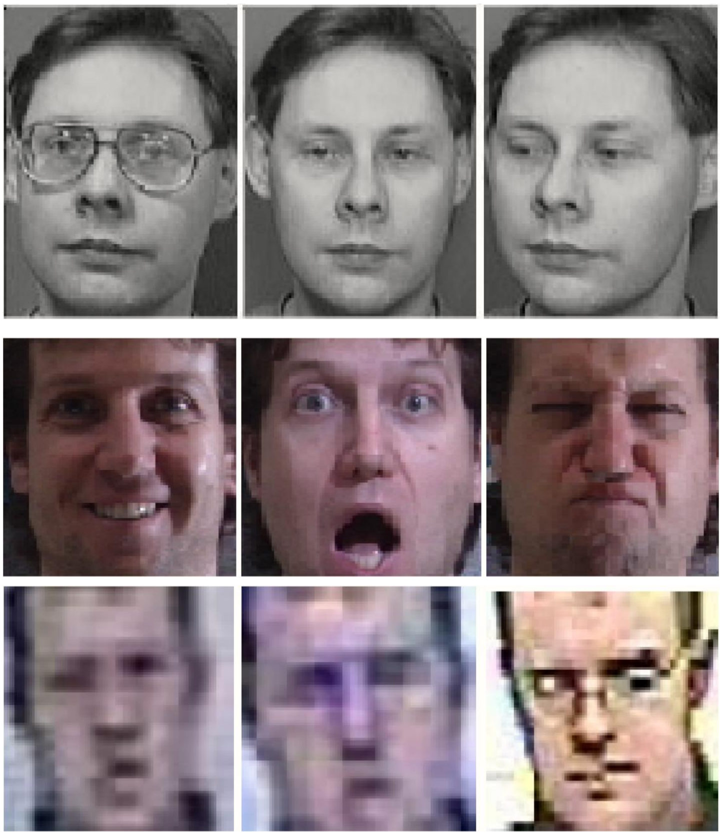
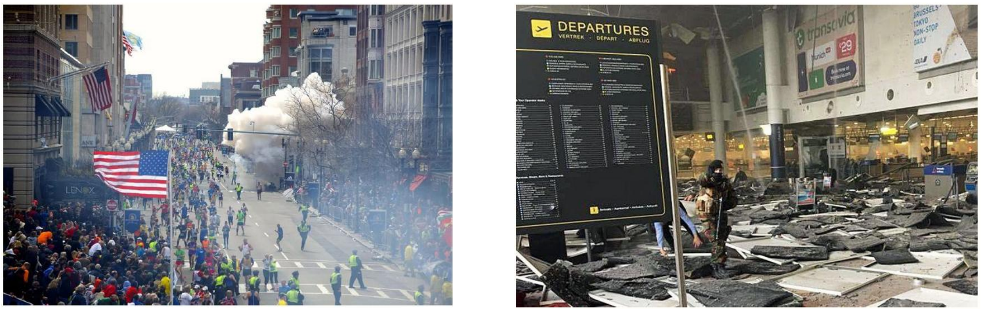


FIGURE 1.1: Representative images for constrained (ORL database [15]) (1st row), semi-constrained (CMU-MultiPIE [8]) (2nd row) and unconstrained scenarios (SCface database [3]) (3rd row) for face recognition.

face is not large enough for modern state of the art systems to accurately identify the individual. The FBI introduced a next generation [17] identification system where face would be used as the foremost biometric trait for identification. As of June 2016, the FBI has a background database of 411 million face images [18] from driver license, visa, passport databases and so on. Such a system could also be used by private enterprises for background searches other than law enforcement. When we consider 24-h surveillance, ensuring equivalent illumination conditions throughout is not possible. Since face recognition is heavily affected by illumination, such cameras switch to infrared (IR) mode at night. A major challenge in surveillance applications is identification from IR images captured in very low visible illumination. Very often security agencies might have a watch-list photograph of the suspect in the form of a high resolution visible spectrum image. This has to be matched with a surveillance video in real time as shown in Fig.2. Thus, the problem becomes all the more challenging due to fact that a cross-spectral face matching adds to the already existing problem of matching surveillance quality low resolution images with high quality enrolled gallery images.

Although a few methods [19–21] exist for such applications, a truly unconstrained environment



Boston Bombings (2013)

Brussels Airport Attack (2016)

FIGURE 1.2: Images showing aftermath of bombings that took place in Boston and Brussels in 2013 and 2016 respectively. Recently, due to rising instances of terrorist activities and miscreants, CCTV cameras have been installed at many public places to equip the law enforcement agencies against such activities.

will challenge any existing surveillance system. State of the art systems for multiple face tracking and recognition using surveillance cameras fail to produce good detection and recognition performance in real world surveillance situations. VeriLook Surveillance SDK, a commercial video surveillance system, can detect, track and recognize multiple faces at the same time in live video. Though it is considered as one of the state of art tools, it has several operational constraints [18] regarding video quality (minimum 10 frames/second), camera resolution for face enrollment (minimum 1 megapixel), face posture (about 180 degrees of roll, 20 degrees of pitch and 45 degrees of yaw), and minimal facial expression, occlusion and illumination variations. If these constraints are not met, facial recognition by this commercial matcher becomes a challenge. Other unconstrained settings like pose, illumination, expression, occlusion and disguise would make the problem even more difficult and challenging. It is a prevalent concept that addressing such challenges also depend heavily on tackling several system level issues [20] as well. Quality of cameras, focal length, quality of NIR illuminators and shutter speed (for tackling motion blur) are some of the system level issues that have to be addressed. Moreover face detection, tracking and dealing with uncooperative subjects also are the challenges encountered with respect to data acquisition from subjects. Developing methods for tackling each of these issues is challenging when it comes to real world unconstrained environments. Therefore, it is our belief that there is a huge scope and prospect for developing new face recognition methods for such applications. The problem of 24-hour surveillance has led the researchers to explore spectrum other than the visible spectrum. In this chapter, we review and discuss the issues of face recognition for surveillance applications. We also provide insights into existing state of the art methods and study the gaps in the existing literature. We conclude with some research directions along with and overview of the salient contributions of this thesis for providing suitable solutions to this problem.

1.1 Challenges for Face Recognition in Unconstrained Scenarios

Facial recognition for frontal and near frontal images has been massively investigated for the last three decades. The research community has attempted to address mainly three challenges, namely pose, illumination and expression [7]. Such applications have been widely used and deployed for access control, authentication, and so on. Recently, as explained in the previous section, due to the insurgence of security threats all over the world, face recognition from surveillance cameras has become an inevitability. In such a scenario there are several challenges that are faced both from system and algorithmic points of view. Since face is probably the most intrusive biometric trait that can be acquired over a distance, it is highly desirable to develop quality methods to tackle this problem.

The main challenges for face recognition in surveillance scenarios are discussed below :

1.1.1 Image Acquisition Issues

There are several issues at the system level for face recognition especially in surveillance conditions. A typical video surveillance system is comprised of an elaborate setup comprising of several cameras, a system to process and match the images along with a repository of gallery images. Several of the system issues discussed here also arise in a closely related problem of face recognition at a distance [22].

The quality of images captured using the cameras has a huge impact on face recognition systems [23, 24]. In a typical surveillance setting the camera is expected to be several meters away from the potential subject, who may be moving away or towards the camera, which mimics a perfect non-cooperative image acquisition scenario. Very often these systems are used for covert surveillance, where the subject is not aware of the fact that he is being photographed. Very often such images are noisy, out of focus and the problem is aggravated by introduction of uncontrolled pose, expression and illumination. Figure 1.3 shows an image of a man caught on the surveillance camera punching a toddler on his face, the image is of very low quality. This is an actual case reported recently in Bakersfield, USA where the surveillance cameras helped the police catch an offender.

Low Resolution: CCTV cameras are typically kept at a distance which helps in getting a wide angle of view over a public area and hence serves the purpose of surveillance, but hampers the facial recognition performance due to several reasons [22]. The inherent distance of the surveillance camera from the subject results in the acquisition of a low resolution image of the subject. On the other hand gallery images are acquired in controlled sceanrios with a high quality camera. This results in a problem for the facial recognition system to match it with a high resolution gallery image. Thus better cameras with very high resolution may solve the

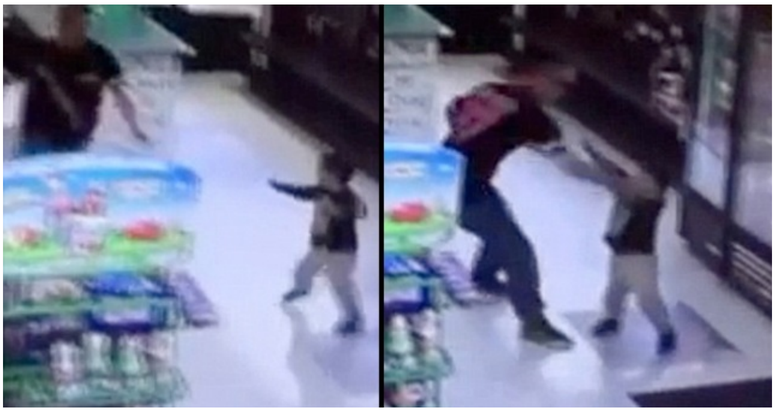


FIGURE 1.3: Low quality image due to rapidly moving subject with uncontrolled pose and illumination. Image source: Youtube

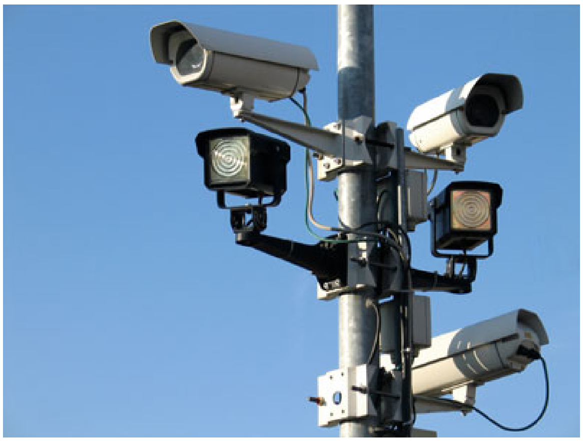


FIGURE 1.4: A typical multiple setup with several NIR illuminators to cover a large field of view. Image source: bbc.com, Image creator: Getty Images.

problem to some extent. Although using very high resolution images may decrease the speed of face detection.

Out of Focus: Although the focus is conceptually a point, most cameras have a small extent within which the focus lies. This region is known as the *blur circle* [22]. In most cases of FRAD, the subject is situated outside this region, which results in image blur. The subjects can be at various distances from the surveillance camera, due to which the camera may find it difficult to adjust the focus quickly for rapidly moving subjects. Moreover, the subject can be well beyond the focal limit of the camera, which will result in a low quality and blurry image.

Height of the Camera: The height of the camera is a major issue in surveillance systems. In order to cover most user's heights, the camera needs to be at a particular height. The issue here is that the facial image of tall people will be different than that of short people when the height of the camera is fixed. Multiple camera systems can also be used as shown in Fig. 1.4.

In such cases there two options for the face recognition systems to effectively use the multiple

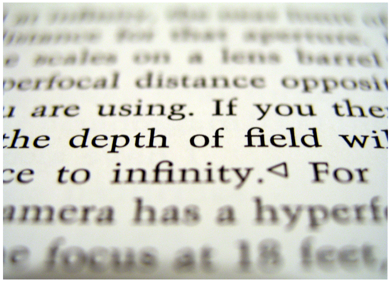

e scales on a lens barrel
perfocal distance opposi
u are using. If you the
the depth of field wi
ce to infinity.◁ For
amera has a hyperf
e focus at 18 feet,

FIGURE 1.5: The area within the DOF appears sharp clear. Image Source: Handbook of Remote Biometrics.[22]

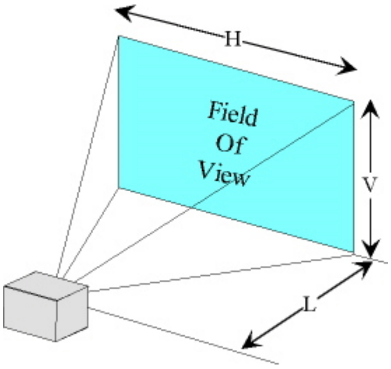


FIGURE 1.6: Field of view. Image source: photographytalk.com

cameras. The first option is to merge the images from the multiple cameras which is a preprocessing task. The second option is to use multiple images independently. The second way has been used in most papers [25] to tackle the problem. This method [25] is an extension of [26], by updating the objective function to incorporate multiple frames. This can be carried forward to create a generalization for multiple cameras. The image formation model used in the objective function takes into account the blurring function of the camera which is a point spread function. For multiple cameras, different blurring functions are taken into account. Each image also would have different regularization parameters depending on which camera was used to acquire that image.

Depth of Field and Field of View: Depth of field (DOF) in a scene is the range between the nearest and farthest object between which objects will appear sharp in an image. It is the area around the focal point where objects appear sharp and clear as shown in Fig 1.5. The DOF for a lens is not symmetric and has separate formulations for the DOF in front of the focal plane and behind the focal plane.

On the other hand the field of view (FOV) does not depend on the distance, rather is dependent on the focal length and the effective resolution of the camera. FOV is defined as the magnitude of the observable scene that can be captured (Fig 1.6) by the camera in good quality. A large FOV is necessary especially for non-cooperative subjects and allows to get more frames on the target as they cross the FOV of the camera.

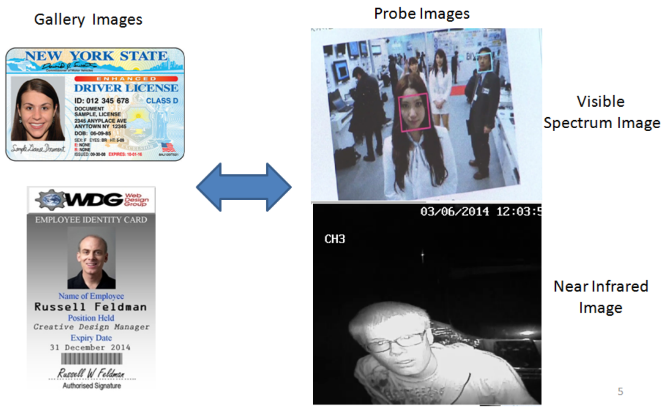


FIGURE 1.7: VIS-NIR face matching problem. Images are crowd-sourced from the internet.

1.1.2 Cross-Spectral Face Matching

Existing legacy datasets have the images acquired in visible spectrum under controlled illumination conditions. These images are used to enroll a user in a biometric identification/authentication system. Probe/Querry images however may be captured under completely different illumination conditions. Very often Near-Infrared (NIR) images are acquired by cameras when the amount of visible illumination is not good enough. These NIR images, would now have to be matched with visible (VIS) light images which were acquired during user enrollment. It has been observed that VIR-NIR face matching is a problem [27] due to different spectral properties of VIS and NIR images and thus it is important to develop efficient methods for doing the same. This problem becomes more challenging when probe images are of a lower resolution in addition to being in NIR spectrum.

1.1.3 Improper Alignment

Face detection and tracking is challenging problem in low resolution images. Detecting facial fiducial points also becomes an increasingly difficult problem in such low quality images. Thus it results in inaccurate face registration and alignment, which severely affects the facial recognition performance.

1.1.4 Dimensional and Quality Mismatch

The gallery images for most facial recognition systems are of high quality acquired in controlled imaging conditions in high resolution. The probe images when acquired in surveillance conditions are expected to be of much lower quality and resolution. Matching higher resolution and good quality images to low quality images is an extremely challenging problem. In Figure 1.9 we can see a representative illustration of the problem. Due to such mismatch feature extraction

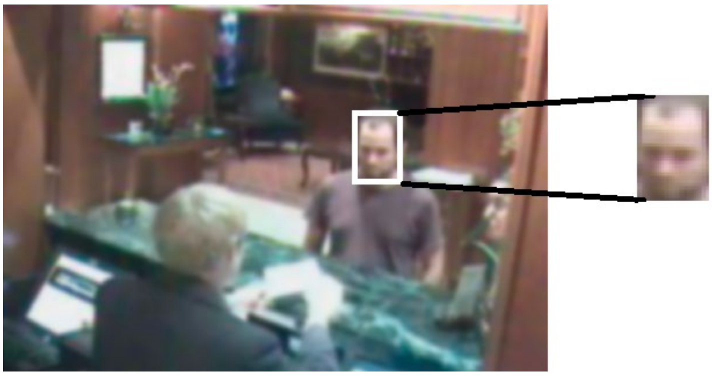


FIGURE 1.8: Out of focus problem: Facial features are not adequate for recognition. Image source: [22]

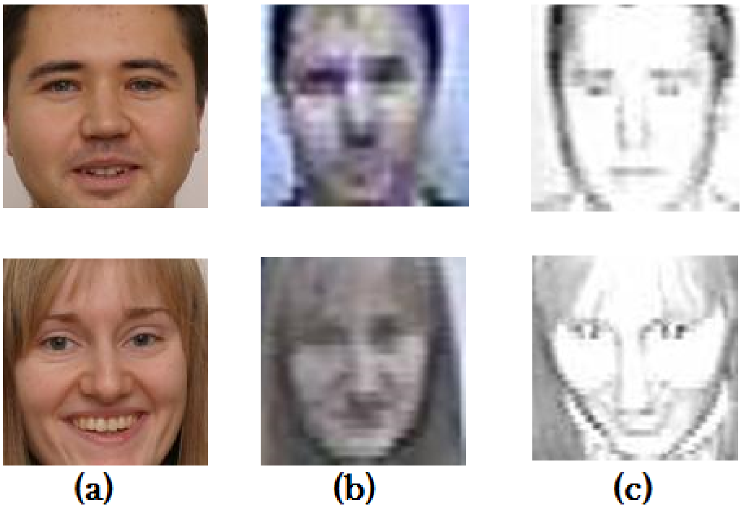


FIGURE 1.9: Dimensional and quality mismatch: (a) high resolution gallery image, (b) low resolution probe in visible spectrum (c) low resolution near-infrared probe image. Images sourced from the SCface dataset [3]

is difficult. Features that work for high quality images [28–31] might not be good enough for low quality images. Thus resolution robust features have been developed for such images. This is still an open research problem and much more needs to be done to solve it.

1.2 Solutions for FR in Unconstrained Scenarios

Face recognition in an unconstrained/survelliance scenario is compounded by several inherent challenges such as low resolution and quality, cross-spectral matching, motion blur, out of focus and so on. Additionally, due to the non-cooperative nature of the users, problems such as pose, expression and illumination are also introduced. This results in images of poor quality, low resolution and in variable spectrums. Most methods introduced for face recognition in surveillance scenarios have approached it from mainly two problem setups, namely, cross-resolution face recognition and cross-spectral face recognition. The covariates of pose, illumination, expression and image quality is implied in both of these problem setups. Since face is probably the most intrusive biometric trait that can be acquired over a distance, it is highly desirable to develop quality methods to tackle this problem.

Thus, it can be inferred that the problem of face recognition in such a complex scenario consisting of multiple covariates at a time, cannot be tackled at a single level rather several approaches at multiple levels are required to mitigate these challenges. In view of the current understanding of this problem, we divide the methods of tackling the challenges of this problem in the following levels.

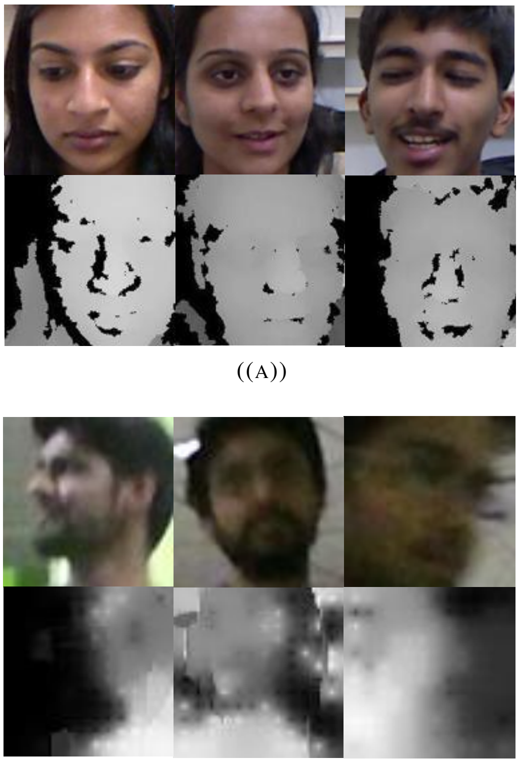
1.2.1 Image Acquisition Level for Heterogeneous FR

In order to have a system which gives good recognition performance on face images acquired in-the-wild, several sensor level issues needs to be addressed. The most prolific issue in Face Recognition at a Distance (FRAD) is image resolution and quality. The cameras frequently used in such tasks are known as PTZ (Pan-Tilt-Zoom) cameras, which has a large field of view. Such a large field of view is essential for covering as much of the scene as possible, especially for deployment in public places. The drawback is that, the effective resolution of the faces captured are very low. A sensor level solution to this problem is using a higher resolution camera, however in such a case the process of face detection would be slower, since a larger number of pixels needs to be processed. Another major issue in surveillance cameras is the 'out of focus' problem. The sensor level solution is to use a camera which has a large focal range and blur circle. Lastly, motion blur is another sensor level issue in surveillance cameras. Since such cameras encompass a completely non-coperative imaging environment, very often the subjects are in motion and thus motion blur may occur. One of the remedies for this is to use a camera which can capture videos at a higher frame rate. If the camera is only capable of taking images, then a higher shutter speed may decrease the chances of motion blur for heavily moving subjects.

The above are some of the commonly used measures for mitigating the system level challenges. Another way of addressing the same would be to use an additional channel of data such as depth maps. The following subsection illustrates how this additional data may help us in improved face recognition performance in unconstrained scenarios.

RGB-D Face Imaging

In unconstrained scenarios, covariates like pose, illumination, expression, distance, resolution, occlusion and so on are introduced. Most face recognition methods utilize 2D images. It has



((a))

((b))

FIGURE 1.10: RGB and depth images: (a) in controlled conditions (IIITD RGB-D database [35]) and (b) with large standoff distance and uncontrolled conditions (Kasparov database [36]).

been shown that in the presence of covariates like pose, illumination, expression and occlusion 3D images yield enhanced recognition performance [32–34] than their 2D counterpart. 3D images can capture facial features in great detail which contributes to high recognition performance. 3D images can be captured in two ways namely passive and active sensing. Passive sensing involves developing a 3D model of a scene by reconstructing shape information from several 2D images captured by multiple cameras from different viewpoints. The need of multiple cameras for developing a 3D model makes it an expensive option when compared to acquiring 2D images. On the other hand active sensing involves using sensors to measure the time taken for the illumination pattern to reflect back to develop an accurate 3D model of the scene. In most 3D cameras it uses a laser beam and also requires nanosecond level precision to work with reasonable accuracy. Most commercial 3D cameras which work with passive sensing, the range of accurate sensing is small and does not work well if the subjects are on the move in the scene. In a surveillance scenario both the distance and the movement of subjects would be unconstrained. In addition to this the cost of the acquisition process also needs to be lower if it is to be applied on a large scale. A 3D model is processed and stored in the form of a polygon mesh. The size of such data is much larger than a 2D image. Storage and processing of such large quantities of data would be a challenge for real time applications like surveillance.

An alternative to 3D is to use pseudo-3D information in the form of RGB-D images. It does not provide a true 3D mesh, but the depth information is provided in the form of a depth map which gives per-pixel depth value of the scene. This depth map however does not provide a very accurate distance of each pixel from the sensor but gives more discriminative information when used along with the RGB image of the same scene. It has been shown [37] that, this depth information, can enhance the performance of face recognition algorithms. This has led to the development of several RGB-D based face recognition methods [38–44] in constrained scenarios. Fig. 2.1(a) shows some RGB images and their corresponding depth maps acquired in constrained scenarios.

This showcases the potential of RGB-D images for usage in unconstrained environments as well. However, in unconstrained scenarios the quality of depth information acquired by low cost sensors like the Microsoft Kinect would be of a much poorer quality than acquired in constrained scenarios. Thus using RGB-D images in unconstrained environments is challenging and requires development of new algorithms that can utilize the low quality depth information. Most RGB-D face recognition algorithms, needs both the RGB and the depth image during testing. However, as shown in Fig 2.1(b) when face images are acquired from a larger standoff distance (more than 2 meters) the depth image is of very poor quality. In such cases all the conventional RGB-D face recognition algorithms would yield poor performance as they depend on the quality of the depth image while testing. In some cases the depth image is not acquired at all. In such cases conventional RGB-D face recognition algorithms would not be able to perform any testing. We present an algorithm that does not need depth image during testing. Our proposed method learns a shared representation from RGB (in grayscale) and depth data and uses only RGB (in grayscale) data during testing. This is an advantage over conventional RGB-D recognition algorithms.

Recently, RGB-D images have also been utilized in surveillance scenarios for applications such as person tracking [45] and person re-identification [46, 47]. The advent of new sensors like Microsoft Kinect [48] and Asus Xtion PRO LIVE [49] has presented a cost effective way of obtaining RGB-D images and thus presents a scalable solution for deployment in surveillance scenarios. These sensors do not explicitly record the depth information using a depth sensor, rather it uses an IR (Infrared) camera and an IR projector for capturing the depth information of a scene. A speckle dot pattern is cast by the IR projector on to the scene, which is captured by the IR camera in the form of reflected IR speckles. This method of capturing depth is known as structured-light 3D surface imaging [50].

However, RGB-D face recognition in unconstrained scenarios like surveillance is challenging. Surveillance cameras need to operate under varying illumination conditions. The standoff distance of the subject from the camera is usually large. In addition to this, other covariates like pose, expression and occlusion are also introduced. Thus, RGB-D face recognition in surveillance scenarios is a difficult problem and largely unexplored in literature. Recently the KaspAROV [36] database was used to perform RGB-D face recognition [51] which is the first RGB-D database captured in unconstrained and surveillance conditions. As shown in Fig. 2.1(b) when an RGB-D face image is acquired from a larger standoff distance (more than 2 meters) the depth map acquired is of very poor quality. Due to this the performance of conventional RGB-D face recognition algorithms on such RGB-D images would be severely affected. It has been shown in [51] that a conventional RGB-D face recognition algorithm [35] which gives excellent performance on constrained RGB-D images (Fig. 2.1(a)) performs poorly when applied on RGB-D images (acquired under surveillance conditions) of the KaspAROV database (Fig. 2.1(b)).

In addition to face recognition, RGB-D images have been used in face detection [52], object detection and segmentation [53], object recognition [54], 3D modeling [55], gender recognition [56], object discovery [57], human action recognition [58], head pose estimation [59], face anti-spoofing [60] and so on.

1.2.2 Image Level Solutions for Heterogeneous FR

At the image level the only measure that can be taken to improve face recognition in an unconstrained scenario is to improve the quality of the image. Some of the classical methods that have been utilized to improve image quality is superresolution and hallucination. There have been a plethora of methods proposed for general image superresolution, however all of them would not be helpful in this case, due to the fact that the minute discriminative information present in the face image needs to be restored during the process of quality/resolution improvement. Recently, with the advent of Generative Adversarial Networks (GANs), improved image to image translation algorithms have been proposed which showcases excellent quality and resolution improvement at the image level. Next, we illustrate the effect of resolution on the performance of a typical face recognition system. thereafter we discuss the popular methods for increasing the resolution and quality of images using superresolution/hallucination and GAN based methods.

The Effect of Resolution on FR Performance A digital image is made of of several pixels. Spatial resolution of an image is known as the number of pixels per unit area. At times it is also measured by the total number of pixels in an image. Higher resolution images are expected to encompass much more detail about the scene it captures. A typical face recognition application is expected to have a high resolution gallery image, the probe image can be of a much lower resolution, depending on the camera which is used for acquisition of the probe image. Alike other image recognition applications, face recognition is expected to produce much better results with a higher resolution probe image. The imaging resolution is limited by two major factors, the camera sensor and the standoff distance of the subject from the camera. Most modern sensors are

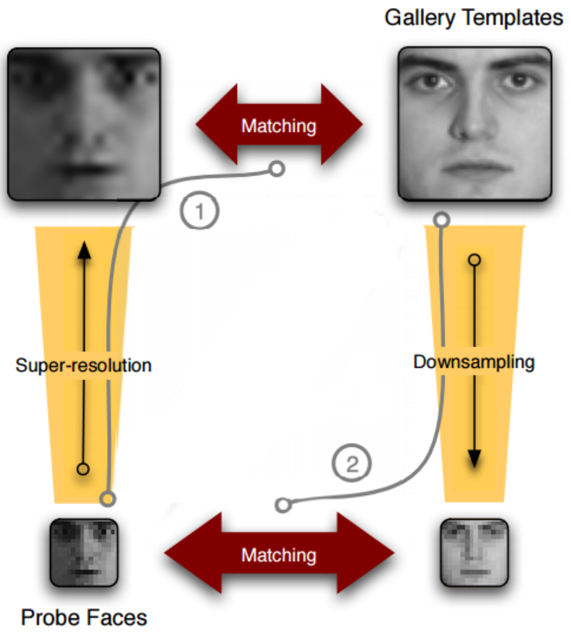


FIGURE 1.11: Matching high resolution gallery image with low resolution probe by adjusting the resolution. Image source: [26]

either a Charge-Coupled Device(CCD) or a CMOS active-pixel sensor. The size of the sensor or the number of sensing elements per unit area determine the spatial resolution of the image.

Due to some of these limitations the acquired pixel may be of a lower resolution than the gallery image of the subject enrolled in the system. In such a case we face the problem of matching a lower resolution probe image with a higher resolution gallery image. In the context of face recognition this problem is known as cross resolution face matching. The use case scenarios where such problems might be faced has already been discussed in the earlier sections. In this section we would discuss how to tackle and mitigate this problem.

Superresolution: One of the ways of tackling this problem is to bring both the gallery and the probe image to the same resolution. As shown in Figure. 1.11, this can be done in two ways, either by applying super-resolution on the probe image to increase its resolution and make it at par with the gallery images, or downsample the gallery image. Several methods have been proposed in literature for the same. We will however restrain our discussion mostly on those methods that are suitable for face images.

Traditional Methods for Super-resolution: There have been a lot of methods proposed to learn the correspondence between the low resolution image and its high resolution version, by using a training set which contained a set of low resolution and their corresponding high resolution versions. Face hallucination proposed by Baker and Kanade [61], was specifically proposed for super-resolution of face images. It used such a training set and learned a MAP based conditional probability objective function. The probability of the high resolution image given the low resolution image was computed by learning the likelihood and the prior from the training image pairs. The priors used were known as gradient based priors. Since then there has been several methods proposed for face hallucination.

Recognition oriented Super-resolution for Low Resolution Face Recognition: The first method which proposed a simultaneous super resolution and recognition technique was by Hennings-Yeomans *et al.* [26]. The objective function for finding a high resolution version of a given low resolution probe image has two components. Firstly reducing the error between the given low resolution probe and the simulated low resolution image of the corresponding super resolved high resolution image. Secondly it takes into account the difference between the ideal representation for the class to which the probe image belongs and the features of the super resolved probe image.

Deep Learning Methods for Image Super-Resolution: The first method to use a convolutional neural network model for the problem of image super-resolution was proposed in [62]. The paper is probably the first one to propose a deep learning solution for image super-resolution. It uses a convolutional neural network to learn an end to end mapping between low resolution and high resolution images. It contains three stages namely, patch extraction and representation, non linear mapping and reconstruction. The first stage is analogous to the convolution operation, the second stage is the pooling operation. The network proposed in the paper contains one convolutional and pooling layer, however then can be multiple layers used for superior performance. The paper also shows that with a few training iterations the average test PSNR values are much higher than bicubic interpolation and sparse coding based methods for super-resolution. There have been several other works which has used CNNs for image superresolution.

Image to Image Translation using Generative Adversarial Networks (GANs) Recently with the advent of GANs, generating synthetic data have been performed with high reliability and excellent performance. One of the foremost applications of such algorithms is to train a deep learning model for image to image translation, which can be utilized to transform low resolution image into their high resolution counterparts.

A GAN [63] is a deep network which consists of a generator G and a discriminator D, where θ_g and θ_d are the parameters of the generator and discriminator respectively. The generator network G produces a synthetic image from a noise distribution p_z, and the discriminator network D is trained to distinguish between a synthetic image $G(z)$ and a real image x sampled from the distribution of real images p_{data}. It is formulated as follows:

$$min_{\theta_g} max_{\theta_d} [\mathbb{E}_{x \in p_{data}} \log D_{\theta_d}(x) + \mathbb{E}_{z \in p_z} \log(1 - D_{\theta_d}(G_{\theta_g}(z)))] \tag{1.1}$$

The G and D models are trained alternatively by updating the parameters. The parameter update of the generator is given by

$$\nabla_{\theta_g} \frac{1}{m} \sum_{i=1}^{m} \left[\log \left(1 - D_{\theta_d} \left(G_{\theta_g} \left(z^{(i)} \right) \right) \right) \right] \tag{1.2}$$

and that of the discriminator is given by

$$\nabla_{\theta_d} \frac{1}{m} \sum_{i=1}^{m} \left[\log D_{\theta_d} \left(x^{(i)} \right) + \log \left(1 - D_{\theta_d} \left(G_{\theta_g} \left(z^{(i)} \right) \right) \right) \right] \tag{1.3}$$

where, *m* is the size of the minibatch used. The initial model of GANs [63] allowed the generation of synthetic images from noise. The applicability of such models is in data augmentation which has resulted in generation of large amount of training data for better and robust training of classifiers.

Image Translation using Conditional GANs: The original GAN model was extended to Conditional GANs [64], in order to utilize GANs for translating images from one domain to another. It allows both the generator and the discriminator to be conditioned on an extra information y which can be either the class label of the sample or any extra auxiliary information. Such models can be formulated as:

$$min_{\theta_g} max_{\theta_d} [\mathbb{E}_{x \in p_{data}} \log D_{\theta_d}(x|y) + \mathbb{E}_{z \in p_z} \log(1 - D_{\theta_d}(G_{\theta_g}(z|y)))] \tag{1.4}$$

The conditional GAN model is formulated for image to image translation by Isola *et al.* [65]. It can be modeled as a mapping function $G : \{\mathbf{p}, \mathbf{z}\} \rightarrow \mathbf{q}$ where $\mathbf{p}$ and $\mathbf{q}$ are sets of images of the source and target modalities respectively and $\mathbf{z}$ is a noise vector. It can be formulated as follows:

$$\varphi_{GAN}(G, D) = [\mathbb{E}_{\mathbf{p},\mathbf{k} \in p_{data}} \log D_{\theta_d}(\mathbf{p}, \mathbf{k}) + \mathbb{E}_{\mathbf{z} \in p_z; \mathbf{p},\mathbf{x} \in p_{data}} \log(1 - D_{\theta_d}(\mathbf{x}, G_{\theta_g}(\mathbf{p}, \mathbf{z})))] \tag{1.5}$$

The generator can also be conditioned on an extra constraint $\varphi_{l_1}(G)$ which is the l_1 difference of the input images $\mathbf{p}$ and output images $\mathbf{q}$. The final objective function of the above model is as follows:

$$G^* = min_{\theta_g} max_{\theta_d} \left[\varphi_{GAN}(G, D) + \varphi_{l_1}(G) \right] \tag{1.6}$$

The l_1 constraint makes the output images (from the generator) consistent with the structural properties of the target image. The drawback of such a model is that it is trained only on the goal of transforming the image from one modality into another. As a byproduct of such a transformation, discriminative information in the images (which are important for classifying the image) might get distorted. Thus, a supervised version of the above model is proposed and is utilized to train a GAN for translating face images from the source to the target modality. It ensures that the class of the generated image is consistent with that of the input image of the source domain.

There have been significant advances in the investigation of Generative Adversarial Networks and are applied in various applications such as image generation [66–69], image to image translation [65, 70, 71], social network analysis [72, 73] image superresolution [74], 3D shape modeling [75], text to image synthesis [76], image style transfer [77, 78], and texture synthesis [79]. In this paper, the focus is particularly on image to image translation, which is one of the most popular applications of GANs. Most image to image translation methods using GANs [65, 70, 71, 80] can transform an image from the source domain to the target domain effectively. Isola *et al.* [65] proposed one of the first image to image translation based methods using GANs. They used a generator which received a random noise and the source domain image as input. The discriminator was given a pair of images (real and fake) as input and it had to discriminate between a real-real and a real-fake pair. This was extended in [70] by introducing a cycle consistency loss for unpaired image to image translation. Yi *et al.* [71] used a dual-learning [81] based formulation to train a GAN model for unpaired image to image translation. These methods only focus on the domain transfer, and do not take into account that important discriminative information may be distorted during the process of domain translation.

1.2.3 Feature Level Solutions for Heterogeneous FR

At the feature level, algorithms have been proposed to train effective models which can produce discriminative features in the embedding space of the model. Prior to the popularity of deep learning, such methods were mostly based on discriminative learning and feature transformation based methods. Such methods, broadly known as structure based methods, aim to project both the low resolution/quality probe and the high resolution/quality gallery into a common subspace, where they can be efficiently matched. The popularity of deep learning contributed to this advancement by the development of novel and effective loss functions that is focused primarily on the feature space. These methods, known as Deep Metric Learning (DML), update the parameters of the model such that it produces more feature rich representations in the embedding space of the model. In case of face recognition in unconstrained scenarios, these loss functions provide a huge impetus for the improvement of algorithms for the same. Deep Metric Learning (DML) algorithms allows us to train discriminative classifiers even on databases containing insufficient training data.

In addition to traditional structure based methods, this section establishes DML as an effective technique for training highly discriminative models, for face recognition from low quality and/or NIR images captured in surveillance scenarios.

Structure Based Methods: The most traditional class of methods for face recognition from low quality images, are structure based methods. These approaches map the high resolution gallery images and the low resolution probe images into a common feature space, where they can be matched. Matching the low resolution and the high resolution images directly in the input feature space is not possible due to difference in dimensionality and/or spectrum of the images. Moreover, due to considerable difference in quality such matching would not yield acceptable results. These methods learn this kind of mapping which is then used to map both the HR and LR images and then matching is performed in the transformed space.

Bhatt et. al. [23] proposed a co-transfer learning framework for cross-resolution face recognition problem. In a transfer learning framework there is notion of a source and a target domain. For cross-resolution face recognition the source domain consists of high resolution face images and the target domain has low resolution face images. It is not difficult to have a well trained classifier in the source domain due to the abundance of labeled high resolution face images, but doing the same for low resolution images is a difficult task. The scarcity of labeled low resolution images makes it difficult to have a well trained classifier on such images. This paper was the one of first ones to propose a framework which uses both transfer learning and co-training to transfer knowledge from the source to the target domain. The base classifier used is support vector machines. This method uses two views (features) of the data for utilizing the co-training on the unlabelled probe data instances. It utilizes unlabelled target domain probe data instances by classifying them using two ensemble prediction functions (one for each view), which uses a weighted combination of the source and the target domain classifier to predict a pseudo-label. These data instances are then used to retrain the target domain classifier.

Two notable methods. [82, 83] uses multidimensional scaling to embed both the HR gallery images and LR probe images into a space where their distances would closely approximate the situation if both the images were in high resolution. This method does not require to bring HR and LR images to the same resolution for matching. Since HR and LR images are of different dimensions two different weight matrices (for LR and HR) needs to be learnt. The training data consists of HR and corresponding LR images. The objective function for learning the transformation functions tries to minimize the error between the distance between the representation of the images in the transformed space and the representation of the images in the images if they were in high resolution. The objective function also contains a class separability function so as to improve discriminability of the images in the transformed space. The class separability term ensures that distance between the transformed feature representations of the same subject is small compared to that of different subjects. The iterative majorization algorithm is used to minimize this objective function.

A similar method [84], was proposed which can match face images across resolution, pose and illumination. It uses SIFT [30] descriptors from 15 spatial fiducial keypoints as features to learn the transformation of HR frontal gallery images and LR arbitrary-pose-illuminated images into a common subspace using Multidimensional scaling. During training the features are extracted from facial keypoints which are extracted by STASM [85]. During testing SIFT features can be extracted from the entire face image and the learned transformation function is used to map the features into the transformed space, where the stereo cost is computed between transformed gallery and probe images by using the algorithm by Criminisi *et al.* [86].

Recently in [87] an approach for matching LR images to HR images was proposed, where coupled mappings were learnt to map both the LR and the HR images to a unified latent subspace so that they can be matched efficiently. A local optimization problem is formulated, which contains three components, namely, consistency of LR/HR images, interclass distance and intraclass distance. The coupled mappings are solved by solving an eigen-decomposition. Then, a formulation is proposed to combine all the local geometries into a global structure as shown in [87].

Deep Metric Learning (DML): Conventional deep learning models like Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) are trained with data samples along with their corresponding labels, so that they can correctly predict the class/label of an input sample during testing. However, deep metric learning algorithms train a model with the objective of distinguishing between a pair of data samples whether they belong to the same class/category or not. During training, a vanilla deep metric learning loss function would update the weights of the model so that it produces embeddings/features (unlike class labels in conventional deep learning models) of data samples that belong to the same class close to each other, and that of different classes away from each other in the output embedding space of the model. In order to train such a model, we need large quantities of data samples during training. Since a discriminative model is being trained, it may be tested/evaluated on classes that are not encountered by the model during training. This flexibility makes deep metric learning models a popular choice for building real world recognition systems.

The most seminal work in deep metric learning was by Chopra *et al.* [88] where the contrastive loss was proposed. It optimized a Siamese network for matching a pair of images, by the same optimization goal as illustrated above. Thereafter, several research works [89, 90] have utilized this optimization technique using a deep-CNN network as the backbone model, before a new loss function, known as the triplet loss was proposed by Schroff *et al.* [91]. This was extended in 2017 , by a new loss function known as the quadruplet loss [92]. An N-pair loss metric [93] was also proposed which uses an N-tuple as a training data sample. Several variants of these loss functions have been proposed, and some of them cater to scenarios of small sample learning.

1.2.4 Sample Mining in DML

One of the most important and heavily discussed issues in the development of an efficient DML method is sample mining. In order to train a deep network using loss functions such as the triplet loss, which is the most popular among them, we need to prepare triplets (or 3-tuples) using the data samples available for training. Given N training classes and K samples for each class, the total number of triplets that can be prepared for training is upper bounded by $N(N-1)K^2(K-1)$. Therefore, the number of training samples (each triplet is treated as a training sample) increases from $O(N.K)$ (available for conventional deep learning algorithms) to $O(N^2.K^3)$ which is a very large sample space. For quadruplets or N-pair loss functions, this space would be ever larger. This increased sample space is extremely useful for learning a model with a DML algorithm on a database that has a small number data samples. It also makes DML algorithms a natural choice where the amount of training data is insufficient to learn a conventional classifier. However, on large databases this enormous input sample space may hinder efficient learning due to several reasons. One of the reasons is that, after several epochs of training, the model would have learnt to solve most of the data samples, each of which is a triplet/quadruplet. Thus, fewer samples would be useful for the model to continue learning and make significant weight updates. During this phase it is required to provide only useful triplets/quadruplets, in other words mine those triplets/quadruplets which are hard (which is still not correctly classified by the model) in order to continue learning the model. This technique known as hard mining, has been extensively explored [94–98] in the last few years for DML methods.

1.3 Research Contributions

The key contribution of this dissertation is heterogeneous face recognition (As explained in the previous sections and Figure 1.12) which caters to face recognition in real world unconstrained environments. Face recognition in such unconstrained scenarios is of paramount importance to surveillance and non-cooperative authentication scenarios. This dissertation focuses on improving the performance of heterogeneous face recognition by learning discriminative representations which are invariant to heterogeneity. **We present five different contributions each of which caters to the central idea of heterogeneous face recognition, although the primary use case for them is face surveillance. Experiments and evaluations for each chapter has been illustrated accordingly.** In order to achieve this, first, we present a shared representation learning based approach which can be trained to generate such representations for RGB-D images acquired under unconstrained scenarios. This can be utilized to match with high quality gallery images for face recognition. Thereafter we propose a supervised image to image translation algorithm which uses generative modelling to achieve the same. Further, this dissertation

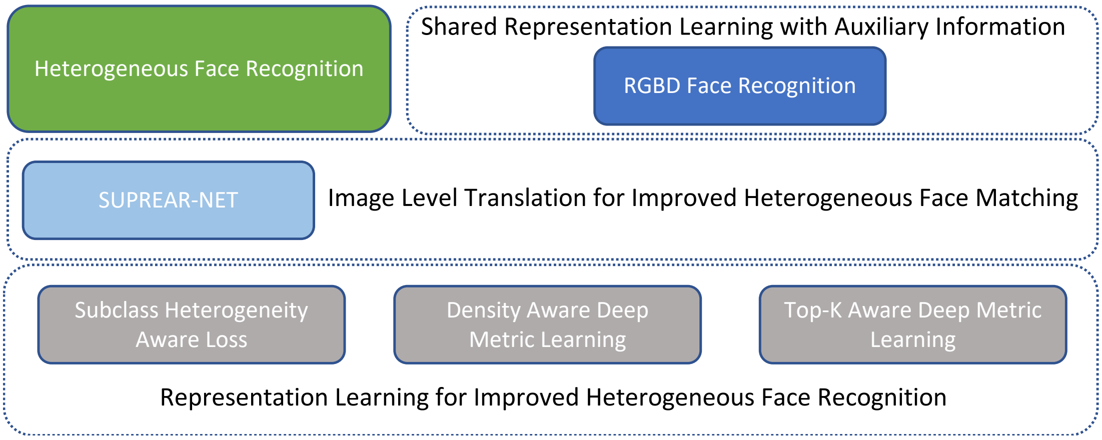

Heterogeneous Face Recognition
Shared Representation Learning with Auxiliary Information
RGBD Face Recognition
SUPREAR-NET
Image Level Translation for Improved Heterogeneous Face Matching
Subclass Heterogeneity Aware Loss
Density Aware Deep Metric Learning
Top-K Aware Deep Metric Learning
Representation Learning for Improved Heterogeneous Face Recognition

FIGURE 1.12: Illustration of the research contributions

also contributes in proposing novel algorithms for training heterogeneity aware discriminative models using novel deep metric learning methods for face recognition. The key contributions of this dissertation are as follows:

The first contribution of this dissertation is aimed at learning a shared representation for unconstrained face images utilizing an auxiliary information in the form of RGB-D images.

Shared Representation Learning for RGB-D Face Recognition Low cost time-of-flight based depth sensors such as Kinect have opened new avenues for their usage in video surveillance scenarios. RGB-D images obtained from such sensors have shown their utility in improved face recognition capabilities. Generally, existing RGB-D face recognition methods fuse the depth information with RGB information which results in enhanced recognition performance. However, in the real world surveillance scenarios, cameras are placed at a distance too large for low cost depth sensors to capture good quality depth information. Such poor quality depth information may not contribute significantly to face recognition. In this paper, we present a novel representation learning algorithm by learning shared representation of RGB and depth information using a reconstruction based deep neural network. The proposed network, once trained in the offline mode, can generate a shared representation of RGB and depth data using only the RGB image. This feature rich representation is then utilized for face identification. This allows the framework to be used in scenarios where low quality or no depth image is captured. Experiments on multiple real world databases show the effectiveness of the proposed approach.

The second contribution of this dissertation contributes in translating poor quality low resolution face images at the image level using generative modelling.

Supervised Resolution Enhancement and Recognition Network: Heterogeneous face recognition is a challenging problem where the probe and gallery images belong to different modalities such as, low and high resolution, visible and near-infrared spectrum. A Generative Adversarial Network (GAN) enables us to learn an image to image transformation model for enhancing the resolution of a face image. Such a model would be helpful in a heterogeneous face recognition scenario. However, unsupervised GAN based transformation methods in their native formulation might alter useful discriminative information in the transformed face images. This affects the performance of face recognition algorithms when applied on the transformed images. We propose a Supervised Resolution Enhancement and Recognition Network (SUPREAR-NET), which does not corrupt the useful class-specific information of the face image and transforms a low resolution probe image into a high resolution one, followed by effective matching with the gallery using a trained discriminative model. We show the results for cross-resolution face recognition on three datasets including the FaceSurv face dataset, containing poor quality low resolution videos captured at a standoff distance up to 10 meters from the camera. On the FaceSurv, NIST MEDS and CMU MultiPIE datasets, the proposed algorithm outperforms recent unsupervised and supervised GAN algorithms.

The next three contributions are focused towards learning discriminative embeddings for heterogeneous face recognition utilizing novel Deep Metric Learning methods.

Subclass Heterogeneity Aware Loss: One of the most challenging scenarios of face recognition is matching images in presence of multiple covariates such as cross-spectrum and cross-resolution. In this work, we propose a Subclass Heterogeneity Aware Loss (SHEAL) to train a deep convolutional neural network model such that it produces embeddings suitable for heterogeneous face recognition, both single and multiple heterogeneities. The performance of the proposed SHEAL function is evaluated on four databases in terms of the recognition performance as well as convergence in time and epochs. We observe that SHEAL not only yields state-of-the-art results for the most challenging case of Cross-Spectral Cross-Resolution face recognition, it also achieves excellent performance on homogeneous face recognition.

Density Aware Deep Metric Learning Deep metric learning algorithms have been utilized to learn discriminative and generalizable models which are effective for classifying unseen classes. In this work, a novel noise tolerant deep metric learning algorithm is proposed. The proposed method, termed as Density Aware Deep Metric Learning, enforces the model to learn embeddings that are pulled towards the most dense region of the clusters for each class. It is achieved by iteratively shifting the estimate of the center towards the dense region of the cluster thereby leading to faster convergence and higher generalizability. In addition to this, the approach is robust to noisy samples in the training data, often present as outliers. Detailed experiments and analysis on two challenging cross-modal face recognition databases and two popular object recognition databases exhibit the efficacy of the proposed approach. It has superior convergence, requires lesser training time, and yields better accuracies than several popular deep metric learning methods.

Top-K Aware Deep Metric Learning: Recently, in addition to optimizing for the classification accuracy, the top-k accuracy has also gained considerable attention from machine learning practitioners. Optimizing the overall classification accuracy of a network does not always lead to best top-k accuracy. This behavior is often observed in cases where multiple classes are close to each other in the embedding space and trained classifiers may not retrieve the correct class due to class ambiguity. This work presents an elegant solution for enhancing the top-k matching performance. The algorithm first uses a clustering algorithm to identify *superclusters*, which are made of classes that are similar and are mapped close to each other in the embedding space. The compactness of these *superclusters* is then enhanced while protecting the discriminative properties of the individual classes, which results in improved top-k matching performance during testing. Experimental results on STL-10, CIFAR-10, CIFAR-100, Stanford Online Products, CARS196, and SCface databases demonstrate the efficacy of the proposed approach.

So, to summarize, the contributions are catering to the problem of Heterogeneous face Recognition, which is to perform face recognition in the presence multiple covariates like resolution, spectrum, pose and so on. These scenarios are more prevalent in surveillance scenarios and each chapter will illustrate experiments on at least on one face dataset captured in an unconstrained setup. Figure1.12 illustrates the different types of problems that this dissertation is aimed at under the umbrella of Heterogeneous Face Recognition.

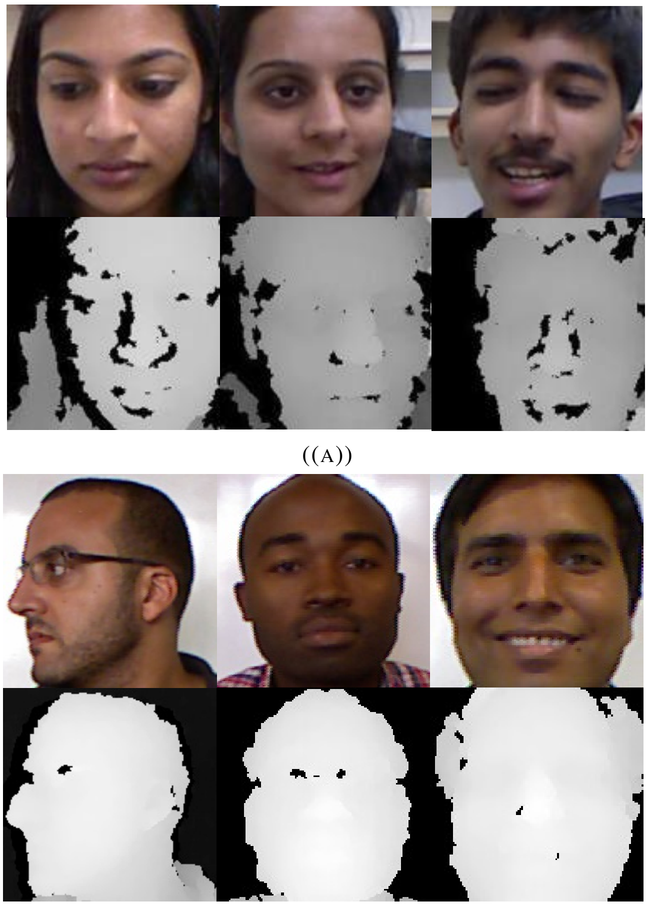
Chapter 2

RGB-D Face Recognition using Reconstruction based Shared Representation

2.1 Introduction

Innate capabilities of human mind to recognize familiar faces has motivated researchers to mimic and build next generation algorithms. In the last two decades, face recognition has been one of the most investigated topics in the area of computer vision and artificial intelligence. In general, face recognition methods utilize 2D images. However, it has been shown that in the presence of covariates such as pose, illumination, expression and occlusion, 3D images yield enhanced recognition performance than their 2D counterpart [32–34, 39]. 3D images can capture facial features in great detail which contribute to high recognition performance. However, the requirement of multiple cameras for developing a 3D model and the high cost of range sensors makes it an expensive option when compared to acquiring 2D images.

RGB-D cameras such as Microsoft Kinect [48], Asus Xtion PRO LIVE [49] and Xtion 2 [100] offer an attractive alternative to expensive 3D imaging. These cost-effective sensors provide RGB-D images that may be used in controlled surveillance scenarios. Generally, these cameras do not provide a true 3D mesh, but the depth information is provided in the form of a depth map which gives per-pixel depth value of the scene. It uses an IR (Infrared) camera and an IR projector for capturing the depth information of a scene. A speckle dot pattern is cast by the IR projector on to the scene, which is captured by the IR camera in the form of reflected IR speckles. This method of capturing depth is known as structured-light 3D surface imaging [50]. The presence of an extra depth sensor in a sensor allows capturing information from both RGB and depth sensor simultaneously [101]. Although this depth map does not provide a very accurate



((A))

((B))

FIGURE 2.1: RGB and depth images (upper row: RGB images, lower row: depth images) in controlled conditions (a) IIITD RGB-D database [35] and (b) Eurecom dataset [99]).

distance of each pixel from the sensor but researchers have shown [37, 102] that, it can be combined with RGB information to enhance the performance of face recognition algorithms. This has led to the development of several RGB-D based face recognition methods in constrained scenarios [38–44, 103–105]. Recently, a detailed study of representative fusion schemes have been performed by Cui *et al.* [106]. Figure 2.1 shows some RGB images and their corresponding depth maps acquired in constrained scenarios. In addition to face recognition, RGB-D images have also been utilized for several other applications such as face detection [52], object detection and segmentation [53], object recognition [54], 3D modeling, gender recognition [56], object discovery [57], human action recognition [58], head pose estimation [59, 107], and face anti-spoofing [60].

Li et al. [38] presented one of the first methods for RGB-D face recognition. They utilized the depth map to crop the face from the RGB-D image captured by the Kinect sensor. A set of preprocessing operations like pose correction and filling of missing data using facial symmetry is performed. Thereafter, Discriminant Color space transform [108] is applied and three color

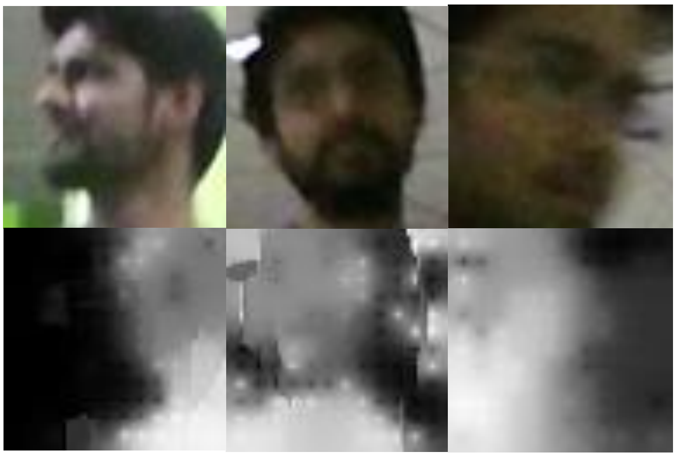


FIGURE 2.2: RGB-D images (upper row: RGB images, lower row: depth images) with large standoff distance in surveillance scenarios (Kasparov database [114])

channels are concatenated into a single vector. Face matching is performed using a Sparse Representation classifier [10]. Ciaccio et al. [109] proposed an RGB-D face recognition algorithm where images of different poses were rendered from the RGB-D face images. Thus, multiple face images were generated even though the gallery contained only one RGB-D image per subject. Face representations were generated using a combination of Local Binary Patterns (LBP) and the covariance descriptor [110]. A probabilistic integration scheme was proposed for face matching. Segundo et al. [111] proposed an RGB-D face recognition algorithm for continuous face authentication. They used the iterative closest point (ICP) method to normalize the face followed by Histogram of Oriented Gradients (HOG) features for face matching. Goswami et al. [39] proposed an RGB-D image descriptor based on saliency and entropy. Entropy map is extracted from the depth image and both entropy and saliency map are extracted from RGB image. Thereafter, HOG features were computed on them and face identification was done using a random decision forest classifier.

Hsu et al. [112] proposed a 3D reconstruction based approach followed by sparse representation based classifier for RGB-D face recognition. The proposed 3D reconstruction method works in two stages. The first stage handles the corrupted depth map captured. In the second stage, an iterative face surface estimation approach was proposed. Hayat et al. [44] proposed a block based covariance matrix representation for RGB-D face recognition which is used to model images in a subset on Riemannian manifold (Lie group). SVM is used for classification on each subset of the Lie group. The results from all these subsets are combined using a fusion algorithm. Xu et al. [113] formulated RGB-D face recognition as a distance metric learning task with privileged information. The depth images are treated as privileged information during the training process. Using this approach, a Mahalanobis distance is learned, which is used for face verification.

Li et al. [43] proposed Multi-channel Discriminant Transform (MDT) which was applied on

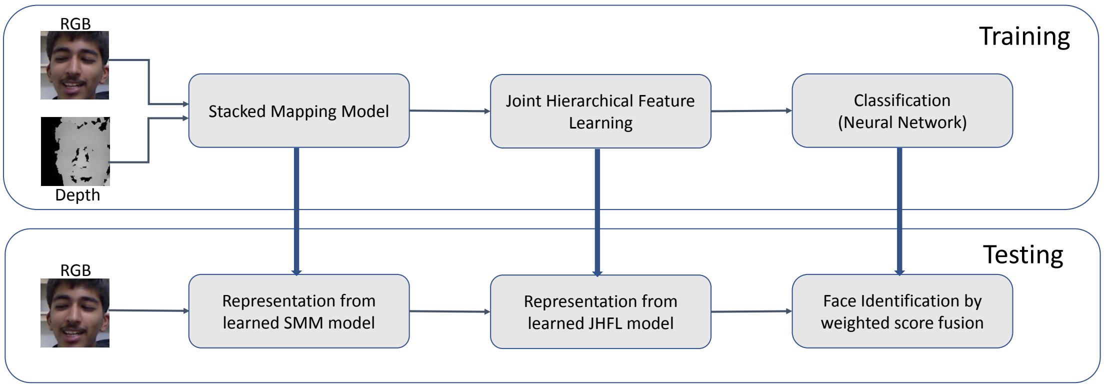

RGB
Depth
Training
Stacked Mapping Model
Joint Hierarchical Feature Learning
Classification (Neural Network)
RGB
Testing
Representation from learned SMM model
Representation from learned JHFL model
Face Identification by weighted score fusion

FIGURE 2.3: Block diagram illustrating the steps of the proposed algorithm. In the training block, the representations are learned using the Stacked Mapping Model (SMM) and Joint Hierarchical Feature Learning (JHFL) Model. In the testing phase only the RGB image is used to generate representations from the SMM and the JHFL models which are used for identification by trained neural network classifier.

both the RGB image and the depth map for face recognition. A Multi-channel Weighted Sparse Coding (MWSC) method computes a weight mask over multiple data channels for finding the invariance property of the channels to imaging conditions. Chowdhury et al. [51] presented an RGB-D face recognition algorithm which uses a neural network to reconstruct depth images from RGB images. It utilizes the reconstructed depth images to train a neural network classifier for face identification. Recently, Jiang et al. [115] proposed an approach for training a deep learning model for RGB-D face recognition utilizing facial features in an attribute-aware loss function.

A large number of algorithms have been proposed for RGB-D face recognition in constrained environments. However, most of the real world scenarios where face recognition systems are deployed are unconstrained environments. In the last few years, face recognition systems have been heavily deployed [116] for surveillance applications and CCTV cameras have been installed in several public places to facilitate it. However, RGB-D face recognition in unconstrained surveillance scenario is challenging. Surveillance cameras need to operate under varying illumination conditions. The standoff distance of the subject from the camera is usually large. In addition to this, other covariates such as pose, expression and occlusion, are also introduced. As shown in Figure 2.2 when an RGB-D face image is acquired from a larger standoff distance (more than 2 meters) the depth map acquired is of very poor quality. Due to this, the performance of conventional RGB-D face recognition algorithms on such RGB-D images would be severely affected. It has been shown in [51] that a conventional RGB-D face recognition algorithm [35] which gives excellent performance on constrained RGB-D images (Figure 2.1(a)) performs poorly when applied on RGB-D images (acquired under surveillance conditions) of the Kasparov database [114] (Figure 2.1(b)).

In this research, as shown in Figure 2.3, we present a novel RGB-D face recognition algorithm which learns a mapping from RGB to depth data using a neural network based mapping model termed as the Stacked Mapping Model (SMM). This mapping model learns the representation of RGB and depth data in its hidden layers. We learn another level of representation on the features learned by the mapping model using the proposed Joint Hierarchical Feature Learning (JHFL). This enables us in combining representations learned by the SMM into a shared representation of RGB and depth during training. During testing we use only RGB images to generate the shared representation of RGB and depth. Neural network classifier is then utilized for face identification. Since the proposed algorithm does not use the depth image during testing, the poor quality depth map acquired in a surveillance scenario would not affect its performance. We evaluate the proposed algorithm using two real world MS Kinect based RGB-D databases, IIITD RGB-D and Kasparov. The research contributions of this work can be summarized as:

- We present RGB-D face recognition algorithm which does not need depth images while testing. This makes our algorithm suitable to surveillance scenarios where the captured depth image by sensors like Kinect [48] are expected to be of poor quality.
- We present a stacked mapping model, which maps RGB data to depth data. We show that such a mapping network can be utilized to learn shared representation of RGB and depth data by putting together the representation from different layers of this mapping model into a hierarchical representation learning architecture.
- We achieve state-of-the-art (SOTA) identification accuracies in two RGB-D face recognition datasets, namely the IIITD RGB-D [35] and Kasparov [114] datasets.

2.2 Stacked Mappping Model (SMM)

The proposed Stacked Mapping Model (SMM), is a multi layer neural network model which learns a mapping from RGB to depth images. In this mapping process, the model learns a *fused* representation of both the modalities in its hidden layers. Once trained, the weights of the SMM is utilized to extract a representation which encodes both RGB and depth information.

2.2.1 SMM with One Hidden Layer

The proposed SMM is composed of an input layer, one or more hidden layers and an output layer. For simplicity, we first formulate the model with one-hidden-layer which is then extended for multiple layers. Figure 2.4 shows the functioning of SMM model with single hidden layer. Let X_R and X_D be the unlabeled training data from two different (but registered) modalities,

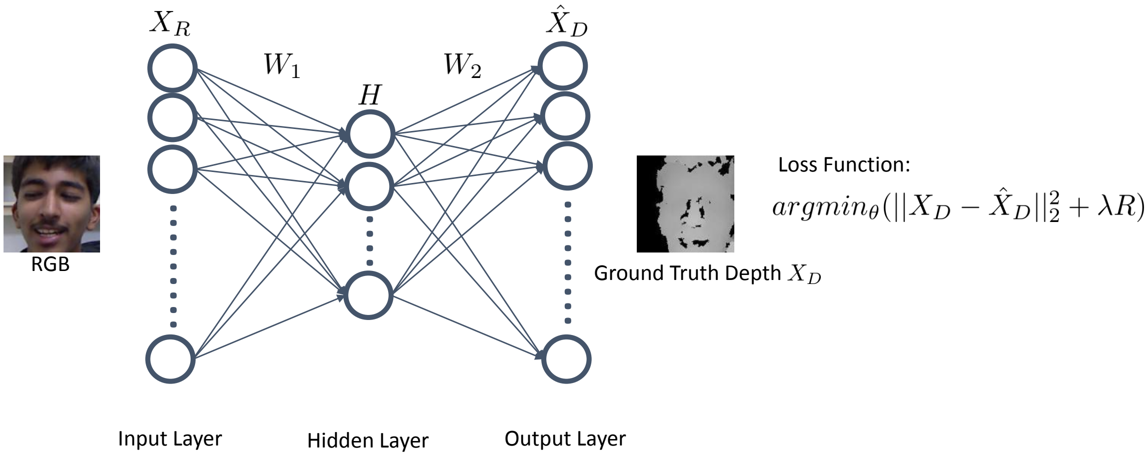


FIGURE 2.4: The Stacked Mapping Model (SMM) with one hidden layer

namely RGB and depth given, respectively. Let $X_R = \left\{x_R^{(1)}, x_R^{(2)}, x_R^{(3)}, ...x_R^{(n)}\right\}$ be the n training RGB image samples and $X_D = \left\{x_D^{(1)}, x_D^{(2)}, x_D^{(3)}, ...x_D^{(n)}\right\}$ be the n depth image samples (ground truth). Using these training samples we train the proposed stacked mapping model to learn a mapping $\mathcal{R} : X_R \longrightarrow X_D$ with a help of hidden layer representation H. Let $\{W_1, b_1\}$ be the weights and bias of network between input X_R and hidden layer and $\{W_2, b_2\}$ be the weights and bias between hidden layer and output $\hat{X}_D$. The hidden layer representation H can be expressed as,

$$H = \phi(W_1.X_R + b_1) \tag{2.1}$$

where, ϕ is the activation (sigmoid) function. In the output layer, depth map $\hat{X}_D$ is estimated as follows:

$$\begin{aligned} \hat{X}_D &= \phi(W_2.H + b_2) \\ &= \phi(W_2.\phi(W_1.X_R + b_1) + b_2) \end{aligned} \tag{2.2}$$

The optimization of entire network is governed by minimizing the difference between estimated depth map $\hat{X}_D$ and actual depth map X_D using following loss function,

$$argmin_\theta(||X_D - \hat{X}_D||_2^2 + \lambda R) \tag{2.3}$$

Expanding Equation 2.3 using Equation 2.2,

$$argmin_\theta(||X_D - \phi(W_2.\phi(W_1.X_R + b_1) + b_2)||_2^2 + \lambda R) \tag{2.4}$$

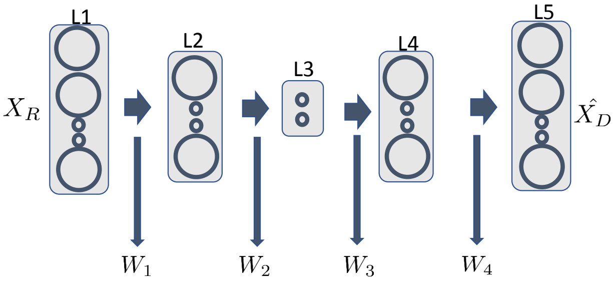


FIGURE 2.5: The Stacked Mapping Model (SMM) with multiple hidden layer.

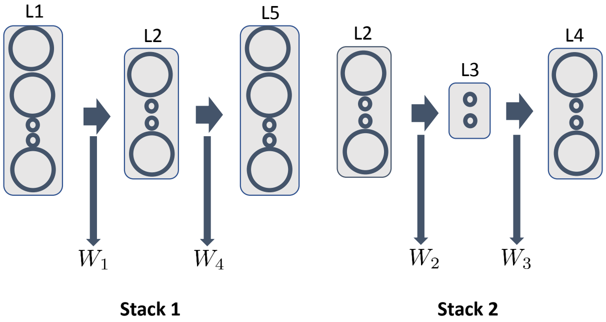


FIGURE 2.6: Training the Stacked Mapping Model (SMM): Training the SMM in (a) with 3 hidden layers (L2, L3 and L4) in a stacked fashion.

where, λ is the regularization parameter, R is the regularizer, and θ is the set of parameters $\{W_1, W_2, b_1, b_2\}$. Using the proposed mapping model, weights W_1 provides RGB (texture) rich information, whereas, weights W_2 provides features which are *depth* rich.

2.2.2 SMM with Multiple Hidden Layers

The proposed SMM model can be extended to multiple layers by adopting the stacked training approach which is similar to Stacked Denoising Autoencoders (SDAE) [117]. A multi-layer SMM with i hidden layers will contain k stacked networks where $i = 2k - 1$. We illustrate the training of an SMM with 3 hidden layers ($i = 3$) which contains two stacks ($k = 2$). As shown in Figure 2.6, a network with 3 hidden layers ($L2$, $L3$, $L4$) can be trained in two steps, training one stack in each step. Each stack can be viewed as a separate network with one input layer, one output layer and one hidden layer.

Training Stack 1: The hidden layer for the first stack is given by

$$H_1 = \phi(W_1.X_R + b_1) \tag{2.5}$$

wherein, ϕ is the sigmoid function. As shown in Figure 2.6, W_1 and b_1 are the weights and biases between the input layer ($L1$) and hidden layer ($L2$) (for the first stack), respectively. The output layer ($L5$ in Figure 2.6) for the first stack can be written as,

$$\begin{aligned}\hat{X}_D &= \phi(W_4.H_1 + b_4)\\ &= \phi(W_4.\phi(W_1.X_R + b_1) + b_4)\end{aligned} \tag{2.6}$$

such that

$$argmin_\theta(||X_D - \hat{X}_D||_2^2 + \lambda R_1) \tag{2.7}$$

expanding Equation 2.7 using Equation 2.6 we get,

$$argmin_\theta(||X_D - \phi(W_4.\phi(W_1.X_R + b_1) + b_4)||_2^2 + \lambda R_1) \tag{2.8}$$

Thus λ is the regularization parameter, R_1 is the regularizer, and θ is the set of parameters $\{W_1, W_4, b_1, b_4\}$.

The hidden layer for the second stack is given by

$$H_2 = \phi(W_2.H_1 + b_2) \tag{2.9}$$

where, ϕ is the sigmoid function. As shown in (Figure 2.6), W_2 and b_2 are the weights and biases between the layer L2 and layer L3 (for the second stack), respectively. The output layer (given by $L4$ in Figure 2.6) for the second stack is given by,

$$\begin{aligned}\hat{H}_1 &= \phi(W_3.H_2 + b_3)\\ &= \phi(W_3.\phi(W_2.H_1 + b_2) + b_3)\end{aligned} \tag{2.10}$$

such that

$$argmin_\theta(||H_1 - \hat{H}_1||_2^2 + \lambda R_2) \tag{2.11}$$

expanding Equation 2.11 using Equation 2.10 we get,

$$argmin_\theta(||H_1 - \phi(W_3.\phi(W_2.H_1 + b_2) + b_3)||_2^2 + \lambda R_2) \tag{2.12}$$

In the above equation θ represents the set of parameters given by $\{W_2, W_3, b_2, b_3\}$, λ represents the regularization parameter and R_2 is the regularizer. Thus, the second stack learns a reconstruction of H_1 by training the weight matrices $\{W_2, W_3, b_2, b_3\}$.

We have applied both $l_1 - norm$ and $l_2 - norm$ regularization on the weight matrix for learning the SMM stacks. The outer layers are learned in the first stack directly on RGB and depth data. Thus, in the first stack the weights W_1 and W_4 learn richer features. On the other hand the inner layers learned in the second stack are learned on the first layer representation (H_1) of the SMM. The $l_1 - norm$ regularization is applied during training of the second stack (training W_2 and W_3) which enforce sparsity ($R_1 = ||W||_1$) in these trained weight matrices. On the outer layers (the first stack) we apply l_2 regularization on the weights ($R_2 = ||W||_2$) which prevents overfitting (by penalizing large weights) by performing weight decay. This results in a more robust representation.

2.2.3 The Joint Hierarchical Feature Learning (JHFL) Model

As discussed in Section 2.2.2, for an input RGB image, following representations can be obtained using SMM:

- The representation obtained at the first hidden layer is given by H_1 (equation 2.5)
- In the second hidden layer, the representation obtained is given by H_2 (equation 2.9)
- The representation in the third hidden layer is given by $\hat{H}_1$ (equation 2.10)
- The representation (mapped depth from RGB) obtained at the output layer is given by $\hat{X}_D$ (equation 2.6)

The JHFL model serves the purpose of combining these features from the different layers of the SMM into a combined representation. The most straightforward way to do that is to combine (concatenate) all the four representations (H_1, H_2, $\hat{H}_1$ and $\hat{X}_D$) and learn a classifier on it. A disadvantage of this is that the dimensionality of the combined representation will be too large for the classifier. In this research, we combine the features in a hierarchical learning manner. As shown in Figure 2.7, we concatenate the representations obtained from first two ([H_1, H_2]= N_1) and the last two layers ([$\hat{H}_1$,$\hat{X}_D$]=N_2), separately, and learn autoencoders on each of them.

An autoencoder (AE) is learned on N_1 and another autoencoder is learned on N_2. The first autoencoder trained on N_1 has the following objective function.

$$argmin_\theta(||N_1 - \hat{N}_1||_2^2 + \lambda_1||\mathcal{W}_1||_2 + \lambda_2||\mathcal{W}_1||_*) \tag{2.13}$$

and the second autoencoder is trained on N_2 which has the objective function

$$argmin_{\theta}(||N_2 - \hat{N}_2||_2^2 + \lambda_1||\mathcal{W}_2||_2 + \lambda_2||\mathcal{W}_2||_*) \tag{2.14}$$

where, $||.||_2$ is the l_2 norm and $||.||_*$ is the trace norm on the weights $\mathcal{W}$. The regularization parameters λ_1 and λ_2 are used for the l_2 and the trace norm regularizations, respectively. Further, the trace norm or nuclear norm is defined as $||\cdot||_* = \sum_k \sigma_k$, where σ_k is the k^{th} singular value of the input matrix.

The training of the autoencoder is done by stochastic gradient descent where the weight update term is determined using $\frac{\partial \mathcal{L}}{\partial \mathcal{W}}$ where $\mathcal{L}$ is the loss function given by equations 2.13 and 2.14. The partial derivative of the first two terms of equations 2.13 and 2.14 is straightforward. The third term, the trace norm term, is not differentiable. Hence a sub-gradient is used in the gradient descent weight update process. The sub-gradient of $||\mathcal{W}||_*$ with respect to $\mathcal{W}$ is given by

$$\frac{\partial||\mathcal{W}||_*}{\partial \mathcal{W}} = \mathcal{W}.(\mathcal{W}^T\mathcal{W})^{-1/2} \tag{2.15}$$

A more stable formulation of the above subgradient has also been formulated by using the singular value decomposition (SVD) of W which can be expressed as

$$\mathcal{W} = U\Sigma V^T \tag{2.16}$$

where $W \in \mathbb{R}^{m\times n}$ with $m \geq n$. The sub-differential [118] of $||\mathcal{W}||_*$ can be expressed as

$$\frac{\partial||\mathcal{W}||_*}{\partial \mathcal{W}} = U_{1:m,1:r}V_{1:n,1:r}^T \tag{2.17}$$

where r is the rank of the matrix $\mathcal{W}$.

We have used a combination of tracenorm and l_2 regularization in equations 2.13 and 2.14. The tracenorm based regularization results in low ranked features. It has been attributed in several other research papers [119, 120] that this kind of regularization is well suited for data which has missing features. Since the inner layer (second stack of the SMM in Figure 2.6) is regularized using l_1 norm on the weights (from Section 2.2.1), most of these features will be very close to zero. In such cases, the tracenorm based regularizer is effective since such sparse features lie

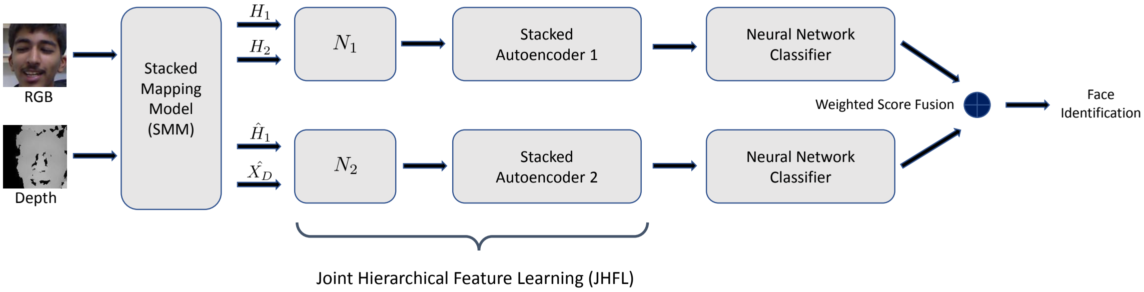


FIGURE 2.7: Proposed RGB-D face recognition approach using information fusion across feature and score levels of SMM and the JHFL models.

on a much lower dimensional manifold. This enables the JHFL model to learn a more compact representation.

2.2.4 Fusion and Classification

As formulated in Section 2.2.3, we utilize the representations H_1, H_2, $\hat{H}_1$ and $\hat{X}_D$ (representations from the proposed SMM) to further learn features (AE on N_1 and AE on N_2) in the JHFL framework. Features learned by the AE on N_1 and AE on N_2 are denoted as F_{N_1} and F_{N_2}, respectively. As shown in Figure 2.7, the features learned by the two autoencoders on the hidden layer representations of the SMM (denoted as F_{N_1} and F_{N_2}), are individually utilized to train neural network classifiers for identification. The scores of these classifiers are combined using weighted score fusion, i.e. $S_{N_1+N_2}$.

In addition to the features learned by the SMM, single-hidden-layer autoencoders (AE) are learned on RGB and depth data separately, known as the auxiliary RGB and depth models in our framework. These models learn features/representations directly from the RGB and depth data and complement the features learned by the JHFL framework. Individual classifiers are trained on the representations learned by the autoencoders on raw RGB and depth (F_{rgb} and F_d, respectively). The scores of these classifiers are combined (denoted as S_{rgb+d}) using weighted score fusion. Finally, the scores obtained by weighted fusion of the scores S_{rgb+d} and $S_{N_1+N_2}$ are utilized for face identification. In other words, the scores obtained from classifiers trained on the JHFL features and the scores from classifiers trained on the features obtained from the auxiliary models are fused at multiple levels using weighted score fusion.

2.3 Database and Experimental Protocol

This section presents the databases and experimental protocols used to show the efficacy of the proposed algorithm along with implementation details.

TABLE 2.1: Details of databases used and protocol for RGB-D face recognition.

Database	**Classes**	**Images**	**Image size**		**Protocol**	
			RGB	**Depth**	**Training**	**Testing**
Eurecom [99]	52	728	256×256	256×256	728	-
IIITD RGB-D [35]	106	23025	variable	variable	9210	13815
Kasparov [114]	108	62120	variable	variable	31060	31060

2.3.1 Databases

The results are demonstrated on different databases. The database characteristics are described below and the statistics are summarized in Table 2.1.

The **IIITD RGB-D database** [35] contains RGB and depth images of 106 subjects. The total number of images are 4605, and are captured in two sessions using two different sensors namely the Microsoft Kinect 1 and OpenNI SDK. The resolution of each image is 640×480. The number of images per subject varies between 11 and 254.

The **Kasparov database** [114] is collected in surveillance scenarios using the Kinect version 1 and 2 cameras. The subjects are captured at varying distances from 1 to 5 meters under semi-controlled illumination, uncontrolled occlusion, pose and expression. Detected and cropped face images and depth maps from the video frames of 108 subjects are provided in the dataset. For our experiments we have used the frames of the videos captured using the Kinect v1 sensor since it is relatively easier to detect and align faces in these videos. These RGB videos have a resolution of 1920×1080 and that of the depth videos is 512×424. The total number of images used in our experiments are $62, 120$.

The **EURECOM database** [99] is prepared with the Kinect version 1 sensor and comprises RGB and depth images of 52 subjects (38 males and 14 females). The cameras are placed at a fixed distance of 1 meter from the subject. The data is acquired in 9 different variations of occlusion, illumination, pose and expression. As shown in Table 2.1, this database is only used for training the models.

2.3.2 Experimental Protocol

The protocol and implementation details are divided into three different parts: (i) training the Stacked Mapping Model (SMM), (ii) training the Joint Hierarchical Feature Learning model (JHFL), and (iii) training classifiers for face identification.

Learning the SMM: The SMM model is trained using the complete EURECOM [99] dataset along with the training partition of the IIITD RGB-D [35] dataset. These two databases contain

728 and 9210 RGB and depth (training) images, respectively. For effective training, data is augmented with image flipping and intensity variations of the RGB images. After augmentation, the training dataset size increases to $728 \times 3 + 9210 \times 3 = 29814$. For performing experiments on the Kasparov [114] dataset, we use 50% of the data (pairs of RGB and depth images) for fine-tuning the SMM model. Since the SMM model is pre-trained using the training data of the IIITD RGB-D dataset, we do not fine-tune it for evaluation on this dataset.

Along with SMM model, additional autoencoders from RGB and depth data are also trained to learn two auxiliary models as illustrated in Section 2.2.4. These auxiliary models are pre-trained on the $29,814$ images (as explained above) and are fine-tuned on the training images of the respective datasets (IIITD RGB-D or Kasparov) on which evaluation is performed.

Learning JHFL Model: The $29,814$ pairs of RGB and depth images that are used in learning the SMM are also utilized to learn the next level of representations in the JHFL model. However, only RGB images are required to learn the representations in the JHFL model. For the Kasparov dataset [114], we use 50% of the data (RGB images) for fine-tuning the JHFL model. The fine tuning is done in an unsupervised manner.

Training Classifiers for Identification: Neural network classifiers are trained on the features obtained on the JHFL and the auxiliary models. In order to train these classifiers we utilize the training data (Table 2.1) for the IIITD RGB-D [35] and the Kasparov [114] databases, respectively. As illustrated in Section 2.2.4, the scores from these classifiers are then fused for face identification.

Testing Protocols: To report testing performance, the (unseen) test partition of each dataset is utilized. We need only RGB images during testing. The detailed training and testing split for each database is outlined in Table 2.1.

2.3.3 Implementation Details

As illustrated in Section 2.2.2, the Stacked Mapping Model has 3 hidden layers. To reduce the dimensionality of the input, the RGB images are converted to grayscale. For training the SMM, the RGB and the depth images (of the data that is used to train the SMM) are resized to 64×64. Thus, the dimensionality of the input and output layers of the SMM are 4096 each. The dimensionality of the hidden layers are 2048, 1024, and 2048, respectively. The sigmoid activation function is used for training the SMM. The value of the regularization coefficient (λ) for training stack 1 and stack 2 (Figure. 2.6) of the SMM are 0.26 and 0.12, respectively. The JHFL model comprises of two autoencoders (AE) which are trained on the representations of the SMM. Each of the AEs has 1 hidden layer, which is trained using the $tanh$ activation function.

TABLE 2.2: Comparison with SOTA methods for Rank 1 identification accuracies (%) of RGB-D face recognition

Method		**Database**	
		IIITD RGB-D [35]	**Kasparov [114]**
Goswami et al. [35]		98.74	52.38
Chowdhury et al. [51]		98.71	67.77
Multimodal Learning [121]	Multimodal 1	97.63	41.95
	Multimodal 2	94.59	39.32
Proposed		**99.61**	**72.98**

The two regularization coefficients (λ_1 and λ_2) used in the JHFL are 0.22 and 0.08, respectively. The neural network classifiers for face identification have 2 hidden layers and are trained using the stochastic gradient descent optimizer.

2.4 Experiments

The performance of the proposed approach is compared with existing state-of-the-art (SOTA) algorithms [35, 51, 102, 103], for RGB-D face identification, and deep multimodal learning [121]. On the IIITD RGB-D database [35], RISE [35] features have shown excellent results and for the Kasparov database [114], the results presented by Chowdhury et al. [51] are the best reported in the literature. In addition to this, we compare the performance of the proposed algorithm with Cui et al. [102] and Zhang et al. [103], which are recent Convolutional Neural Network (CNN) based RGB-D face recognition algorithms.

Since the proposed method can be considered as a multimodal fusion algorithm, the results are also compared with the deep multimodal learning technique by Ngiam et al. [121]. This algorithm has the scope of testing using data of only one modality. It uses an autoencoder to learn shared representation of two different modalities. There are two different ways in which this algorithm can be trained/tested. The first method (multimodal 1 in Table 2.2) is by using both the modalities during training, and testing using data of both modalities for each sample. The second method (multimodal 2 in Table 2.2) is using half of the training data to train with both the modalities and using the other half for training with only one modality. This is done so that during testing even if one of the modalities is missing, classification can be performed. Along with evaluating the performance of the proposed algorithm, the effectiveness of individual components of the SMM and JHFL are also evaluated through an elaborate ablation study.

2.4.1 Results and Analysis

According to the experiments performed, the results are divided into two sections. First, as shown in Table 2.2, we illustrate the comparative results of the proposed algorithm with other

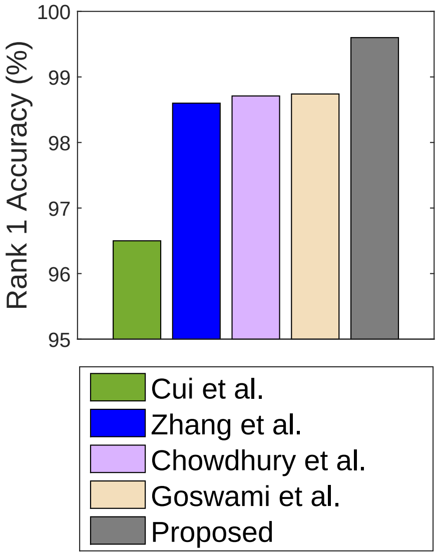


FIGURE 2.8: Comparing the proposed algorithm with other methods, including recent CNN based architectures on IIITD RGB-D [35] database.

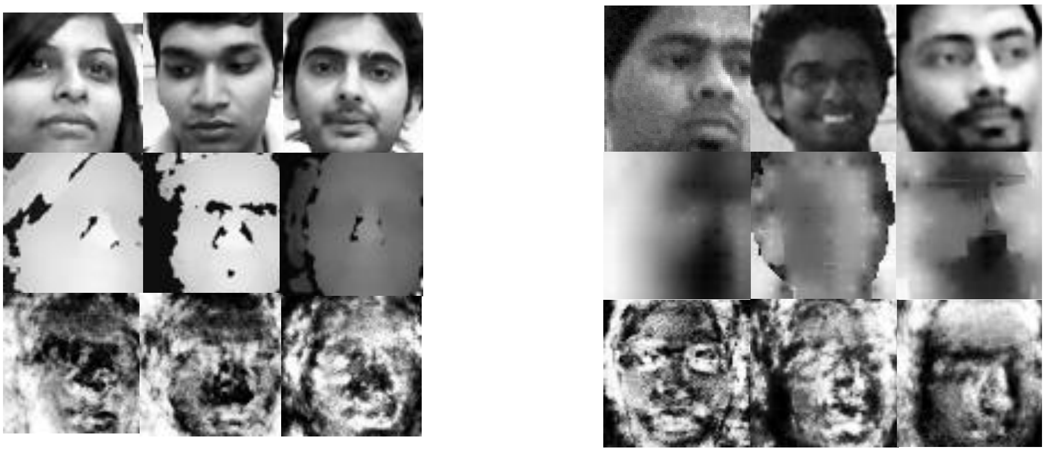
RGB-D face recognition and multimodal learning algorithms. Next, Table 2.3 shows the results obtained by using features from each component of the SMM and JHFL models and fusion of their components. The results are discussed below.

Comparison with existing RGB-D Face Recognition and Multimodal Learning Algorithms: We have compared the results of the proposed algorithm with Goswami et al. [35] and Chowdhury et al. [51], which have reported the best results on the IIITD RGB-D and Kasparov databases, respectively (Table 2.2). In addition to this, we have also compared the results of the proposed algorithm on the IIITD RGB-D database with recent CNN based methods [102, 103]. The algorithms used for comparison have been trained and tested using consistent protocols as discussed before. Figure 2.8 shows the comparative results of the proposed algorithms with SOTA RGB-D face recognition algorithms, where it can be observed that the proposed algorithm outperform the existing RGB-D face recognition algorithms including recent CNNs based methods (Cui et al. [102] and Zhang et al. [103]).

Table 2.2 also shows the relative performance of multimodal deep learning [121] and the proposed JHFL model. This comparison is performed to showcase the effectiveness of our approach with respect to Ngiam et al. [121], which was one of the first methods on multimodal deep learning. They proposed autoencoder based architectures for learning a shared representation from audio and video data. The proposed method is also presented as an algorithm for learning a shared representation from RGB and depth data. The proposed algorithm outperforms both the variants of multimodal learning (illustrated in Section 2.4) by a significant margin, especially on the Kasparov database.

TABLE 2.3: Rank 1 identification accuracies (%) of RGB-D face recognition on individual components of the SMM and JHFL models.

Model	**Features**	**Databases**	
		IIITD RGB-D	**Kasparov**
Raw Data	Raw RGB	95.03	53.69
	Raw Depth	89.20	18.90
Autoencoder	AE on RGB (F_{rgb})	98.46	64.04
	AE on Depth (F_d)	95.75	19.43
	Fusion of F_{rgb} and F_d (S_{rgb+d})	99.16	65.10
Stacked Mapping Model	Layer 1 (H_1)	98.62	66.32
	Layer 2 (H_2)	97.86	65.38
	Layer 3 ($\hat{H}_1$)	98.42	63.24
	Full Reconstruction ($\hat{X}_D$)	97.14	63.82
Joint Hierarchical Feature Learning	AE on N_1 (F_{N_1})	99.38	68.51
	AE on N_2 (F_{N_2})	97.02	53.36
	Fusion of F_{N_1} and F_{N_2} ($S_{N_1+N_2}$)	99.44	69.14
Proposed	Fusion of S_{rgb+d} and $S_{N_1+N_2}$	99.61	72.98



((A))

((B))

FIGURE 2.9: Visualizations of different representations used in the proposed method, (a) IIITD RGB-D database [35], (b) KaspAROV database [114], where row 1: RGB image (in grayscale), row 2: Captured depth image, row 3: Visualization of mapped depth (full reconstruction), shows the properties of the reconstructed depth.

Ablation Study on Individual Components of SMM and JHFL: Face identification is performed with features extracted from different layers of the SMM and JHFL, individually. The representations obtained from the SMM (H_1, H_2, $\hat{H}_1$ and $\hat{X}_D$ in Section 2.2.2) are used as features and a neural network classifier is trained for face identification. Similarly, we perform face identification on the representations obtained from the JHFL model as well. Furthermore, we perform fusion of JHFL features with features obtained from raw RGB and depth data, as outlined in section 2.2.4.

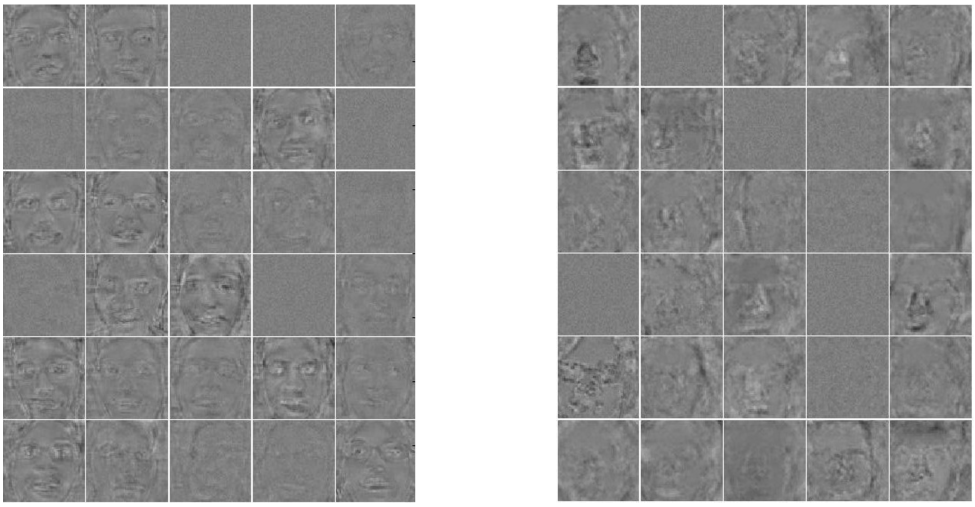
As shown in Table 2.3 it may be observed that the individual layers of the SMM have identification accuracies lower than that of the JHFL model. This demonstrates that the JHFL model is able to effectively combine the representations provided by the layers of the SMM. Features from the JHFL model are fused ($S_{N_1+N_2}$ in Table 2.3) and the rank 1 identification accuracies on the same are 99.44% and 69.14% for IIITD RGB-D and Kasparov databases, respectively. We also perform face identification using the representations obtained from the auxiliary models trained using autoencoders (AE) on RGB (F_{rgb} in Table 2.3) and depth (F_d in Table 2.3) data. Finally, the proposed algorithm which combines the features learned from both JHFL and auxiliary models (Fusion of S_{rgb+d} and $S_{N_1+N_2}$), yields best results on both the databases. It yields rank-1 identification accuracy of 99.61% on the IIITD RGB-D database, and 72.98% on the Kasparov database.

2.4.2 Analysis of SMM and JHFL

In this section we analyze and discuss the individual components of the proposed algorithm.

Analyzing Results of SMM: Figure 2.4 illustrates the architecture of the SMM, which is a neural network that maps RGB (in grayscale) data to depth. This model is trained on pairs of RGB and depth images captured from the same camera position. As illustrated in Section 2.2.2, SMM can be used to extract different representations of RGB (grayscale) data (Figure. 2.9) and classifiers trained on these layers are used for identification. From the results reported in Table 2.3, it can be inferred that:

- The accuracies obtained from the representation of the different layers of the SMM are higher than those obtained by training on raw RGB or depth.
- Analyzing the visualizations of the SMM weights W_1 and W_4 in Figure 2.10, shows that W_1 exhibits RGB like features and W_4 looks more like features learned from depth. This gives an intuition that different layers of the SMM learn different kinds of features. The weights of the first and second layers learn RGB-like features and the weights of the layers further away (near the output layer) learns depth like features. This gives an indication that the initial layers of the SMM learns representation of RGB information while the later layers learns the same from depth information. As the layers of the SMM seem to learn different kinds of features, this gives us a rationale of combining these features from different layers of the SMM.
- The results of face identification on the features of the individual layers of the SMM (Table 2.3) exhibit the comparative performance of its different layers. It is commonly believed (also evident from the accuracies from raw RGB and depth features in the same table) that RGB images contain much more discriminative information than depth images. It is observed that the representation obtained from the initial layers of the SMM yields higher accuracy than that from the last layer across both databases. Since the initial layers provide RGB like features, the classification accuracies achieved from them is



((A)) ((B))

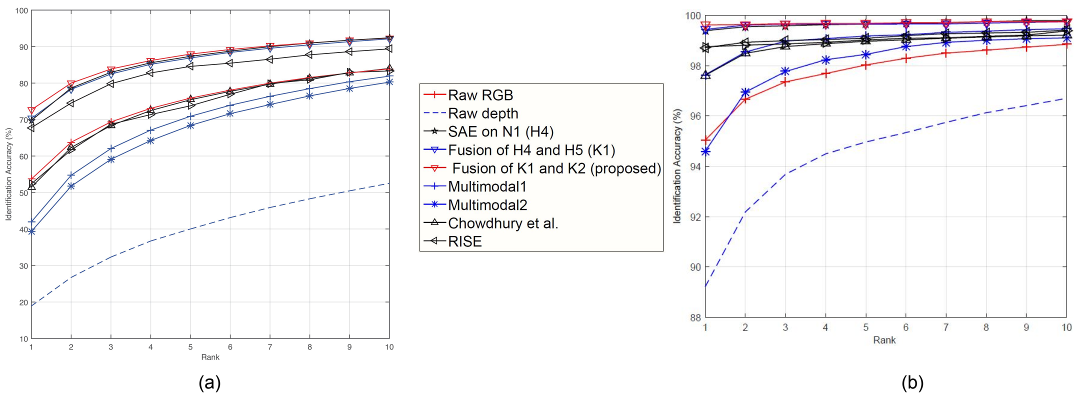
FIGURE 2.10: Visualization of (a) first layer weights (W_1) , (b) last layer weights (W_4) of the SMM. The first layer weights have visually detectable RGB like features and the last layer has depth like features

higher than the last layer which provides depth like features. In addition to the visualizations (Figure 2.10), this also confirms our previous observation that layer 1 features have more properties of RGB and that of the last layer have more properties of depth.

Analysis of Results obtained from the JHFL model: The SMM layers learn different kind of features and combining them should be useful for face identification. The rationale for using JHFL model was also given in the beginning of Section 2.2.3. To follow up on that, we present the analysis of results from the JHFL models as follows:

- From Table 2.3, it can be seen that combining the features from the SMM layers and score fusion with other auxiliary AE models on RGB and depth results in higher identification accuracies. This also shows that indeed the different layers of the SMM provide “additional” information which results in improved performance after fusion.

- The JHFL model illustrated in Figure 2.7 can also be used for databases that have both RGB and depth images. However, during testing only RGB images have been used from the RGB-D datasets. It also shows that the proposed algorithm is not heavily dependent on the presence of depth images and that RGB images alone is sufficient for the algorithm to perform face recognition in real-world settings.



(a)

(b)

FIGURE 2.11: CMC curves for face identification experiments on the (a) IIITD RGB-D Database and (b) Kasparov database.

2.5 Summary

This paper presents a novel representation learning algorithm which can learn features from RGB and depth data and can be tested using only RGB images. The proposed algorithm combines heterogeneous data (RGB and depth) and features from different modalities and learns a shared representation for recognition. Visualization and results of our algorithm show that the model learns shared representation from the two modalities. The proposed algorithm is evaluated on RGB-D face recognition in both controlled and surveillance scenarios. It is shown that the proposed algorithm produces state-of-the-art results on RGB-D databases without using any depth image during testing.

Chapter 3

SUPREAR-NET: Supervised Resolution Enhancement and Recognition Network

3.1 Introduction

Digital surveillance has become an integral part of law enforcement. For instance, in 2018, the British Security Industry Association estimated that there are around 6 million surveillance cameras [16] installed in the United Kingdom and an average person is caught in such cameras around 70 times a day [122]. Similarly, in the United States, the FBI uses an automated facial recognition system as a part of their Next Generation Identification (NGI) system [17], which processes the images captured by the surveillance cameras throughout the country. These surveillance cameras are installed in public places, government and private buildings, railway stations, bus terminals and so on. They are placed at a large standoff distance from the subjects and have high field of view. Since these cameras cover a large area, the effective resolution and quality of the captured face images are very low (Fig. 3.1). In terms of face recognition, the query images captured from the surveillance cameras are known as probes and the background database to which the probe images are matched is known as the gallery. In this scenario, the gallery images are usually of high resolution with proper illumination captured under constrained settings, collected from government documents such as passport, driver's license and so on. On the other hand, the probe images captured using the surveillance cameras are of much lower resolution and quality. As shown in Fig. 3.1, matching such low resolution probe images to a high resolution gallery, is known as cross-resolution face recognition.

Cross-Resolution face recognition can be majorly approached at three levels: (1) Image level, (2) Feature level and (3) Classifier level. At the image level, researchers have generally attempted

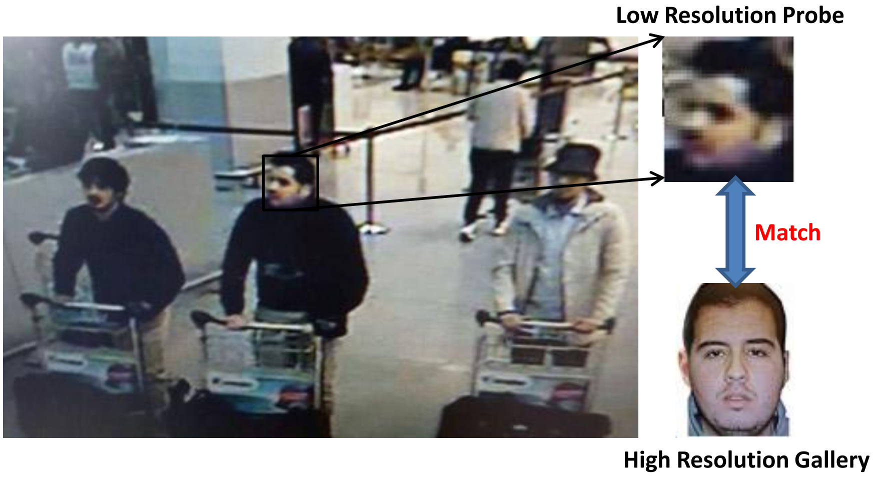


FIGURE 3.1: CCTV image of the suspects of the 2016 Brussels bombing attack, right frame: the modality gap of the low resolution probe and the high resolution gallery image of the suspect.

super-resolution [123–126] (Fig 3.12) and face hallucination based techniques [61, 127–129]. Such approaches primarily attempt to improve the image quality and resolution in order to improve the recognition performance. Feature level approaches optimize the parameters of a mapping function in order to project the features of low resolution probe images and high resolution gallery images to a common subspace, where the distance between images of the same subjects/classes are lower than the distances between images of different subjects/classes. Most of these algorithms have utilized discriminative learning [82, 83, 130–132] for the same. Classifier level approaches [23, 133, 134] have also been designed to match faces in cross-resolution scenarios. Such approaches learn the parameters of a classifier or create an optimum ensemble of classifiers so that cross-resolution matching can be performed efficiently.

Recent success of Generative Adversarial Networks (GANs) [63, 135] for noise to image generation have motivated researchers to use it for image to image translation and impressive results have been obtained [65, 70, 71, 136]. A vanilla image to image translation framework [65] using GANs consist of a *generator* which is provided with an image of the *source domain* (low resolution probe image in our case) to transform it into an image of the *target domain* (high resolution gallery). While these methods [65, 70] can produce photo-realistic and visually appealing images, they do not take into consideration that the class of the generated samples should remain consistent with the one that was given as input to the *generator*. This makes such unsupervised image to image translation methods unsuitable for improving the resolution and quality of low resolution probe images without distorting the discriminative information in those images captured by surveillance cameras for face recognition. Some recent research on semi-supervised GANs [137, 138] have shown that using the discriminator as a classifier (in addition to its usual task) yields better results than a vanilla GAN. However, such studies are not utilized for image to image translation, where, as shown by our approach, an additional classifier gives superior results compared to a multi-task discriminator, as far as the recognition of the generated images are concerned.

3.1.1 Research Contributions

This research focuses on generating high resolution images corresponding to the low resolution ones, which can then be utilized for matching with a high resolution gallery. Since this task requires two specific objectives namely, (1) Enhance the image resolution and quality, and (2) Retain the discriminative information present in the original low resolution image, we propose a combination of image level and feature level framework for cross resolution face recognition.

At the image level, a Supervised Resolution Enhancement-GAN (SURE-GAN) (Fig. 4.4) is proposed which contains a *classifier* in addition to the *generator* and the *discriminator*. The *generator* takes in an image of the *source domain* (low resolution probe) and outputs another image in the *target domain* (high resolution). The *classifier* provides feedback to the *generator* in the form of cross-entropy loss on the output image, so that in the process of translation, important discriminative information in the *source domain* image is not compromised. Thereafter, at the feature level we utilize a heterogeneous quadruplet loss (HQL) based algorithm to learn a discriminative model for cross-resolution face recognition. During testing, the *generator* of the SURE-GAN is used to synthesize high resolution face images from low resolution ones, which are fed into the discriminative model trained by the heterogeneous quadruplet loss (HQL) for face recognition. Since this component directly learns a discriminative embedding space for matching, it implicitly works at both the feature and the classifier level. Fig 4.4 illustrates the proposed SUPREAR-NET. The key contributions of this work are as follows:

1. A Supervised Resolution Enhancement-GAN (SURE-GAN) is proposed for image to image translation. The proposed SURE-GAN can transform images from the *source domain* (low resolution) into the *target domain* (high resolution) without corrupting the discriminative information in the *source domain* (low resolution) images.

2. A heterogeneous quadruplet loss metric is proposed which is utilized to train a discriminative model for cross resolution face recognition using the synthesized images from the trained *generator* of the SURE-GAN. Thus, we propose an end to end Supervised Resolution Enhancement and Recognition (SUPREAR-NET) network which works both at the image level and at the feature level.

3. The proposed SUPREAR-NET is utilized for generating high resolution images from low resolution ones followed by matching them with the gallery on three publicly available face databases, namely FaceSurv [4], CMU MultiPIE [139], and NIST MEDS [1]. The proposed SUPREAR-NET outperforms other existing supervised [140] and unsupervised [65, 141] GAN based image to image translation methods on all three databases.

3.2 Related work

This section is divided into two heads: (1) Cross-Resolution Face recognition and (2) GAN based image to image translation.

3.2.1 Cross-Resolution Face Recognition

As illustrated in Section 3.1, cross-resolution face recognition may be approached at different levels. Superresolution based approaches can be utilized for an image level transformation of the low resolution probe images for cross-resolution matching. Zhao *et al.* [123] proposed a superresolution algorithm using wavelet-domain hidden Markov tree. Niu and Nguyen [124] proposed an approach for the same using support vector regression. Yang *et al.* [128] proposed a face hallucination algorithm using sparse coding via Non-negative Matrix Factorization. Jiang *et al.* [129] utilized a locality-constrained representation based approach for translating low resolution face images into a higher resolution image. Kang *et al.* [12] and Maeng *et al.* [142] proposed a manifold learning based image denoising algorithm to improve the quality of the low resolution images for cross-resolution face recognition. Recently, Yang *et al.* [143] used an attribute-embedded upsampling network which consists of a discriminative network and an autoencoder for face super-resolution. Singh *et al.* [144] proposed a Synthesis via Hierarchical Sparse Representation (SHSR) algorithm for generating high resolution face images from low resolution ones for recognition. Several algorithms have also been proposed for face recognition in surveillance scenarios [114, 145–147].

Feature level approaches have been one of the most popular algorithms for cross-resolution face recognition. Lee *et al.* [148] proposed a support vector data description method for low resolution face recognition. Li *et al.* [130] formulated an algorithm for projecting low and high resolution face images into a common subspace using a coupled locality preserving mapping based approach for matching cross-resolution face images. Biswas *et al.* proposed algorithms [82, 83] for recognition of low resolution faces based on discriminative learning via a multidimensional scaling based approach. Zou and Yuen [131] proposed a relationship based superresolution algorithm for very low resolution face recognition problem, for face images which are smaller than 16×16. Mudunuri and Biswas [84] utilized a coupled discriminative dictionary based algorithm for the same. Klare *et al.* [134] proposed a discriminant analysis based classification approach using kernel prototype similarities for heterogeneous face recognition. Ghosh *et al.* [149] used fusion of features from Restricted-Boltzmann Machines and DSIFT for face identification from low resolution probe images.

Some classifier level approaches have also been proposed for cross-resolution matching of face images. Bhatt *et al.* [23, 133] formulated a co-transfer based learning approach where the knowledge learnt from high resolution images is transfered for matching low resolution images to high resolution gallery. The algorithm combined co-learning and transfer learning to update the classifier's decision boundary for cross-resolution face matching. Lu *et al.* [150] proposed a coupled ResNet model, consisting of a trunk network and two branch networks for low resolution face recognition. A discriminative multidimensional scaling based algorithm was proposed by Yang *et al.* [151] for the same.

3.2.2 Image to Image Translation using GANs

There have been significant advances in the investigation of Generative Adversarial Networks and are applied in various applications such as image generation [66–69], image to image translation [65, 70, 71], social network analysis [72, 73] image superresolution [74], 3D shape modeling [75], text to image synthesis [76], image style transfer [77, 78], and texture synthesis [79]. In this paper, the focus is particularly on image to image translation, which is one of the most popular applications of GANs. Most image to image translation methods using GANs [65, 70, 71, 80] can transform an image from the source domain to the target domain effectively. Isola *et al.* [65] proposed one of the first image to image translation based methods using GANs. They used a generator which received a random noise and the source domain image as input. The discriminator was given a pair of images (real and fake) as input and it had to discriminate between a real-real and a real-fake pair. This was extended in [70] by introducing a cycle consistency loss for unpaired image to image translation. Yi *et al.* [71] used a dual-learning [81] based formulation to train a GAN model for unpaired image to image translation. Recently, a fusion based approach [152] was proposed for combining infrared and visible spectrum images using a GAN based framework. These methods only focus on the domain transfer, and do not take into account that important discriminative information may be distorted during the process of domain translation.

Some supervised and semi-supervised GANs have been proposed. Salimans *et al.* [137] introduced a semi-supervised GAN based algorithm, so that the class of the translated images can be kept consistent with the source domain images. They used a multitask based discriminator model, which also computed a cross-entropy based supervised loss in addition to the adversarial loss. This loss was back-propagated to the generator during training. Recently, Zhang *et al.* [153] have proposed a similar approach. Zhao *et al.* [140] proposed a supervised GAN formulation for translating face images across different pose variations. A BEGAN [67] style discriminator (using an autoencoder) was used in this model which was multi-tasked to produce both the supervised and the adversarial loss. The supervised loss was computed from an intermediate representation of the discriminator, the gradient of which was then back-propagated to

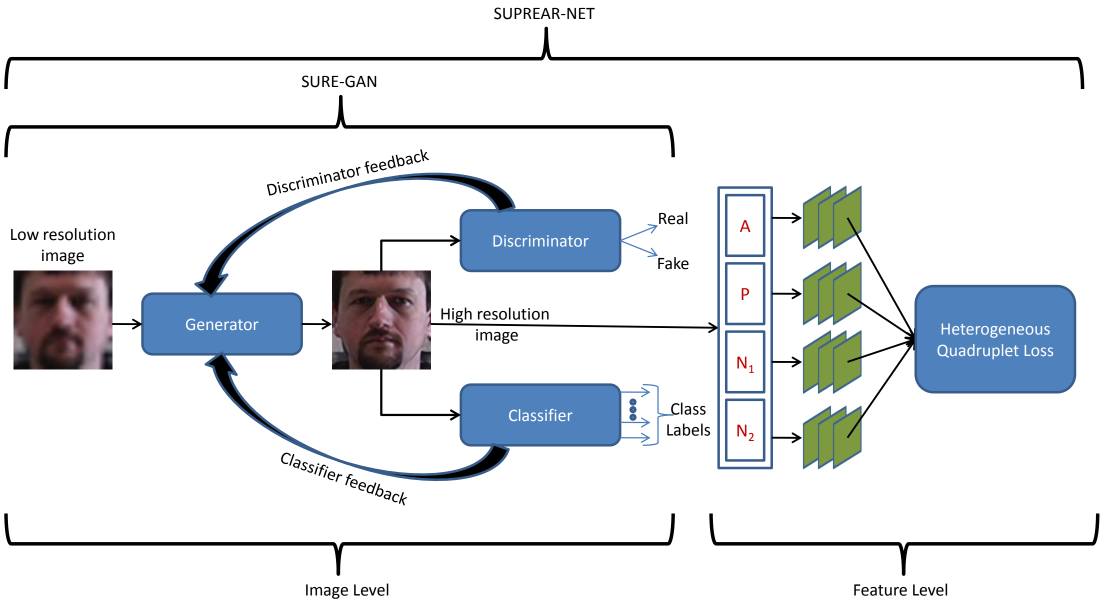


FIGURE 3.2: The proposed SUPREAR-NET for cross resolution face recognition using the SURE-GAN followed by a heterogeneous quadruplet loss based discriminative model. The SURE-GAN contains a classifier in addition to the generator and the discriminator, which prevents the generator from corrupting useful discriminative information while performing the image to image translation (resolution enhancement) task. The transformed probe images from the trained generator is used to match with the high resolution gallery images using the discriminative model trained by the heterogeneous quadruplet loss. (best viewed in color)

the generator during training. Park *et al.* [154] presented a conditional image synthesis framework which takes in a semantic segmentation mask and produces a corresponding real world scene as output. Mo *et al.* [155] developed an instance aware image to image translation GAN which takes in a set of instance attributes in addition to the source image. They proposed a context preserving loss which forces the network to focus on target instances in addition to the general scene that is to be transformed. On similar lines, Tang *et al.* [156] proposed an unsupervised image to image translation GAN framework using an attention guided training mechanism which disentangles the semantic object and the unwanted part of the image by using an attention mask and a content mask during training.

3.3 Proposed Algorithm

In this section, we explain the proposed SUPREAR-NET for cross-resolution face recognition using supervised image to image translation with the SURE-GAN. The first step is to learn the SURE-GAN model for supervised image to image translation, followed by learning a discriminative Convolutional Network (CNN) model for cross-resolution face recognition. We begin by briefly introducing the basic formulation of GANs.

3.3.1 Background

A GAN [63] consists of a generator G and a discriminator D, where θ_g and θ_d are the parameters of the generator and discriminator respectively. The generator G produces a synthetic image from a noise distribution p_z, and the discriminator D is trained to distinguish between a synthetic image $G(z)$ and a real image x sampled from the distribution of real images p_{data}. It is formulated as follows:

$$min_{\theta_g} max_{\theta_d} [\mathbb{E}_{x \in p_{data}} \log D_{\theta_d}(x) + \mathbb{E}_{z \in p_z} \log(1 - D_{\theta_d}(G_{\theta_g}(z)))] \tag{3.1}$$

The G and D models are trained alternatively by updating the parameters. The parameter update of the generator is given by

$$\nabla_{\theta_g} \frac{1}{m} \sum_{i=1}^{m} \left[\log \left(1 - D_{\theta_d} \left(G_{\theta_g} \left(z^{(i)} \right) \right) \right) \right] \tag{3.2}$$

and that of the discriminator is given by

$$\nabla_{\theta_d} \frac{1}{m} \sum_{i=1}^{m} \left[\log D_{\theta_d} \left(x^{(i)} \right) + \log \left(1 - D_{\theta_d} \left(G_{\theta_g} \left(z^{(i)} \right) \right) \right) \right] \tag{3.3}$$

where, *m* is the size of the minibatch used. The initial model of GANs [63] allowed the generation of synthetic images from noise. The applicability of such models is in data augmentation which has resulted in generation of large amount of training data for better and robust training of classifiers.

Conditional GAN [64] is proposed as an extension of the original GAN model. It allows both the generator and the discriminator to be conditioned on an extra information y which can be either the class label of the sample or any extra auxiliary information. Such models can be formulated as:

$$min_{\theta_g} max_{\theta_d} [\mathbb{E}_{x \in p_{data}} \log D_{\theta_d}(x|y) + \mathbb{E}_{z \in p_z} \log(1 - D_{\theta_d}(G_{\theta_g}(z|y)))] \tag{3.4}$$

The conditional GAN model is formulated for image to image translation by Isola *et al.* [65]. It can be modeled as a mapping function $G : \{\mathbf{p}, \mathbf{z}\} \rightarrow \mathbf{q}$ where $\mathbf{p}$ and $\mathbf{q}$ are sets of images of the source and target modalities respectively and $\mathbf{z}$ is a noise vector. It can be formulated as follows:

$$\begin{aligned} \varphi_{GAN}(G, D) = [\mathbb{E}_{\mathbf{p},\mathbf{k} \in p_{data}} \log D_{\theta_d}(\mathbf{p}, \mathbf{k}) + \\ \mathbb{E}_{\mathbf{z} \in p_z; \mathbf{p},\mathbf{x} \in p_{data}} \log(1 - D_{\theta_d}(\mathbf{x}, G_{\theta_g}(\mathbf{p}, \mathbf{z})))] \end{aligned} \tag{3.5}$$

The generator can also be conditioned on an extra constraint $\varphi_{l_1}(G)$ which is the l_1 difference of the input images $\mathbf{p}$ and output images $\mathbf{q}$. The final objective function of the above model is

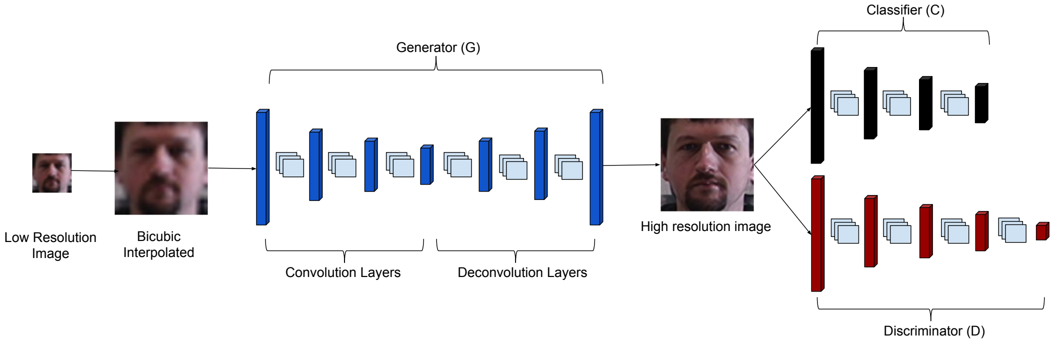


FIGURE 3.3: Pipeline of the proposed SURE-GAN. The discriminator model contains convolution and deconvolution layers. The vertical bars denote the intermediate output of the layers and the light blue squares (in between the vertical bars) denote the filters. The classifier takes in the synthetic image produced by the generator and backpropagates the loss back to the generator. The discriminator is a conventional one, contains convolutional layers which helps the generator to produce realistic images. (best viewed in color)

as follows:

$$G^{*} = min_{\theta_g} max_{\theta_d} \left[\varphi_{GAN}\left(G, D\right) + \varphi_{l_1}\left(G\right)\right] \tag{3.6}$$

The l_1 constraint makes the output images (from the generator) consistent with the structural properties of the target image. The drawback of such a model is that it is trained only on the goal of transforming the image from one modality into another. As a byproduct of such a transformation, discriminative information in the images (which are important for classifying the image) might get distorted. Thus, a supervised version of the above model is proposed and is utilized to train a GAN for translating face images from the source to the target modality. It ensures that the class of the generated image is consistent with that of the input image of the source domain.

3.3.2 Proposed Supervised Resolution Enhancement GAN (SURE-GAN)

As shown in Fig 4.4, we present a Supervised Resolution Enhancement GAN (SURE-GAN) for translating a low resolution image to high resolution. To generate high quality images preserving the identity information, the proposed SUPREAR-NET architecture comprises a two step framework which not only operates at the image level but also at the feature level. The proposed SURE-GAN is illustrated as follows.

Let $\mathbf{p} = \{(p_1, c_1), (p_2, c_2), \dots (p_n, c_n)\}$ and $\mathbf{q} = \{(q_1, j_1), (q_2, j_2), \dots (q_m, j_m)\}$ be the labeled data (labeled images) from the source (low resolution) and target (high resolution) domains, respectively, where c_i and j_i are the class labels of images p_i and q_i respectively. The model has a generator G and a discriminator D with parameters θ_g and θ_d respectively. A classifier network C is introduced with parameters θ_c. The generator G takes in as input a low resolution image p_i and a random noise z_i. The output of the generator is a high resolution synthesized image, denoted by $G(p_i, z_i)$. Input to the classifier C is the high resolution synthetic image $G(p_i, z_i)$ produced by the generator. The classifier is also given the real images from the target modality $\mathbf{q}$. The objective function of the classifier can be defined as:

$$C(G) = \mathbb{E}\left[\log\left(P\left(C_{\theta_c}\left(G_{\theta_g}(\mathbf{p},\mathbf{z})\right)\right) = \mathbf{c}\right)\right] + \mathbb{E}\left[\log\left(P\left(C_{\theta_c}(\mathbf{q})\right) = \mathbf{j}\right)\right] \tag{3.7}$$

where $\mathbf{p}$, $\mathbf{q}$ and $\mathbf{z}$ are sets of low and high resolution images and the corresponding noise vectors in a minibatch. The proposed supervised GAN for resolution enhancement is formulated as:

$$\varphi_{GAN}^{S}(G, D, C) = \varphi_{GAN}(G, D) + C(G) \tag{3.8}$$

An l_1 constraint on the images generated by the generator (with respect to the high resolution images) is applied. This ensures that the generator learns to produce images which resemble the target data distribution. It may be expressed as:

$$\varphi_{l_1}(G) = \mathbb{E}_{\mathbf{p},\mathbf{q}\in p_{data};\mathbf{z}\in p_z}\left\|\mathbf{q} - G(\mathbf{p},\mathbf{z})\right\|_1 \tag{3.9}$$

Combining Equations 3.8 and 3.9, the final objective of the proposed model is:

$$G_s^* = min_{\theta_g} max_{\theta_d} min_{\theta_c}\left[\varphi_{GAN}^{S}(G, D, C) + \varphi_{l_1}(G)\right] \tag{3.10}$$

The above formulation thus has three models namely G, D and C which are trained as shown in Fig 3.3. The stochastic gradient descent is used for training all three models. The parameter update for the generator is:

$$\begin{aligned}\nabla_{\theta_g}\frac{1}{m}\sum_{i=1}^{m}[[\log(1 - D_{\theta_d}(x^{(i)}, G_{\theta_g}(p^{(i)}, z^{(i)})))]+ \\ [\log P(C_{\theta_c}(G_{\theta_g}(p^{(i)}, z^{(i)}))) = c^{(i)}] + [||q^{(i)} - G(p^{(i)}, z^{(i)})||_1]]\end{aligned} \tag{3.11}$$

and that of the discriminator is given by:

$$\nabla_{\theta_d}\frac{1}{m}\sum_{i=1}^{m}[\log D_{\theta_d}(p^{(i)}, k^{(i)}) + \log(1 - D_{\theta_d}(x^{(i)}, G_{\theta_g}(p^{(i)}, z^{(i)})))] \tag{3.12}$$

and that of the classifier is:

$$\begin{aligned}\nabla_{\theta_c}\frac{1}{m}\sum_{i=1}^{m}\left[\log\left(P\left(C_{\theta_c}\left(G_{\theta_g}\left(p^{(i)}, z^{(i)}\right)\right)\right) = c^{(i)}\right)\right] + \\ \mathbb{E}\left[\log\left(P\left(C_{\theta_c}\left(q^{(i)}\right)\right) = j^{(i)}\right)\right]\end{aligned} \tag{3.13}$$

To summarize, the proposed SURE-GAN is a supervised GAN model that uses a classifier in addition to the generator and the discriminator. The loss of the classifier on the images produced by the generator is back-propagated to the generator and to the classifier. In addition to this, the loss of the discriminator is also used to update the generator weights. This ensures that the generator produces realistic looking images of high resolution without corrupting useful discriminative information present in the low resolution images. Algorithm 1 summarizes the training of the proposed method.

Algorithm 1: Training for Supervised Image to Image Translation GAN

```
Input: p (low resolution training data), q (high resolution training data), $G_{\theta_g}$ (generator
       network), $D_{\theta_d}$ (discriminator network), $C_{\theta_c}$ (classifier)
Output: $G_{\theta_g}$ ( trained generator network),$D_{\theta_d}$ (trained discriminator network),$C_{\theta_c}$ (trained
        classifier network)
Parameters: $n_1$ (epochs for pretraining ), $n_2$ (epochs for end to end training ),$\theta_g$
  (parameters of G), $\theta_d$ (parameters of D), $\theta_c$ (parameters of C), $\lambda_1$ (coefficient for
  adversarial loss), $\lambda_2$ (coefficient for classifier's loss), $\lambda_3$ (coefficient for $L_1$ loss), $m$ (size
  of minibatch), $k$ (number of batches)
Pre-Training of the Classifier:
for Epoch=1 to $n_1$ do
    for every minibatch, $M_b$=0 to $k-1$ do
        Forward pass $q\{M_b.m....M_b.m+m\}$ through $C_{\theta_c}$
        Calculate Gradient and Update weights:
        $\triangle\theta_c = \nabla_{\theta_c}\frac{1}{m}\sum_{i=1}^{m}\left[\log\left(P\left(C_{\theta_c}\left(q^{(i)}\right)\right)=j^{(i)}\right)\right]$
        Update weights of $C_{\theta_c}$ using $\triangle\theta_c$
    end
end
End to end training of the Generator, Discriminator and the Classifier:
for Epoch=1 to $n_2$ do
    for every minibatch, $M_b$=0 to $k-1$ do
        Forward pass $p\{M_b.m....M_b.m+m\}$ through $C_{\theta_c}$
        Calculate Gradient and Update weights:
        $L_{adv} = [\log(1 - D_{\theta_d}(x^{(i)}, G_{\theta_g}(p^{(i)}, z^{(i)})))]$
        $L_C = \left[\log P\left(C_{\theta_c}\left(G_{\theta_g}\left(p^{(i)}, z^{(i)}\right)\right)\right)=c^{(i)}\right]$
        $L_C^s = \mathbb{E}\left[\log\left(P\left(C_{\theta_c}\left(q^{(i)}\right)\right)=j^{(i)}\right)\right]$
        $L_1(G) = \left[\left\|q^{(i)} - G\left(p^{(i)}, z^{(i)}\right)\right\|_1\right]$
        $\triangle\theta_g = \nabla_{\theta_g}\frac{1}{m}\sum_{i=1}^{m}[\lambda_1.L_{adv} + \lambda_2.L_C + \lambda_3.L_1(G)]$
        $\triangle\theta_d = \nabla_{\theta_d}\frac{1}{m}\sum_{i=1}^{m}\left[\log D_{\theta_d}\left(p^{(i)}, k^{(i)}\right) + L_{adv}\right]$
        $\triangle\theta_c = \nabla_{\theta_c}\frac{1}{m}\sum_{i=1}^{m}[L_C + L_C^s]$
        Update weights of $G_{\theta_g}$ using $\triangle\theta_g$
        Update weights of $D_{\theta_d}$ using $\triangle\theta_d$
        Update weights of $C_{\theta_c}$ using $\triangle\theta_c$
    end
end
```

3.3.3 SUPREAR-NET for Cross-Resolution Face Recognition

The SURE-GAN presents an image level solution for cross-resolution face recognition. The trained generator of the SURE-GAN can be utilized to translate low resolution probe images to high resolution ones. These translated images can be used to match with high resolution gallery images. In order to learn a discriminative embedding space where both the high resolution gallery images and low resolution probe images can be projected, a feature level solution is proposed. Inspired from the quadruplet loss [92], a heterogeneous quadruplet loss (HQL) is proposed to train a discriminative model for cross-resolution face recognition.

In order to have an effective and discriminative model for matching, Chen et *al.* [92] proposed the quadruplet loss method of training a deep-CNN for face matching. In this method, each training sample is a quadruplet $(\vec{X}_a, \vec{X'}_a, \vec{X}_n, \vec{X}_k)$, where $\vec{X}_a$ is a set of images known as the anchor, $\vec{X'}_a$ is the positive set and $\vec{X}_n$ and $\vec{X}_k$ are the negative sets of images. The anchor $\vec{X}_a$ are images of a particular subject a, $\vec{X'}_a$ is another set of images of the same subject a, $\vec{X}_n$ are images of any other subject n and $\vec{X}_k$ belongs to a different subject k where $a \neq n \neq k$. The loss function is designed to ensure that the distance of the anchor images $\vec{X}_a$ from the positive samples $\vec{X'}_a$ is lesser than the distance of the anchor from the sets of negative samples $\vec{X}_n$ and $\vec{X}_k$.

$$
\begin{aligned}
L = \left[\left\| f(\vec{X}_a) - f(\vec{X}_p) \right\|_2^2 - \left\| f(\vec{X}_a) - f(\vec{X}_n) \right\|_2^2 + \alpha_1 \right]_+ \\
+ \left[\left\| f(\vec{X}_a) - f(\vec{X}_p) \right\|_2^2 - \left\| f(\vec{X}_n) - f(\vec{X}_k) \right\|_2^2 + \alpha_2 \right]_+ \quad (3.14) \\
\forall (\vec{X}_a, \vec{X}_p, \vec{X}_n, \vec{X}_k) \in \P
\end{aligned}
$$

where, $[z]_+ = max(0, z)$, f is the discriminative model being trained, $\{\alpha_1, \alpha_2\}$ are the margin parameters and ¶ is the set of quadruplets.

Let $\vec{p}$ and $\vec{q}$ denote sets of images from the target and source domains respectively where, $\vec{p}_a$ and $\vec{q}_a$ be the images of a subject a from the source and target domains respectively. In order to train a discriminative model for matching images from the source and target domains, we generate 4-tuples (known as quadruplets) from the training set. These 4-tuples are used for training, comprising of high resolution anchor images $\vec{p}_a$, low resolution positive images $\vec{q}_a$ of the same subject/class "a", and two sets of low resolution negative images $\vec{q}_b$ and $\vec{q}_c$ of two different subjects "b" and "c", where $a \neq b \neq c$. The loss function for the heterogeneous quadruplet loss is as follows:

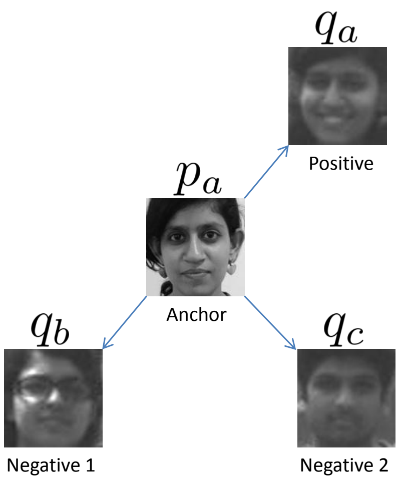


FIGURE 3.4: Illustration of the heterogeneous quadruplet loss. Images are taken from the FaceSurv database [4].

$$\begin{aligned} L = \Big[\|f(\vec{p}_a) - f(\vec{q}_a)\|_2^2 - \|f(\vec{p}_a) - f(\vec{q}_b)\|_2^2 + \alpha_1 \Big]_+ \\ + \Big[\|f(\vec{p}_a) - f(\vec{q}_a)\|_2^2 - \|f(\vec{p}_a) - f(\vec{q}_c)\|_2^2 + \alpha_2 \Big]_+ \\ \forall (\vec{p}_a, \vec{q}_a, \vec{q}_b, \vec{q}_c) \in \P \end{aligned} \tag{3.15}$$

where ¶ is the set of quadruplets that can be prepared from $\vec{p}$ and $\vec{q}$. The first part of the loss function $[||f(\vec{p}_a) - f(\vec{q}_a)||_2^2 - ||f(\vec{p}_a) - f(\vec{q}_b)||_2^2]$ minimizes the distance of the high resolution anchor images $\vec{p}_a$ and low resolution positive images $\vec{q}_a$ of the same subject "a", and maximizes the distance between the same set of anchor images and the low resolution negative images $\vec{q}_b$ of another subject "b". The second part of the loss function $[||f(\vec{p}_a)-f(\vec{q}_a)||_2^2-||f(\vec{p}_a)-f(\vec{q}_c)||_2^2]$ performs the same by introducing a second set of negative images $\vec{q}_c$. This is done in order to improve the generalizable of the model f by increasing the inter-class distance with a different negative class "c". The loss function is illustrated in Fig. 3.4.

Thus, the model f trained using this heterogeneous quadruplet loss is used to match synthesized probe images (using the trained generator of the SURE-GAN) to high resolution gallery for cross-resolution face recognition. This depicts a solution for cross-resolution face recognition which works both at the image (SURE-GAN) and feature level (HQL model).

3.4 Databases and Protocol

This section outlines the datasets, experimental protocol and implementation details for evaluating the performance of the proposed SUPREAR-NET.

TABLE 3.1: Details of protocol used for the experiments.

Database	Probe Resolution	Gallery Resolution	Number of Images		
			Training	Testing	
				Gallery	Probe
FaceSurv [4]	128x128	128 x 128	108	576	3066
	96 × 96				2261
	64 × 64				3209
CMU MultiPIE [139]	48 × 48, 32 × 32, 24 × 24	128 × 128	17121	238	32889
NIST MEDS [1]	48 × 48, 32 × 32, 24 × 24	128 × 128	247	271	787

3.4.1 Datasets

The proposed framework is evaluated on three real world face datasets namely FaceSurv [4], CMU MultiPIE [1, 139], and NIST MEDS []. The details of the databases and their corresponding protocols used for the experiments are as follows:

FaceSurv: This database [4] contains videos of 240 subjects in both day and night captured under surveillance scenarios, where each video contains at most three and at least one subject. There are 368 daytime and 365 nighttime videos. For our experiments we have used the videos that are captured during the day. In all the videos, the standoff distance of the subjects from the camera vary from 1m to 10m. Every subject has three high resolution gallery images which are captured under controlled settings in daytime from a distance of less than 1m.

The images in the database are divided into train and test sets (subject-disjoint) according to a predefined protocol of the database (Table 3.1). The training set contains both low resolution and high resolution images. For preparing the probe set, we divide each probe video into 3 partitions, namely when the subject is at a distance of 1-4m, 4-7m and 7-10m. Faces are detected in each frame by the Viola Jones face detector [157]. The images in each partition are kept at a resolution of 128 × 128, 96 × 96, and 64 × 64 respectively. The gallery images are in high resolution (128 × 128).

CMU MultiPIE: The CMU MultiPIE database [139] contains 750,000 images of 337 subjects captured with variations of pose, illumination, and expression. A subset of the database is used which contains 50,247 images of all the 337 subjects with variations in illumination and expression only.

For training, images pertaining to 100 subjects are used and the images of the rest of the subjects are utilized for testing. In the test set, one high resolution image of each subject is kept as gallery. The probes are divided into 3 resolution variants namely 24 × 24, 32 × 32, and 48 ×

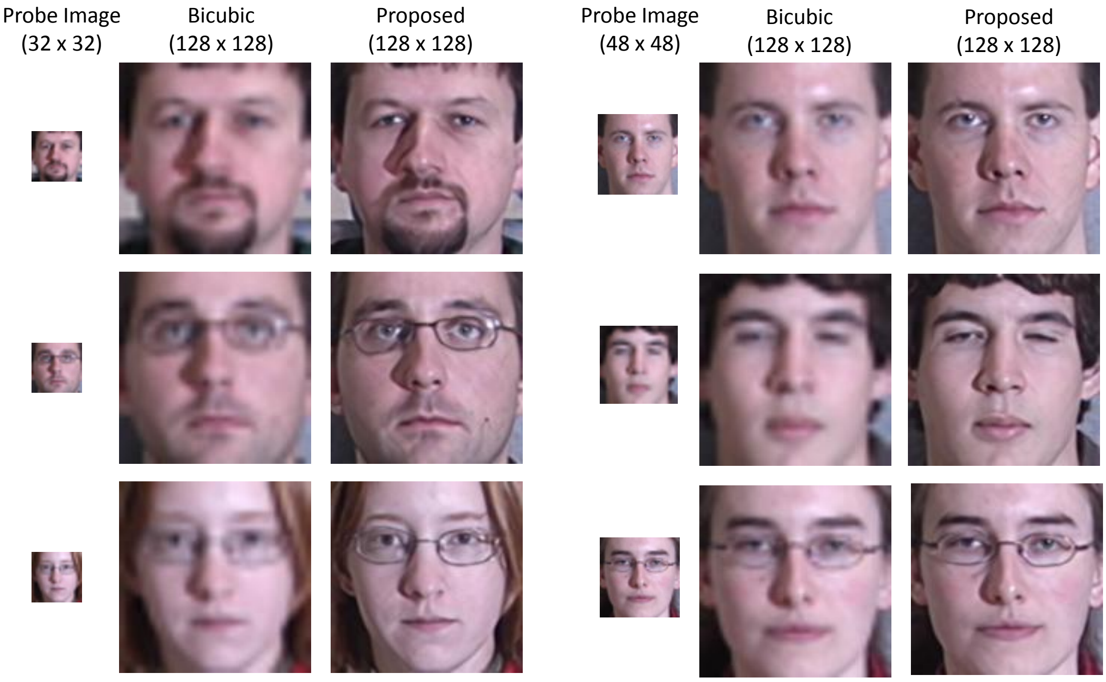


FIGURE 3.5: Illustration of the enhancement of resolution and quality using the proposed SURE-GAN on the images of CMU MultiPIE [139] database.

48 respectively, which are created by sub-sampling the high resolution images present in the database. The gallery images are in high resolution (128 × 128).

NIST MEDS: The Multiple Encounter Dataset (MEDS) [1] contains images of individuals with prior multiple encounters. There are 1306 images of 518 subjects in this database, out of which 271 subjects have more than 1 image.

The probes are divided into 3 resolution variants namely 24×24, 32×32, and 48×48 respectively, which are created by sub-sampling the high resolution images present in the database. The gallery images are in high resolution (128×128).

3.4.2 Implementation Details

After training the SURE-GAN and the discriminative model (using the Heterogeneous Quadruplet loss (HQL)) on the training set (pairs of low and high resolution images), the trained generator of the SURE-GAN is used to transform low resolution face images of the test set (probes) into their high resolution counterparts. Thereafter, we use these transformed probe images for face identification and verification experiments using the discriminative model. Various aspects of implementation for this entire process are discussed as follows.

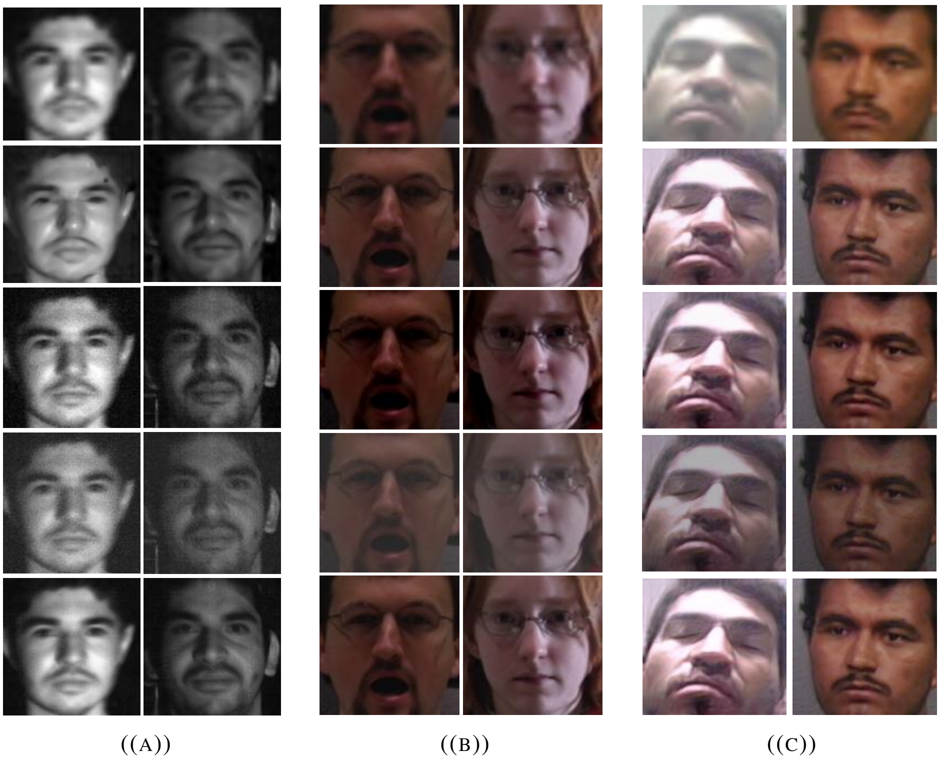


FIGURE 3.6: Translation of probe images from low to high resolution, showing enhancement in image quality, where (a) Images of FaceSurv database, (b) Images of CMU MultiPIE database and (c) Images of NIST MEDS database and first row: bicubic interpolated low resolution images, second row: images generated by Isola *et al.* [65], third row: images generated by Zhao *et al.* [140], fourth row: images generated by Choi *et al.* [141], fifth row: images generated by our proposed method. These translated images are then used for face recognition experiments.

3.4.2.1 Model Architectures and Parameters

In order to train the generator and discriminator of the SURE-GAN, we use the well known encoder decoder style architecture [77, 158–160]. The encoder part of the generator contains 3 layers of convolutions with 64, 128, and 256 filters respectively. The decoder part of the generator contains 3 layers of deconvolutions with 512 filters in each layer. The encoder layers downsample the image by a factor of 2 and the deconvolution layers upsample the image by a factor of 2. After the last layer of the decoder there is a layer of convolution which maps the last layer output into the number of channels of the output image, followed by a *tanh* activation function. In case of MultiPIE and NIST MEDS, the number of output channels is 3 and for the FaceSurv database it is 1. The discriminator has 4 layers of convolutions with 64, 128, 256 and 512 filters in each layer respectively. The last layer of the discriminator maps the output to a single value which is 0 for a fake image and 1 for a real image. The ResNet [161] classifier (ResNet50) is used for supervising the training of the proposed SURE-GAN model. The ELU [162] activation function is utilized for the generator and discriminator whereas ReLU is used for the classifier. The Adam optimizer with a learning rate of 0.001 is used for training the generator and the discriminator. The ResNet classifier is trained using stochastic gradient descent with a learning rate of 10^{-2}.

The lightCNN29 [163] model (used to train the discriminative HQL model for face matching) has 29 convolutional and 4 pooling layers. The dimensionality of the embedding from the last layer is 256 which is used to calculate the HQL. The model is pretrained on the MS-Celeb-1M [164] dataset. In order to train with the HQL function, we resize each image to 128×128 before giving the images as input to the model. The 4-tuples required to train with the HQL is produced at every epoch and is not generated offline. No hard mining is used to generate these 4-tuples. The training is performed using the Adam optimizer. The learning rate is initially kept at 10^{-5} and gradually decreased to 10^{-6}.

3.4.2.2 Training of the SURE-GAN

The proposed SURE-GAN model is pretrained on the CASIA NIR-VIS 2.0 [165] database. This database contains cropped face images in both visible and near infrared spectrum. We create pairs using all 5094 visible spectrum images where each pair has a low resolution (32×32) and a high resolution (128×128) image. The pretraining is done in an unsupervised manner, without using the classifier.

After the SURE-GAN model is pretrained, it is trained (finetuned) on each of the testing datasets (Table 3.1). Each dataset is divided into train and sets. The subjects/classes in the train and the test set are non-overlapping. The train set (Table 3.1) for each dataset is used to train (finetune) the SURE-GAN. The test set has probe and gallery images. The probe images are given as input to the trained SURE-GAN model for translation. These translated images are utilized (as illustrated in the Section 3.4.2.3) to report results on face identification and verification. The number of images for training (finetuning) and testing is outlined in Table 3.1.

3.4.2.3 Face Recognition using SUPREAR-NET

We use a classifier (LightCNN [163]) for learning a discriminative model using the heterogeneous quadruplet loss (HQL). In order to train the HQL model, we use the training set for each dataset as outlined in Table 3.1. The HQL function (Equation 3.15) requires training data in both the domains (high and low resolution), which are directly synthesized (using the trained SURE-GAN model) from the training folds of each dataset. Thus, for each of the three datasets, a separate model using the HQL is trained. The trained model (for each database) is then used to match the synthesized high resolution probe images (using the SURE-GAN) with the gallery.

TABLE 3.2: Identification and verification accuracies for cross resolution face recognition using the proposed SUPREAR-NET.

Database	Probe Resolution	Rank 1 Accuracy (%)					GAR (%) at 1% FAR				
		Bicubic	Isola et al. [65]	Zhao et al. [140]	Choi et al. [141]	Proposed	Bicubic	Isola et al. [65]	Zhao et al. [140]	Choi et al. [141]	Proposed
FaceSurv [4]	128 x128	83.07	80.11	83.53	85.45	**91.05**	87.05	84.74	86.39	86.92	**95.39**
	96 x 96	71.21	66.46	70.95	75.41	**81.21**	73.44	70.58	73.11	74.04	**84.24**
	64 x 64	49.84	45.09	48.33	50.21	**55.50**	52.26	48.54	51.12	52.78	**56.89**
CMU MultiPIE [139]	48 x 48	97.62	92.26	93.40	94.21	**97.82**	98.81	93.69	94.05	95.11	**98.96**
	32 x 32	89.77	78.79	80.81	87.20	**90.39**	93.62	81.72	82.20	85.40	**94.29**
	24 x 24	87.01	77.68	85.54	84.31	**88.70**	89.98	78.68	86.12	83.21	**90.17**
NIST MEDS [1]	48 x 48	91.74	91.61	91.35	91.70	**93.26**	94.16	94.29	93.78	92.57	**94.92**
	32 x 32	90.59	89.32	89.30	91.74	**93.51**	93.91	92.77	92.89	92.82	**94.80**
	24 x 24	84.24	81.70	82.21	88.20	**89.33**	90.10	89.59	90.86	91.54	**93.02**

3.4.2.4 Additional Face Recognition Experiments

In order to evaluate the effectiveness of SURE-GAN, additional matching experiments are performed using a pretrained model instead of the trained discriminative model (as illustrated in Section 3.4.2.3). A pretrained VGGFace [166] model is used to generate embeddings for the probe (translated using the SURE-GAN) and gallery images. The fully connected layers of the VGGFace model are dropped and a global average pooling layer is used in its place. The dimensionality of the embeddings are 512. Prior giving as input to the model, all images are resized to 128 × 128 so that the embeddings are of the same dimensionality. The cosine distance on the embeddings of the probe and gallery images is used for matching.

3.5 Experimental Results

Face identification and verification experiments are performed for three different type of probe resolution variations across all three databases, using the discriminative model (Section 3.4.2.3) which is trained using the HQL. The proposed method is compared with Isola *et al.* [65], Zhao *et al.* [140] and Choi *et al.* [141] on all the databases and on all the probe resolutions. The method by Isola *et al.* [65] is an algorithm for image to image translation which uses conditional GANs. Zhao *et al.*'s [140] method is proposed to transform the facial pose of an image, using the representation of an autoencoder as a classifier to the generator model in their framework. Choi *et al.* [141] proposed an algorithm for image to image translation using a single generator to translate a given source image into multiple target domains. In order to make a fair comparison, the methods by Isola *et al.* [65], Zhao *et al.* [140] and Choi *et al.* [141] are utilized only for translating the low resolution probes to high resolution. Thereafter the model trained using the HQL is utilized for cross-resolution face recognition.

In addition to achieving impressive image quality and resolution enhancement (as shown in Fig. 3.5), SUPREAR-NET yields significantly better recognition accuracies (Table 3.2) across

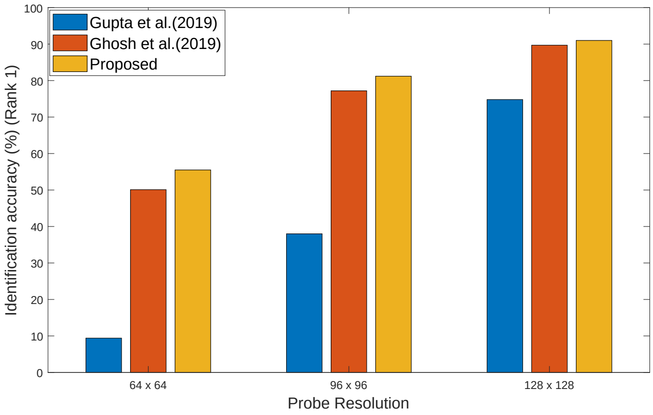
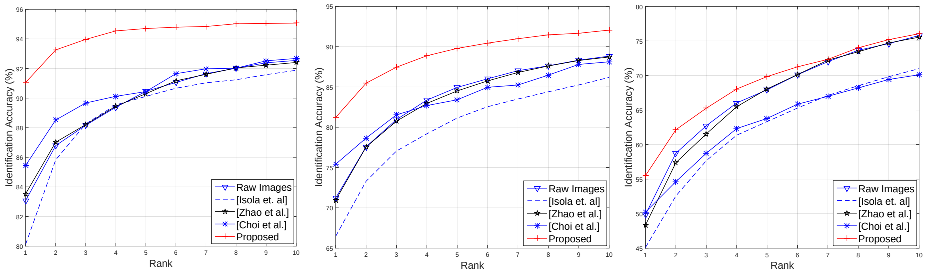


FIGURE 3.7: Comparative face identification (%) (Rank 1) results of the proposed approach with recent methods (Gupta et al. [4] and Ghosh et al. [167]) on the FaceSurv database.

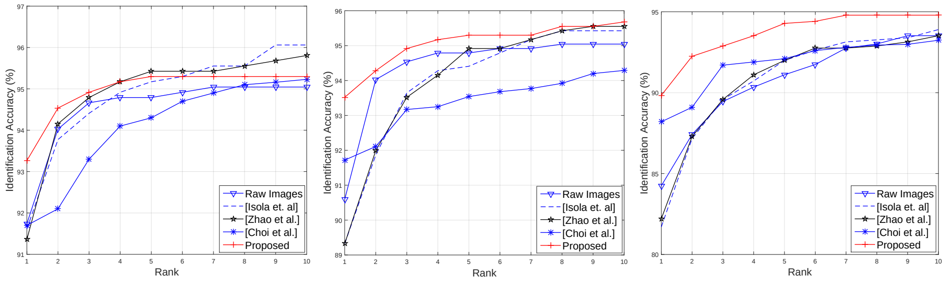


((A)) 128 × 128 probes

((B)) 96 × 96 probes

((C)) 64 × 64 probes

FIGURE 3.8: CMC curves showing identification accuracies for different probe resolutions for the FaceSurv [4] database.



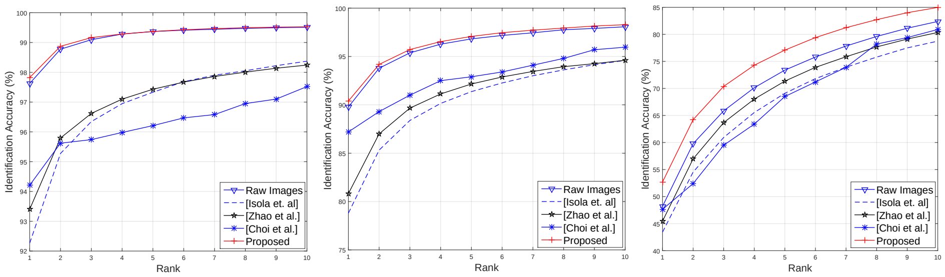
((A)) 48×48 probes

((B)) 32×32 probes

((C)) 24×24 probes

FIGURE 3.9: CMC curves showing identification accuracies for different probe resolutions for the NIST MEDS [1] database.

all the databases compared to Isola *et al.* [65], Zhao *et al.* [140] and Choi *et al.* [141]. On the FaceSurv database, for each of the 4-7m, and 7-10m (equivalent to 96 × 96 and 128 × 128 probe resolutions) probe images, rank 1 identification accuracies are higher by at least 8%, and around 7% for 1-4m probe images, compared to the results by Choi *et al.* [141]. We have also compared the results of the proposed approach on all the three probe resolutions with two recent algorithms namely Gupta et *al.* [4] and Ghosh et *al.* [167] as shown in Figure 3.7. On the CMU MultiPIE database, since the overall image quality is much better (than the FaceSurv database)



((A)) 48×48 probes ((B)) 32×32 probes ((C)) 24 × 24 probes

FIGURE 3.10: CMC curves showing identification accuracies for different probe resolutions for the CMU MultiPIE [139] database.

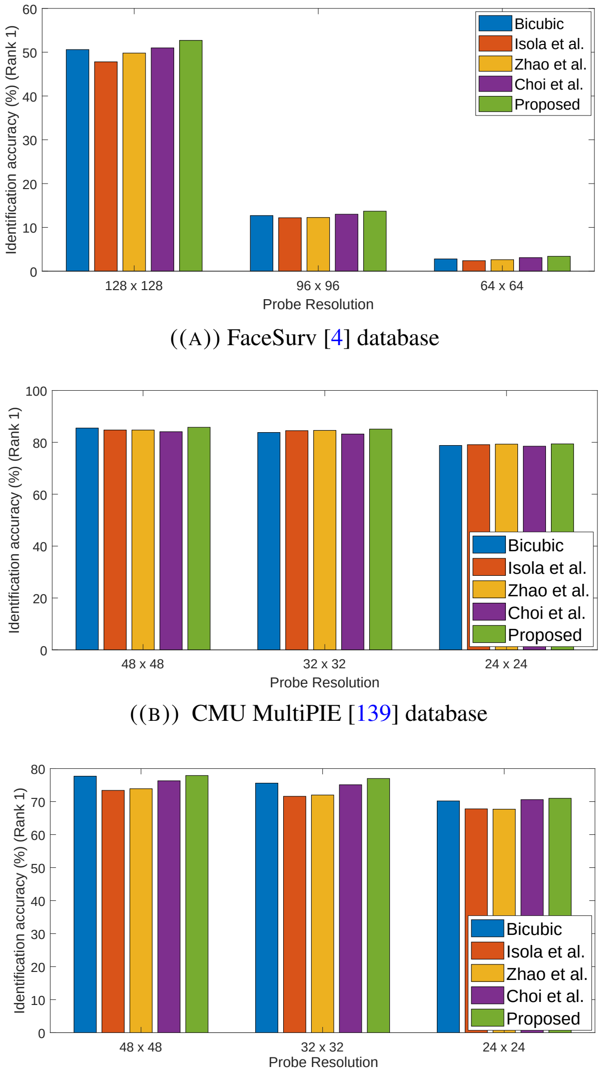
rank 1 identification accuracies are marginally higher for all the probe image resolutions compared to the other GAN based methods. For the NIST MEDS database, identification accuracies (compared to the other methods) for all probe image resolution, are at least 3% higher than the other GAN based methods. As observed in the CMC curves (Fig. 3.8, 3.9 and 3.10) for the FaceSurv, NIST MEDS and CMU MultiPIE databases, SUPREAR-NET produces significantly better results than the other GAN based methods.

3.5.1 Analysis of Results

The proposed algorithm presents an effective way for translating the low resolution probe images to a higher resolution. The results obtained can be analyzed as follows:

1. **Improvement in Image Quality:** The proposed supervised GAN yields images with improved quality. Fig. 3.5 and 3.6 shows the translation of images using the trained generator of the proposed SURE-GAN from low to high resolution. It can be visually inferred that the increased resolution results in improved quality. To accentuate and quantify this observation, we performed no-reference image quality assessment of face images generated by all the five algorithms (bicubic, Zhao *et al.* [140], Isola *et al.* [65], Choi *et al.* [141] and the proposed SURE-GAN) on the NIST MEDS database. We observed that the proposed algorithm yields the lowest BRISQUE [168] values (Bicubic: 25.40, Zhao *et al.* [140]: 20.73, Isola *et al.* [65]: 20.21, Choi *et al.* [141]: 20.98 and Proposed: 19.27). It is to be noted that lower BRISQUE values are considered better, in terms of image quality.

2. **Additional Results using a Pretrained model:** Additional results are also presented by replacing the discriminative model in SUPREAR-NET by a pretrained VGGFace [166] classifier. In all the three databases better results are obtained compared to a recent unsupervised [65] and supervised [140] image to image translational GAN methods. The results are outlined in Fig. 3.11.



((A)) FaceSurv [4] database

((B)) CMU MultiPIE [139] database

((C)) NIST MEDS[1] database

FIGURE 3.11: Bar plots showing identification accuracies for cross-resolution face recognition on different probe resolutions of the three databases. Matching was performed using a pretrained VGGFace [166] model.

TABLE 3.3: Identification and Verification accuracies for the ablation study on the NIST MEDS [1] database.

Probe Resolution	Rank 1 Accuracy (%)				GAR (%) at 1% FAR			
	Triplet Loss [91]	Quadruplet Loss [92]	Heterogeneous Triplet Loss [169]	Proposed	Triplet Loss [91]	Quadruplet Loss [92]	Heterogeneous Triplet Loss [169]	Proposed
48 x 48	86.65	90.88	92.91	**93.26**	75.51	80.71	94.67	**94.92**
32 x 32	86.53	90.85	92.88	**93.51**	75.13	81.85	94.42	**94.80**
24 x 24	76.62	81.19	88.81	**89.33**	57.49	64.85	91.24	**93.02**

3. **Pretrained VGGFace vs Trained HQL model:** We utilized two different ways of performing recognition post the image translation process by the generator of the SURE-GAN. The proposed approach using the HQL model (results in Table 3.2) exhibits much superior recognition performance in terms of both identification and verification accuracies. On the other hand, the improvement in recognition performance over the other methods is marginal using the pre-trained VGGFace model (Fig. 3.11). This showcases the advantage of using a heterogeneous deep metric learning approach for recognition in

TABLE 3.4: Comparison of the proposed approach with a superresolution method (Wang *et al.* [2]) in the NIST MEDS database [1].

Probe Resolution	**Rank 1 Accuracy (%)**		**GAR (%) at 1% FAR**	
	Wang *et al.* [2]	**Proposed**	**Wang *et al.* [2]**	**Proposed**
48 x 48	92.50	**93.26**	89.21	**94.92**
32 x 32	92.37	**93.51**	89.72	**94.80**
24 x 24	84.62	**89.33**	77.28	**93.02**

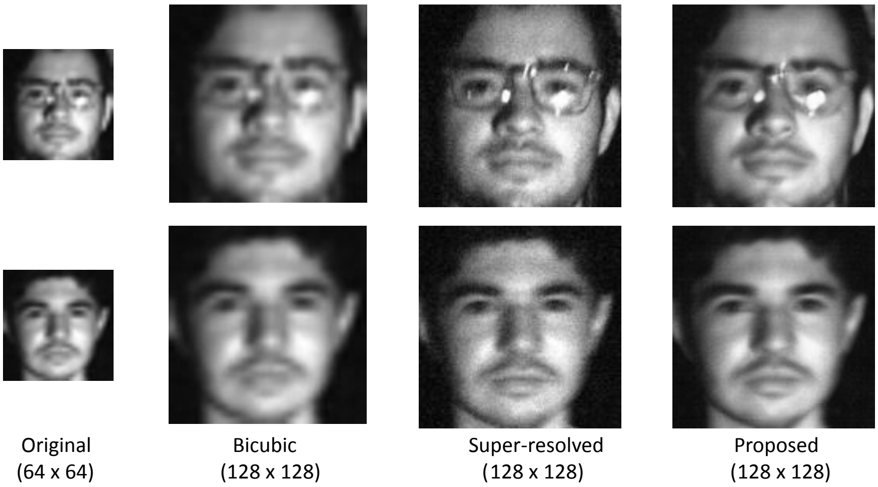


FIGURE 3.12: Comparison of image level approaches for cross-resolution face recognition. The superresolved images (third column) (generated using [2]) has a grainy appearance compared to that of the proposed SURE-GAN (last column). Images are from the FaceSurv [4] database.

conjunction with the proposed SURE-GAN.

4. **Comparison with Superresolution:** Superresolution methods are popular for image to image translation (from low to high resolution). A comparison of the proposed approach (SUPREAR-NET) with a popular deep-learning based superresolution method by Wang *et al.* [2] is performed. This method is used to translate images of the NIST MEDS database [1] for each of the three resolution variations (28 × 28, 32 × 32 and 48 × 48) to the resolution of the gallery (128 × 128). Thereafter, the model trained using the proposed HQL function was used to match the translated probe images with the high resolution gallery. As shown in Table 3.4, the proposed method outperforms the superresolution method both in terms of verification and identification accuracies on all the three probe resolutions.

5. **Ablation Study:** An ablation study (Table 3.3) was performed on the performance of the proposed framework (SUPREAR-NET). The performance of the same is compared with three popular deep metric learning algorithms namely Triplet Loss [91] and Quadruplet Loss [92] and Heterogeneous Triplet Loss [169]. The heterogeneous triplet loss [169] is a heterogeneous deep metric learning algorithm, the loss function for which is given by,

$$L = \left[\| f(p^a) - f(q^a) \|_2^2 - \left\| f(p^a) - f(q^b) \right\|_2^2 + \alpha_1 \right]_+ \tag{3.16}$$

$$\forall (p^a, q^a, q^b) \in \P$$

where the terms have their usual meanings as in Equation 3.15. We observe that the proposed method (SUPREAR-NET) outperforms all the three deep metric learning methods on the NIST MEDS database [1]. Table 3.3 outlines the results of the ablation study.

3.6 Summary

A novel method for training a supervised generative adversarial network is presented for image to image translation which does not corrupt the discriminative information of the source image. This is achieved by backpropagating the loss of a classifier to the generator, in addition to the conventional adversarial loss from the discriminator. The trained generator is used to translate the low resolution probe images into high resolution ones. A discriminative model trained using a heterogeneous quadruplet loss function is used to match the translated probe images to the high resolution gallery. These translated images yield better face recognition performance than the conventional image to image translation GAN. In addition to this, the proposed method also outperforms recent image to image translation GAN methods and a deep-learning based superresolution algorithm. The proposed SURE-GAN is robust to unseen testing data since the subjects/classes in training and testing splits are non overlapping. Thus our method is an useful step for application in face recognition for surveillance scenarios, where resolution of the query images obtained is a major obstacle for the same.

Chapter 4

Subclass Heterogeneity Aware Loss for Cross-Spectral Cross-Resolution Face Recognition

4.1 Introduction

The increasing effectiveness of Deep Convolutional Neural Networks (Deep-CNNs) has led to the emergence of very efficient face recognition algorithms [150, 166, 170, 171]. With this development, various applications ranging from unlocking of mobile phones and laptops to monitoring of public places are now using face recognition technology. These images are usually captured in controlled scenarios and constrained settings. However, the query images may be captured in unconstrained environment by any kind of camera; for instance, surveillance cameras. These cameras are generally placed at a high standoff distance from the subjects and have a large field-of-view [172]. As a result, the effective resolution and quality of the captured face image may be low. In addition to that, when sufficient visible illumination is not available, these cameras operate in the Near-Infrared (NIR) mode and the probe images are captured in NIR spectrum. This results in a heterogeneous image/face matching (recognition) problem between the high resolution visible spectrum gallery and low resolution NIR spectrum probes (Fig. 4.1). The combination of the acquisition environment and the position of the user in relation to the camera location leads to three possible scenarios of heterogeneous face matching.

- Cross-Spectral matching where the visible spectrum face image (gallery) is matched with the NIR spectrum images (probes).
- Cross-Resolution matching where the high resolution face images (gallery) is matched with the low resolution images (probes) obtained from surveillance cameras.

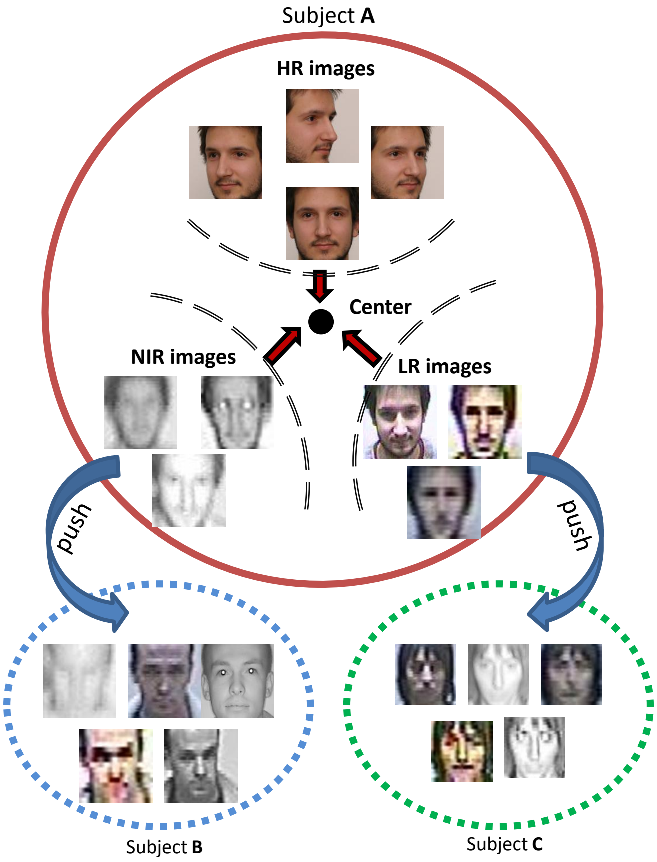


FIGURE 4.1: Visual abstract of the proposed Subclass Heterogeneity Aware Loss (SHEAL). The intraclass distance between the different subclasses, each represented by a particular covariate such as high resolution (HR), low resolution (LR) and NIR images is minimized, while pushing other impostor classes away, in the embedding space of the model. (best viewed in color)

- Cross-Spectral Cross-Resolution matching where low resolution NIR images (probe) are matched with high resolution visible spectrum mugshot images (gallery).

Several researchers have proposed solutions for heterogeneous face recognition. At the core of many of these solutions lies the most fundamental concept of training a face recognition model, which is, train the model such that the intra-class is minimized and the inter-class distance is maximized, both for with intra-view (homogeneous) and inter-view (heterogeneous) data variations [173–175]. However, most of the existing algorithms focus on only one covariate at a time, either cross-resolution or cross-spectral variations, not both together. Given the increasing use of surveillance cameras for security, it is important to address both the covariates together.

In this research, we propose a unified Subclass Heterogeneity Aware Loss (SHEAL) to train a discriminative model which produces face embeddings for accurate classification in the presence of multiple face recognition covariates. A novel subclass based optimization approach is presented, which optimizes the clusters based on different subclasses in the data. As shown

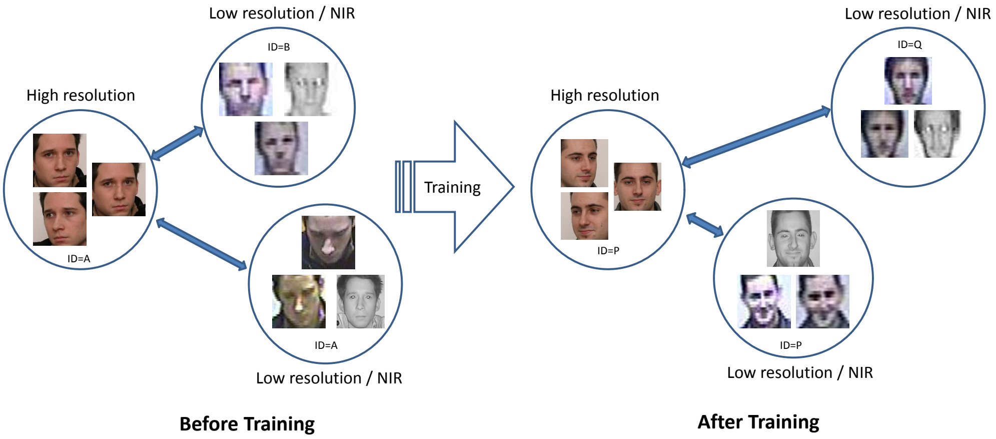


FIGURE 4.2: Illustrating the effect of training with the proposed loss function. The proposed loss function attempts to minimize the distance between the intra-class embeddings compared to the distance between the embeddings of the images belonging to different classes.

in Fig. 4.2, the proposed model learns discriminative embeddings for both high resolution and visible spectrum gallery images and low resolution, NIR spectrum probe images. These learnt embeddings are then matched using Euclidean distance. Experiments on four challenging databases, namely SCface [3], FaceSurv [4], CASIA NIR-VIS 2.0 [165], and Labeled Faces in the Wild [176], demonstrate the efficacy of the proposed approach, not only in the identification performance but also with respect to convergence in terms of training time and epochs.

4.2 Related Work

This paper addresses the problem of cross-spectral cross-resolution face recognition with a novel deep metric learning algorithm. Therefore, the review section first outlines the related work performed on cross-spectral and cross resolution face recognition especially using deep learning methods, followed by the literature on deep metric learning methods for face recognition.

Prior to the emergence of deep learning based face recognition algorithms, several discriminative learning and transfer learning based approaches were proposed for cross-spectral [27, 165, 173, 175, 177–179] and cross-resolution [23, 82, 84, 180] face recognition. Deep learning based algorithms have also been proposed for these tasks. Lu *et al.* [181] learned binary descriptors for heterogeneous face recognition. Yi *et al.* [182] used a shared representation learning based approach using Restricted Boltzman Machines for cross-spectral face recognition. Saxena *et al.* [183] used a metric learning based algorithm to learn a Mahalanobis distance based embedding space for the same. Lezama et *al.* [184] used a low rank embedding based approach for hallucination of NIR to visible spectrum face images for cross-spectral face matching. He et *al.* [185] proposed an algorithm to learn a deep-CNN model where the high level

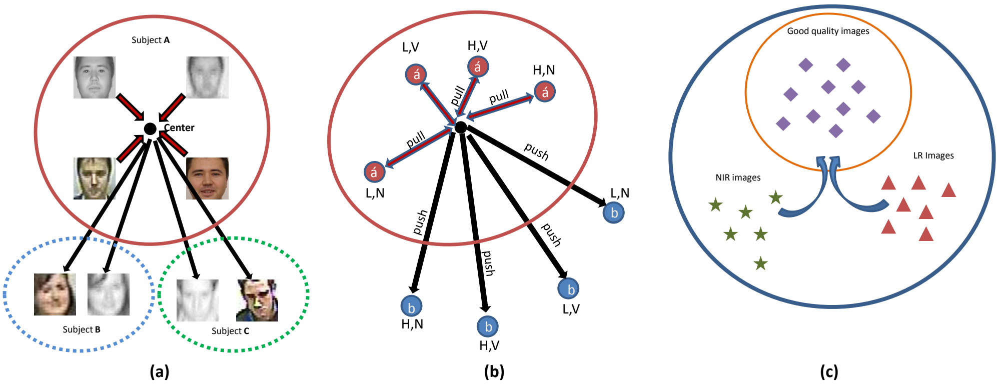


FIGURE 4.3: Representation of the proposed method: (a) Overall motivation of the problem, (b) Illustration of the proposed loss metric which minimizes the intra-class distance and maximizes the inter-class distance (including intra-view and inter-view variations) and (c) Subclass based cluster optimization.

layer is divided into two orthogonal subspaces that learn modality-invariant representation for cross-spectral face recognition. Wu et *al.* [186] used an approximate variational formulation in a coupled deep learning framework for matching NIR face images to a gallery of visible-spectrum face images. Song et *al.* [187] proposed an adversarial discriminative learning algorithm for the same, using an integration of cross-spectral face hallucination and discriminative feature learning. Pereira et *al.* [188] proposed a deep learning approach using a framework that learns domain specific feature detectors for cross-spectral face recognition. Recently, Peng et *al.* [189] proposed a locally linear re-ranking (LLRe-Rank) approach for the same problem. He et *al.* [190] performed face completion by texture inpainting and pose correction using generative modelling for translating NIR face images for efficient matching with visible spectrum images.

Singh *et al.* [144] proposed a Synthesis via Hierarchical Sparse Representation for generating a high resolution face image from a low resolution one, for cross-resolution face recognition. Lu et *al.* [150] utilized a deep coupled end to end CNN consisting of a trunk network and two branch networks for cross-resolution face matching. Lu et *al.* [151] utilized a discriminative multidimensional scaling approach for face recognition from low resolution images. Ge et *al.* [191] proposed an approach using a two-stream CNN for low resolution face recognition. Li et *al.* [192] used a supervised discriminative learning approach for low resolution face recognition. Zangeneh et *al.* [193] proposed a novel nonlinear coupled mapping architecture for face recognition from low resolution images. Abdollahi et *al.* [194] proposed a modified finetuning approach using different variations of the training data for low resolution face recognition. Recently Singh et *al.* [195] utilized a dual directed capsule network for very low resolution face recognition.

The popularity of deep metric learning methods has led to the development of several loss functions [21, 196–204] to train deep neural network models for face recognition. Schroff *et al.* [91] introduced the triplet loss based training method for face verification. Quadruplet loss [92], an extension of triplet loss, adds an extra negative sample to the loss function. This loss function enforces a stricter inter-class distance on the output embedding space of the model being trained. However, both these techniques do not consider any heterogeneity in the data during training. They also require extensive hard-sample mining for effective training. In order to account for heterogeneity in data, Liu *et al.* [169] have proposed a heterogeneous variant of triplet loss. This loss function can take at most one heterogeneity (e.g. cross-resolution) at a time and is not suitable for handling more than one covariate (e.g. cross-resolution and cross-spectral both). In addition, it required exhaustive hard mining prior to the training process. Several modifications [92, 200, 205] to the triplet loss have been proposed for a diverse range of applications such as person-re-identification, matching images of cars, object recognition, patch matching and so on. However, none of these methods addressed scenarios where matching of images with multiple heterogeneity is involved.

4.3 Proposed Algorithm

In this section, we illustrate the proposed algorithm which is utilized to learn a model for face recognition invariant to both spectrum and resolution. First, the framework for a heterogeneous matching problem is illustrated with only one covariate/heterogeneity (resolution and spectrum) across probes and gallery images. The formulation is then extended to include invariance to two covariates, namely resolution and spectrum. It is important to note that while the proposed loss function SHEAL, L_{SHEAL}, optimizes for heterogeneous matching with one or two covariates, it also optimizes for homogeneous matching (no covariates). The first subsection presents the formulation of SHEAL followed by the sub-class based cluster optimization. Finally, the heterogeneous face recognition algorithm is presented which learns a model with a highly discriminative output embedding space for cross-spectral cross-resolution face recognition. Fig. 4.3 illustrates the concept of the proposed Subclass Heterogeneity Aware Loss (SHEAL).

4.3.1 SHEAL: Subclass Heterogeneity Aware Loss

For a heterogeneous face matching problem, the gallery contains images with high resolution visible spectrum while the probe images are captured with different covariates present (for instance low resolution and/or NIR). For simplicity, let us assume only one kind of heterogeneity, e.g. resolution, is available in the data (i.e. gallery of high resolution images and probes are low resolution images). In order to learn a discriminative model for such a task, the loss metric needs to perform two tasks, minimizing (pulling together) and maximizing (pushing away) the intra-class and inter-class distances, respectively in intra-view[1] (homogeneous) settings, and

[1] Intra-view settings refer to the scenario when the gallery and probe are homogeneous in nature, for example, same resolution and spectrum.

performing the same in inter-view[2] (heterogeneous) settings. The proposed heterogeneous loss function is expressed as,

$$L = [||g(X_i^H) - g({X'}_i^H)||_2^2 - ||g(X_i^H) - g(X_j^H)||_2^2 + \alpha_1]_+ \\ + [||g(X_i^H) - g(X_i^L)||_2^2 - ||g(X_i^H) - g(X_k^L)||_2^2 + \alpha_2]_+ \quad (4.1)$$

$$\forall (X_i^H, X_j^H, X_i^L, X_k^L) \in \tau$$

where, H and L signify high and low resolution, respectively. X_i^H is the high resolution anchor image of subject i, ${X'}_i^H$ is another high resolution image of the same subject i, X_i^L is a low resolution image of the subject i, X_j^H is the high resolution image of subject j, X_k^L is a low resolution image of another subject k where, $i \neq j \neq k$ and $[\cdot]_+ = max(\cdot, 0)$.

In a complex (more realistic) scenario, the heterogeneity may be due to two different views, namely resolution and spectrum. For example, the gallery images are in visible spectrum and high resolution, while the probes are in NIR and low resolution. Let the visible spectrum and NIR spectrum be denoted as V and N, respectively, and subscripts i, j, k, l, m represent different subjects/classes. Let the high resolution visible spectrum image of the i^{th} subject (class) be $X_i^{H,V}$. Another image of the same subject in the same setting is denoted as ${X'}_i^{H,V}$. Similarly, $X_i^{H,N}$, $X_i^{L,V}$ and $X_i^{L,N}$ represent the high resolution NIR spectrum image, low resolution visible spectrum image, and low resolution NIR spectrum image of the i^{th} subject, respectively. To accommodate both cross-resolution cross-spectral variations, the proposed loss function is formulated with two cross-views, hence require four separate terms. The first term takes care of the homogeneous matching scenario, the next two terms accommodates for cross-resolution and cross-spectral matching respectively, followed by the last term for cross-spectral cross-resolution matching.

The homogeneous loss term (L_{Ho}) is computed as,

$$L_{Ho} = [||g(X_i^{H,V}) - g({X'}_i^{H,V})||_2^2 - ||g(X_i^{H,V}) - g(X_j^{H,V})||_2^2 + \alpha_1]_+ \quad (4.2)$$

This loss expression is composed of two parts, the former $||g(X_i^{H,V}) - g({X'}_i^{H,V})||_2^2$ minimizes the intra-class distance between the embedding of the anchor image $g(X_i^{H,V})$ and $g({X'}_i^{H,V})$, which is another image of the same subject captured in the same condition. The later part of the expression, $||g(X_i^{H,V}) - g(X_j^{H,V})||_2^2$ maximizes the inter-class distance between $g(X_i^{H,V})$ and $g(X_j^{H,V})$. However, in order to calculate the intra-class loss, we can replace the embedding of the anchor image $g(X_i^{H,V})$ by the center embedding of the i^{th} class (subject) given by

[2] Inter-view settings refer to the scenario when both gallery and probe images are heterogeneous in nature, for example, different resolution or spectrum.

$g_c(X_i^{H,V})$. In addition to that, the inter-class distances are also computed from $g_c(X_i^{H,V})$ instead of $g(X_i^{H,V})$. Therefore, for SHEAL, the loss function for the homogeneous component (L_{Ho}^c) can be written as:

$$L_{Ho}^c = [||g_c(X_i^{H,V}) - g({X'}_i^{H,V})||_2^2 - ||g_c(X_i^{H,V})) - g(X_j^{H,V})||_2^2 + \alpha_1]_+ \tag{4.3}$$

Next, the cross-resolution loss term is expressed as:

$$L_{CR}^c = [||g_c(X_i^{H,V}) - g(X_i^{L,V})||_2^2 - ||g_c(X_i^{H,V}) - g(X_k^{L,V})||_2^2 + \alpha_2]_+ \tag{4.4}$$

This loss expression contains two parts, the former $\|g_c(X_i^{H,V}) - g(X_i^{L,V})\|_2^2$ pertains to the distance between the center embedding of the images of the same subject i in visible spectrum and low resolution. The later term $\|g_c(X_i^{H,V}) - g(X_k^{L,V})\|_2^2$ focuses on maximizing the inter-class distance between the center embedding of another subject k in visible spectrum and low resolution. Similarly, the cross-spectral loss is expressed as,

$$L_{CS}^c = [||g_c(X_i^{H,V}) - g(X_i^{H,N})||_2^2 - ||g_c(X_i^{H,V}) - g(X_l^{H,N})||_2^2 + \alpha_3]_+ \tag{4.5}$$

Along the same lines, the **cross-spectral cross-resolution loss** is computed as,

$$L_{CS-CR}^c = [||g_c(X_i^{H,V}) - g(X_i^{L,N})||_2^2 - ||g_c(X_i^{H,V}) - g(X_m^{L,N})||_2^2 + \alpha_4]_+ \tag{4.6}$$

Equation 4.6 models the most challenging scenario where the intra-class and inter-class distances are evaluated between the high resolution visible spectrum images and images captured in low resolution and NIR. Such probe images differ from the gallery images with respect to both resolution and spectrum. The final loss function combines the homogeneous and heterogeneous losses as follows:

$$L_{SHEAL} = \lambda_1.L_{Ho}^c + \lambda_2.L_{CR}^c + \lambda_3.L_{CS}^c + \lambda_4.L_{CS-CR}^c \tag{4.7}$$

$$\forall({X'}_i^{H,V}, X_j^{H,V}, X_i^{L,V}, X_k^{L,V}, X_i^{H,N}, X_l^{H,N}, X_i^{L,N}, X_m^{L,N}) \in \tau$$

where, τ is the set of 8-tuples. Each such 8-tuple is considered as a training sample and the coefficients $\{\lambda_1, \lambda_2, \lambda_3, \lambda_4\}$ may be used to adjust the weights of each component of the loss function. The gradient of this loss can be utilized to train the parameters of a model using representation learning (e.g. CNN model).

The traditional triplet loss works by *pulling* the embeddings of all samples of the same class towards the anchor and *pushing* the same for the impostor classes away from the anchor. However, this loss is unable to handle a heterogeneous matching problem where a pair of images of different views/modalities are to be matched during testing. In order to approach this problem we are required to train a discriminative model which can generate heterogeneity aware embeddings. To train such a model, the loss function should incorporate different matching scenarios, i.e. both homogeneous and heterogeneous. The proposed loss function (Equation 4.7) has been formulated by combining multiple heterogeneous variations for face matching. To summarize, the salient contributions/novelty of this work are as follows:

- We propose a method to train a discriminative model which can be utilized to match images belonging to more than one covariate. Equation 7 has four loss terms, viz L^c_{Ho}, L^c_{CR}, L^c_{CS} and L^c_{CS-CR}. Each of them contributes a gradient which is used to update the weights of the model $g(.)$ being trained.
- In Equation 4.7, different terms are weighed by adjustable λ parameters. If we want the model to be more often used for cross-spectral-cross-resolution matching then the coefficient of L^c_{CS-CR} can be given a higher value. This allows the model to be tuned for a specific application scenario as well.

The model $g(.)$ is trained using Equation 4.7 which results in disjoint clusters for each class in the output embedding space of the model. These clusters are further optimized using subclass based cluster optimization as illustrated in the next subsection.

4.3.2 Subclass based Cluster Optimization

We optimize the clusters (learned using L_{SHEAL}) in the embedding space of the model using a subclass based loss formulation. As shown in Fig. 3(c), in each cluster which contains embeddings of the images of a particular subject, the embeddings of the good quality images (high resolution visible spectrum) of the respective subject are expected to be very close to each other. On the other hand, the images different from the good quality ones (i.e. low resolution and NIR) are expected to be farther away in the same cluster. Using this as a hypothesis, each cluster is expected to contain two subclasses, one representing the good quality images (homogeneous) and the other for the heterogeneous images. An optimization stage is utilized to create a more compact cluster by bringing these two subclasses closer to each other. The loss function for the cluster optimization stage is expressed as,

$$L_{PP} = \beta_1.[||g_{c_1}(X_i^{H,V}) - g({X'}_i^{H,V})||_2^2 - ||g_{c_1}(X_i^{H,V}) - g(X_j^{H,V})||_2^2 + \alpha_1]_+ + \beta_2.[||g_{c_2}(X_i^{L,N}) - g(X_i^{L,N})||_2^2 - ||g_{c_2}(X_i^{L,N}) - g(X_k^{L,N})||_2^2 + \alpha_2]_+ + \beta_3.[||g_{c_1}(X_i^{H,V}) - g_{c_2}(X_i^{L,N})||_2^2]_+ \quad (4.8)$$

$$\forall (X_i^{H,V}, X_i^{L,N}, X_j^{H,V}, X_k^{L,N}) \in \tau$$

TABLE 4.1: Experimental details to evaluate the performance of the proposed SHEAL.

Experiment	Databases	Spectrum		Resolution	
		Gallery	**Probe**	**Gallery**	**Probe**
Cross-Resolution Face Recognition (CR-FR)	SCface	Visible	Visible	128 x 128	24 x 24, 32 x 32, 48 x 48
	FaceSurv	Visible	Visible	128 x 128	48 x 48, 64 x 64
	LFW	Visible	Visible	128 x 128	32 x 32, 48 x 48
Cross-Spectral Face Recognition (CS-FR)	SCface	Visible	NIR	128 x 128	128 x 128
	CASIA NIR-VIS 2.0	Visible	NIR	128 x 128	128 x 128
Cross-Spectral Cross-Resolution Face Recognition (CSCR-FR)	SCface	Visible	NIR	128 x 128	24 x 24, 32 x 32, 48 x 48
	FaceSurv	Visible	NIR	128 x 128	48 x 48, 64 x 64
	CASIA NIR-VIS 2.0	Visible	NIR	128 x 128	48 x 48, 64 x 64

where, $g_{c_1}(X_i^{H,V})$ and $g_{c_2}(X_i^{L,N})$ are the centers of the subclasses pertaining to the homogeneous (good quality) and the heterogeneous (low resolution and NIR) images, respectively. $\beta_1, \beta_2, \beta_3$ are weights for each component. The first term in Equation 4.8 is similar to the first term of Equation 4.3, which brings the embedding of the good quality (homogeneous) images closer in the output embedding space of the model. The second term brings the embedding of the heterogeneous images closer thus making the subclass of the heterogeneous images (low resolution and NIR) more compact. The third term brings the centers of the two subclasses of the cluster closer to each other. The coefficients $\{\beta_1, \beta_2, \beta_3\}$ are used to adjust the strength of each component of the loss function. At the end of this cluster optimization phase, it is expected that all the images (heterogeneous and homogeneous) of each class must make a compact cluster, thereby enhancing heterogeneous matching performance of the trained model.

4.3.3 Heterogeneous Face Recognition using SHEAL

In order to train a heterogeneity aware model for face recognition, Equation 4.7 followed by Equation 4.8 is utilized. Once the model is trained, the test data is partitioned into probe and gallery according to the protocol of the testing database. For cross-spectral cross-resolution face recognition, the probes are NIR images of low resolution and the gallery images are high resolution visible spectrum images. A probe is given as input to the trained discriminative model to extract the embeddings, and the same is performed to generate the embeddings of the gallery images. The Euclidean distance is used to calculate match scores between the probe and gallery embeddings, which is finally used for face recognition.

4.3.4 Implementation Details

In this section, we outline the implementation details required to reproduce the results.

4.3.4.1 CNN Model

The proposed SHEAL is utilized to train a deep-CNN model for heterogeneous face recognition. The CNN model used is Light-CNN-29 [171] which is one of the popular models for face recognition with 29 convolutional layers and 4 pooling layers. After every convolutional layer a Max-Feature-Map operation is performed. The network is built using 6 blocks and each block contains convolution and Max-Feature-Map layers. The final layer is a Max-Feature-Map layer which gives an embedding of size 256.

4.3.4.2 Preparing Data for Training

In order to prepare training data for the SHEAL metric, each training sample is represented by an 8-tuple. Unlike existing approaches [91, 92, 205] we do not perform any hard mining on the set of 8-tuples, rather, we randomly prepare 500 8-tuples for every epoch. In order to prepare each 8-tuple, the images of a randomly selected subject/class (high resolution and visible spectrum) are used to calculate the center embedding. Other images for the 8-tuples are then chosen randomly from the training set of the database accordingly. We train only one epoch on each set of 500 8-tuples that are created in every iteration. We also performed experiments by running multiple epochs on each set of 8-tuples, but there is a tendency of the model to overfit on the samples that are generated. Thus, each epoch constitutes generating the set of 500 8-tuples and running one iteration of training on it. This keeps the pace of learning stable and effective.

4.3.4.3 Loss Function Parameters

The deep-CNN model is trained by back-propagating the gradient of the proposed SHEAL. The optimization is performed using Adam with a batch size of 20. The learning rate is initially kept at 10^{-3} which is then gradually decreased to 10^{-7}. The criteria for decreasing the learning rate was non-increment of validation accuracy for 20 epochs. The learning rate was decreased in steps of 0.05. The values of the margin variables α_1, α_2, α_3, α_4 are set differently for different databases during training. For the SCface database, we keep $\alpha_1 = 0.2$, $\alpha_2 = 0.4$, $\alpha_3 = 0.4$, and $\alpha_4 = 0.6$. For the FaceSurv database, we keep $\alpha_1 = 0.2$, $\alpha_2 = 0.4$, $\alpha_3 = 0.4$, and $\alpha_4 = 0.8$. For the CASIA NIR-VIS 2.0 database, we keep $\alpha_1 = 0.3$, $\alpha_2 = 0.4$, $\alpha_3 = 0.4$, and $\alpha_4 = 0.8$.

The parameters for the loss function coefficients λ_1, λ_2, λ_3, λ_4 for training are set as follows. For the SCface database, $\lambda_1 = 0.1$, $\lambda_2 = 0.2$, $\lambda_3 = 0.4$, and $\lambda_4 = 0.7$. For the FaceSurv database, $\lambda_1 = 0.4$, $\lambda_2 = 0.5$, $\lambda_3 = 0.5$, and $\lambda_4 = 0.9$. For the CASIA NIR-VIS 2.0 database, $\lambda_1 = 0.1$, $\lambda_2 = 0.4$, $\lambda_3 = 0.6$ and $\lambda_4 = 0.6$. Experiments are performed on a machine with Intel Core i7 CPU, with 32GB of RAM and NVIDIA GTX 1080Ti GPU with a PyTorch implementation.

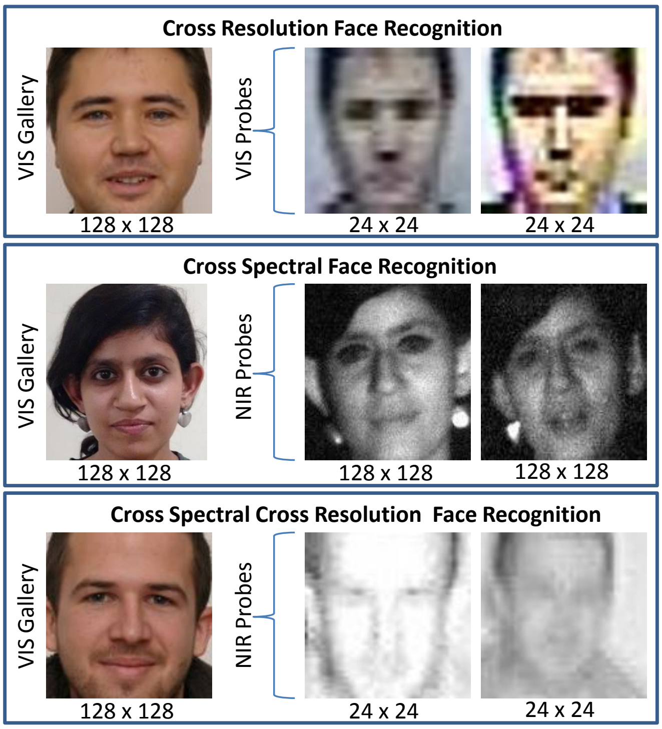


FIGURE 4.4: Three different cases of heterogeneous face recognition considered in this work, including the most challenging case of cross-spectral cross-resolution matching. Images are taken from the SCface [3] and FaceSurv [4] databases.

4.3.4.4 Weights (β) for Subclass Cluster Optimization

The β parameters are used to assign weight of different components (Equation 4.8) in the subclass optimization step of SHEAL. Although these parameters are chosen empirically, we have followed a strategy while selecting the β parameters. As illustrated in Section 4.3.2, β_1 and β_2 are the weights of the subclasses for the visible spectrum high resolution images and the NIR low resolution images, respectively. On the other hand β_3 are weights for bringing the two subclasses closer into a single compact cluster. Since the later is the main motive of this step, β_3 is given a higher value than β_1 and β_2. Using this guideline and some empirical observations, the best β parameters are obtained for the subclass based cluster optimization.

4.4 Experiments and Analysis

To show the efficacy of the proposed approach, we have performed three different heterogeneous experiments on four challenging face databases. Very few papers in the literature have analyzed all three heterogeneous scenarios (a typical scenario of face recognition for video surveillance) using a single algorithm.

4.4.1 Databases and Protocol

As shown in Fig. 4.4 three experiments are performed, namely Cross-Resolution Face Recognition (CR-FR), Cross-Spectral Face Recognition (CS-FR), and Cross-Spectral Cross-Resolution Face Recognition (CSCR-FR). Details of experimental protocol are illustrated in Table 4.1. The details of the databases used for the experiments are as follows.

SCface Database [3] is one of the most popular face datasets that contains real world surveillance quality images. It contains 4160 images of 130 subjects captured using 8 surveillance cameras from three standoff distances namely 1 mt, 2.6 mts and 4.2 mts. The effective resolution of the face images detected from these surveillance images are 24×24, 32×32 and 48×48 for these three distances, respectively. Out of the 8 cameras, 5 operate in the visible spectrum and the remaining capture images in the NIR mode. The gallery images are captured using high resolution cameras and are sub-sampled to a resolution of 128×128. For CSCR-FR, NIR probe images pertaining to the three different resolutions have been matched with the high resolution visible spectrum gallery. For CR-FR, the same matching has been performed with low resolution visible spectrum probe images.

CASIA NIR-VIS 2.0 Database [165] is the largest publicly available dataset for CS-FR. It contains a total of 17,415 visible spectrum and NIR images pertaining to 725 subjects. The images in the training and testing sets are fixed and contain non-overlapping subjects. The database is divided into two views, namely view 1 and 2. The former is a development set and the later is for reporting the results. The gallery set contains one high resolution visible spectrum image for each subject. In order to train the deep CNN model using the proposed loss metric, we need low resolution visible and NIR images, in addition to the high resolution (128×128) visible and NIR images that are already present in the database. The images (both visible and NIR) are subsampled to a resolution of 32×32 to synthetically create low resolution versions of the same. To perform testing for CSCR-FR, the probe images are subsampled to a resolution of 48×48 and 64×64. For CS-FR, the usual protocol of the database (128×128 NIR probes) is utilized.

FaceSurv Database [4] contains videos captured under surveillance conditions in both day-time (in visible spectrum) and night-time (in NIR). The videos contain subjects walking at a standoff distance of 1-10 mts from the camera. The night-time videos have been captured in a completely dark environment using NIR illumination, while the day-time videos have been captured in outdoor settings. Both day-time and night-time videos are captured under uncontrolled illumination, pose and expression variations. The gallery images contain three high resolution (subsampled to 128×128) visible spectrum images for every subject. Images pertaining to 30 subjects are used for training and images of the remaining subjects are used for testing. In terms

TABLE 4.2: Rank 1 identification accuracies on the SCface database.

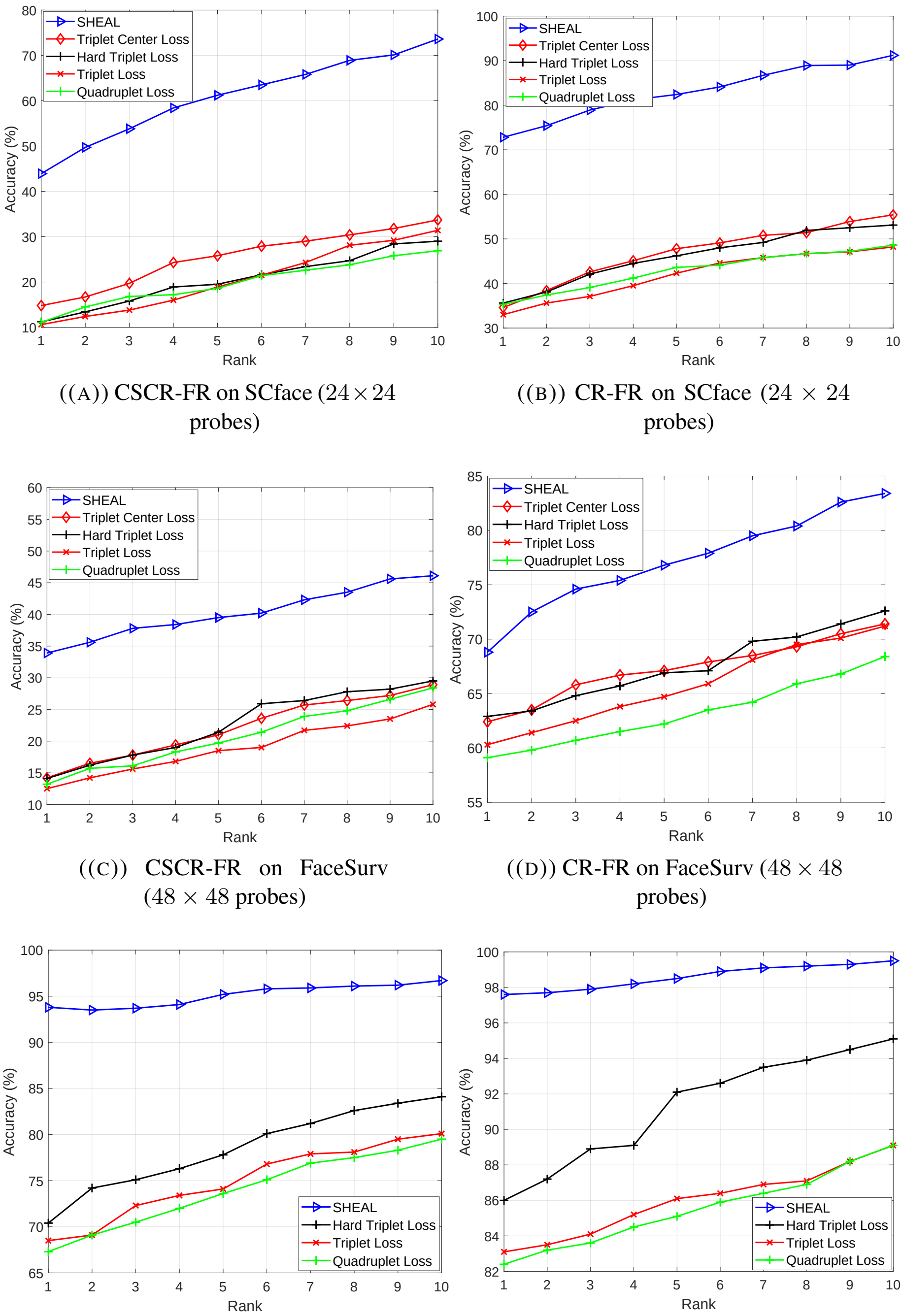
Algorithm	**Cross-Spectral Cross-Resolution**			**Cross-Resolution**		
	24 x 24	**32 x 32**	**48 x 48**	**24 x 24**	**32 x 32**	**48 x 48**
Biswas *et al.* (2013) (Multi-Dimensional Scaling) [82]	-	-	-	64.8	70.4	76.1
Bhatt *et al.* (2014) (Co-Transfer Learning) [23]	-	-	-	70.1	76.2	83.4
Wu *et al.* (2015) (LightCNN29) [171]	8.4	23.7	69.0	33.1	85.5	97.8
COTS (2016) (FaceVacs) [149]	1.7	2.9	6.5	10.3	18.5	35.7
Ghosh *et al.* (2016) (Autoencoder+SIFT) [149]	-	37.0	53.8	-	-	-
Schroff *et al.* (2015) (Triplet loss) [91]	11.1	37.9	67.8	35.3	87.5	97.4
Chen *et al.* (2017) (Quadruplet Loss) [92]	10.6	25.6	70.7	33.0	86.0	97.7
Hermans *et al.* (2017) (Hard Triplet Loss) [206]	11.2	28.9	71.7	35.6	87.7	97.2
He *et al.* (2018) (Triplet Center Loss) [207]	14.8	29.4	72.0	34.6	89.1	97.9
Yang *et al.* (2018) (DMDS) [151]	-	-	-	61.5	67.2	62.9
Yang *et al.* (2018) (LDMDS) [151]	-	-	-	62.7	70.7	65.5
Talreja *et al.* (2019) [208]	-	-	-	44.8	49.6	54.3
Li *et al.* (2019) [192]	-	-	-	20.4	20.8	31.7
Proposed SHEAL	**43.9**	**73.0**	**87.6**	**72.8**	**97.6**	**99.1**

TABLE 4.3: Rank 1 identification accuracies on the CASIA NIR-VIS 2.0 database.

Algorithm	**Cross-Spectral Cross-Resolution**		**Cross-Spectral**
	48 x 48	**64 x 64**	**128 x 128**
Wu *et al.* (2015) (LightCNN29) [171]	62.9	77.4	79.1
Schroff *et al.* (2015) (Triplet loss) [91]	67.3	81.2	82.5
Liu *et al.* (2016) (Transferable Triplet Loss) [169]	-	-	95.7
Lezama *et al.* (2017) (Face Hallucination) [184]	-	-	96.4
He *et al.* (2017) (Invariant Deep Representation) [185]	-	-	97.3
Chen *et al.* (2017) (Quadruplet Loss) [92]	68.5	81.7	83.1
Hermans *et al.* (2017) (Hard Triplet Loss) [206]	70.4	83.8	86.0
Lu *et al.* (2018) (C-SLBFLE) [209]	-	-	86.9
Huo *et al.* (2018) (K-MCMML) [170]	-	-	96.5
Proposed SHEAL	**93.8**	**97.5**	**97.6**

of the number of images, the training and testing sets have 13,617 and 109,131 video frames (of non-overlapping subjects), respectively. In order to perform CSCR-FR, night-time (NIR) probe videos have been divided into two subsets, video frames that are captured at a distance to 5-10 mts (48 × 48 resolution) and frames that are captured at a distance to 1-5 mts (64 × 64 resolution) from the camera. For CR-FR, the same matching has been performed with day-time video frames.

Labeled Faces in the Wild (LFW) Database [176] contains 13,233 images of 5,749 subjects, out of which 1,680 subjects have more than 2 images. The database is divided into views 1 and 2, where view 1 is the development set. View 2, which is the set on which results are reported has 10 folds, each of which contains 300 genuine and 300 impostor pairs. In order to perform cross-resolution face recognition experiments, low resolution images (32 × 32 and 48 × 48) are synthetically prepared (similar to the experiment on the CASIA NIR-VIS 2.0 database) for both training and testing.



((A)) CSCR-FR on SCface (24 × 24 probes)

((B)) CR-FR on SCface (24 × 24 probes)

((C)) CSCR-FR on FaceSurv (48 × 48 probes)

((D)) CR-FR on FaceSurv (48 × 48 probes)

((E)) CSCR-FR on CASIA (48 × 48 probes)

((F)) CS-FR on CASIA (128 × 128 probes)

FIGURE 4.5: CMC curves for Cross-Resolution Face Recognition (CR-FR), Cross-Spectral Face Recognition (CS-FR) and Cross-Spectral Cross-Resolution Face Recognition (CSCR-FR) on the SCface [3], FaceSurv [4] and CASIA NIR-VIS 2.0 [165] databases.

TABLE 4.4: Rank 1 identification accuracies on the FaceSurv database.

Algorithm	Cross-Spectral Cross-Resolution		Cross-Resolution	
	48 x 48	**64 x 64**	**48 x 48**	**64 x 64**
Wu *et al.* (2015) (LightCNN29) [171]	14.0	62.3	62.6	90.4
Schroff *et al.* (2015) (Triplet loss) [91]	13.2	62.5	59.1	90.1
Chen *et al.* (2017) (Quadruplet Loss) [92]	12.5	59.0	60.3	90.2
Hermans *et al.* (2017) (Hard Triplet Loss) [206]	14.1	61.7	62.9	90.0
He *et al.* (2018) (Triplet Center loss) [207]	14.2	59.8	62.4	90.5
Proposed SHEAL	**33.9**	**74.8**	**68.8**	**90.7**

TABLE 4.5: Verification accuracies at 1% False accept rate (FAR) for cross-resolution face recognition on the LFW database, with unrestricted no-outside labeled data protocol.

Algorithm	Cross-Resolution	
	32 x 32	**48 x 48**
Schroff *et al.* (2015) (Triplet Loss) [91]	58.2	87.6
Chen *et al.* (2017) (Quadruplet Loss) [92]	60.5	91.1
Hermans *et al.* (2017) (Hard Triplet Loss) [206]	62.9	92.4
He *et al.* (2018) (Triplet Center Loss) [207]	61.7	90.3
Proposed SHEAL	**64.4**	**94.2**

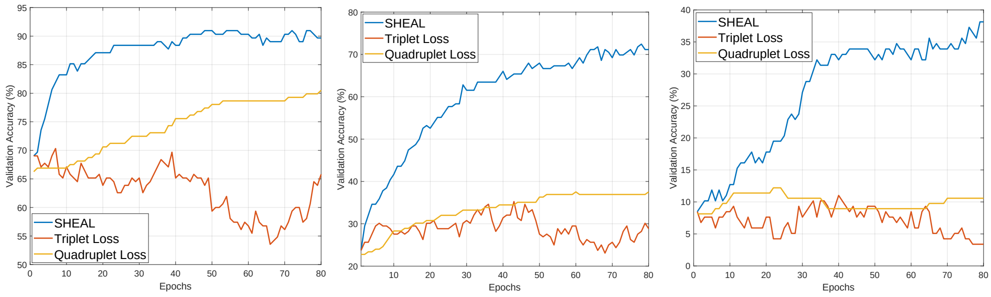
4.4.2 Experimental Results and Analysis

The proposed method is evaluated on four datasets and the results are outlined in Tables 5.1, 4.3, 5.2, and 4.5[3] and Figures 6.6 to 4.8. The experiments are performed to analyze the accuracies along with convergence analysis, ablation study, and visual inspection of results. The results are also compared with without pretraining and comparison with recent state-of-the-art algorithms.

4.4.2.1 Comparison with State-of-the-Art Methods

For the SCface [3], CASIA NIR-VIS 2.0 [165] and FaceSurv [4] databases, extensive comparisons have been performed with recent deep metric learning methods and state-of-the-art heterogeneous face recognition methods. For the SCface [3] database, CS-FR, CR-FR, and CSCR-FR experiments are performed on three resolution variations of the probes, 24 × 24, 32 × 32, and 48 × 48. As shown in Table 5.1, the proposed method achieves state-of-the-art results and outperforms popular deep metric learning and recent heterogeneous face recognition methods on all the resolutions. It can be observed that SHEAL yields larger improvement with low resolution probe images. As shown in Table 4.3, on the CASIA NIR-VIS 2.0 [165] database as well, the proposed SHEAL metric outperforms popular deep metric learning and recent cross-spectral face recognition methods on both CS-FR and CSCR-FR. On the FaceSurv [4] database we have performed CSCR-FR and CR-FR on two different probe resolutions, namely 48 × 48 and 64 × 64. As shown in Table 5.2, the proposed algorithm outperforms both Triplet [91]

[3] For CR-FR or CS-FR, wherever applicable, published results are reported. On the other hand, for CSCR-FR, we have performed the comparisons with existing algorithms using publicly available codes.



((A)) 48 × 48 ((B)) 32 × 32 ((C)) 24 × 24

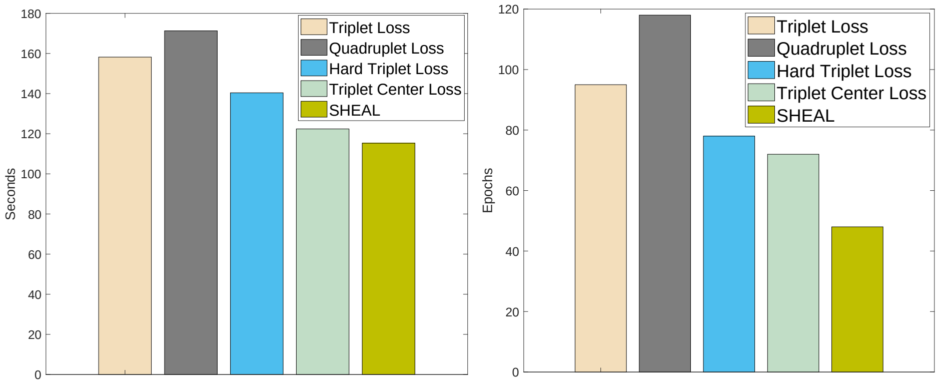
FIGURE 4.6: Convergence analysis of the proposed method on different probe resolutions of the SCface database [3]. It can be observed that the convergence of the proposed method is significantly better than the triplet [91] and the quadruplet loss [92] methods.

and Quadruplet loss [92] based methods (along with their variants), and achieves state-of-the-art results on both CSCR-FR and CR-FR experiments. As outlined in Tables 5.1, 4.3, and 5.2, the proposed SHEAL is among the top performing algorithms on all the probe resolutions. It should be noted that for lower resolutions, such as 24 × 24 in Table 5.1, the accuracy of SHEAL is much higher compared to existing algorithms. The CMC curves showcasing the identification accuracies are shown in Fig. 6.6. In addition, **homogeneous face recognition** experiment is performed on the SCface database using the same model (that is trained using the SHEAL metric). The proposed model achieves rank 1 accuracy of 97.29% on CR-FR for 32 × 32 probes on the SCFace database.

Finally, on the LFW face database [176], CR-FR experiment is performed and the results are documented in Table 4.5. The results show that with gallery images of size 128 × 128 and probe images of 32 × 32 or 48 × 48, the proposed algorithm is at least 1.5% better than other deep metric learning algorithms. On the LFW database, comparisons have been performed with popular deep metric learning methods, and the proposed method outperform them for CR-FR scenario on this database on two different probe resolutions. Note that due to image size variations to conduct CR-FR experiments, we cannot directly compare with reported results on the LFW database.

4.4.2.2 Convergence Analysis

Figure 4.6 shows the rate of convergence of SHEAL, triplet loss, and quadruplet loss on the SCface dataset. It can be observed that the convergence of SHEAL is significantly fast and effective. The validation accuracy of the model trained using SHEAL reaches to 83.23% from 69.03% (on 48 × 48 probes) in just 10 epochs (Fig. 4.6(a)). Compared to SHEAL, the quadruplet and triplet losses converge slowly. Figures 4.7(a) and (b) show the time taken and the number of epochs required to converge, respectively. The number of epochs required by SHEAL to converge is only 48 compared to 95 and 118 epochs required by triplet and quadruplet loss for



((A)) Training time

((B)) Epochs for convergence

FIGURE 4.7: Performance analysis of the proposed method: (a) Time taken to converge (training) and (b) Number of epochs for convergence. It can be observed that the proposed algorithm not only converges rapidly, but also takes much lesser time and epochs for the same. Training is performed on the SCface database.

the same. In terms of total time, SHEAL takes 115.3 seconds against 158.2, 171.3, 140.4 and 122.3 seconds required by triplet loss, quadruplet loss, hard triplet loss and triplet center loss respectively for convergence. These results suggest that the proposed SHEAL converges rapidly, takes lesser time, and exhibits significantly higher face recognition accuracies in heterogeneous settings.

4.4.2.3 Ablation Study

We have performed two separate ablation studies for a thorough understanding of the effect of the loss functions (Equations 4.7 and 4.8) on the trained model's performance on CR-FR and CSCR-FR scenarios. As illustrated in Section 4.3.1, Equation 4.7 is composed of four separate terms: L^c_{Ho}, L^c_{CR}, L^c_{CS} and L^c_{CS-CR}. We have performed an ablation study on Equation 4.7, where we have utilized these specific terms for training the models separately. Each of these terms have a disjoint effect on the trained model which is evident in the results obtained on the SCface database (Table 4.6). The model, when trained only with the L^c_{Ho} loss term yields the worst performance. However, when L^c_{CR} and L^c_{CS} loss terms are used separately for training, the corresponding testing performance (eg. when L^c_{CR} term is used for training the cross-resolution performance is improved during testing) is improved. It can be observed that when training is performed with L^c_{CS-CR}, the results, during testing, are improved considerably for both CS-CR and CR face recognition.

In addition to the above, we have also performed an ablation study on Equation 4.8 (subclass based cluster optimization). As illustrated in Section 4.3.2, Equation 4.8 is composed of three terms. The first term is a homogeneous matching term, the second term makes subclass containing the low resolution and NIR images more compact, and the third terms brings the subclasses

TABLE 4.6: Rank 1 identification accuracies (%) for the ablation study on Equations 4.7 and 4.8 performed on the SCface database.

Loss Term		**CS-CR**			**CR**		
		24 x 24	**32 x 32**	**48 x 48**	**24 x 24**	**32 x 32**	**48 x 48**
Eq. 7	L^c_{Ho}	9.5	29.4	65.1	37.0	58.9	63.4
	L^c_{CR}	18.6	39.4	80.4	70.9	96.4	98.7
	L^c_{CR}	12.5	31.2	69.0	40.5	51.2	64.3
	L^c_{CS-CR}	41.3	72.6	85.9	68.7	94.3	99.0
	$L^c_{Ho} + L^c_{CR} + L^c_{CR}$	38.4	68.4	83.9	64.2	92.6	98.2
Eq. 8	$1^{st}term + 2^{nd}term$	37.2	67.1	78.9	64.3	92.1	85.4
	$1^{st}term + 3^{rd}term$	42.1	72.5	87.3	70.2	96.4	98.4
	Proposed	**43.9**	**73.0**	**87.6**	**72.8**	**97.6**	**99.1**

closer into one compact cluster. As shown in Table 4.6, we observe that the third term is the major contributing factor in the subclass optimization stage. The value of β_3 is also kept higher during this optimization stage, to give more weight to the third term of Equation 4.8.

4.4.2.4 Loss Function Coefficients

For training using SHEAL and the cluster optimization phase (Equations 4.7 and 4.8), we have the loss function coefficients (λ_1, λ_2, λ_3 and λ_4), which can be used to adjust the weight of each component of the loss function. Since homogeneous matching is a less challenging problem compared to CSCR-FR, CS-FR and CR-FR, we kept a considerably lower value for λ_1 than the other weight terms. For the SCface [3] database, we kept a much higher value for λ_4 since CSCR-FR matching is an extremely challenging problem.

4.4.2.5 Without Pretraining

In order to make a fair comparison, the other deep metric learning methods with which we have compared in the paper ([92], [91] and their variants) have been trained on the same data using the weights of the same pretrained model. In addition to this, we trained our method from scratch (on a randomly initialized model), and achieved 90.71% accuracy wheras those obtained by Chen *et al.* [92] and Schroff *et al.* [91] are 82.13% and 80.34% respectively, on 64×64 probes of the CASIA NIR-VIS 2.0 database. It shows that even without pretraining, the proposed method outperforms the most popular deep metric learning algorithms.

4.4.2.6 Visual Inspection of the Results

We also performed visual inspection of the results and some cases are presented in Fig. 4.8. It can be observed that images of the SCface database which have extremely low resolution

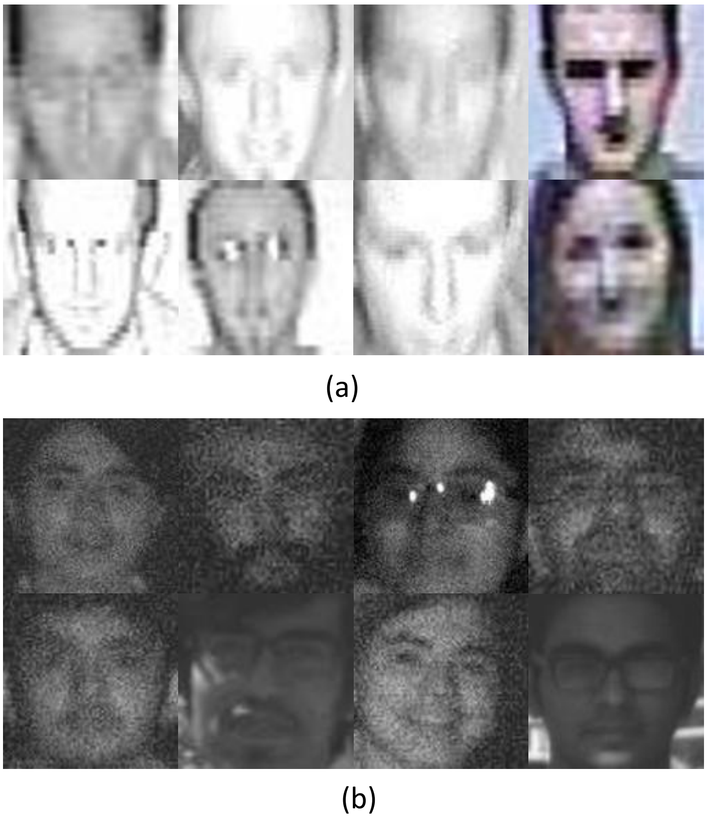


FIGURE 4.8: Sample images of some extremely noisy and poor quality images of (a) SCface [3] and (b) FaceSurv [4] databases that are correctly classified by the model trained with SHEAL, but were incorrectly classified using triplet [91], quadruplet [92] and triplet center loss [207] based methods.

and quality (Fig. 4.8(a)) are correctly classified by the proposed algorithm. On the other hand, the images in the FaceSurv database, which in addition to low resolution suffer from heavy motion blur and poor illumination, are also correctly classified by the proposed algorithm. These results showcase the potential applicability of the proposed algorithm to real world surveillance scenarios.

4.5 Summary

The problem of heterogeneous face recognition is compounded when test data shows multiple heterogeneity. Current deep metric learning approaches generally do not handle such heterogeneous problems and yield poor recognition accuracies. This paper introduces a subclass heterogeneity aware loss function which is utilized to train a discriminative model to generate heterogeneity invariant embeddings. This helps to project a pair of face images of different covariates into an embedding space where matching can be performed efficiently irrespective of the images being captured in very different scenarios. This can be applied in a surveillance application where the data may encompass multiple covariates. In future, we plan to extend the proposed algorithm to include other covariates of face recognition such as disguise and aging along with multiple heterogeneous variations.

Chapter 5

On Learning Density Aware Embeddings

5.1 Introduction

Classification models such as Convolutional Neural Networks (CNN) utilize deep metric learning based loss function for learning discriminative embeddings. The loss function attempts to bring the embeddings of the same classes close to each other in the output manifold. In this embedding space, a direct computation of the distance gives the dissimilarity score between the two images. Several different applications have investigated the use of deep metric learning algorithms such as person re-identification [205, 206, 210], 3D object retrieval [207], biometric recognition [89, 89–91, 211], robot perception [212], patch matching [213, 214], and object recognition [215, 216].

In the literature, very efficient deep metric learning methods have been proposed such as triplet loss [91] and quadruplet loss [92]. However, a major limitation of these loss functions is their heavy dependence on mining of hard samples for training [91, 95, 206, 210]. In the triplet Loss [91], for N training classes and K samples in each class, the total number of triplets for training can be as high as $N(N-1)K^2(K-1)$, which increases the training time on large datasets significantly. Another limitation of these methods is slow convergence, this heavily depends on the appropriate choice of the training curriculum. Further, the presence of outliers (noisy/poor-quality samples) in the training data, and their participation in triplets may hurt the training process. To the best of our knowledge there has been no study to understand the effect of outliers and density distribution of the training data on the performance of deep metric learning algorithms.

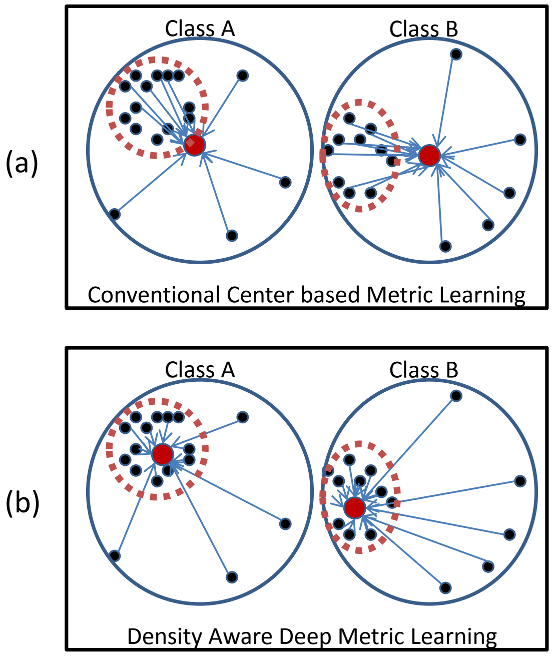

(a)
Class A
Class B
Conventional Center based Metric Learning
(b)
Class A
Class B
Density Aware Deep Metric Learning

FIGURE 5.1: Illustrating the difference in the conventional and proposed metric learning techniques. (a) Conventional center loss based deep metric learning algorithms pull the data of a class towards the centroid of that class. (b) The proposed density aware deep metric learning algorithm pulls the samples of every class towards the most dense region of the respective clusters.

As seen in Figure 5.1(a), conventional center loss based deep metric learning methods [200, 207] generate embeddings of each class that lie closer to the centroid of the samples of that particular class. However, they do not take into account the distribution of the training data. In cases where outliers are present, the convergence of such methods on large databases can be slow and the outliers/noisy training samples can adversely affect the training of a discriminative model. In order to mitigate this challenge, the proposed algorithm minimizes the effect of outliers by calculating the center, taking into account the most dense region of the respective clusters for each class (Figure 5.1(b)). Using the philosophy of the classical mean-shift algorithm [217], the estimate of the mean is shifted to a denser region from the initial estimate of the centroid. This shifted center embedding is used for learning a discriminative model. The research contributions of the paper can be summarized as follows:

- The proposed density aware deep metric learning algorithm provides a generalized framework which can be augmented with any deep metric learning method for effective training especially with noisy data.
- Detailed analysis and comparison with other popular deep metric learning methods on four challenging databases pertaining to face and object images show that the proposed approach gives better recognition accuracies, exhibits superior convergence with reduced training time and is resilient to noisy training data.

5.2 Related Work

Hadsell [88] proposed the contrastive loss, which was one of the first deep metric learning methods for training a discriminative model with a deep neural network. They used a single loss function to pull positive pairs and push negative pairs in the output embedding space of the model. This method of training a discriminative neural network, popularly known as the Siamese Network, resulted in several extensions [89, 90, 215] which produced excellent results on a variety of image recognition problems. Recently, one of the most popular methods for deep metric learning is the triplet loss [91]. The triplet loss enforces the model to learn an embedding space where samples of similar classes are mapped closer to each other and that of other classes are pushed away. Wen [200] used a combination of the softmax and the center loss for face recognition. Later, Chen [92] proposed the quadruplet loss which used an extra negative sample in addition to the anchor, positive and the negative sample that were utilized by the triplet loss. They showed that the extra negative term helps to train a more generalizable model. Thereafter, several methods have attempted to improve upon the triplet and the quadruplet loss based methods. Yuan [95] proposed an ensemble based technique for mining hard examples which are used for training a deep network using the contrastive loss. Hermans [206] proposed a triplet mining technique, by selecting the k hardest positive samples and k hardest negative samples for each anchor image in a batch of N randomly sampled images from the training set. Recently, He [207] proposed the triplet center loss where the center of the set of anchors and the center of the nearest negative cluster were utilized in the loss function of the triplet loss, for person re-identification.

5.3 Density Aware Metric Learning

The proposed method presents a novel contribution to the deep metric learning paradigm by incorporating the density of data in the clusters during training. Before delving into the detailed formulation, a brief illustration of the background is discussed.

5.3.1 Background

In a classical pattern classification scenario, data $\vec{Z}$ from n different classes is available, $\vec{Z} = \{z_1, z'_1, z_2, ..., z_i, z'_i, ...z_n\}$, where z_i and z'_i are two images of the same class i. Let the i^{th} class contain n_i number of training samples. The goal of a deep metric learning algorithm is to learn a function $g_\theta(z) : \mathbb{R}^S \longrightarrow \mathbb{R}^T$ where S is the dimensionality of the source data manifold, T is the dimensionality of the output embedding space of the model g, and θ represents the trainable parameters of the model. For illustration, let $\{x, y\}$ be the pair of points on the embedding

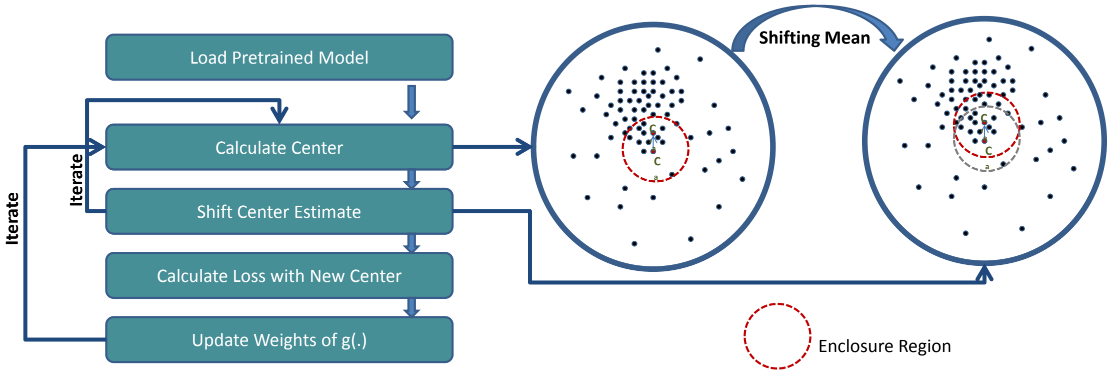


FIGURE 5.2: The proposed algorithm iteratively finds the estimate for the center in the most dense region of the cluster. This center, when used with a deep metric learning algorithm, is expected to provide effective training and better convergence. (best viewed in color).

manifold of the model g. The distance metric function is defined as:

$$D\{x, y\} : \mathbb{R}^T \times \mathbb{R}^T \longrightarrow \mathbb{R} \tag{5.1}$$

In this paper, Euclidean distance is used as the distance metric, which can be defined as:

$$D\{x, y\} = \|g_\theta(x) - g_\theta(y)\|_2^2 \tag{5.2}$$

Inspired from Large Margin Nearest Neighbor Classification [218], a typical deep metric learning loss L can be used which is minimized by pulling intra-class embeddings together into one cluster and pushing the inter-class embeddings. From the training set Z, a 3-tuple is formed using three images, z_a, which is a sample of the class a, a positive sample z'_a, which is another image of the same class a, and a negative sample z_b which is an image of another class b. The loss function can be expressed as:

$$L = \left[D\{\vec{Z}_a, \vec{Z'}_a\} - D\{\vec{Z}_a, \vec{Z}_b\} + \alpha \right]_+ \tag{5.3}$$
$$\forall (\vec{Z}_a, \vec{Z'}_a, \vec{Z}_b) \in \tau$$

where, τ is the set of all 3-tuples in the training data, $[f]_+ = max(f, 0)$, α is the margin parameter and $\vec{Z}_a$, $\vec{Z'}_a$ and $\vec{Z}_b$ are sets of all the anchors, positive and negative samples, prepared from the training set.

5.3.2 Proposed Formulation

In order to present the proposed approach, the standard loss metric (Equation 6.2) is re-formulated where the anchor z_a is replaced with the center of class a. The center embedding C_a is calculated as the mean of all the embeddings of class a. Thus, the loss function may be expressed as:

$$L = \Big[D\{C_a, \vec{Z'}_a\} - D\{C_a, \vec{Z}_b\} + \alpha\Big]_+ \tag{5.4}$$

$$where \qquad C_a = \frac{\sum_{n_a} g(z_a)}{n_a}$$

C_a represents the centroid of the cluster corresponding to class a, containing n_a training samples. However, depending on the density of the cluster (as shown in Figure 5.2), the mean-shift algorithm [217] may be applied to iteratively arrive at the mean driven by non-parametric density estimation of the cluster.

5.3.2.1 Shifting the Mean to a Denser Region

Reiterating, C_a is the initial centroid of the cluster which is calculated by taking the mean of all the embeddings of class a. Now, from C_a, selecting the nearest p points (in the embedding manifold of the model g) out of all the n_a points in the cluster corresponding to class a, we take the mean of only these p points, where $p < n_a$. We term the region of the embedding manifold containing these p points around the centroid (the red dotted circle in Figure 5.2) as the **enclosure region**, and the set of these p points as **enclosure points**. The estimate for the new center C'_a can be calculated as,

$$C'_a = \frac{\sum_{i=1}^{p} g(z_a^i)}{p} \qquad \forall z_a^i \in \{z_a^1, z_a^2 ... z_a^p\} \tag{5.5}$$

where, z_a^i is the i^{th} point inside the enclosure region. Figure 5.2 shows the new mean C'_a, which is expected to be in a denser region of the cluster. The difference of the new mean C'_a and the old mean C_a gives the mean shift vector which can be expressed as:

$$V_a = \left[\frac{\sum_{i=1}^{p} g(z_a^i)}{p} - \frac{\sum_{n_a} g(z_a)}{n_a}\right] \tag{5.6}$$

This process is repeated iteratively until the mean shift is negligible, thus leading to convergence.

5.3.2.2 Weighted Mean Shift

The above calculation of the centroid does not take into account any weightage of the points around the mean that are considered. In order to give importance to the points nearer to the centroid, we can use a weight coefficient W_i for every point i in the *enclosure region*. The k^{th} estimate of the center with respect to weights W_i can be calculated as,

$$C_a^k = \frac{\sum_{i=1}^{p} W_i C_a^{k-1} g(z_a^i)}{\sum_{i=1}^{p} W_i C_a^{k-1}} \quad \forall z_a^i \in \{z_a^1, z_a^2 ... z_a^p\} \tag{5.7}$$

C_a^{k-1} being the $(k-1)^{th}$ estimate of the mean. The corresponding mean shift vector may be expressed as,

$$V_a^k = \left[\frac{\sum_{i=1}^{p} W_i C_a^{k-1} g(z_a^i)}{\sum_{i=1}^{p} W_i C_a^{k-1}} - C_a^{k-1}\right] \tag{5.8}$$

Here, p is the number of *enclosure points* for the k^{th} iteration, and z_a^i is the i^{th} data point for the class a.

5.3.3 Selecting weights using a Kernel Density Estimate (KDE)

In order to select weights W_i for each point i in the cluster represented by the centroid C_a for a particular class a, we can use a kernel density estimate that are generally used by non-parametric density estimation techniques. A uniform kernel for selecting the weights can be expressed as:

$$W_i = \begin{cases} c & if \left\|C_a - z_a^i\right\| < f \\ 0 & otherwise \end{cases} \tag{5.9}$$

where, $\left\|C_a - z_a^i\right\|$ gives the distance of the point z_a^i from the cluster centroid C_a for class a. The uniform kernel assigns a weight c to the point z_a^i if it is within the *enclosure region*. The *enclosure region* has a radius of f, thus all the points which are at a distance of f or less from the centroid C_a are assigned the same weight c. Instead of directly using a parameter for the radius of the *enclosure region*, p nearest *enclosure points* can also be considered out of all the points in the cluster for class a. Algorithm 6.3 outline the steps of the proposed approach using the triplet loss.

5.4 Density Aware Deep Metric Learning in Triplet and Quadruplet Loss

The proposed Density Aware Metric Learning is a generic formulation and can be incorporated into any deep metric learning loss function. Here, we present the formulations of triplet and

Algorithm 2: Density Aware Triplet Loss.

Input: CNN model g_θ, training data $\{\vec{Z}\}$
Output: Trained model g_θ
Parameters: e (epochs), θ (parameters of g), m (batch size), k (number of batches) p (number of enclosure points) s (mean shift iterations), t^p (threshold for hard positive selection), t^n (threshold for hard negative selection), f (radius of enclosure region)

```
for Epoch=1 to e do
    Generate Triplets:
    Initialize: $X = \{\}$ (empty set of selected Hard Triplets)
    Initialize: $Pool = \{\}$ (empty pool of samples)
    for every class $a = 1$ to $n$ do
        Select b images randomly from class $a$
        $Pool = Pool \cup$ selected images
    end
    for each image $z_a^i$ of each class $a$ in $Pool$ do
        Select $z_a^i$ as the anchor image
        for each image $z_l^y$ in $Pool$ such that $z_l^y \neq z_a^i$ do
            if $a = l$ and $D\{z_a^i, z_l^y\} > t^p$ then $X = X \cup z_l^y$
            if $a \neq l$ and $D\{z_a^i, z_l^y\} < t^n$ then $X = X \cup z_l^y$
        end
    end
    Calculate Center:
    $C_a = \frac{\sum_{n_a} g(z_a)}{n_a}$
    Shift Center:
    for every class $a$ do
        for k=1 to s do
            $W_i = K_u(C_a^{k-1} - z_a^i) = \begin{cases} c & if \left\| C_a^{k-1} - z_a^i \right\| < f \\ 0 & otherwise \end{cases}$
            $C_a^k = \frac{\sum_{i=1}^{p} W_i C_a^{k-1} g(z_a^i)}{\sum_{i=1}^{p} W_i C_a^{k-1}} \quad \forall z_a^i \in \{z_a^1, z_a^2 ... z_a^p\}$
            $V_a^k = \left[ \frac{\sum_{i=1}^{p} W_i C_a^{k-1} g(z_a^i)}{\sum_{i=1}^{p} W_i C_a^{k-1}} - C_a^{k-1} \right]$
        end
    end
    Generate embeddings
    for every batch of size $m$ do
        Forward pass through $g$ to find $g_\theta(Z_a), g_\theta(Z'_a), f_\theta(Z_b)$
        Calculate loss $L$
        $L = \sum_m \left[ \left\| C_a^s - g(\vec{Z'}_a) \right\|_2^2 - \left\| C_a^s - g(\vec{Z}_b) \right\|_2^2 + \alpha \right]$
        Calculate gradient
        $\triangle W = \nabla_\theta \frac{1}{m} \sum_m L$
        Update weights of $g_\theta$ using $\triangle W$
    end
end
```

quadruplet loss based density aware metric learning:

5.4.1 Density Aware Triplet Loss (DATL)

Schroff [91] proposed the triplet loss based deep metric learning technique where the loss L is minimized by the same philosophy as discussed in Section 5.3.1. From the training set Z, a triplet is formed using an anchor z_a, which is an image of the class a, a positive sample z'_a, which is another image of the same class a, and a negative sample z_b which is an image of another class b. The loss function is expressed as,

$$L = \left[\left\| g(\vec{Z}_a) - g(\vec{Z'}_a) \right\|_2^2 - \left\| g(\vec{Z}_a) - g(\vec{Z}_b) \right\|_2^2 + \alpha \right]_+ \tag{5.10}$$

$$\forall (\vec{Z}_a, \vec{Z'}_a, \vec{Z}_b) \in \tau$$

where $\vec{Z}_a$, $\vec{Z'}_a$ and $\vec{Z}_b$ are sets of all anchors, positive and negative samples, respectively, and τ is the set of all triplets in the training data. Using the proposed approach, the anchor is replaced with the center which is iteratively determined (C_a or C'_a, and so on) with an appropriate kernel density estimate. The loss function for the Density Aware Triplet Loss (DATL) is as follows:

$$L = \left[\left\| C_a - g(\vec{Z'}_a) \right\|_2^2 - \left\| C_a - g(\vec{Z}_b) \right\|_2^2 + \alpha \right]_+ \tag{5.11}$$

5.4.2 Density Aware Quadruplet Loss (DAQL)

The triplet loss is extended by Chen [92] as the quadruplet loss where a second negative image z_c is introduced. The loss function for the same in the proposed density aware paradigm can be expressed as,

$$\begin{aligned} L = & \left[\left\| C_a - g(\vec{Z'}_a) \right\|_2^2 - \left\| C_a - g(\vec{Z}_b) \right\|_2^2 + \alpha_1 \right]_+ \\ & + \left[\left\| C_a - g(\vec{Z'}_a) \right\|_2^2 - \left\| C_a - g(\vec{Z}_c) \right\|_2^2 + \alpha_2 \right]_+ \end{aligned} \tag{5.12}$$

$$\forall (\vec{Z}_a, \vec{Z'}_a, \vec{Z}_b, \vec{Z}_c) \in \vartheta$$

where ϑ is the set of all the quadruplets prepared from the training set.

5.4.3 Experimental Setup and Implementation

The deep CNN architecture by Wu [163] is utilized to learn a discriminative model with the proposed loss function. The weights are initialized from a network that is pretrained on the MS-Celeb 1M dataset. The model has 17 convolutional layers, along with 10 Max-Feature-Map layers. The network has two fully connected layers at the end, producing embeddings of dimensionality 256. Training is performed using the Adam optimizer. The batch size is kept at 60 and the learning rate of 10^{-3} is used which is decreased gradually till 10^{-7}. Hard mining is performed (steps 7-11 of Algorithm 6.3) for all the variants of triplet and quadruplet losses according to the *Batch Hard* scheme proposed by Hermans [206]. The hard mining is only performed at the end of training (once the learning plateaus) to accelerate the training process. All the codes are implemented using the Pytorch platform on a machine with Intel Core i7 CPU, 64GB RAM and NVIDIA GTX 1080Ti GPU.

5.5 Experiments

The proposed algorithm is evaluated on the SCface [3] and FaceSurv [4] datasets for cross-modal face matching, and on the CIFAR10 [5] and STL-10 [6] datasets for object recognition.

5.5.1 Datasets

Details of the databases and experimental protocols are described in this section.

SCface [3] is a face dataset containing poor quality face images captured from surveillance cameras in indoor environment. The database contains 4160 images of 130 subjects. The images are captured with eight different cameras, out of which two cameras operated in night-vision mode and one camera is operated in Near-Infrared mode. The images are taken from three different stand-off distances namely 4.2 mts, 2.6 mts, and 1 mt. Out of the 130 subjects, images of 50 subjects are used for training and the remaining are used for testing. The classes/subjects in the train and test set are non overlapping.

FaceSurv [4] is a video face database where the subjects walk towards the camera from a distance of about 10 mts. It has 396 daytime and 365 nighttime videos of 240 subjects. The nighttime videos are captured in complete darkness with an NIR illuminator. Each video has about 200 frames on an average. Each subject has three gallery images which are captured in controlled scenarios from a standoff distance of 1 mt. Videos of only 39 subjects are used for training and the remaining are used for testing. The classes/subjects in train and test sets are disjoint.

CIFAR-10 [5] is a popular object recognition dataset consisting 60,000 images of 10 classes. The resolution of the images is 32×32. The training set contains 50,000 images (5,000 images of each class) and 10,000 images (1,000 per class) comprise the testing set.

STL-10 [6] contains 113,000 images of 10 different objects. The resolution of the images is 96×96. The total number of images for training and testing are 5,000 (500 per class) and 8,000 (800 per class), respectively. The remaining images are unlabeled and have not been used for the experiments.

5.5.2 Evaluation Criteria

In this work, two different kinds of experiments have been performed: cross-modal face recognition (identification) and object recognition (retrieval). For face identification, the test set is partitioned into probe (query images) and gallery (reference set/database) sets. For every image of the probe set, matching is performed with each image of the gallery by a forward pass through the learned model (g_θ), followed by computing Euclidean distance between the embeddings of the probe and the gallery images to calculate the match score. Rank $\mathcal{K}$ accuracy is the ratio (multiplied by 100 to get a percentage) of the number of times the correct class is among the top $\mathcal{K}$ matches to the number of matching attempts (once for each probe image). Recall @ $\mathcal{K}$ is the average recall score for all the query images. Following the definition by Song [219], the recall score is one if the relevant class is retrieved in the top $\mathcal{K}$ matches with the gallery/database set, and is zero otherwise.

5.6 Results

The experiments have been performed by partitioning each database into the train and test sets. The CNN model is trained on the train set using the proposed density aware deep metric learning, i.e. the Density Aware Triplet Loss (DATL) and the Density Aware Quadruplet Loss (DAQL). Comparisons have been performed with the vanilla triplet and quadruplet losses, (their variants for cross-modal matching are implemented for the SCface and FaceSurv databases). In addition, the proposed algorithm is also compared with hard triplet loss [206] and recently proposed triplet center loss [207]. The former is a variant of the vanilla triplet loss using a moderate hard mining approach. Triplet center loss is a formulation which mimics the conventional center based triplet loss previously discussed (Equation 5.4).

For face recognition, Rank 1 accuracies for three different probe resolutions, namely 48×48, 32×32 and 24×24 are reported. As shown in table 5.1 on 32×32 and 24×24 resolutions, the proposed algorithm produces state-of-the-art results for the SCface database. It outperforms the vanilla triplet and the quadruplet losses and their variants on the FaceSurv database (Table 5.2) as well. Moreover, on the SCface database, we report published results from different cross-modal face recognition methods. As shown in Table 5.1, it can be observed that the proposed algorithm outperforms these existing algorithms, specifically for lower resolution levels.

For the object retrieval task, experiments are performed on the CIFAR-10 and STL-10 datasets. As shown in Table 5.3, on the CIFAR-10 dataset, the proposed algorithm outperforms both the triplet and the quadruplet losses and their variants for recall @ 1 and recall @ 10. However, for recall @ 100 it produces competitive accuracy with respect to the other algorithms. As shown

TABLE 5.1: Summarizing the results of face identification on the SCface [3] Database.

Method		**Identification (%) (Rank 1)**		
		24 x 24	**32 x 32**	**48 x 48**
MDS [82]		64.87	70.48	76.14
Co-Transfer Learning [23]		70.14	76.29	83.47
Res-Net [200]		36.30	81.80	94.30
Coupled Res-Net [150]		73.30	93.50	98.00
VGGFace [150]		41.30	75.50	88.80
Coupled VGGFace [150]		62.30	91.00	94.80
Coupled Light-CNN [150]		50.50	85.00	94.00
Triplet loss [169]		70.69	95.42	97.02
Quadruplet loss [92]		74.00	96.57	98.41
Hard triplet loss [206]		72.65	96.12	98.05
Triplet Center Loss [207]		75.45	96.10	**98.50**
Discriminative MDS [151]		62.70	65.50	70.70
Proposed	DATL	**76.24**	**96.87**	98.09
	DAQL	**77.25**	**96.58**	98.14

TABLE 5.2: Summarizing the results of face identification on the FaceSurv [4] database.

Method		**Identification (%) (Rank 1)**		
		24 x 24	**32 x 32**	**48 x 48**
Triplet loss [169]		18.0	38.5	72.5
Quadruplet loss [92]		16.4	38.5	77.9
Hard triplet loss [206]		17.8	40.8	78.9
Triplet Center Loss [207]		18.4	41.5	82.7
Proposed	DATL	**20.4**	**42.8**	**85.9**
	DAQL	**21.3**	**48.6**	**85.6**

TABLE 5.3: Summarizing the results of object retrieval on the CIFAR-10 [5] database.

Method		**Recall @ $\mathcal{K}$ (%)**		
		$\mathcal{K}$ **= 1**	$\mathcal{K}$ **= 10**	$\mathcal{K}$ **= 100**
Triplet loss [169]		76.24	94.78	97.21
Quadruplet loss [92]		78.35	95.40	**98.99**
Siamese+Triplet [212]		78.62	92.57	97.19
Hard triplet loss [206]		78.51	93.87	97.41
Triplet Center Loss [207]		79.41	96.10	95.78
Proposed	DATL	**80.34**	**96.68**	97.84
	DAQL	**80.81**	**96.12**	97.58

in Table 5.4, on the STL-10 dataset, the proposed algorithm outperforms the vanilla triplet and quadruplet losses along with their variants on recall @1, 10 and 100.

5.7 Analysis and Discussion

This section analyzes the performance of the proposed algorithm with respect to training with noisy data, convergence, training time, and parameters.

TABLE 5.4: Summarizing the results of object retrieval on the STL-10 [6] database.

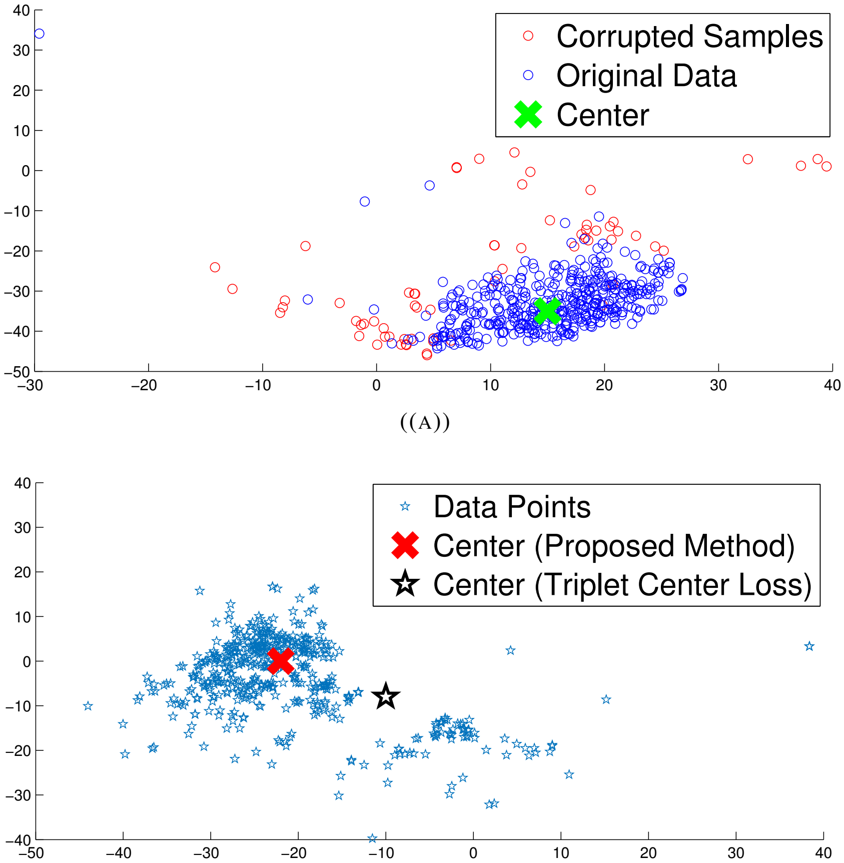
Method		**Recall @ $\mathcal{K}$ (%)**		
		$\mathcal{K}$ = 1	**$\mathcal{K}$ = 10**	**$\mathcal{K}$ = 100**
Triplet loss [169]		72.47	78.54	80.41
Quadruplet loss [92]		73.98	78.71	81.77
Siamese+Triplet [212]		73.62	77.15	81.34
Hard triplet loss [206]		72.95	76.08	81.90
Triplet Center Loss [207]		74.61	77.98	81.59
Proposed	DATL	**75.27**	**79.08**	**82.38**
	DAQL	**75.84**	**80.17**	**83.74**

TABLE 5.5: Results on the STL-10 database after adding noisy training data for every class.

Method		**Resolution of Noisy Samples**			
		24 x 24		**32 x 32**	
		Recall @ $\mathcal{K}$ (%)			
		$\mathcal{K}$=1	**$\mathcal{K}$=10**	**$\mathcal{K}$=1**	**$\mathcal{K}$=10**
Triplet Loss [91]		54.12	58.00	68.45	72.58
Quadruplet Loss [92]		56.51	59.77	68.74	73.01
Hard Triplet Loss [206]		58.29	60.41	65.37	72.91
Triplet Center Loss [207]		62.76	65.40	68.76	73.16
Proposed	DATL	**66.52**	69.45	**69.87**	73.21
	DAQL	65.47	**69.80**	68.52	**75.40**

5.7.1 Effect of Noisy Data during Training

One of the primary properties of the proposed method is the ability to ignore outliers during training. Such outliers may often be represented by noisy data (low resolution/quality and poor illumination). As shown in Figure 5.3(a), these noisy data samples affect the training process of conventional deep metric learning based algorithms. Since the proposed method computes the cluster center only by using the points inside the enclosure region, the outliers are effectively ignored. On the other hand, conventional deep metric learning algorithms would consider all the points (including outliers) which may lead to unnecessary jitter in the convergence during training. An experiment is performed on the STL-10 database by replacing 15% of samples from each class by low resolution variants (32 × 32 and 24 × 24 as two separate experiments) of the same. Such training samples are expected to be outliers and thus may have potential to hurt the training process. We use a no reference image quality score (BRISQUE [168]) (a lower score implies better image quality) for the original training samples of STL-10 which is 33.90 (average for training set). For the noisy samples, the score is 45.14 (for 24 × 24) and 41.83 (for 32 × 32). This infers that the low resolution data are of lower quality. As shown in Table 5.5, the proposed methods perform better than conventional deep metric learning techniques when noisy data is introduced in the training process. It also exhibits that performance improvement is greater for the experiment where higher amount of data corruption (adding 24 × 24 images) is performed.



((A))

((B))

FIGURE 5.3: (a) tSNE Visualization of noisy samples (for a particular class), which shows that most of the noisy samples are outliers (b) Center computed by the proposed method is in the dense region of the class while the conventional center is away from the dense region. Visualization is on the STL-10 database, one particular class is shown for illustrative brevity (best viewed in color).

5.7.2 Size of the Enclosure Region

One important parameter of the proposed algorithm is the size of the enclosure region. For implementation, the enclosure region is determined by taking the nearest k% points from the current center embedding point. A region of 20% signifies that the nearest 20% points (with respect to all the points of the particular cluster) from the current center are considered to be inside the enclosure region. Figure 5.4 shows the results for the proposed Density Aware Triplet Loss on the STL-10 and CIFAR-10 databases for four different enclosure regions. It can be seen that an enclosure region of 17% yields optimal results while larger or smaller enclosure region results in reduced accuracy on both the databases.

5.7.3 Training Time

Owing to better convergence properties of the proposed DATL, total training time required is much less as compared to the vanilla triplet loss and its variants. As shown in Figure 5.5(a),

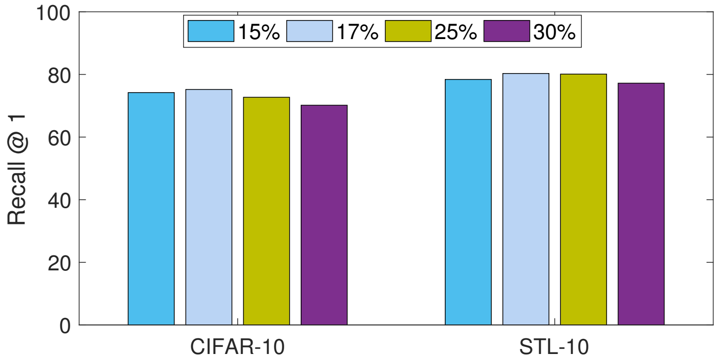
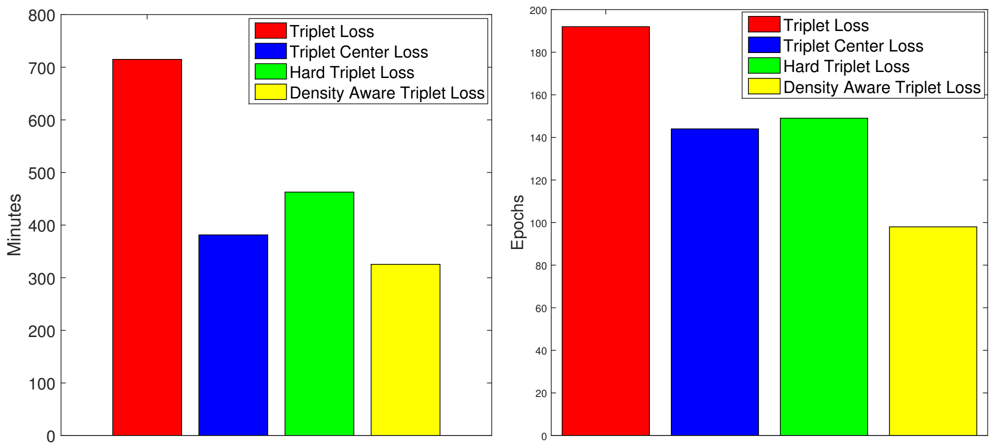


FIGURE 5.4: Effect of the size of the enclosure region on the proposed Density Aware Triplet Loss on the STL-10 and CIFAR-10 databases (best viewed in color).



((A)) ((B))

FIGURE 5.5: Total (a) training time and (b) epochs for training on the STL-10 database (best viewed in color).

the total time needed to train the proposed algorithm is 325.4 minutes. On the other hand, the vanilla triplet loss, triplet center loss, and the hard triplet loss requires 714.9, 381.4, and 462.7 minutes, respectively on the STL-10 dataset. In terms of the number of epochs as well, the proposed density aware triplet loss requires much lesser number of epochs (98 epochs), whereas the triplet loss, triplet center loss, and the hard triplet loss takes 192, 144, and 149 epochs, respectively.

5.7.4 Convergence

The foremost advantage of the proposed density based deep metric learning approach is its ability to converge quickly as compared to the vanilla triplet and quadruplet loss methods. In addition, the proposed algorithm also converges much faster with respect to the triplet center loss [207] which uses the centroid of the cluster in the loss function (Equation 5.4). The proposed method avoids outliers and thus updates the weights of the model in such a way, so as to

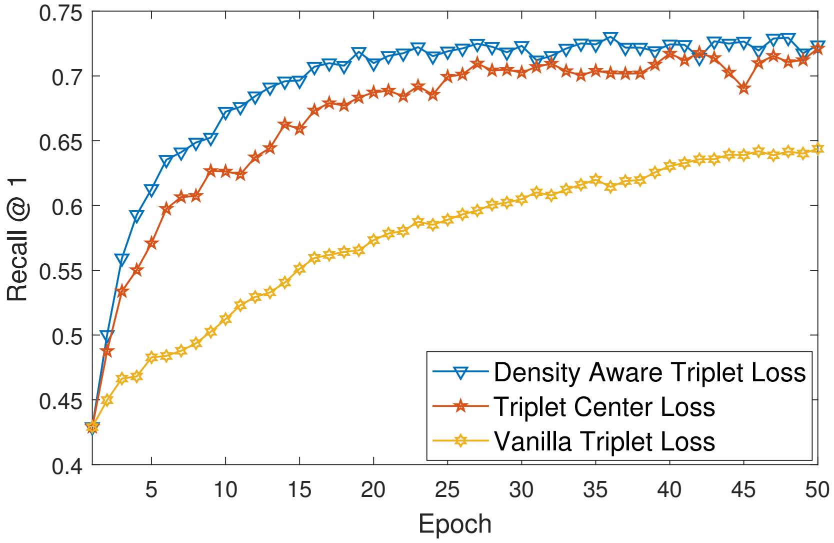


FIGURE 5.6: Convergence of the proposed DATL compared with the vanilla triplet loss and triplet center loss on the STL-10 dataset.

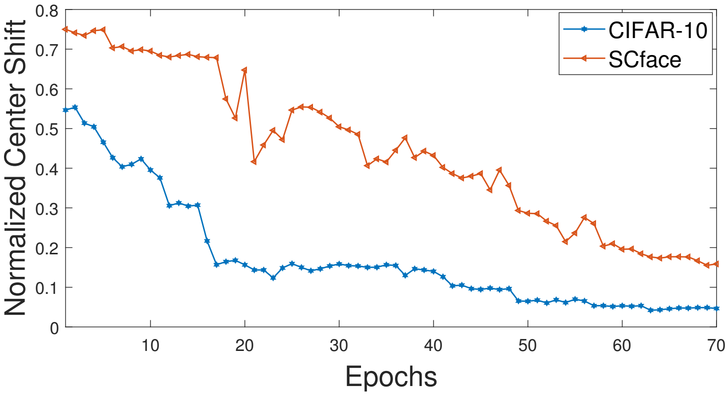


FIGURE 5.7: Normalized center shift epoch-wise for the CIFAR-10 and SCface database (best viewed in color).

create embeddings in the most dense region of the clusters. This avoids large weight updates as the embeddings need not be shifted away from the dense region. As shown in Figure 5.6, the convergence of the proposed density aware triplet loss (on a validation set which is prepared by randomly selecting 10% samples from the training set) is superior than the vanilla triplet loss and the triplet center loss. For all the methods, convergence is defined as the stage when the validation accuracy does not improve for 50 epochs at a stretch.

5.7.5 Shifting of the Center

The proposed approach iteratively evaluates the center towards the most dense region of each class (Figure 5.3(b)). An analysis is performed showing the magnitude of center shift for each epoch. The average of the center shift for all the classes is used to plot the graph in Figure 5.7 which shows that the magnitude of the center shift is much higher for the SCface database which has a very large number of noisy training samples. It can be also observed that the magnitude of the center shift decreases as the training progresses, thereby producing more compact clusters.

5.8 Summary

This chapter presents an elegant approach for density aware deep metric learning. The proposed approach can be augmented with any deep metric learning technique such as triplet and quadruplet loss, and its variants. It results in superior convergence and accuracies, thus providing an important enhancement in current deep metric learning strategies. The proposed DATL and DAQL have also shown to be resilient to noisy training data compared to other deep metric learning methods. Extensive experiments on four datasets showcase the superiority of the proposed DATL and DAQL over existing deep metric learning techniques.

This chapter involves training a model which is robust to noise and data quality using a classifier level algorithm. The model learning algorithm is designed such that it is robust to noise and outliers, which is suitable for surveillance applications where the probe data is expected to be noisy and of poor quality. In the next chapter we will delve into another important requirement of surveillance applications, to optimize for the top-k accuracy, where we will discuss a classifier level technique so that the trained model is more suited towards that requirement.

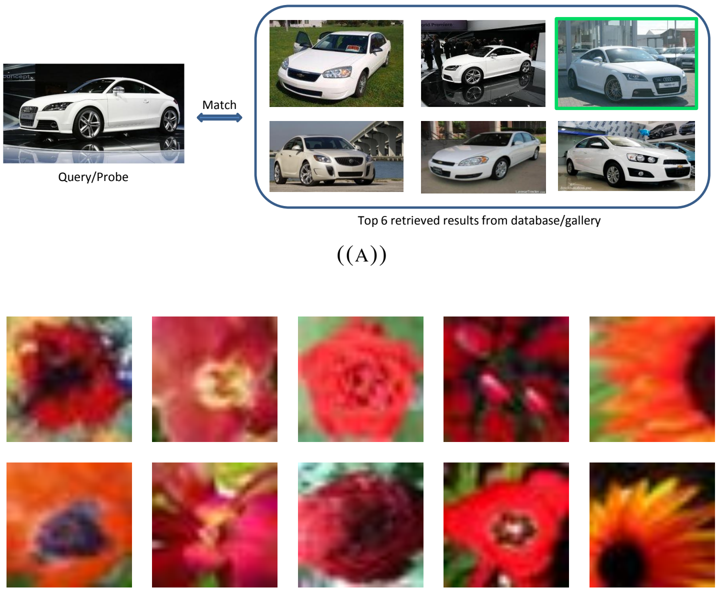
Chapter 6

Top-k Aware Deep Metric Learning

6.1 Introduction

Conventional deep learning algorithms are generally trained for optimizing the classification accuracy or top-1 accuracy, which is reflected by the performance of the model for obtaining the most likely class label of a test image. However, in problems such as object retrieval and biometrics, improving the top-k accuracy has received paramount attention [220–222]. In these applications where the number of classes is extremely large, ensuring very high performance for the top match might not be feasible. In such cases, the top-k best matching results obtained from the algorithm can be further analyzed to ascertain the correct class of the test image. Formally, top-k accuracy is defined as probability of the trained model for predicting the correct class among the *k-most* likely classes for a given test image.

Considering an example of object retrieval, the algorithm matches a query/probe image to a database/gallery and retrieves the most likely matches. The most desirable scenario is that the top match belongs to the relevant class (the same class as the probe image). As shown in Figure. 6.1(a), the query image when matched with the database may retrieve samples that are very similar to each other. In this example, the correct match is present at rank 3 and the matches at previous ranks are not correct matches (does not belong to the same class of the query). The reason is that all these images (top 6) are extremely similar in terms of image properties which leads to class ambiguity [223–226]. Similarly, Figure 6.1(b) shows several classes from the CIFAR-100 database which are similar in terms of features and properties, may lie close to each other in the output embedding space of the model being used for matching.



((A))

((B))

FIGURE 6.1: Motivation for enhancing the top-k matching performance. (a) A typical object retrieval scenario. It shows that the correct class (with green border) is retrieved among the top-6 matches from the gallery, while the other matches although look similar to the probe are actually incorrect matches. (b) Illustration of class confusion in the CIFAR-100 database. Images of different classes of *flowers*, namely orchids, poppies, roses, sunflowers and tulips are shown. It is evident that several samples across different classes in each of the superclasses have visually similar color, texture and appearance, which may lead to class confusion while training a classifier (best viewed in color).

6.1.1 Related Work

Several notable research contributions have been made for enhancing the top-k matching performance of classifiers for databases where the number of classes are large. Gupta *et al.* [227] investigated the behaviour of classifiers on databases with two kinds of class ambiguity, namely label noise and overlapping classes. They proposed an approach to avoid class confusion by ignoring such classes and focusing the classifier on those classes that are more discriminative. Lapin *et al.* [224] proposed a loss function with a tight convex upper bound on the top-k error. Further, they presented a detailed analysis [228] of multi-label classification techniques and top-k based algorithms in the context of large scale databases. Yan *et al.* [229] utilized multi-modal feature fusion for top-k classification by using several classifiers which are allowed to participate in the process. Chu *et al.* [221] showcased a semi-smooth Newton algorithm which improves the training time for top-k classification significantly. Recently Berrada *et al.* [220] introduced a family of smooth loss functions designed for top-k optimization. Their work also exhibits better top-k performance on large scale datasets in the presence of label noise.

One of the popular ways to build image recognition systems is to train a deep learning model using a loss function which minimizes intra-class distance and maximizes inter-class distance in the output embedding space of the model. Such approaches, categorized as deep metric learning methods, have produced impressive results on several image recognition problems

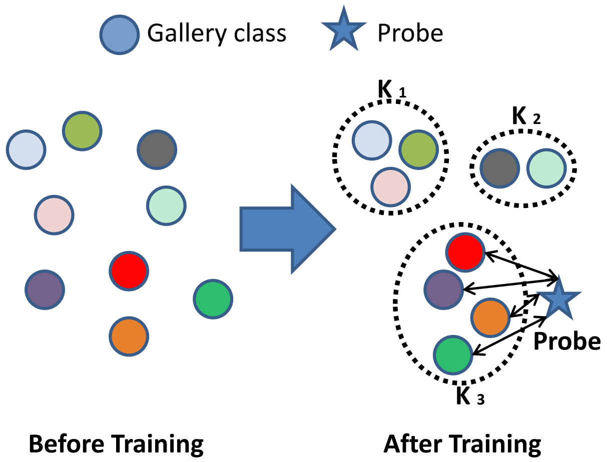


FIGURE 6.2: Illustrating the high level working of the proposed approach. It brings similar classes (represented as smaller circles) closer in the embedding space to form compact *super-cluster* (dotted circle), which results in increased top-k matching performance during testing.

[91, 94, 205, 207, 210, 213, 215], unlike conventional classifiers which output the class label for an input image. Models trained using deep metric learning methods output the embedding (feature representation) of an image that is given as input to the model. Further, unlike conventional classifiers, deep metric learning models are often tested on classes which are not encountered during training. As a result, existing top-k approaches [220–222, 224, 229] built on conventional classifiers such as support vector machines and neural networks are not suitable for the deep metric learning paradigm and they cannot be directly utilized in unseen train-test classification/identification scenarios.

6.1.2 Research Contributions

In this research, we propose an algorithm for training a Convolutional Neural Network (CNN) model using a deep metric learning loss function, with the objective of enhancing the top-k matching performance. The proposed method achieves this objective by bringing together classes into *superclusters*. As illustrated in Figure 6.2, this is done by using a loss metric which brings together similar classes in the output embedding space of the model being trained. During testing, when a query/probe is matched with the database/gallery, it is likely to be located close to the *supercluster* containing the relevant class. As a result, the probe is expected to contain the correct class in one of the top-k matches, where k is approximately equal to the number of classes in the *supercluster* containing the class. The salient contributions of this research are as follows.

1. A deep metric learning formulation is proposed for enhancing the top-k matching performance of a CNN model. The formulation is presented as independent of the application and can be used in its generic form for enhancing the top-k accuracy. To the best of our

knowledge, this is the first work on enhancing the top-k performance of a classifier for deep metric learning. The proposed approach may also be utilized in scenarios where classes encountered during testing are not seen during training.

2. Extensive experiments on four popular object recognition datasets namely STL-10 [6], CIFAR-10 [5], CIFAR-100 [5], CARS196 [230] and a challenging heterogeneous face recognition dataset SCface [3] unravel the effectiveness of the proposed approach.

6.2 Preliminaries

Consider a classical pattern recognition scenario, where data $\{X_1^1, X_1^2, X_1^3, ...X_i^1, ...X_n^1\}$ is available from n classes. X_i^1 and X_i^2 represent two data samples belonging to the i^{th} class. A deep metric learning algorithm applied to this data attempts to learn a projection function $f_\theta(x) : \mathbb{R}^S \longrightarrow \mathbb{R}^T$, where S is the dimensionality of the data and T is the dimensionality of the projection of the input data. In the purview of deep metric learning, the function f_θ is often realized as a deep-CNN model f with trainable parameters (weights of the model) θ. After the model f maps the input data to their respective embeddings, a distance metric function is utilized to calculate the distances between the data samples in the output embedding space of the model f. The distance metric function is defined as,

$$D\{p, q\} : \mathbb{R}^T \times \mathbb{R}^T \longrightarrow \mathbb{R} \tag{6.1}$$

where, D is the distance metric function which outputs the distance between two data samples p and q in the embedding space. This distance may be calculated between two images p and q using the Euclidean measure $D\{p, q\} = \|f_\theta(p) - f_\theta(q)\|_2^2$. Similarly, $D(.)$ may also be used to evaluate a distance vector between two sets of images.

Inspired by Large Margin Nearest Neighbor Classification [218], the triplet loss [91] is defined as,

$$\mathcal{L} = \Big[D\{\vec{X}_i^a, \vec{X}_i^b\} - D\{\vec{X}_i^a, \vec{X}_j^c\} + \alpha\Big]_+ \tag{6.2}$$

$$\forall(\vec{X}_i^a, \vec{X}_i^b, \vec{X}_j^c) \in \tau$$

where, $\vec{X}_i^a$ and $\vec{X}_i^b$ are set of images of class i and $\vec{X}_j^c$ is the set of images of class j where $i \neq j$, $[f]_+ = max(f, 0)$, α is the margin parameter and τ is the set of all 3-tuples which are generated from the training set. In order to minimize $\mathcal{L}$, $D\{\vec{X}_i^a, \vec{X}_i^b\}$ which represents the intraclass distance, is minimized and $D\{\vec{X}_i^a, \vec{X}_j^c\}$, which represents the interclass distance, is maximized. The gradient of this loss $\mathcal{L}$ is used to update the weights of the model f, which is trained as a discriminative function.

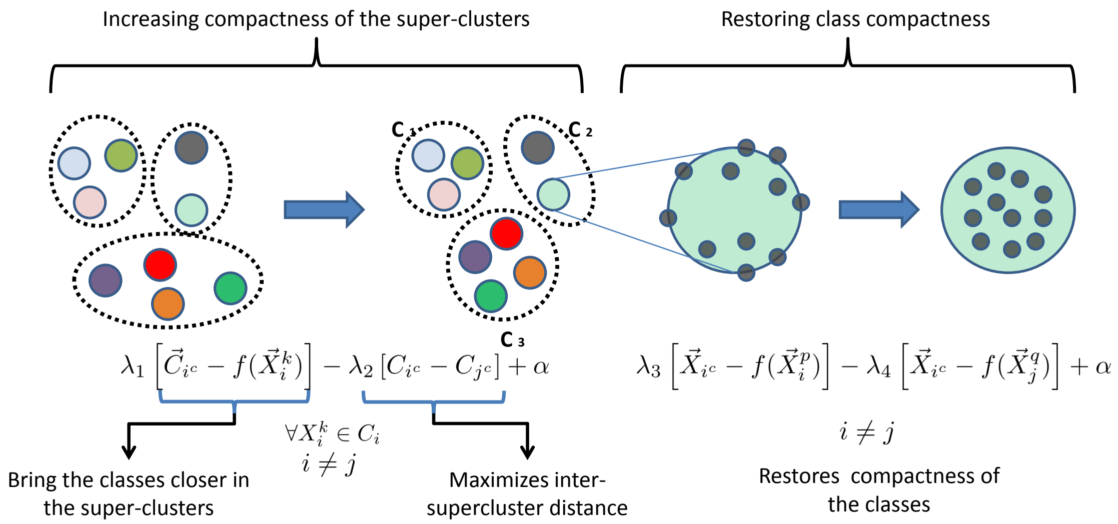

Increasing compactness of the super-clusters
Restoring class compactness
C 1
C 2
C 3
$\lambda_1 \left[\vec{C}_{i^c} - f(\vec{X}_i^k)\right] - \lambda_2 \left[C_{i^c} - C_{j^c}\right] + \alpha$
$\lambda_3 \left[\vec{X}_{i^c} - f(\vec{X}_i^p)\right] - \lambda_4 \left[\vec{X}_{i^c} - f(\vec{X}_j^q)\right] + \alpha$
$\forall X_i^k \in C_i$
$i \neq j$
$i \neq j$
Bring the classes closer in the super-clusters
Maximizes inter-supercluster distance
Restores compactness of the classes

FIGURE 6.3: Illustration of the proposed top-k aware deep metric learning algorithm.

6.3 Top-k Aware Deep Metric Learning

In this section, we present the proposed deep metric learning algorithm which is used to train a classifier for enhancing the top-k matching performance. In order to enhance the top-k matching performance, we introduce the concept of *superclusters*, which are a collection of classes. A pretrained deep model can be initially used to generate embeddings of the training data. From these embeddings, a density based clustering algorithm is used to create superclusters, each of which is a set of similar classes in the output embedding space of the pretrained model. Thereafter, the intra-supercluster distance is minimized and the inter-supercluster distance is maximized, thereby making the superclusters more compact. This may result in some of the classes losing their compactness, since some of the classes may not have their data allocated to a single supercluster, rather may be allocated to different superclusters. In order to mitigate this effect, a generic deep metric learning algorithm is utilized to restore the compactness and separation of the classes in the embedding space. The proposed algorithm is illustrated in Figure 6.3. In this section we illustrate the proposed algorithm in detail, along with the intuition and significance of each step of the process.

6.3.1 Preparing Superclusters

The first step of the proposed method is clustering of the embeddings of the input training data obtained using a pretrained deep-CNN model. As shown in Figure 6.3, the clustering step is used to categorize the input data into k clusters, where each cluster may contain data from multiple ground truth classes, hence called a supercluster. A density based clustering method DBSCAN [231] is applied on the embeddings of n-class training data $\{ X_1^1, X_1^2, X_1^3, ...X_i^1, ...X_n^1\}$.

This algorithm attempts to cluster the embeddings using density (number of points in the neighborhood of a point) as a criterion. It identifies each data sample as one of the three types, namely core point, border point and noise point. Core points are those which are located in a dense neighborhood. The neighborhood of a point is dense, if it has at least c points in its ϵ neighborhood, where ϵ is the radius of the neighborhood with respect to the core point and c is a threshold parameter. A border point is one, which is not a core point but it is located in the ϵ neighborhood of a core point. Points which are neither core nor border points are known as noise points. DBSCAN performs a density traversal from an initial core point p_1 to other core points $\{p_2, p_3..p_i, p_{i+1}..p_m\}$ such that p_{i+1} is density reachable from p_i. A point k is density reachable from l if l is a core point and it is present in the ϵ neighborhood of k. Thus the set $\{p_1, p_2, p_3....p_m\}$ forms a supercluster with m points. This process is repeated until all the points in the training data have been allocated to a supercluster except the noise points which are ignored in this process. As shown in Figure 6.3, C_1, C_2 and C_3 are three superclusters, each containing multiple classes. It should also be noted that the data of a particular class may not be located in a single supercluster, rather distributed across superclusters.

6.3.2 Compaction of the Superclusters

Each supercluster, denoted by C_i, has a centroid C_{i^c}, which can be calculated as a mean of all the embeddings of the data points in the supercluster. The objective function (loss) for compaction of the superclusters is expressed as,

$$\mathcal{L}_1 = \left[\lambda_1 \left[\vec{C}_{i^c} - f(\vec{X}_i^k)\right] - \lambda_2 \left[\vec{C}_{i^c} - \vec{C}_{j^c}\right] + \alpha\right]_+ \tag{6.3}$$

$$\forall X_i^k \in C_i \quad \text{and} \quad i \neq j$$

wherein, X_i^k is an element of the i^{th} class that has been allocated to the supercluster C_i. The first term $[C_{i^c} - f(X_i^k)]$ computes the intra-supercluster distance (in the embedding space), which is minimized, resulting in a more compact supercluster. The second term $[C_{i^c} - C_{j^c}]$ gives the inter-supercluster distance between two superclusters i and j, which is maximized as a result of the minimization of $\mathcal{L}_1$. Thus, Equation 6.3 results in more compact superclusters, which is the primary goal of the proposed approach.

6.3.3 Restoring Compactness of the Classes

The process of compactness of the superclusters may distort the compactness of the classes. This is because all the data samples of a particular class may not be assigned to the same supercluster. In such a case, as per the process outlined in Section 6.3.1 the data points pertaining to the same class may be separated further apart. To ensure that the compactness of the classes is restored,

Algorithm 3: top-k Aware Deep Metric Learning

```
Input: Training Data $\mathbf{X} = \{X_1^1, X_1^2, X_1^3, ....X_i^1, ....X_n^1\}$ ,
  $f_\theta$ (Pretrained Deep-CNN model)
Output: $f'_\theta$ ( trained Deep-CNN model)
Parameters: $\epsilon$, $c$, $\alpha_1$, $\alpha_2$, $\lambda_1$, $\lambda_2$, $\lambda_3$, $\lambda_4$, $t^p$ (threshold for hard positive selection), $t^n$
  (threshold for hard negative selection), $k$ (batches) and $e$ (epochs)
Identify Superclusters:
$Core = \{\}$
for $X_i \in X$ do
    if $X_i$ is a core point then $Core = Core \cup X_i$
end
for $Cr_i \in Core$ do
    find the Core points $Cr_i^j$ that are density connected to $Cr_i$
end
----------------------------------------------------------------
for $Epoch$=1 to $e$ do
    Compaction of Superclusters:
    for $Epoch_a$=1 to $e_1$ do
        for each supercluster $C_i$ do
            Evaluate $C_{ic}$ as the center
            Generate embeddings:
            for every batch of size $m$ do
                Forward pass through $f$ to find $f_\theta(C_{i^c}), f_\theta(X_i^k), f_\theta(C_{i^c})$
                Calculate loss $\mathcal{L}_1$
                $\mathcal{L}_1 = \left[\lambda_1 \left[\vec{C}_{i^c} - f(\vec{X}_i^k)\right] - \lambda_2 \left[\vec{C}_{i^c} - \vec{C}_{j^c}\right] + \alpha_1\right]_+$
                Calculate gradient
                $\triangle W = \nabla_\theta \frac{1}{m} \sum_m \mathcal{L}_1$
                Update weights of $f_\theta$ using $\triangle W$
            end
        end
    end
    ------------------------------------------------------------
    Restore Compactness of Classes:
    for $Epoch_b$=1 to $e_2$ do
        Generate embeddings:
        for every batch of size $m$ do
            Forward pass through $f$ to find $f_\theta(\vec{X}_{i^c}), f_\theta(\vec{X}_i^p), f_\theta(\vec{X}_j^q)$
            Calculate loss $\mathcal{L}_2$
            $\mathcal{L}_2 = \left[\lambda_3 \left[\vec{X}_{i^c} - f(\vec{X}_i^p)\right] - \lambda_4 \left[\vec{X}_{i^c} - f(\vec{X}_j^q)\right] + \alpha_2\right]_+$
            Calculate gradient
            $\triangle W = \nabla_\theta \frac{1}{m} \sum_m \mathcal{L}_2$
            Update weights of $f_\theta$ using $\triangle W$
        end
    end
end
```

we use the triplet loss with respect to the centroids of the classes. The loss function can be written as,

$$\mathcal{L}_2 = \left[\lambda_3 \left[\vec{X}_{i^c} - f(\vec{X}_i^p)\right] - \lambda_4 \left[\vec{X}_{i^c} - f(\vec{X}_j^q)\right] + \alpha\right]_+ \qquad (6.4)$$
$$\forall(\vec{X}_{i^c}, \vec{X}_i^p, \vec{X}_j^q) \in \tau$$

where, τ is the set of all triplets and $\vec{X}_{i^c}$ is a vector prepared by replicating the center embedding of a class m times. X_i^p and X_j^q denote the p^{th} and q^{th} data samples of classes i and j respectively and m is the size of the triplet batch.

6.3.4 Optimization Process

Summarizing the algorithm, the first step is to run a density based clustering method such as DBSCAN, on the embeddings of the training data (using a pretrained model) which identifies the initial superclusters in the embedding space. Thereafter, Equation 6.3 is utilized for enforcing compactness in the superclusters, followed by the triplet loss (Equation 6.4) to restore the compactness of the original classes. It may be noted that Equation 6.3 and 6.4 are not optimized serially, rather in an alternating fashion. The reason for this is that, optimizing Equation 6.4 after Equation 6.3 in a serial fashion may corrupt the structure of the superclusters which was optimized by Equation 6.3. The gradient of these losses is utilized to update the weights of the deep-CNN model. This alternating optimization process is repeated until convergence. The detailed algorithm is outlined in Algorithm 3.

6.4 Databases, Protocols and Evaluation

The proposed approach is utilized for two kinds of experiments, namely object retrieval and heterogeneous face recognition. For object retrieval, we have performed experiments on CIFAR-10 [5], STL-10 [6], CIFAR-100 [5] and CARS196 [230] databases, and for heterogeneous face recognition under surveillance scenarios, we have used the SCface [3] dataset.

6.4.1 Databases and Protocols

STL-10 [6] database contains 113000 images (of resolution 96 × 96) of 10 object categories of resolution 96 × 96. The training set contains 5000 images and 8000 images are in the testing set. Each class has 500 images for training and 800 images for testing. The rest of the images are unlabelled and have not been used in our experiments.

CIFAR-10 [5] and **CIFAR-100 [5]** contains 60000 images (of resolution 32 × 32) of 10 and 100 object categories respectively out of which 50000 images are for training and the rest are

TABLE 6.1: Face identification accuracies (%) on the SCface [3] database with varying probe resolution.

Method	**24 x 24**		**32 x 32**		**48 x 48**	
	Rank 5	**Rank 10**	**Rank 5**	**Rank 10**	**Rank 5**	**Rank 10**
Triplet Loss [91]	78.2	83.7	97.2	98.7	98.2	98.6
Quadruplet Loss [92]	79.4	82.4	97.5	98.9	98.7	99.1
Triplet Center Loss [207]	80.1	83.8	98.4	99.1	98.8	99.2
Hard Triplet Loss [206]	79.1	82.9	97.4	98.5	99.1	99.3
Proposed	**81.7**	**85.7**	**98.8**	**99.4**	**99.3**	**99.7**

TABLE 6.2: Retrieval results (%) on the STL-10 and CIFAR-10 databases.

Database	**Method**	**Recall@1**	**Recall@10**	**Recall@100**
STL-10	Triplet Loss [91]	72.4	78.5	80.4
	Quadruplet Loss [92]	73.9	78.7	81.7
	Triplet Center Loss [207]	**74.6**	77.9	81.5
	Hard Triplet Loss [206]	72.9	76.1	81.9
	Proposed	68.9	**78.9**	**82.3**
CIFAR-10	Triplet Loss [91]	76.2	94.7	97.2
	Quadruplet Loss [92]	78.3	95.4	98.9
	Triplet Center Loss [207]	**79.4**	96.1	95.7
	Hard Triplet Loss [206]	78.5	93.8	97.4
	Proposed	75.7	**96.8**	**99.2**

for testing. In CIFAR-10, the training and testing sets contain 5000 and 1000 images for each class respectively. For CIFAR-100 the classes are grouped into 20 superclasses, each of which contains multiple classes. For example, the superclass *fish* contains classes such as aquarium fish, flatfish, ray, shark and trout.

CARS196 [230] dataset contains 16185 images of cars belonging to 196 categories, each corresponding to a class. The training set contains 8054 images of 98 classes and the test set contains 8131 images of the same number of classes.

SCface [3] contains images captured using 8 surveillance cameras of 130 non-cooperative subjects under uncontrolled imaging scenarios. Images are captured from three different standoff distances, namely 4.2m, 2.6m and 1m, which are used as probes and are resized into 24×24, 32×32 and 48×48 respectively. Gallery images are captured in high resolution under controlled illumination pose and expression. Out of 130 subjects, images pertaining to 50 subjects are used in training and rest of the images are used for testing. The gallery images are resized to a resolution of 72×72. The subjects (classes) in the train and test set are disjoint.

6.4.2 Evaluation Criteria

For object retrieval the results are demonstrated in terms of the average recall@k calculated on the images of the testing set. Recall@k is defined [98, 219] as the ratio of the number of images correctly retrieved (that belongs to the class of the query image) for top k retrievals

TABLE 6.3: Retrieval results (%) on the CARS196 database.

Method	**Recall@1**	**Recall@4**	**Recall@8**
Triplet Loss [91]	51.5	73.5	82.4
Quadruplet Loss [92]	52.4	71.6	81.9
Triplet Center Loss [207]	52.9	73.6	83.7
Hard Triplet Loss [206]	51.6	70.2	83.4
Lifted Structures [98]	52.9	76.0	84.2
Discriminative Loss [232]	68.3	85.2	91.1
Proposed	56.7	**89.1**	90.4

(top-k matches). If a sample of the query image's class is retrieved among the top k (k nearest neighbor) matches, we assign a score of 1 to the retrieval, we assign a score of 0 otherwise. The final recall@k score is calculated by averaging these scores for all the query images. In case of heterogeneous face recognition, identification experiments are performed. The training set is partitioned into probes (query images) and gallery (database). Each probe is matched to the images in the gallery. The matching performance is evaluated by rank-k accuracy. It is defined as the ratio of the number of times an image of the relevant class is retrieved in the top k matches to the number of matching attempts (number of probes).

It is worthwhile to note that the proposed algorithm does not optimize for a particular value of k. The loss functions does not specifically optimize for the top-k accuracy, rather it optimizes the embedding space in such a way so that the top-k matching performance is enhanced during testing.

6.4.3 Implementation Details

The proposed method uses the LightCNN-29 [163] which has 29 convolutional and 4 pooling layers. It uses the Max-Feature-Map (MFM) activation after every convolutional layer. The network is organized into 6 blocks, each containing convolutional, pooling and MFM layers. The size of the final embedding from the model is 256. The model is trained by backpropagating the gradient of the loss terms (as given by Equations 6.3 and 6.4). The optimization is performed by Adam with different batch sizes for different databases. The learning rate is initially kept at 10^{-4} and is gradually reduced to 10^{-8}. In order to run DBSCAN efficiently on the training data, the value of ϵ and c are determined empirically. A lower value of ϵ and higher value of c would enable DBSCAN to search for denser clusters and vice versa otherwise. It will also result in more number of noise and border points. In order to balance this effect, we chose ϵ as 3.8 for all the databases, with the exception of the CIFAR-100 database. The c parameter is chosen as 55 for CIFAR-10, 32 for STL-10, 30 for CARS196 and 20 for the SCface database. It has been observed that higher number of training samples require a higher value for the c parameter for optimal performance of the clustering process. The experiments are performed on a machine with Intel Core $i7$ CPU, 32GB RAM and NVIDIA GTX 1080Ti GPU. We would be releasing the source code for the proposed algorithm.

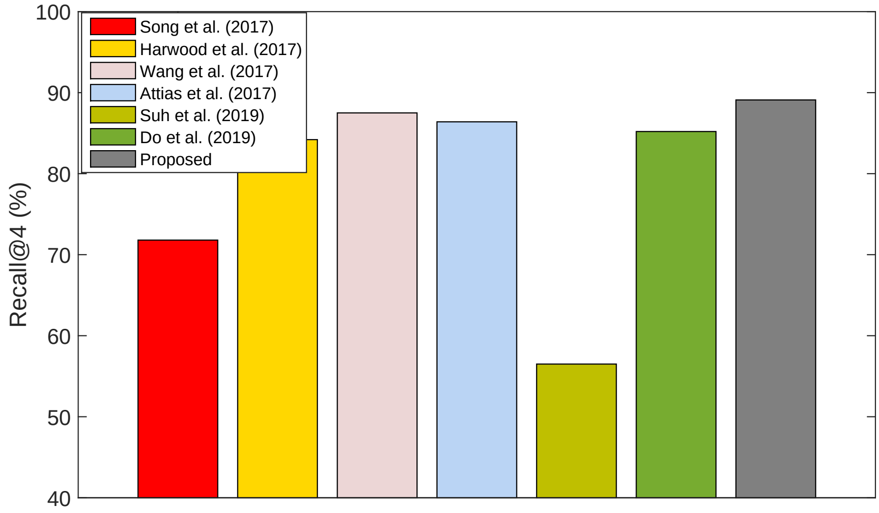
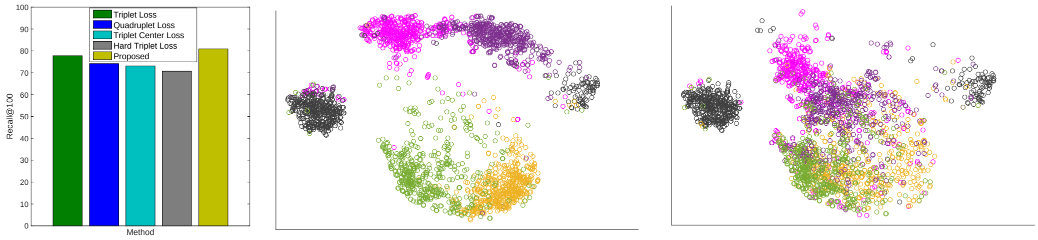


FIGURE 6.4: Bar Graph showcasing image retrieval accuracies (recall@4) on recent deep metric learning methods on the CARS196 database.

6.5 Results and Observations

The results of the proposed approach have been compared with widely used deep metric learning algorithms [91, 92, 96–98, 206, 207, 232–234]. We outline the results of object retrieval experiments followed by an illustration on face identification performed on a challenging face recognition dataset.



((A)) ((B)) ((C))

FIGURE 6.5: Evaluation on the CIFAR 100 database (a) Object retrieval accuracies (%) on the CIFAR-100 database, (b) t-SNE [235] visualization of embeddings (for the flower superclass) containing five classes of the CIFAR-100 [5] dataset (each distinct color represents a class) with the pretrained LightCNN model before training, and (c) visualization after training. It shows that the proposed algorithm results in a compact superclass, however the discriminative properties of the embeddings are not completely done away with. This culminates in a higher top-k accuracy during testing, while maintaining an appreciable top-1 matching performance (best viewed in color).

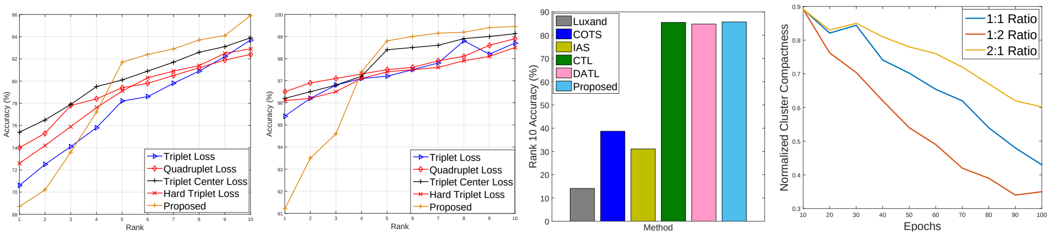
The results of object retrieval on the STL-10 and CIFAR-10 databases are presented in Table 6.2 and results on CARS196 database are summarized in Table 6.3. It may be observed that on the CIFAR-10 and STL-10 databases, the proposed method yields the highest accuracies for recall@10 and recall@100. On the CARS196 database, recall@4 is highest for the proposed method, however for recall@1 the accuracy is lower for the proposed method. We have also compared (Figure 6.4) the performance of the proposed method on the CARS196 database with recent deep metric learning methods. On the CIFAR-100 database, the proposed algorithm gives the best matching accuracies (Figure 6.5(a)) for recall@100.

Law enforcement agencies often rely on automated image matching systems. These systems retrieve the top-k matches for a probe (face) image acquired from a surveillance camera. In order to showcase the effectiveness of the proposed algorithm for enhancing the top-k matching accuracy, we perform face identification experiments on the SCface [3] dataset. It may be observed (Table 6.1) that rank 5 and rank 10 accuracy for the proposed method on all three probe resolutions is higher compared to the other deep metric learning methods. We have also compared (Figure 6.6(c)) the performance of face identification on 24 x 24 probes of this database with other state-of-the-art heterogeneous face matching methods, including Co-Transfer Learning [23], Identity Aware Synthesis [144] and Density Aware Triplet Loss (DATL) [167]. Comparisons have also been performed with two commercial systems including Luxand, which is a CNN based face recognition system. The results in Figure 6.6(c) show that the proposed approach is able to enhance the top-k matching performance for a real world face surveillance application.

There are several interesting perspectives of the proposed algorithm which have been extensively discussed in this section. One of the interesting observations is that the algorithm brings together similar classes in the embedding space, and this has been shown through several experiments and visualizations (Figures 6.5 and 6.6). We assert that it is this aspect of the proposed algorithm that helps in enhancing the top-k matching performance.

6.5.1 Effect of ϵ and c Parameters

As illustrated in Section 6.3.1 the clustering process uses two important parameters namely ϵ and c, where ϵ gives the size of the neighborhood of a point, and c is the minimum number points that should be present in the ϵ neighborhood of the point to qualify it as a core point. To understand the parameter sensitivity of ϵ and c, we performed experiments with by keeping $\epsilon = 3.8$ constant and varying the value of c while training. On the SCface database for values of c=10, 15, 20, 25, 30 the rank 10 accuracies are 84.25%, 84.65%, 85.7%, 85.14% and 85.06% respectively. Since different values of c results in different supercluster formation, we inferred that both high and low values affect the testing performance and the optimal value should be chosen carefully, probably with the help of empirical testing on a small validation set data. We do not use test data to finetune these parameters.



((A)) ((B)) ((C)) ((D))

FIGURE 6.6: Face recognition on the SCface [3] database (a) CMC curves for 24 x 24 probes, (b) CMC curves for 32 x 32 probes and (c) Bar graph for rank 10 accuracies (%) on 24 x 24 probes, comparing the results with Luxand (Commercial CNN based system), COTS (Commercial Off-The-shelf System), Identity Aware Synthesis (IAS) [144], Co-Transfer Learning (CTL) [23], and Density Aware Triplet Loss (DATL) [**?**], and (d) Ablation study showing normalized cluster compactness for every 10 epochs during training, for three different variants of the training algorithm. (best viewed in color).

6.5.2 Individual Contribution of the Losses

To understand the effectiveness of individual losses, we performed multiple experiments with individual losses on the SCface database. In the first experiment, we used the first (supercluster compaction) loss (Equation 6.3) to train and used the second loss (Equation 6.4) for a very brief training spell (about 10 epochs). During testing, on 24 x 24 probes, at rank 5 and 10 we observed identification accuracies of 75.3% and 78.8% respectively. In another experiment we ran the second loss for a longer spell (about 50 epochs) after the first loss and obtained accuracies of 80.1% and 84.9% on the same ranks. It shows that the second loss performs a balancing act of maintaining the structure of the superclusters and restoring the compactness of the individual classes.

6.5.3 Analysis of Compaction of Superclusters

The proposed algorithm works on the strategy of compaction of the superclusters. The initial process of the density based clustering points out the key superclusters. In a database such as CIFAR-100 these superclusters are represented by the superclasses. Each of these superclasses have multiple classes within them, some of which are visually similar to each other (Figure 6.1). As shown in Figure 6.5 the proposed algorithm results in compaction of these superclasses. However, as formulated in Section "Compaction of Superclusters" (with Equation 6.4), the compaction of the superclasses are not overdone. The reason is, if the superclusters are compacted beyond a certain level, the individual classes might lose their dicriminative properties in the embedding space. This might result in extremely poor top match accuracies. Figures 6.5(b) and 6.5(c) present the evidence of this process, where it may be observed that compaction of the (a particular superclass in the CIFAR-100 database) superclass takes place, but the constituent classes still continue to hold their separability with each other to an appreciable extent. In addition to this, it is observed that the images belonging to the classes of the superclusters (of the training set) appear together in the retrieved list after matching. This helps in obtaining a higher top-k matching accuracy since the probe/query image's class has a greater chance of appearing in one of the top-k matches.

6.5.4 Ablation Study on Cluster Quality

As shown in Algorithm 3, Equation 6.3 and 6.4 are optimized in an alternating fashion. The algorithm also has the flexibility to calibrate the number of epochs for optimization of Equation 6.3 and 6.4 in each alternating step. This can be done by setting appropriate values of variables e_1 (in Step 5) and e_2 (in Step 13) in the algorithm. Figure 6.6(d) shows normalized values of cluster compactness, which is defined as the average intra-class distance across all classes in the training set. We observe that the best retrieval/recognition results are achieved when the number of epochs $e_1 = e_2$, shown as 1:1 ratio in Figure 6.6(d). To perform this ablation study we also train using $e_1 = 2 * (e_2)$ (1:2 ratio) and $2 * (e_1) = e_2$ (2:1 ratio). It may be observed that having a higher value of e_2 improves cluster quality considerably, but affects the top-k identification accuracy. Thus, it may be inferred that a right balance of the two optimization steps (Equation 6.3 and 6.4) is required for optimal top-k matching/recognition/retrieval accuracy.

6.5.5 Selection of Noise Points

As discussed previously, the DBSCAN clustering method classifies every point into either of the three categories namely core point, border point, and noise point. The clusters are made out of core and border points whereas the noise points are ignored. Selection of the noise points impact further learning of the deep model, as these points do not participate in this process. The c (threshold for the number of points) and ϵ (neighborhood distance threshold) parameters of the DBSCAN algorithm directly impact the amount of noise points selected. With a higher value of ϵ, the number of noise points are expected to decrease, since the core points would include more points inside its ϵ neighborhood. In the CIFAR-100 database, the amount of class confusion was the highest among all the five databases. A direct implication is the increase in the number of noise points, which is 6.78% of the training data. To address this, a higher value of ϵ is required while training on this database. On the other hand, SCface, CIFAR-10, and the STL-10 databases have 3.98%, 1.72% and 2.57% noise points respectively.

6.6 Summary and Future Work

This paper presents a top-k aware deep metric learning algorithm. It works by initially clustering the data into superclusters to first identify the sets of similar classes. Thereafter, these superclusters are compacted though a two-step deep metric learning algorithm which results in similar classes being mapped closer to each other in the output embedding space of the model. This results in a higher top-k matching performance during testing. Extensive experiments and analysis have been carried out for object retrieval and face identification in surveillance scenarios, which show that the proposed approach yields enhanced top-k matching performance when compared to popular and recent deep metric learning algorithms. It is observed that enhancing both top-1 and the top-k accuracies is a trade-off and the proposed algorithm may result in slightly lower top-1 accuracy in order to enhance the top-k accuracy. As part of our future work, we plan to extend the proposed algorithm to optimize both top-1 and top-k accuracies.

Chapter 7

Conclusion and Future Work

In recent years, face recognition is one of the most challenging and relevant problems in computer vision and artificial intelligence. In order to protect their citizens from attack and public disorder, governments around the world has been investing heavily in developing heterogeneous face recognition systems which can be used for video surveillance. This helps the law enforcement agencies in two ways, firstly, it can prevent such attacks and secondly it presents a chance to identity the suspects if at all such an event takes place. However, a true surveillance scenario presents a completely unconstrained setup for face image acquisition and recognition. The problems of such scenarios that has been discussed in this thesis can be solved at several levels. Although this thesis attempts to solve the problem at several of these levels, the most promising contributions are at the image level, feature level and classifier levels.

RGB-D Face Recognition: The first contribution is focused on learning feature rich representations for face recognition from RGB and depth data. Hierarchical layers of features are learnt on the representation learned by a stacked mapping model which is a mapping function from RGB to depth data. This allows the model to be utilized for RGBD face recognition where depth data is not available during testing. Thus it presents a framework for learning a shared representation of RGB and depth for face recognition.

Supervised Resolution Enhancement and Recognition Network: In the second contribution, an image level transformation from low quality/resolution to high quality/resolution is proposed to achieve improved face recognition results in unconstrained scenarios. However, during such transformation important discriminative information in the source domain images (low quality/resolution images) may be distorted. We utilize a highly effective GAN based framework which learns an image to image transformation from low to high resolution images. The proposed framework uses a classifier which backpropagates a supervised signal (gradient) into the generator, which prevents the generator of the proposed Supervised Resolution Enhancement-GAN (SURE-GAN) in distorting important discriminative information during the process of transformation. Experiments on several face datasets show that the proposed approach significantly enhances the recognition accuracy of low resolution face images.

The next three contributions are primarily aimed towards learning discriminative representations for heterogeneous face recognition.
Subclass Heterogeneity Aware Loss Other than the image level, one of the most significant level at which the problem of face recognition in unconstrained scenarios can be addressed is the feature level. At this level we can utilize the embedding space of the training set to formulate a loss function to train a deep CNN model, which can produce heterogeneity aware embeddings for effective heterogeneous face matching performance. We propose a Density Aware Deep Metric Learning algorithm which trains a deep CNN model considering the density distribution of the embeddings of the different classes in the training set. We have shown that this results in significantly better convergence and lower training time. We also propose a Subclass Heterogeneity Aware Deep Metric learning formulation to train a discriminative model which produces face embeddings for accurate classification in the presence of multiple face recognition covariates. In a true surveillance scenario, a Cross-Spectral Cross-Resolution matching is required. We show that this approach produces excellent results on several challenging face datasets including the FaceSurv and SCface datasets, which were acquired in real world surveillance scenarios.
Density Aware Deep metric Learning: The work presents an elegant approach for density aware deep metric learning. The proposed approach can be augmented with any deep metric learning technique such as triplet and quadruplet loss, and its variants. It results in superior convergence and accuracies, thus providing an important enhancement in current deep metric learning strategies. The proposed DATL and DAQL have also shown to be resilient to noisy training data compared to other deep metric learning methods. Extensive experiments on four datasets showcase the superiority of the proposed DATL and DAQL over existing deep metric learning techniques.
Top-k Aware Deep Metric Learning Optimizing the overall classification accuracy of a network does not always lead to best top-k accuracies. This behavior is often observed in cases where multiple classes are close to each other in the embedding space and trained classifiers may not retrieve the correct class due to class ambiguity. In problems such as object retrieval and biometrics, improving the top-k accuracy has received paramount attention. In these applications where the number of classes is extremely large, ensuring very high performance for the top match might not be feasible. The fifth contribution presents a novel deep metric learning algorithm that is formulated for optimizing the embedding space (feature level), but its effect is observed in the rank level performance for face recognition. In this research, we propose an algorithm for training a Convolutional Neural Network (CNN) model using a deep metric learning loss function, with the objective of enhancing the top-k matching performance. The proposed method achieves this objective by bringing together classes into *superclusters*. Experimental

results on several popular image classification datasets demonstrate the efficacy of the proposed approach.

7.1 Future Research Directions

Heterogeneous face recognition is an active area of computer vision research since the last two decades. Development of this area has provided impetus to several real world applications in surveillance and non-cooperative authentication environments. Although, this dissertation contributes to this area of research by proposing several methods each of which solves an important problem, there are useful future research directions in this area.

Heterogeneous Noise-resilient Representation learning: In a typical real world surveillance scenario, the probe images that are captured from CCTV cameras are of extremely poor quality. In one of the contributions of this dissertation, we have proposed a Density Aware Deep Metric Learning method which can learn a robust model in the presence of such noise. However this method does take into account the heterogeneity of the data while learning the model. In future we can extend this method so that it not only helps us train a model robust to noise, but also yield a model which is invariant to heterogeneity. Further, this method does not take into account the specific kind of noise, which if imbibed in the train model would lead to an improved representation for heterogeneous face recognition.

Resilience to Adversarial Attacks: Heterogeneous face recognition finds its applicability in covert video surveillance scenarios. Watch list surveillance systems are aimed at recognition of miscreants and criminals. Such systems may be prone to adversarial attacks by sophisticated criminals and this is one of the most important areas where further research should be carried out.

Handling Multiple Heterogeneities: One of the contributions of this thesis was to propose a method for training a model which is invariant to more than one heterogeneity simultaneously. However, in a real world scenario there can be several other covariates that we might need to handle, other than resolution and spectrum of image acquisition. Some of them are disguise, occlusion, pose, expression and so on. Future work may be carried out in this direction so that we may be able to train model for several heterogeneities at the same time. We could also utilize **Domain Adaptation** techniques for handling multiple heterogeneities. Data for training models for poor quality images is limited. Since CCTV cameras in most countries is mostly operated by law enforcement agencies, data from those cameras cannot be made publicly available, without compromising privacy of individuals. Thus training very large deep learning models would be a challenge in such cases. To tackle this problem, novel domain adaptation methods could be proposed so that even if large amount of labelled data is not available, high performing models may be trained by adapting models already trained on controlled face images.

Bibliography

[1] Andrew P Founds, Nick Orlans, Whiddon Genevieve, and Craig I Watson. NIST special databse 32-multiple encounter dataset. *NIST Interagency/Internal Report (NISTIR)-7807*, 2011.

[2] Zhaowen Wang, Ding Liu, Jianchao Yang, Wei Han, and Thomas Huang. Deep networks for image super-resolution with sparse prior. In *IEEE International Conference on Computer Vision*, pages 370–378, 2015.

[3] Mislav Grgic, Kresimir Delac, and Sonja Grgic. SCface–surveillance cameras face database. *Springer Multimedia Tools and Applications*, 51(3):863–879, 2011.

[4] S Gupta, N Gupta, S Ghosh, M Singh, S Nagpal, R Singh, and M Vatsa. A benchmark video dataset for face detection and recognition across spectra and resolutions. In *IEEE International Conference on Automatic Face and Gesture Recognition*, 2019.

[5] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, Technical Report, University of Toronto, 2009.

[6] Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *International Conference on AI and Statistics*, pages 215–223, 2011.

[7] Wenyi Zhao, Rama Chellappa, P Jonathon Phillips, and Azriel Rosenfeld. Face recognition: A literature survey. *ACM computing surveys (CSUR)*, 35(4):399–458, 2003.

[8] R Gross, I Matthews, J Cohn, T Kanade, and S Baker. The CMU multi-pose, illumination, and expression (multi-pie) face database. Technical report, TR-07-08, Technical Report, 2007.

[9] Xiaozheng Zhang and Yongsheng Gao. Face recognition across pose: A review. *Pattern recognition*, 42(11):2876–2896, 2009.

[10] John Wright, Allen Y Yang, Arvind Ganesh, S Shankar Sastry, and Yi Ma. Robust face recognition via sparse representation. *IEEE transactions on pattern analysis and machine intelligence*, 31(2):210–227, 2008.

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

[12] Dongoh Kang, Hu Han, Anil K Jain, and Seong-Whan Lee. Nighttime face recognition at large standoff: Cross-distance and cross-spectral matching. *Pattern Recognition*, 47(12): 3750–3766, 2014.

[13] Aleix M Martínez. Recognizing imprecisely localized, partially occluded, and expression variant faces from a single sample per class. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (6):748–763, 2002.

[14] Tejas I Dhamecha, Aastha Nigam, Richa Singh, and Mayank Vatsa. Disguise detection and face recognition in visible and thermal spectrums. In *International Conference on Biometrics*, pages 1–8, 2013.

[15] Ferdinando S Samaria and Andy C Harter. Parameterisation of a stochastic model for human face identification. In *IEEE Workshop on Applications of Computer Vision*, pages 138–142, 1994.

[16] Sarah OBeirne. *What CCTV users need to know about GDPR*, (accessed July 2019). https://tinyurl.com/yahhmnua.

[17] K. J. D. Greco. *Law Enforcement's Use of Facial Recognition Technology*, (accessed June 2019). https://tinyurl.com/yaysxaou.

[18] K. J. D. Greco. *Fbi's face-recognition system searches 411 million photos*, (accessed June 2019). https://money.cnn.com/2016/06/16/technology/fbi-facial-recognition/index.html.

[19] Daniel A Vaquero, Rogerio S Feris, Duan Tran, Lisa Brown, Arun Hampapur, and Matthew Turk. Attribute-based people search in surveillance environments. In *Workshop on Applications of Computer Vision*, pages 1–8, 2009.

[20] Himanshu S Bhatt, Richa Singh, and Mayank Vatsa. On recognizing faces in videos using clustering-based re-ranking and fusion. *IEEE Transactions on Information Forensics and Security*, 9(7):1056–1068, 2014.

[21] Junlin Hu, Jiwen Lu, and Yap-Peng Tan. Discriminative deep metric learning for face verification in the wild. In *IEEE conference on computer vision and pattern recognition*, pages 1875–1882, 2014.

[22] Massimo Tistarelli, Stan Z Li, and Rama Chellappa. *Handbook of remote biometrics*, volume 1. Springer, 2009.

[23] Himanshu S Bhatt, Richa Singh, Mayank Vatsa, and Nalini K Ratha. Improving cross-resolution face matching using ensemble-based co-transfer learning. *IEEE Transactions on Image Processing*, 23(12):5654–5669, 2014.

[24] BJ Boom, GM Beumer, Luuk J Spreeuwers, and Raymond NJ Veldhuis. The effect of image resolution on the performance of a face recognition system. In *International Conference on Control, Automation, Robotics and Vision*, pages 1–6, 2006.

[25] Pablo H Hennings-Yeomans, Simon Baker, and BVK Vijaya Kumar. Recognition of low-resolution faces using multiple still images and multiple cameras. In *IEEE International Conference on Biometrics: Theory, Applications and Systems*, pages 1–6, 2008.

[26] Pablo H Hennings-Yeomans, Simon Baker, and BVK Vijaya Kumar. Simultaneous super-resolution and feature extraction for recognition of low-resolution faces. In *IEEE International Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–8, 2008.

[27] Zhen Lei and Stan Z Li. Coupled spectral regression for matching heterogeneous faces. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1123–1128, 2009.

[28] Laurenz Wiskott, Jean-Marc Fellous, Norbert Kruger, and Christoph von der Malsburg. Face recognition by elastic bunch graph matching. *Intelligent biometric techniques in fingerprint and face recognition*, 11(5):355–396, 1999.

[29] Timo Ojala, Matti Pietikäinen, and Topi Mäenpää. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (7):971–987, 2002.

[30] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[31] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2005.

[32] Hassen Drira, Boulbaba Ben Amor, Anuj Srivastava, Mohamed Daoudi, and Rim Slama. 3D face recognition under expressions, occlusions, and pose variations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(9):2270–2283, 2013.

[33] Yueming Wang, Jianzhuang Liu, and Xiaoou Tang. Robust 3D face recognition by local shape difference boosting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(10):1858–1870, 2010.

[34] J Kittler, A Hilton, M Hamouz, and J Illingworth. 3D assisted face recognition: A survey of 3D imaging, modelling and recognition approaches. In *IEEE International Conference on Computer Vision and Pattern Recognition Workshops*, pages 114–114, 2005.

[35] Gaurav Goswami, Samarth Bharadwaj, Mayank Vatsa, and Richa Singh. On RGB-D face recognition using kinect. In *IEEE Conference on Biometrics Theory, Applications and Systems*, pages 1–6, 2013.

[36] KaspAROV kinect video dataset. http://iab-rubric.org/resources/Kasparov.html/, 2016.

[37] Jungong Han, Ling Shao, Dong Xu, and Jamie Shotton. Enhanced computer vision with microsoft Kinect sensor: A review. *IEEE Transactions on Cybernetics*, 43(5):1318–1334, 2013.

[38] Billy YL Li, Ajmal Mian, Wanquan Liu, and Aneesh Krishna. Using kinect for face recognition under varying poses, expressions, illumination and disguise. In *IEEE Workshops on Applications of Computer Vision*, pages 186–192, 2013.

[39] Gaurav Goswami, Mayank Vatsa, and Richa Singh. RGB-D face recognition with texture and attribute features. *IEEE Transactions on Information Forensics and Security*, 9(10): 1629–1640, 2014.

[40] S Elaiwat, Mohammed Bennamoun, Farid Boussaid, and A El-Sallam. A curvelet-based approach for textured 3D face recognition. *Elsevier PR*, 48(4):1235–1246, 2015.

[41] Huibin Li, Di Huang, Jean-Marie Morvan, Yunhong Wang, and Liming Chen. Towards 3D face recognition in the real: A registration-free approach using fine-grained matching of 3D keypoint descriptors. *International Journal on Computer Vision*, 113(2):128–142, 2015.

[42] Yue Ming. Robust regional bounding spherical descriptor for 3D face recognition and emotion analysis. *Elsevier IVC*, 35:14–22, 2015.

[43] Billy YL Li, Mingliang Xue, Ajmal Mian, Wanquan Liu, and Aneesh Krishna. Robust RGB-D face recognition using Kinect sensor. *Neurocomputing*, 214:93–108, 2016.

[44] Munawar Hayat, Mohammed Bennamoun, and Amar A El-Sallam. An RGB–D based image set classification for robust face recognition from Kinect data. *Neurocomputing*, 171:889–900, 2016.

[45] Suraj Raghuraman, Kanchan Bahirat, and Balakrishnan Prabhakaran. Evaluating the efficacy of rgb-d cameras for surveillance. In *IEEE ICME*, pages 1–6, 2015.

[46] Federico Pala, Riccardo Satta, Giorgio Fumera, and Fabio Roli. Multimodal person reidentification using rgb-d cameras. *IEEE Transactions on Circuits and Systems for Video Technology*, 26(4):788–799, 2016.

[47] Javier Oliver, Alberto Albiol, Antonio Albiol, and José M Mossi. Using latent features for short-term person re-identification with rgb-d cameras. *Springer PAA*, 19(2):549–561, 2016.

[48] Kinect for Windows Sensor. https://msdn.microsoft.com/en-us/library/hh855355.aspx. [Online; accessed 14-March-2017].

[49] Xtion PRO LIVE, howpublished = "https://www.asus.com/3D-Sensor/Xtion_PRO_LIVE/", note = "[online; accessed 14-march-2017]".

[50] Jason Geng. Structured-light 3D surface imaging: a tutorial. *Advances in Optics and Photonics*, 3(2):128–160, 2011.

[51] Anurag Chowdhury, Soumyadeep Ghosh, Richa Singh, and Mayank Vatsa. RGB-D face recognition via learning-based reconstruction. In *IEEE Conference on Biometrics Theory, Applications and Systems*, pages 1–7, 2016.

[52] RI Hg, Petr Jasek, Clement Rofidal, Kamal Nasrollahi, Thomas B Moeslund, and Gabrielle Tranchet. An RGB-D database using Microsoft's Kinect for windows for face detection. In *IEEE SITIS*, pages 42–46, 2012.

[53] S Gupta, R Girshick, P Arbeláez, and J Malik. Learning rich features from RGB-D images for object detection and segmentation. In *European Conference on Computer Vision*, pages 345–360, 2014.

[54] Ajmal S Mian, Mohammed Bennamoun, and Robyn Owens. Three-dimensional model-based object recognition and segmentation in cluttered scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(10):1584–1601, 2006.

[55] Peter Henry, Michael Krainin, Evan Herbst, Xiaofeng Ren, and Dieter Fox. RGB-D mapping: Using depth cameras for dense 3D modeling of indoor environments. In *ISER*, volume 20, pages 22–25, 2010.

[56] Tri Huynh, Rui Min, and Jean-Luc Dugelay. An efficient LBP-based descriptor for facial depth images applied to gender recognition using RGB-D face data. In *IEEE Asian Conference on Computer Vision Workshops*, pages 133–145, 2012.

[57] Andrej Karpathy, Steven Miller, and Li Fei-Fei. Object discovery in 3D scenes via shape analysis. In *IEEE IRCA*, pages 2088–2095, 2013.

[58] Jiang Wang, Zicheng Liu, Jan Chorowski, Zhuoyuan Chen, and Ying Wu. Robust 3D action recognition with random occupancy patterns. In *IEEE European Conference on Computer Vision*, pages 872–885. 2012.

[59] Gabriele Fanelli, Matthias Dantone, Juergen Gall, Andrea Fossati, and Luc Van Gool. Random forests for real time 3D face analysis. *International Journal on Computer Vision*, 101(3):437–458, 2013.

[60] Nesli Erdogmus and Sébastien Marcel. Spoofing in 2D face recognition with 3D masks and anti-spoofing with Kinect. In *IEEE Conference on Biometrics Theory, Applications and Systems*, pages 1–6, 2013.

[61] Simon Baker and Takeo Kanade. Hallucinating faces. In *IEEE Conference on Automatic Face and Gesture Recognition*, pages 83–88, 2000.

[62] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *European conference on computer vision*, pages 184–199, 2014.

[63] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Neural Information Processing Systems*, pages 2672–2680, 2014.

[64] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[65] P Isola, J-Y Zhu, T Zhou, and A A Efros. Image-to-image translation with conditional adversarial networks. In *IEEE International Conference on Computer Vision*, 2017.

[66] Alexey Dosovitskiy and Thomas Brox. Generating images with perceptual similarity metrics based on deep networks. In *Neural Information Processing Systems*, pages 658–666, 2016.

[67] David Berthelot, Tom Schumm, and Luke Metz. Began: Boundary equilibrium generative adversarial networks. *arXiv preprint arXiv:1703.10717*, 2017.

[68] Junbo Zhao, Michael Mathieu, and Yann LeCun. Energy-based generative adversarial network. *arXiv preprint arXiv:1609.03126*, 2016.

[69] Emily L Denton, Soumith Chintala, Rob Fergus, et al. Deep generative image models using a laplacian pyramid of adversarial networks. In *Neural Information Processing Systems*, pages 1486–1494, 2015.

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

[71] Zili Yi, Hao Zhang, Ping Tan, and Minglun Gong. Dualgan: Unsupervised dual learning for image-to-image translation. In *IEEE International Conference on Computer Vision*, 2017.

[72] Yize Chen, Yishen Wang, Daniel Kirschen, and Baosen Zhang. Model-free renewable scenario generation using generative adversarial networks. *IEEE Transactions on Power Systems*, 33(3):3265–3275, 2018.

[73] Agrim Gupta, Justin Johnson, Li Fei-Fei, Silvio Savarese, and Alexandre Alahi. Social gan: Socially acceptable trajectories with generative adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[74] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. *arXiv preprint arXiv:1609.04802*, 2016.

[75] Jiajun Wu, Chengkai Zhang, Tianfan Xue, Bill Freeman, and Josh Tenenbaum. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. In *Neural Information Processing Systems*, pages 82–90, 2016.

[76] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. *arXiv preprint arXiv:1605.05396*, 2016.

[77] Xiaolong Wang and Abhinav Gupta. Generative image modeling using style and structure adversarial networks. In *European Conference on Computer Vision*, pages 318–335, 2016.

[78] Chuan Li and Michael Wand. Precomputed real-time texture synthesis with markovian generative adversarial networks. In *European Conference on Computer Vision*, pages 702–716, 2016.

[79] Dmitry Ulyanov, Vadim Lebedev, Andrea Vedaldi, and Victor S Lempitsky. Texture networks: Feed-forward synthesis of textures and stylized images. In *International Conference on Machine Learning*, pages 1349–1357, 2016.

[80] Matteo Fabbri, Guido Borghi, Fabio Lanzi, Roberto Vezzani, Simone Calderara, and Rita Cucchiara. Domain translation with conditional gans: from depth to RGB face-to-face. In *IAPR International Conference on Pattern Recognition*, 2018.

[81] Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tieyan Liu, and Wei-Ying Ma. Dual learning for machine translation. In *Neural Information Processing Systems*, pages 820–828, 2016.

[82] Soma Biswas, Gaurav Aggarwal, Patrick J Flynn, and Kevin W Bowyer. Pose-robust recognition of low-resolution face images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(12):3037–3049, 2013.

[83] Soma Biswas, Kevin W Bowyer, and Patrick J Flynn. Multidimensional scaling for matching low-resolution face images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(10):2019–2030, 2012.

[84] Sivaram Prasad Mudunuri and Soma Biswas. Low resolution face recognition across variations in pose and illumination. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(5):1034–1040, 2016.

[85] Stephen Milborrow and Fred Nicolls. Locating facial features with an extended active shape model. In *European conference on computer vision*, pages 504–513, 2008.

[86] Antonio Criminisi, Andrew Blake, Carsten Rother, Jamie Shotton, and Philip HS Torr. Efficient dense stereo with occlusions for new view-synthesis by four-state dynamic programming. *International Journal of Computer Vision*, 71(1):89–110, 2007.

[87] Jingang Shi and Chun Qi. From local geometry to global structure: Learning latent subspace for low-resolution face image recognition. *IEEE Signal Processing Letters*, 22 (5):554–558, 2014.

[88] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *IEEE International Conference on Computer Vision and Pattern Recognition Workshops*, pages 1735–1742, 2006.

[89] Yi Sun, Xiaogang Wang, and Xiaoou Tang. Deep learning face representation from predicting 10,000 classes. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 1891–1898, 2014.

[90] Yi Sun, Yuheng Chen, Xiaogang Wang, and Xiaoou Tang. Deep learning face representation by joint identification-verification. In *Neural Information Processing Systems*, pages 1988–1996, 2014.

[91] Florian Schroff, Dmitry Kalenichenko, and James Philbin. FaceNet: A unified embedding for face recognition and clustering. In *IEEE International Conference on Computer Vision and Pattern Recognition Workshops*, pages 815–823, 2015.

[92] Weihua Chen, Xiaotang Chen, Jianguo Zhang, and Kaiqi Huang. Beyond triplet loss: a deep quadruplet network for person Re-identification. In *IEEE International Conference on Computer Vision and Pattern Recognition Workshops*, 2017.

[93] Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. In *Neural Information Processing Systems*, pages 1857–1865, 2016.

[94] Abhinav Shrivastava, Abhinav Gupta, and Ross Girshick. Training region-based object detectors with online hard example mining. In *IEEE International Conference on Computer Vision and Pattern Recognition Workshops*, pages 761–769, 2016.

[95] Yuhui Yuan, Kuiyuan Yang, and Chao Zhang. Hard-aware deeply cascaded embedding. In *IEEE International Conference on Computer Vision*, 2017.

[96] Yumin Suh, Bohyung Han, Wonsik Kim, and Kyoung Mu Lee. Stochastic class-based hard example mining for deep metric learning. In *IEEE International Conference on Computer Vision and Pattern Recognition Workshops*, pages 7251–7259, 2019.

[97] Ben Harwood, BG Kumar, Gustavo Carneiro, Ian Reid, Tom Drummond, et al. Smart mining for deep metric learning. In *IEEE International Conference on Computer Vision*, pages 2821–2829, 2017.

[98] Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. Deep metric learning via lifted structured feature embedding. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 4004–4012, 2016.

[99] Rui Min, Neslihan Kose, and Jean-Luc Dugelay. KinectfaceDB: A kinect database for face recognition. *IEEE SMC*, 44(11):1534–1548, Nov 2014.

[100] Xtion 2. https:https://www.asus.com/3D-Sensor/Xtion-2, Online; accessed 12-June-2019.

[101] Bahador Khaleghi, Alaa Khamis, Fakhreddine O Karray, and Saiedeh N Razavi. Multisensor data fusion: A review of the state-of-the-art. *Information Fusion*, 14(1):28–44, 2013.

[102] Jiyun Cui, Hao Zhang, Hu Han, Shiguang Shan, and Xilin Chen. Improving 2d face recognition via discriminative face depth estimation. *IAPR International Conference on Biometrics*, pages 1–8, 2018.

[103] Hao Zhang, Hu Han, Jiyun Cui, Shiguang Shan, and Xilin Chen. RGB-D face recognition via deep complementary and common feature learning. In *IEEE Face and Gesture Recognition*, pages 8–15, 2018.

[104] Wen Li, Lin Chen, Dong Xu, and Luc Van Gool. Visual recognition in RGB-D images and videos by learning from RGB-D data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(8):2030–2036, 2018.

[105] Nastaran Nourbakhsh Kaashki and Reza Safabakhsh. RGB-D face recognition under various conditions via 3D constrained local model. *Journal of Visual Communication and Image Representation*, 52:66–85, 2018.

[106] Jiyun Cui, Hu Han, Shiguang Shan, and Xilin Chen. RGB-D face recognition: A comparative study of representative fusion schemes. In *Springer CCBR*, pages 358–366, 2018.

[107] David Joseph Tan, Federico Tombari, and Nassir Navab. Real-time accurate 3D head tracking and pose estimation with consumer RGB-D cameras. *International Journal on Computer Vision*, 126(2-4):158–183, 2018.

[108] Su-Jing Wang, Jian Yang, Na Zhang, and Chun-Guang Zhou. Tensor discriminant color space for face recognition. *IEEE Transactions on Image Processing*, 20(9):2490–2501, 2011.

[109] Cesare Ciaccio, Lingyun Wen, and Guodong Guo. Face recognition robust to head pose changes based on the rgb-d sensor. In *IEEE International Conference on Biometrics: Theory, Applications, and Systems*, pages 1–6, 2013.

[110] Oncel Tuzel, Fatih Porikli, and Peter Meer. Region covariance: A fast descriptor for detection and classification. *European Conference on Computer Vision*, pages 589–600, 2006.

[111] Mauricio Pamplona Segundo, Sudeep Sarkar, Dmitry Goldgof, Luciano Silva, and Olga Bellon. Continuous 3D face authentication using RGB-D cameras. In *IEEE International Conference on Computer Vision and Pattern Recognition Workshops*, pages 64–69, 2013.

[112] Gee-Sern Jison Hsu, Yu-Lun Liu, Hsiao-Chia Peng, and Po-Xun Wu. Rgb-d-based face reconstruction and recognition. *IEEE Transactions on Information Forensics and Security*, 9(12):2110–2118, 2014.

[113] Xinxing Xu, Wen Li, and Dong Xu. Distance metric learning using privileged information for face verification and person re-identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(12):3150–3162, 2015.

[114] Pawas Chhokra, Anurag Chowdhury, Gaurav Goswami, Mayank Vatsa, and Richa Singh. Unconstrained kinect video face database. *Information Fusion*, 44:113–125, 2018.

[115] Luo Jiang, Juyong Zhang, and Bailin Deng. Robust RGB-D face recognition using attribute-aware loss. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 2215–2226, 2019.

[116] FBI launches a face recognition system. http://money.cnn.com/2014/09/16/technology/security/fbi-facial-recognition/index.html. [Online; accessed 6-June-2017].

[117] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research*, 11: 3371–3408, 2010.

[118] G Alistair Watson. Characterization of the subdifferential of some matrix norms. *Linear algebra and its applications*, 170:33–45, 1992.

[119] Shuiwang Ji and Jieping Ye. An accelerated gradient method for trace norm minimization. In *International Conference on Machine Learning*, pages 457–464, 2009.

[120] Afshin Rostamizadeh, Alekh Agarwal, and Peter L Bartlett. Learning with missing features. In *Uncertainty in Artificial Intelligence*, pages 312–319, 2011.

[121] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. Multimodal deep learning. In *International Conference on Machine Learning*, pages 689–696, 2011.

[122] Online Spy Shop. *Do Facial Recognition Cameras In Public Places Infringe On Our Privacy?*, 2017. URL: https://tinyurl.com/y7lju5c5.

[123] Shubin Zhao, Hua Han, and Silong Peng. Wavelet-domain hmt-based image super-resolution. In *IEEE International Conference on Image Processing*, volume 2, pages 949–953, 2003.

[124] Karl S Ni and Truong Q Nguyen. Image superresolution using support vector regression. *IEEE Transactions on Image Processing*, 16(6):1596–1610, 2007.

[125] Kassem Al Ismaeil, Djamila Aouada, Bruno Mirbach, and Björn Ottersten. Enhancement of dynamic depth scenes by upsampling for precise super-resolution (UP-SR). *Computer Vision and Image Understanding*, 147:38–49, 2016.

[126] Arnav V Bhavsar and Ambasamudram N Rajagopalan. Range map superresolution-inpainting, and reconstruction from sparse data. *Computer Vision and Image Understanding*, 116(4):572–591, 2012.

[127] Ce Liu, Heung-Yeung Shum, and William T Freeman. Face hallucination: Theory and practice. *International Journal of Computer Vision*, 75(1):115–134, 2007.

[128] Jianchao Yang, Hao Tang, Yi Ma, and Thomas Huang. Face hallucination via sparse coding. In *IEEE International Conference on Image Processing*, pages 1264–1267, 2008.

[129] Junjun Jiang, Ruimin Hu, Zhongyuan Wang, and Zhen Han. Noise robust face hallucination via locality-constrained representation. *IEEE Transactions on Multimedia*, 16(5): 1268–1281, 2014.

[130] Bo Li, Hong Chang, Shiguang Shan, and Xilin Chen. Low-resolution face recognition via coupled locality preserving mappings. *IEEE Signal processing letters*, 17(1):20–23, 2010.

[131] Wilman WW Zou and Pong C Yuen. Very low resolution face recognition problem. *IEEE Transactions on Image Processing*, 21(1):327–340, 2012.

[132] Sumit Shekhar, Vishal M Patel, and Rama Chellappa. Synthesis-based robust low resolution face recognition. *arXiv preprint arXiv:1707.02733*, 2017.

[133] Himanshu S Bhatt, Richa Singh, Mayank Vatsa, and Nalini Ratha. Matching cross-resolution face images using co-transfer learning. In *IEEE International Conference on Image Processing*, pages 1453–1456, 2012.

[134] Brendan F Klare and Anil K Jain. Heterogeneous face recognition using kernel prototype similarities. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(6): 1410–1422, 2013.

[135] Yang-Jie Cao, Li-Li Jia, Yong-Xia Chen, Nan Lin, Cong Yang, Bo Zhang, Zhi Liu, Xue-Xiang Li, and Hong-Hua Dai. Recent advances of generative adversarial networks in computer vision. *IEEE Access*, 7:14985–15006, 2019.

[136] Cunjian Chen and Arun Ross. Matching thermal to visible face images using a semantic-guided generative adversarial network. *arXiv preprint arXiv:1903.00963*, 2019.

[137] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training GANs. In *Neural Information Processing Systems*, pages 2234–2242, 2016.

[138] Kamran Ghasedi Dizaji, Xiaoqian Wang, and Heng Huang. Semi-supervised generative adversarial network for gene expression inference. In *International Conference on Knowledge Discovery and Data Mining*, pages 1435–1444, 2018.

[139] Terence Sim, Simon Baker, and Maan Bsat. The CMU pose, illumination, and expression (PIE) database. In *IEEE Conference on Face and Gesture Recognition*, pages 53–58, 2002.

[140] Jian Zhao, Lin Xiong, Karlekar Jayashree, Jianshu Li, Fang Zhao, Zhecan Wang, Sugiri Pranata, Shengmei Shen, Shuicheng Yan, and Jiashi Feng. Dual-agent gans for photorealistic and identity preserving profile face synthesis. In *Neural Information Processing Systems*, pages 267–275, 2017.

[141] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 8789–8797, 2018.

[142] Hyunju Maeng, Shengcai Liao, Dongoh Kang, Seong-Whan Lee, and Anil K Jain. Nighttime face recognition at long distance: Cross-distance and cross-spectral matching. In *Asian Conference on Computer Vision*, pages 708–721, 2012.

[143] Xin Yu, Basura Fernando, Richard Hartley, and Fatih Porikli. Super-resolving very low-resolution face images with supplementary attributes. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 908–917, 2018.

[144] Maneet Singh, Shruti Nagpal, Mayank Vatsa, Richa Singh, and Angshul Majumdar. Identity aware synthesis for cross resolution face recognition. In *IEEE International Conference on Computer Vision and Pattern Recognition Workshops*, pages 479–488, 2018.

[145] Miguel De-la Torre, Eric Granger, Paulo VW Radtke, Robert Sabourin, and Dmitry O Gorodnichy. Partially-supervised learning from facial trajectories for face recognition in video surveillance. *Information Fusion*, 24:31–53, 2015.

[146] Jiayi Ma, Pengwei Liang, Wei Yu, Chen Chen, Xiaojie Guo, Jia Wu, and Junjun Jiang. Infrared and visible image fusion via detail preserving adversarial learning. *Information Fusion*, 54:85–98, 2020.

[147] Paulo VW Radtke, Eric Granger, Robert Sabourin, and Dmitry O Gorodnichy. Skew-sensitive boolean combination for adaptive ensembles–an application to face recognition in video surveillance. *Information Fusion*, 20:31–48, 2014.

[148] Sang-Woong Lee, Jooyoung Park, and Seong-Whan Lee. Low resolution face recognition based on support vector data description. *Pattern Recognition*, 39(9):1809–1812, 2006.

[149] Soumyadeep Ghosh, Rohit Keshari, Richa Singh, and Mayank Vatsa. Face identification from low resolution near-infrared images. In *IEEE International Conference on Image Processing*, pages 938–942, 2016.

[150] Ze Lu, Xudong Jiang, and Alex ChiChung Kot. Deep coupled ResNet for low-resolution face recognition. *IEEE Signal Processing Letters*, 2018.

[151] Fuwei Yang, Wenming Yang, Riqiang Gao, and Qingmin Liao. Discriminative multidimensional scaling for low-resolution face recognition. *IEEE Signal Processing Letters*, 25(3):388–392, 2018.

[152] Jiayi Ma, Wei Yu, Pengwei Liang, Chang Li, and Junjun Jiang. Fusiongan: A generative adversarial network for infrared and visible image fusion. *Information Fusion*, 48:11–26, 2019.

[153] Xiao-Yu Zhang, Haichao Shi, Xiaobin Zhu, and Peng Li. Active semi-supervised learning based on self-expressive correlation with generative adversarial networks. *Neurocomputing*, 345:103–113, 2019.

[154] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2337–2346, 2019.

[155] Sangwoo Mo, Minsu Cho, and Jinwoo Shin. Instagan: Instance-aware image-to-image translation. In *International Conference on Learning Representations*, 2019.

[156] Hao Tang, Dan Xu, Nicu Sebe, and Yan Yan. Attention-guided generative adversarial networks for unsupervised image-to-image translation. *arXiv preprint arXiv:1903.12296*, 2019.

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

[158] Donggeun Yoo, Namil Kim, Sunggyun Park, Anthony S Paek, and In So Kweon. Pixel-level domain transfer. In *European Conference on Computer Vision*, pages 517–532, 2016.

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

[160] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 2536–2544, 2016.

[161] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[162] Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. Fast and accurate deep network learning by exponential linear units (ELUs). *arXiv preprint arXiv:1511.07289*, 2015.

[163] Xiang Wu, Ran He, Zhenan Sun, and Tieniu Tan. A light CNN for deep face representation with noisy labels. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(11):2884–2896, 2018.

[164] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *European Conference on Computer Vision*, pages 87–102, 2016.

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

[166] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, et al. Deep face recognition. In *British Machine Vision Conference*, volume 1, page 6, 2015.

[167] Soumyadeep Ghosh, Richa Singh, and Mayank Vatsa. On learning density aware embeddings. In *IEEE conference on computer vision and pattern recognition*, pages 4884–4892, 2019.

[168] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, 21 (12):4695–4708, 2012.

[169] Xiaoxiang Liu, Lingxiao Song, Xiang Wu, and Tieniu Tan. Transferring deep representation for NIR-VIS heterogeneous face recognition. In *IAPR International Conference on Biometrics*, pages 1–8, 2016.

[170] Jing Huo, Yang Gao, Yinghuan Shi, Wanqi Yang, and Hujun Yin. Heterogeneous face recognition by margin-based cross-modality metric learning. *IEEE Transactions on Cybernetics*, 48(6):1814–1826, 2018.

[171] Xiang Wu, Ran He, Zhenan Sun, and Tieniu Tan. A light CNN for deep face representation with noisy labels. *arXiv preprint arXiv:1511.02683*, 2015.

[172] Sarah OBeirne. What cctv users need to know about gdpr, 2017. URL: http://www.fmj.co.uk/cctv-users-need-know-gdpr/.

[173] 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*, 9(3):501–514, 2014.

[174] Shruti Nagpal, Maneet Singh, Richa Singh, Mayank Vatsa, Afzel Noore, and Angshul Majumdar. Face sketch matching via coupled deep transform learning. In *IEEE International Conference on Computer Vision*, pages 5429–5438. 2017.

[175] 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 *IAPR International Conference on Pattern Recognition*, pages 1788–1793, 2014.

[176] Erik Learned-Miller, Gary B Huang, Aruni RoyChowdhury, Haoxiang Li, and Gang Hua. Labeled faces in the wild: A survey. In *Springer AFDFIA*, pages 189–248. 2016.

[177] Meina Kan, Shiguang Shan, Haihong Zhang, Shihong Lao, and Xilin Chen. Multi-view discriminant analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(1):188–194, 2016.

[178] 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*, 14(4):886–896, 2018.

[179] Felix Juefei-Xu, Dipan K Pal, and Marios Savvides. NIR-VIS heterogeneous face recognition via cross-spectral joint dictionary learning and reconstruction. In *IEEE International Conference on Computer Vision and Pattern Recognition Workshops*, pages 141–150, 2015.

[180] Sivaram Prasad Mudunuri and Soma Biswas. A coupled discriminative dictionary and transformation learning approach with applications to cross domain matching. *Pattern Recognition Letters*, 71:38–44, 2016.

[181] 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*, 37(10):2041–2056, 2015.

[182] Dong Yi, Zhen Lei, and Stan Z Li. Shared representation learning for heterogenous face recognition. In *IEEE FG*, volume 1, pages 1–7, 2015.

[183] Shreyas Saxena and Jakob Verbeek. Heterogeneous face recognition with CNNs. In *European Conference on Computer Vision*, pages 483–491, 2016.

[184] José Lezama, Qiang Qiu, and Guillermo Sapiro. Not afraid of the dark: NIR-VIS face recognition via cross-spectral hallucination and low-rank embedding. In *IEEE International Conference on Computer Vision and Pattern Recognition Workshops*, pages 6807–6816, 2017.

[185] Ran He, Xiang Wu, Zhenan Sun, and Tieniu Tan. Learning invariant deep representation for NIR-VIS face recognition. In *AAAI Conference on Artificial Intelligence*, volume 4, pages 7–16, 2017.

[186] Xiang Wu, Lingxiao Song, Ran He, and Tieniu Tan. Coupled deep learning for heterogeneous face recognition. In *AAAI Conference on Artificial Intelligence*, pages 1679–1686, 2018.

[187] Lingxiao Song, Man Zhang, Xiang Wu, and Ran He. Adversarial discriminative heterogeneous face recognition. In *AAAI Conference on Artificial Intelligence*, pages 7355–7362, 2018.

[188] Tiago de Freitas Pereira, André Anjos, and Sébastien Marcel. Heterogeneous face recognition using domain specific units. *IEEE Transactions on Information Forensics and Security*, 14(7):1803–1816, 2018.

[189] 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*, 28(9):4553–4565, 2019.

[190] Ran He, Jie Cao, Lingxiao Song, Zhenan Sun, and Tieniu Tan. Cross-spectral face completion for NIR-VIS heterogeneous face recognition. *arXiv preprint arXiv:1902.03565*, 2019.

[191] Shiming Ge, Shengwei Zhao, Chenyu Li, and Jia Li. Low-resolution face recognition in the wild via selective knowledge distillation. *IEEE Transactions on Image Processing*, 28(4):2051–2062, 2018.

[192] Pei Li, Loreto Prieto, Domingo Mery, and Patrick J Flynn. On low-resolution face recognition in the wild: Comparisons and new techniques. *IEEE Transactions on Information Forensics and Security*, 14(8):2000–2012, 2019.

[193] Erfan Zangeneh, Mohammad Rahmati, and Yalda Mohsenzadeh. Low resolution face recognition using a two-branch deep convolutional neural network architecture. *Expert Systems with Applications*, pages 112–121, 2019.

[194] Omid Abdollahi Aghdam, Behzad Bozorgtabar, Hazim Kemal Ekenel, and Jean-Philippe Thiran. Exploring factors for improving low resolution face recognition. In *IEEE International Conference on Computer Vision and Pattern Recognition Workshops*, pages 1158–1166, 2019.

[195] Maneet Singh, Shruti Nagpal, Richa Singh, and Mayank Vatsa. Dual directed capsule network for very low resolution image recognition. In *IEEE International Conference on Computer Vision*, pages 340–349, 2019.

[196] Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In *IEEE International Conference on Computer Vision and Pattern Recognition Workshops*, volume 1, pages 539–546, 2005.

[197] Yueqi Duan, Jiwen Lu, Jianjiang Feng, and Jie Zhou. Deep localized metric learning. *IEEE Transactions on Circuits and systems for Video Technology*, pages 1212–1222, 2017.

[198] Junlin Hu, Jiwen Lu, Yap-Peng Tan, Junsong Yuan, and Jie Zhou. Local large-margin multi-metric learning for face and kinship verification. *IEEE Transactions on Circuits and systems for Video Technology*, pages 1254–1265, 2017.

[199] Jiwen Lu, Junlin Hu, and Yap-Peng Tan. Discriminative deep metric learning for face and kinship verification. *IEEE Transactions on Image Processing*, 26(9):4269–4282, 2017.

[200] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *European Conference on Computer Vision*, pages 499–515, 2016.

[201] Xiao Zhang, Zhiyuan Fang, Yandong Wen, Zhifeng Li, and Yu Qiao. Range loss for deep face recognition with long-tailed training data. In *IEEE International Conference on Computer Vision*, pages 5409–5418, 2017.

[202] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 212–220, 2017.

[203] Yutong Zheng, Dipan K Pal, and Marios Savvides. Ring loss: Convex feature normalization for face recognition. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 5089–5097, 2018.

[204] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 5265–5274, 2018.

[205] De Cheng, Yihong Gong, Sanping Zhou, Jinjun Wang, and Nanning Zheng. Person Re-identification by multi-channel parts-based CNN with improved triplet loss function. In *IEEE International Conference on Computer Vision and Pattern Recognition Workshops*, pages 1335–1344, 2016.

[206] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person Re-identification. *arXiv preprint arXiv:1703.07737*, 2017.

[207] Xinwei He, Yang Zhou, Zhichao Zhou, Song Bai, and Xiang Bai. Triplet-center loss for multi-view 3D object retrieval. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2018.

[208] Veeru Talreja, Fariborz Taherkhani, Matthew C Valenti, and Nasser M Nasrabadi. Attribute-guided coupled gan for cross-resolution face recognition. *arXiv preprint arXiv:1908.01790*, 2019.

[209] Jiwen Lu, Venice Erin Liong, and Jie Zhou. Simultaneous local binary feature learning and encoding for homogeneous and heterogeneous face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(8):1979–1993, 2018.

[210] Hailin Shi, Yang Yang, Xiangyu Zhu, Shengcai Liao, Zhen Lei, Weishi Zheng, and Stan Z Li. Embedding deep metric for person Re-identification: A study against large variations. In *European Conference on Computer Vision*, pages 732–748, 2016.

[211] Rishabh Garg, Yashasvi Baweja, Richa Singh, Mayank Vatsa, and Nalini Ratha. Heterogeneity aware deep embedding for mobile periocular recognition. In *IEEE International Conference on Biometrics: Theory, Applications, and Systems*, 2018.

[212] BG Kumar, Gustavo Carneiro, Ian Reid, et al. Learning local image descriptors with deep siamese and triplet convolutional networks by minimising global loss functions. In *IEEE International Conference on Computer Vision and Pattern Recognition Workshops*, pages 5385–5394, 2016.

[213] Xufeng Han, Thomas Leung, Yangqing Jia, Rahul Sukthankar, and Alexander C Berg. MatchNet: Unifying feature and metric learning for patch-based matching. In *IEEE International Conference on Computer Vision and Pattern Recognition Workshops*, pages 3279–3286, 2015.

[214] Sergey Zagoruyko and Nikos Komodakis. Learning to compare image patches via convolutional neural networks. In *IEEE International Conference on Computer Vision and Pattern Recognition Workshops*, pages 4353–4361, 2015.

[215] Jonathan Masci, Davide Migliore, Michael M Bronstein, and Jürgen Schmidhuber. Descriptor learning for omnidirectional image matching. In *RRIV*, pages 49–62. Springer, 2014.

[216] Paul Wohlhart and Vincent Lepetit. Learning descriptors for object recognition and 3D pose estimation. In *IEEE International Conference on Computer Vision and Pattern Recognition Workshops*, pages 3109–3118, 2015.

[217] Dorin Comaniciu and Peter Meer. Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(5):603–619, 2002.

[218] Kilian Q Weinberger and Lawrence K Saul. Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research*, 10:207–244, 2009.

[219] Hyun Oh Song, Stefanie Jegelka, Vivek Rathod, and Kevin Murphy. Deep metric learning via facility location. In *IEEE International Conference on Computer Vision and Pattern Recognition Workshops*, 2017.

[220] Leonard Berrada, Andrew Zisserman, and M Pawan Kumar. Smooth loss functions for deep top-k classification. In *International Conference on Learning Representations*, 2018.

[221] Dejun Chu, Rui Lu, Jin Li, Xintong Yu, Changshui Zhang, and Qing Tao. Optimizing top-k multiclass SVM via semismooth newton algorithm. *IEEE Transactions on Neural Networks and Learning Systems*, (99):1–12, 2018.

[222] Bin Wang, Rui Zhu, Xiaochun Yang, and Guoren Wang. Top-k representative documents query over geo-textual data stream. *World Wide Web Conference*, 21(2):537–555, 2018.

[223] Rong Jin and Zoubin Ghahramani. Learning with multiple labels. In *Neural Information Processing Systems*, pages 921–928, 2003.

[224] Maksim Lapin, Matthias Hein, and Bernt Schiele. Top-k multiclass SVM. In *Neural Information Processing Systems*, pages 325–333, 2015.

[225] Shuang-Hong Yang, Hongyuan Zha, and Bao-Gang Hu. Dirichlet-bernoulli alignment: A generative model for multi-class multi-label multi-instance corpora. In *Neural Information Processing Systems*, pages 2143–2150, 2009.

[226] Zhi-Hua Zhou and Min-Ling Zhang. Multi-instance multi-label learning with application to scene classification. In *Neural Information Processing Systems*, pages 1609–1616, 2007.

[227] Maya R Gupta, Samy Bengio, and Jason Weston. Training highly multiclass classifiers. *Journal of Machine Learning Research*, 15(1):1461–1492, 2014.

[228] Maksim Lapin, Matthias Hein, and Bernt Schiele. Analysis and optimization of loss functions for multiclass, top-k, and multilabel classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(7):1533–1554, 2018.

[229] Caixia Yan, Minnan Luo, Huan Liu, Zhihui Li, and Qinghua Zheng. Top-k multi-class SVM using multiple features. *Information Sciences*, 432:479–494, 2018.

[230] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3D object representations for fine-grained categorization. In *IEEE International Conference on Computer Vision Workshops*, pages 554–561, 2013.

[231] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *ACM*, number 34, pages 226–231, 1996.

[232] Thanh-Toan Do, Toan Tran, Ian Reid, Vijay Kumar, Tuan Hoang, and Gustavo Carneiro. A theoretically sound upper bound on the triplet loss for improving the efficiency of deep distance metric learning. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 10404–10413, 2019.

[233] Yair Movshovitz-Attias, Alexander Toshev, Thomas K Leung, Sergey Ioffe, and Saurabh Singh. No fuss distance metric learning using proxies. *arXiv preprint arXiv:1703.07464*, 2017.

[234] Jian Wang, Feng Zhou, Shilei Wen, Xiao Liu, and Yuanqing Lin. Deep metric learning with angular loss. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 2593–2601, 2017.

[235] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

Dissemination of Research Contributions

Publications Related to the Thesis

- S. Ghosh, M. Vatsa, R. Singh, On Learning Discriminative Embeddings for Optimized Top-k Matching, Pattern Recognition, (*under major revision*), 2022.
- S. Ghosh, M. Vatsa, R. Singh, SUPREAR-NET: Supervised Resolution Enhancement and Recognition Network, IEEE Transactions on Biometrics, Behavior and Identity Science, 2022.
- S. Ghosh, R. Singh, M. Vatsa, A. Noore, RGB-D Face Recognition using Reconstruction based Shared Representation, IEEE International Conference on Automatic Face and Gesture Recognition (pp. 1-8), 2021.
- S. Ghosh, M. Vatsa, R. Singh, Subclass Heterogeneity Aware Loss for Unconstrained Face Recognition, IEEE Transactions on Biometrics, Behavior and Identity Science, 2.3 (2020): 245-256.
- S. Ghosh, R. Singh , M. Vatsa, N. Ratha, , V. M. Patel, Domain adaptation for visual understanding. In Domain Adaptation for Visual Understanding (pp. 1-15), 2020.
- S. Ghosh, M. Vatsa, R. Singh, On Learning Density Aware Embeddings, IEEE Conference on Computer Vision and Pattern Recognition, pp. 4884-4892, 2019.
- S. Gupta, N. Gupta, S. Ghosh, M. Singh, S. Nagpal, R. Singh, M. Vatsa, A Benchmark Video Dataset for Face Detection and Recognition Across Spectra and Resolutions, In IEEE International Conference on Automatic Face and Gesture Recognition, pp. 1-7, 2019.
- A. Chowdhury, S. Ghosh, R. Singh, and M. Vatsa, RGB-D Face Recognition via Learning-based Reconstruction, In IEEE International Conference on Biometrics: Theory, Applications and Systems, pp. 1-7, 2016. (Received the Best Poster Award)
- S. Ghosh, R. Keshari, R. Singh, M. Vatsa, Face Identification from Low Resolution Near-Infrared Images, In IEEE International Conference on Image Processing, pp. 938-942, 2016.

Other Publications

Journal and Conference Articles

- T.I. Dhamecha , S. Ghosh, M. Vatsa, and R. Singh, Kernelized Heterogeneity-Aware Cross-View Face Recognition. Frontiers in Artificial Intelligence, p.68-78., 2021.
- P. Tripathy, R. Keshari, S. Ghosh, M. Vatsa, R.Singh, Auto-G: Gesture Recognition In the Crowd for Autonomous Vehicles, IEEE International Conference on Image Processing, 2019.
- Ramya Y. S., S. Ghosh, M. Vatsa, and R. Singh, Face Sketch Image Colorization via Supervised GANs, IAPR International Conference On Biometrics, pp 1-6, 2019.
- R. Garg, Y. Baweja, S. Ghosh, R. Singh, M. Vatsa, N. Ratha, Heterogeneity Aware Deep Embedding for Mobile Periocular Recognition, In IEEE International Conference on Biometrics: Theory, Applications and Systems, 2018.
- M. Singh, S. Nagpal, N. Gupta, S. Gupta, S. Ghosh, R. Singh, and M. Vatsa, Cross-Spectral Cross-Resolution Video Database for Face Recognition, In IEEE International Conference on Biometrics: Theory, Applications and Systems, pp 1-6, 2016.
- R. Keshari, S. Ghosh, A. Agarwal, R. Singh, M. Vatsa, Mobile Periocular Matching with Pre-Post Cataract Surgery, In IEEE International Conference on Image Processing, 2016.
- S. Ghosh, T. I. Dhamecha, R. Keshari, R. Singh, and M. Vatsa, Feature and Keypoint Selection for Visible to Near-infrared Face Matching, In IEEE International Conference on Biometrics: Theory, Applications and Systems, pp 1-6, 2015.
- A. Sankaran, A. Agarwal, R. Keshari, S. Ghosh, A. Sharma, M. Vatsa, and R. Singh, Latent Fingerprint from Multiple Surfaces: Database and Quality Analysis, In IEEE International Conference on Biometrics: Theory, Applications and Systems, pp 1-6, 2015.

Other Publications

- D. Sundriyal, S. Ghosh, M. Vatsa, and R. Singh, Semi-supervised Learning via Triplet Network Based Active Learning, AAAI Student Abstract , 2021.
- R.Keshari, S. Ghosh, S.Chhabra, M. Vatsa, and R. Singh, Unravelling Small Sample Size Problems in the Deep Learning World, IEEE Sixth International Conference on Multimedia Big Data, 2020.
- S. Ghosh, Can artificial intelligence create jobs?, The Financial Express, 2016. DOI https://tinyurl.com/yckfyct9