2021

Deep Learning Architectures for Heterogeneous Face Recognition

Seyed Mehdi Iranmanesh
West Virginia University, seiranmanesh@mix.wvu.edu

2021

Deep Learning Architectures for Heterogeneous Face Recognition

Seyed Mehdi Iranmanesh

Deep Learning Architectures for Heterogeneous Face Recognition

Seyed Mehdi Iranmanesh

Dissertation submitted to the
Benjamin M. Statler College of Engineering and Mineral Resources
at West Virginia University

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Computer Science

Nasser M. Nasrabadi, Ph.D., Chair
Donald A. Adjeroh, Ph.D.
Jacqueline A. Speir, Ph.D.
Matthew C. Valenti, Ph.D.
Jeremy Dawson , Ph.D.

Lane Department of Computer Science and Electrical Engineering

Morgantown, West Virginia
2021

Keywords: Heterogeneous Face Recognition, Deep Learning, Adversarial Learning, Privileged Information, Multi-task Learning, Auxiliary Data, Facial Attribute, Facial Landmark

Abstract

Deep Learning Architectures for Heterogeneous Face Recognition

Seyed Mehdi Iranmanesh

Face recognition has been one of the most challenging areas of research in biometrics and computer vision. Many face recognition algorithms are designed to address illumination and pose problems for visible face images. In recent years, there has been significant amount of research in Heterogeneous Face Recognition (HFR). The large modality gap between faces captured in different spectrum as well as lack of training data makes heterogeneous face recognition (HFR) quite a challenging problem. In this work, we present different deep learning frameworks to address the problem of matching non-visible face photos against a gallery of visible faces.

Algorithms for thermal-to-visible face recognition can be categorized as cross-spectrum feature-based methods, or cross-spectrum image synthesis methods. In cross-spectrum feature-based face recognition a thermal probe is matched against a gallery of visible faces corresponding to the real-world scenario, in a feature subspace. The second category synthesizes a visible-like image from a thermal image which can then be used by any commercial visible spectrum face recognition system. These methods also beneficial in the sense that the synthesized visible face image can be directly utilized by existing face recognition systems which operate only on the visible face imagery. Therefore, using this approach one can leverage the existing commercial-off-the-shelf (COTS) and government-off-the-shelf (GOTS) solutions. In addition, the synthesized images can be used by human examiners for different purposes.

There are some informative traits, such as age, gender, ethnicity, race, and hair color, which are not distinctive enough for the sake of recognition, but still can act as complementary information to other primary information, such as face and fingerprint. These traits, which are known as soft biometrics, can improve recognition algorithms while they are much cheaper and faster to acquire. They can be directly used in a unimodal system for some applications. Usually, soft biometric traits have been utilized jointly with hard biometrics (face photo) for different tasks in the sense that they are considered to be available both during the training and testing phases. In our approaches we look at this problem in a different way. We consider the case when soft biometric information does not exist during the testing phase, and our method can predict them directly in a *multi-tasking* paradigm.

There are situations in which training data might come equipped with additional information that can be modeled as an *auxiliary view* of the data, and that unfortunately is not available during testing. This is the LUPI scenario. We introduce a novel framework based on deep learning techniques that leverages the auxiliary view to improve the performance of recognition system. We do so by introducing a formulation that is general, in the sense that can be used with any visual classifier.

Every use of auxiliary information has been validated extensively using publicly available benchmark datasets, and several new state-of-the-art accuracy performance values have been set. Examples of application domains include visual object recognition from RGB images and from depth data, handwritten digit recognition, and gesture recognition from video.

We also design a novel aggregation framework which optimizes the landmark locations directly using only one image without requiring any extra prior which leads to robust alignment given arbitrary face deformations. Three different approaches are employed to generate the manipulated faces and two of them perform the manipulation via the adversarial attacks to fool a face recognizer. This step can decouple from our framework and potentially used to enhance other landmark detectors. Aggregation of the manipulated faces in different branches of proposed method leads to robust landmark detection.

Finally we focus on the generative adversarial networks which is a very powerful tool in synthesizing a visible-like images from the non-visible images. The main goal of a generative model is to approximate the true data distribution which is not known. In general, the choice for modeling the density function is challenging. Explicit models have the advantage of explicitly calculating the probability densities. There are two well-known implicit approaches, namely the Generative Adversarial Network (GAN) and Variational AutoEncoder (VAE) which try to model the data distribution implicitly. The VAEs try to maximize the data likelihood lower bound, while a GAN performs a minimax game between two players during its optimization. GANs overlook the explicit data density characteristics which leads to undesirable quantitative evaluations and mode collapse. This causes the generator to create similar looking images with poor diversity of samples. In the last chapter of thesis, we focus to address this issue in GANs framework.

Dedication

This thesis is dedicated to my parents for their love, support, and sacrifice. Also to all of my teachers during my student life for their encouragement and motivation.

Acknowledgments

I want to especially thank my PhD advisor, Dr. Nasser M. Nasrabadi, for his great support and all of my labmates.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Problem Definition

In the past few years, computer vision and biometrics technology has reached the attention of the masses because of the widespread use of different cameras, from cellphone to surveillance cameras to more advanced imaging sensors. Computer vision and biometrics deals with acquiring, processing, and understanding images in order to solve different tasks.

This proposal focuses on the Heterogeneous Face Recognition (HFR) task. Face recognition has been one of the most challenging areas of research in biometrics and computer vision. Many face recognition algorithms are designed to address illumination and pose problems for visible face images. In recent years, there has been significant amount of research in HFR [11]. The main issue in HFR is to match the visible face image to a face image that has been captured in another domain such as the infrared spectrum [12, 13], polarimetric [14], or millimeter wave [15] due to the significant phenomenological difference as well as the lack of training data. Infrared images are categorized into two major groups of reflection and emission. The reflection category, which contains near infrared (NIR) and shortwave infrared (SWIR) bands, is more informative about the facial details and it is very similar to the visible imagery. Due to this reflective phenomenology of NIR and SWIR, there has been a significant performance on the NIR-to-visible face recognition accuracy [16, 12] and to some extent for SWIR-to-visible face recognition accuracy [17, 13].

In this thesis, we also focus on additional information such as facial information that can be collected and used in training. In addition, since one of the main challenges in HFR is synthesizing a visible like image from a non-visible modality, we try to improve the generative adversarial networks which are very

powerful tool for synthesizing images.

1.2 Motivation and Challenges

Due to the significant differences between the phenomenology of thermal and visible imagery, matching a thermal face against a gallery of visible faces becomes a challenging task. However, thermal-to-visible recognition is highly demanding because in the thermal data no active illumination is needed at night-time or low-light environments since the thermal imagery is based on the emission originating from the underlying skin and depends on the individual's physiology.

Moreover, face sketch recognition is an important problem when the photo of a suspect is not available or is captured with very poor quality. A face sketch is usually drawn by a forensic artist [18] or facial software [19] based on the information provided by a victim, or an eye-witness. Therefore, the generated sketch using the provided description of the victim is the only clue to identify the victim. An automatic matching method is necessary to identify a suspect accurately via searching the law enforcement face database or surveillance cameras using only the sketch of the suspect. The sketch recognition problem has been extensively studied in recent years [20]. Due to the large phenomenological gap between sketch and photo domains, sketch recognition problem still remains a challenging task.

1.2.1 Heterogeneous Face Recognition

In recent years, there has been a growing research on thermal-to-visible face recognition [21, 22, 23, 24] and thermal-to-visible detection [25]. Visible images contain rich textural and geometrical details with the edges of the key facial structures (i.e, mouth, eyes, and nose) which clearly observable. However, in the conventional thermal images some edges around the eye, eyebrows do appear but they suffer from significant lack of details compared to the corresponding visible images, thus highlighting the large domain gap.

Recently, via an emerging technology [26], the polarization state information of thermal emission has been exploited to provide additional geometrical and textural details, especially around the nose and the mouth, which complements the textural details of the conventional intensity-based thermal images. This additional information is not available in the conventional intensity-based thermal imaging [26], and is utilized in recent algorithms to enhance the cross thermal-to-visible face recognition [26, 27, 28].

1.2.2 Heterogeneous Face Recognition using Facial Attributes

Forensic or composite sketches contain limited information such as a rough spatial topology of the suspect face and lack of some complementary information such as skin color, ethnicity, or hair color are noticeable. In addition, sketch recognition problems mainly focus on single sketch which can be unreliable in real-world situations. This unreliability can lead to a false identification [29]. In forensics investigation multiple sources of information such as verbal description of multiple witnesses or the verbal description and poor video surveillance can be utilized to enhance the performance of suspect identification [30, 31].

In general there are two classical ways to solve the sketch recognition problem. First approach namely generative methods transfer one of the modalities (either sketch or photo) to the other before matching [32, 33]. In the second approach, the discriminative methods utilize feature descriptors such as the scale-invariant feature transform (SIFT) [34], Weber's local descriptor (WLD) [35], and multi-scale local binary pattern (MLBP) [36]. The main drawbacks of these feature descriptors is that they might not be the optimal features for the task of sketch-photo recognition. To compensate for this, some other methods in the literature propose to extract modality-invariant features [23, 37].

Recently, in the literature soft biometric traits have been utilized jointly with hard biometrics (face photo) for different tasks such as person identification or face recognition [38]. In fact, using facial attributes in conjunction with sketch would be more advantageous since some attributes such as eye color, hair color, skin color, and ethnicity do not exist in sketch and could be considered as the complementary information. Moreover, some attributes such as wearing a hat or eyeglasses can be utilized as an auxiliary information to narrow down the suspect in the databases more accurately.

1.2.3 Soft Biometric as a Privileged Data to Improve Deep Face Recognition

Biometric systems can recognize different identities based on various physical, behavioral features and characteristics [39]. In general, unimodal recognition systems, which utilize the data from a single biometric trait of the individual, are prone to failures arising from distorted biometric traits due to sensor noise and limitations of feature extractors. They are also vulnerable to inter-class similarities, especially in the case of large population, which make them less reliable to acquire enough distinctive features from a single modality [40]. A multimodal biometric system with multiple modalities, such as face, fingerprint, and iris, is expected to be more reliable and accurate due to utilization of different sources of information. However,

acquisition of this information is costly and a tedious task which can affect the popularity and ease of using multimodal biometric recognition systems. On the other hand, there are some informative traits, such as age, gender, ethnicity, race, and hair color, which are not distinctive enough for the sake of recognition, but still can act as complementary information to other primary information, such as face and fingerprint. These traits, which are known as soft biometrics, can improve recognition algorithms while they are much cheaper and faster to acquire.

The main question here is can we train a visual recognition model that is more robust to main data view changes in testing when we are given additional information during training? The answer is "yes". It is possible to learn a shared embedding space using the given views (information) in training which is more robust to main data view changes in testing. Typically, this problem is addressed by processing the available data with classifiers trained on the same modalities. However, the missing modality at testing time can be seen as additional information, available only during training. This additional information is an *auxiliary* data of the image/video sample.

The problem of using auxiliary information has been studied within the concept of learning using privileged information (LUPI) as it has been introduced by Vapnik in [41]. The existence of the privileged (auxiliary) information can help to improve the performance of prediction and classification tasks during the testing phase, and increase the rate of convergence in the training phase. In this section, we consider to develop a classifier for a more general and complex scenario in which the privileged information (soft biometrics) and the primary data (hard biometrics) are true heterogeneous modalities.

The lack of auxiliary information during the testing phase, enforces some limitations on how to utilize the auxiliary data during the training phase. The paradigm needs to relate the main modality and the available auxiliary information, and exploit this relationship to increase the convergence rate or learn features which are more discriminative. Sharmanska et. al. [42], have confirmed that the privileged information can distinguish between hard and easy samples in the training set. They made the assumption that the privileged and the primary data share similar informative information, and consequently, the "easy to classify" or "hard to classify" samples are the same in the primary and privileged data. However, in our work, LUPI paradigm is employed to justify that the soft biometrics information can act as a complementary information to the primary data.

1.2.4 Robust Facial Landmark Detection via Aggregation on Geometrically Manipulated Faces

A common approach to facial landmark detection problem is to leverage deep features from ConvNets. These facial features and regressors are trained in an end-to-end manner utilizing a cascade strategy to update the landmark locations progressively [43, 44]. Yu et al. [45] integrate geometric constraints within CNN architecture using a deep deformation network. Lev et al. [46] propose a deep regression framework with two-step re-initialization to avoid the initialization issue. Zhu et al. [5] also tried to deal with poor initialization utilizing a coarse search over a shape space with variant shapes. In another work, Zhu et al. [44], overcome the extreme head poses and rich shape deformations exploiting cascaded regressors.

Another category of landmark detection approaches leverages the end-to-end training from ConvNets frameworks to learn robust heatmaps for landmark detection task [47, 48, 49]. Balut et al. [48] utilized the residual framework to propose a robust network for facial landmark detection. Newell et al. [49] and Wei et al. [47] consider the coordinate of the highest response on the heatmaps as the location of landmarks for human pose estimation task.

In a more general definition, this problem can also be viewed as learning structural representation. Some studies [50, 51], disentangle visual content into different factors of variations such as camera viewpoint, motion and identity to capture the inherent structure of objects. However, the physical parameters of these factors are embedded in a latent representation which is not discernible. Some methods can handle [52, 53] conceptualize structures in the multi-tasking framework as auxiliary information (*e.g.*, landmarks, depth, and mask). Such structures in these frameworks are designed by humans and need supervision to learn.

1.2.5 Generative Adversarial Networks and Avoiding mode Collapse Issue

Generative models have extensively grown in recent years. The main goal of a generative model is to approximate the true data distribution which is not known. Generative models are based on finding the model parameters that maximize the likelihood of the training data. This is equivalent to minimizing the Kullback-Leibler (KL) divergence ($D_{KL}(p_{data}||p_{model})$) between the data distribution p_{data} and model distribution p_{model}. Although this objective spans multiple modes of the data, it leads to generating vague and undesirable samples [54]. There are other approaches that minimize $D_{KL}(p_{model}||p_{data})$ which are usually referred to as the reverse KL divergence [55] and this is the main idea behind the generative adversarial

networks. Although these models generate sharp images, minimizing the reverse KL divergence causes the model distribution to focus on a single mode of the data and ignore the other modes. This is known as the mode collapse in the generative adversarial models [56]. This happens because the reverse KL divergence measures the dissimilarity between two distributions for the fake samples, and there is no penalty on the fraction of the model distribution that covers the data distribution [57]. To address this problem, the authors in [58] suggested the Wassertein distance which has the weakest convergence among existing GAN metrics. This new metric is a powerful tool to avoid mode collapse. However, they used weight clipping to approximate the Wassertain distance which causes a pathological behavior [56].

In general, the choice for modeling the density function is challenging. There are two ways to estimate the density function namely, implicit methods and explicit methods. Implicit approaches tend to calculate the model parameters without the need for the analytical form of p_{model}. Explicit models have the advantage of explicitly calculating the probability densities. There are two well-known implicit approaches, namely the Generative Adversarial Network (GAN) and Variational AutoEncoder (VAE) which try to model the data distribution implicitly. The VAEs try to maximize the data likelihood lower bound, while a GAN performs a minimax game between two players during its optimization in which for an optimal discriminator, the algorithm tries to find a generator that minimizes the Jensen-Shannon divergence (JSD). The JSD minimization has been proven empirically to behave more similar to the reverse KL divergence rather than the KL divergence [59, 56]. This behavior leads to the aforementioned problem of mode collapse in GAN models, which causes the generator to create similar looking images with poor diversity of samples.

1.3 Contributions and Thesis Structure

In this thesis, we introduce algorithms for each of the problems we explained in the previous section. Chapter 2 addresses the heterogeneous face recognition problem. Chapter 3 proposes a heterogeneous face recognition systems which utilize some soft biometric traits (facial attributes) in addition to the main face images (hard biometric traits). Chapter 4 addresses the LUPI problem. Chapter 5 tries to add an adaptive margin to angular softmax loss. This leads to a more discriminative embedding space. Chapter 6 addresses the problem of facial landmark detection utilizing aggregation on manipulated faces which are generated using only one given image. This method is very useful for careful augmentation and also self-supervised learning. Chapter 7 focuses on the theoretical aspect of generative adversarial networks and attempts to

avoid mode collapse issue in them.

1.3.1 Chapter 2

The large modality gap between faces captured in different spectrum makes heterogeneous face recognition (HFR) quite a challenging problem. In this chapter, we present two methods based on deep networks. The first method is a coupled deep neural network to find global discriminative features in a nonlinear embedding space to relate the polarimetric thermal faces to their corresponding visible faces. In the second approach a coupled generative adversarial network (CpGAN) is employed to address the problem of matching non-visible face photos against a gallery of visible faces.

Our CpGAN architecture consists of two sub-networks one dedicated to the visible spectrum and the other sub-network dedicated to the non-visible spectrum. Each sub-network consists of a generative adversarial network (GAN) architecture. Inspired by a *dense network* which is capable of maximizing the information flow among features at different levels, we utilize a densely connected encoder-decoder structure as the generator in each GAN sub-network. The proposed CpGAN framework uses multiple loss functions to force the features from each sub-network to be as close as possible for the same identities in a common latent subspace. To achieve a realistic photo reconstruction while preserving the discriminative information, we also added a perceptual loss function to the coupling loss function. An ablation study is performed to show the effectiveness of different loss functions in optimizing the proposed method. Moreover, the superiority of the model compared to the state-of-the-art models in HFR is demonstrated using multiple datasets.

1.3.2 Chapter 3

In this chapter we proposed two methods. In the first approach we present a deep coupled framework to address the problem of matching sketch image against a gallery of mugshots. Face sketches have the essential information about the spatial topology and geometric details of faces while missing some important facial attributes such as ethnicity, hair, eye, and skin color. We propose a coupled deep neural network architecture which utilizes facial attributes in order to improve the sketch-photo recognition performance. The proposed Attribute-Assisted Deep Convolutional Neural Network (AADCNN) method exploits the facial attributes and leverages the loss functions from the facial attributes identification and face verification tasks in order to learn rich discriminative features in a common embedding subspace. The facial attribute identification task increases the inter-personal variations by pushing apart the embedded features extracted

from individuals with different facial attributes, while the verification task reduces the intra-personal variations by pulling together all the features that are related to one person. The learned discriminative features can be well generalized to new identities not seen in the training data. The proposed architecture is able to make full use of the sketch and complementary facial attribute information to train a deep model compared to the conventional sketch-photo recognition methods. Extensive experiments are performed on composite (E-PRIP) and semi-forensic (IIIT-D semi-forensic) datasets.

We also present a facial attribute-guided deep coupled learning framework to address the problem of matching polarimetric thermal face photos against a gallery of visible faces. The coupled framework contains two sub-network one dedicated to the visible spectrum and the second sub-network dedicated to the polarimetric thermal spectrum. Each sub-network is made of a generative adversarial network (GAN) architecture. We propose a novel Attribute-Guided Coupled Generative Adversarial Network (AGC-GAN) architecture which utilizes facial attributes in order to improve the thermal-to-visible face recognition performance. The proposed AGC-GAN method exploits the facial attributes and leverages multiple loss functions in order to learn rich discriminative features in a common embedding subspace. To achieve a realistic photo reconstruction while preserving the discriminative information we add a perceptual loss term to the coupling loss function.

1.3.3 Chapter 4

We present a novel framework to exploit privileged information for face recognition which is provided only during the training phase. Here, we focus on face classification/verification task, where RGB face images are provided as the main view and soft biometric traits (age, ethnicity, etc) are provided as the privileged (auxiliary) data. We assume that the soft biometric traits are only available during the training phase and consider two different deep architectures in order to make full use of this additional privileged information. In the first approach, a coupled deep neural network (Cp-DNN) architecture is proposed in which a pair of networks are employed during the training phase to find a common latent feature space between the main and privileged data. These two networks are simultaneously trained using both the main and auxiliary data during the training phase. The Cp-DNN learns the complementary information from the privileged data to improve the recognition of the main data. In the second approach, a multi-task learning (MTL) framework is proposed to perform learning using privileged information (LUPI) during the training phase. We refer to this second approach as MTL-LUPI. In this approach, network tries to simultaneously

predict the privileged soft biometric information and perform the classification/verification task by using only the main data. The network learns the common features in a multi-task learning paradigm. Learning the privileged tasks, helps the network to refine its features and enhance the main task performance. Extensive experiments are performed on four different datasets and the results show the superiority of our method compared to the state-of-the-art LUPI models in the face recognition task.

1.3.4 Chapter 5

In this chapter, we demonstrate that more discriminative feature space can be learned by enforcing a deep network to adjust adaptive margins between classes utilizing attributes. This tight constraint also effectively reduces the class imbalance inherent in the local data neighborhood, thus carving more balanced class boundaries locally and using feature space more efficiently. Extensive experiments are performed on five different datasets and the results show the superiority of our method compared to the state-of-the-art models in both tasks of face recognition and person re-identification.

1.3.5 Chapter 6

In this work, we present a practical approach to the problem of facial landmark detection. The proposed method can deal with large shape and appearance variations under the rich shape deformation. To handle the shape variations we equip our method with the aggregation of manipulated face images. The proposed framework generates different manipulated faces using only one given face image. The approach utilizes the fact that small but carefully crafted geometric manipulation in the input domain can fool deep face recognition models. We propose three different approaches to generate manipulated faces in which two of them perform the manipulations via adversarial attacks and the other one uses known transformations. Aggregating the manipulated faces provides a more robust landmark detection approach which is able to capture more important deformations and variations of the face shapes. Our approach is demonstrated its superiority compared to the state-of-the-art method on benchmark datasets AFLW, 300-W, and COFW.

1.3.6 Chapter 7

In this chapter, we present a simple approach to train Generative Adversarial Networks (GANs) in order to avoid a *mode collapse* issue. Implicit models such as GANs tend to generate better samples compared to explicit models that are trained on tractable data likelihood. However, GANs overlook the explicit data

density characteristics which leads to undesirable quantitative evaluations and mode collapse. To bridge this gap, we propose a autoregressive generative adversarial network (AutoGAN) for which we can enforce data density estimation via an autoregressive model and support both adversarial and likelihood framework in a joint training manner which diversify the estimated density in order to cover different modes. We propose to use an adversarial network to *transfer knowledge* from an autoregressive model (teacher) to the generator (student) of a GAN model.

1.4 Related work

Heterogeneous Face Recognition. Algorithms for thermal-to-visible face recognition can be categorized as cross-spectrum feature-based methods, or cross-spectrum image synthesis methods. In cross-spectrum feature-based face recognition a thermal probe is matched against a gallery of visible faces corresponding to the real-world scenario [14], in a feature subspace. The second category synthesizes a visible-like image from a thermal image which can then be used by any commercial visible spectrum face recognition system.

Recently, almost all the state-of-the-art techniques in face recognition task have applied deep convolutional neural networks (DCNN) trained on large datasets to build up a compact discriminative feature subspace. This approach also has been applied in other cross-modal applications such as pedestrian detection [25], and cross-modal retrieval [60] to find a representative embedding subspace. In [61], the authors trained a network on a private dataset containing 4.4 million labeled images of 4,030 different subjects. They also fine-tuned their network with a Siamese network [62] for a face verification task. They also extended their work with an expanded dataset which contained 500 million images related to 10 million subjects. Sun et al. [63, 64, 65, 66] studied a deep neural network architecture employing a joint verification-identification loss function and Bayesian metrics in their works. They used two different datasets, namely, CelebFaces [63] (202,599 images of 10,177 different subjects) and WDRef [67] (99,773 images of 2,995 subjects) to train their deep networks. Schroff et al. [68] also trained a deep network using 200 million images of 8 million different subjects. This network gained the best performance on Labeled Faces in the Wild (LFW) [69] dataset, which is a standard unconstrained face recognition benchmark.

Researchers have also investigated a variety of approaches to exploit the polarimetric LWIR thermal images to improve the cross-spectrum face recognition [26, 27, 70, 71]. One of the first methods developed for polarimetric thermal-to-visible extracted the histogram of oriented gradients (HOG) features from S_0, S_1, and, S_2 and combined them together and performed a one-versus-all support vector machine (SVM)

classifier to do the face recognition [72]. Another work utilized similar approach to extract features [73]. However, they used partial least square (PLS), on top of the extracted features and learned a one-vs-all PLS discriminant analysis classifier.

Recent cross-spectrum feature based approaches learn a function to explicitly map the visible and polarimetric thermal features. Riggan et al. [70, 74] employed a deep perpetual mapping (DPM) and the coupled neural network (CpNN) for polarimetric thermal-to-visible face recognition. The DPM technique [75] learns a direct mapping between the scale invariant feature transform (SIFT) features from the thermal imagery and the corresponding visible SIFT feature subspace by using a multilayer neural network, and then reconstructs the visible image from these features. In contrast, CpNN [74] performs an indirect mapping between thermal and visible SIFT features. The authors in [74] developed a method to jointly learn two mappings in order to extract the shared latent features. The authors also added one-vs-all PLS classification on top of CpNN or DPM to enhance the recognition accuracy. These two approaches are referred to as PLSoDPM [14] and PLSoCpNN [70].

The second category of approaches attempt to synthesize a visible-like image from another modality such as NIR, thermal, or polarimetric thermal image input. These methods also beneficial in the sense that the synthesized visible face image can be directly utilized by existing face recognition systems which operate only on the visible face imagery. In [76], the authors developed a method to synthesize a visible-like face image from the polarimetric input. In order to perform synthesizing, they utilized DPM to map SIFT features to the corresponding SIFT features in visible domain, and then reconstructed the visible images from the mapped SIFT features. The authors extended their work in [77] where they employed a multi-region based approach to jointly optimize the global and local spatial information during the reconstruction. In contrast to the two-step process of Riggan et al. [76, 77], Zhang et al. [78] proposed a generative adversarial network (GAN) based approach to reconstruct a more photo-realistic image using multiple loss functions. Although, in the literature the feature-based cross-spectrum face recognition system has shown a better performance compared to the synthesis-based methods. But, with the emergence of new GAN architectures and deep generative models, it is expected that synthesis based methods will proceed to outperform the feature-based cross-spectrum matching methods.

Recently, deep learning methods have been widely utilized in face recognition and other classification problems [68, 79, 80, 81, 82, 83, 84] instead of classical methods [85, 86]. These methods, can also be employed for the task of sketch-photo recognition problem by learning the relationship between the two

modalities. However, the problem of sketch recognition is more challenging compared to the classical face recognition problem from the deep learning point of view. The reason behind this lies not only in the heterogeneous nature of sketch and photo modalities but also the lack of large databases in order to avoid over-fitting and local minima. For example, most of the datasets contain only one sketch per subject which makes it very challenging for a deep model to learn the robust features [87]. To avoid this, many deep techniques utilize relatively shallow model or train the network only on the photo modality [88].

Heterogeneous Face Recognition with Facial Attributes. Recent approaches on sketch recognition problem have mainly focused on closing the gap between the two domains of sketch and photo and use of soft biometrics has not been investigated adequately. In [89] an approach was proposed to directly use facial attributes in suspect identification without using the sketch. [90] used race and gender to narrow down the galley of mugshots for faster and more accurate matching. Mittal et. al. [1] fused multiple sketches of a suspect to increase the accuracy of their algorithm. They also employed some soft biometric traits such as gender, ethnicity, and skin color to reorder the ranked list of the suspects. Ouyang et al. [91] introduced a framework to combine the facial attributes with low-level features to fill the gap between sketch and photo modalities.

There are some informative traits, such as age, gender, ethnicity, race, and hair color, which are not distinctive enough for the sake of recognition, but still can act as complementary information to other primary information, such as face and fingerprint. These traits, which are known as soft biometrics, can improve recognition algorithms while they are much cheaper and faster to acquire. They can be directly used in a unimodal system for some applications [38]. Soft biometric traits also have been utilized jointly with hard biometrics (face photo) for different tasks such as person identification or face recognition. However, the soft biometric traits are considered to be available both during the training and testing phases. Our approach looks at this problem in a different way. We consider the case when soft biometric information does not exist during the testing phase, and our method can predict them directly in a *multi-tasking* paradigm.

Learning Using Priviledged Information. Design of a face recognition system, comprises of two major phases, namely training and testing. However, in some cases, there is an extra information which is only available during the training phase and is missing during the testing phase. In other words, the training data is augmented with some extra auxiliary information. For example, in object recognition, the labeled images maybe annotated with texts which provide semantic information about the object, or any other extra knowledge, such as the boundary information which determines the exact location of a specific object [92].

This extra information can be regarded as an auxiliary to the primary modality of the data. Unlike the domain adaptation and transfer learning problems in which the data is similar in both the source and target domains but statistically different [93, 94], here, the available data in the source domain has an extra modality which is not available in the target domain. This makes the task much more similar to the multi-task and multiview problems [95, 96, 97]. However, in comparison with the two aforementioned problems, the absence of auxiliary data in the testing phase makes our problem more challenging.

Facial Landmark Detection There has been a wide range of approaches to solve the problem of landmark detection, starting with methods such as active shape model [98], and active appearance models [99] which are related to PCA-based shape constraint.

Many of these approaches utilize a cascade strategy to integrate prediction modules and update the landmark locations in a progressive manner [100, 101]. Cascade regression networks which are designed for landmark localization [43], or human body pose estimation [102] have made improvements by tackling the problem level at coarse to fine levels. However, requiring careful design and initialization for such frameworks and the absence of learned geometric relationships are the main challenges of these architectures.

Recently, with the onset of convolutional neural networks (ConvNets) in feature representation [103], a common approach in facial landmark detection is to extract features from the facial appearance using ConvNets, and afterward learn a model typically a regressor to map the features to the landmark locations [100, 104, 105, 106]. Despite the excellent performance of the ConvNets in different applications, it has been shown [107, 108] that they can be very sensitive and vulnerable to a small perturbation in the input domain which can lead to a drastic change of the output domain, *e.g.*, predicted landmarks.

Generative Adversarial Networks. In contrast to VAE models which implicitly compute the likelihood of the data space, autoregressive models have the advantage of tractable likelihood and can generate diverse samples. The basic idea of these models is to use the autoregressive connections to model an image pixel by pixel. In fact, autoregressive approaches model the joint distribution of pixels in the image as the product of conditional distributions [109]. PixelCNN++ [110] is the most recent autoregressive method that provides a tractable likelihood for the data distribution and generates images with diverse samples. However, these models suffer from a slow synthesis when compared to GANs.

The lack of explicit density function in GANs is problematic for two main reasons. Many applications in deep generative models are based on the density estimation. For instance, the count-based exploration methods [111] rely on density estimation have achieved state-of-the-art performance on reinforcement learning

environments [112]. The second reason is that the quantitative evaluation of the generalization performance of such models is challenging. Since GANs typically are able to generate sharp samples by memorizing the training data, the evaluation criteria based on ad-hoc sample quality metrics [113] does not capture the mode collapse issue.

Recently some approaches have been trying to solve the mode collapse issue by improving the GAN training process. In [113], the authors utilize the mini-batch discrimination trick to allow the discriminator to detect samples that are unusually similar to the other generated samples. This heuristic helps to generate visually more appealing samples at the cost of more computational time. Therefore, this method is usually used in the last hidden layer of the discriminator. Another method is to unroll the optimization of discriminator to make a surrogate objective function in order to help optimizing the generator [114]. Although their model is robust to mode collapse but it is not clear whether this happens at the cost of losing image quality or not. In another approach the author used many generators to discover all the modes of the data [115]. There are some other approaches that attempt to use autoencoders as regularizers or additional losses to penalize the missing modes [116, 117]. In [118] authors used an LSTM-based autoregressive model in their discriminator function and considered the reconstruction loss as the penalty for fake data. However, in their GAN model they trained their discriminator only on the true data as it becomes unbounded for the fake data synthesized by the generator.

Chapter 2

Heterogeneous Face Recognition

2.1 Introduction

Motivated by recent advances in face recognition algorithms using deep approaches and generative models, in this chapter we propose two different methods for cross-spectrum face recognition, namely, a novel Coupled Generative Adversarial Network (CpGAN) , which utilizes non-visible modalities to perform a cross-spectrum face recognition task and a Coupled Deep Convolutional Neural Network (CpDCNN) . In [119], authors used a coupled CNN-based architecture for their face recognition system. However, they evaluated their framework only for near infrared which is very close to visible. Here, we evaluate the proposed algorithm on different bands of electromagnetic spectrum from NIR to the more challenging modalities such as midwave and longwave infrared. We compare our proposed framework against several different state-of-the-art techniques in the literature such as DPM [75], coupled neural network (CpNN) [74], PLS [24], PLS∘DPM and PLS∘CpNN [14, 70]. We present a thorough evaluation using multiple datasets: Wright State (WSRI), Notre Dame X1 (UND X1), Night Vision (NVESD), Polarimetric thermal, and Casia

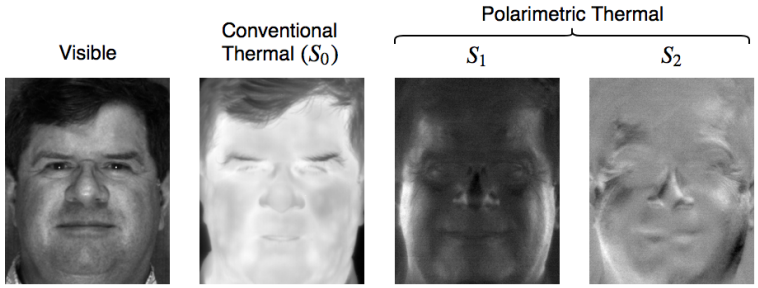


Figure 2.1: Visible spectrum and its corresponding conventional thermal (S_0), and polarimetric state information (S_1 and S_2) of a thermal image of a subject.

NIR-VIS 2.0 datasets. Our results show that our proposed CpGAN could outperform the existing methods for heterogeneous face recognition.

2.2 Polarimetric Thermal Imagery

In comparison to the conventional thermal imaging that captures intensity-only in the midwave infrared (MWIR) or longwave infrared (LWIR) bands, polarimetric thermal considers the polarization state information in the thermal infrared spectrum. Polarization states are characterized using the Stokes parameters S_0, S_1, S_2, and S_3 that are captured from a face, see Fig. 7.1. The polarimetric measurement is done using a series of linear and circular polarizers. The four mentioned Stokes parameters which completely define the polarization states are:

$$S_0 = I^{\circ}_0 + I^{\circ}_{90} \ , \tag{2.1}$$

$$S_1 = I^{\circ}_0 - I^{\circ}_{90} \ , \tag{2.2}$$

$$S_2 = I^{\circ}_{45} + I^{\circ}_{-45} \ , \tag{2.3}$$

$$S_3 = I^{\circ}_R + I^{\circ}_L \ , \tag{2.4}$$

where I°_0, I°_{90}, I°_{45}, and I°_{-45} describe the measured intensity of the light after passing through a linear polarizer with angle of 0°, 90°, 45°, and -45° related to horizontal axes, respectively. I_R and I_L are the intensity of the light after passing through right and left circularly polarization filters. Since there is no artificial illumination in passive imaging, there is almost no circularly polarized information in LWIR or MWIR spectrum. Therefore, S_3 is considered to be zero for most of the applications. To quantify the portion of electromagnetic radiation that is linearly polarized, the Degree of Linear Polarization (DoLP), is computed with the linear combination of the Stokes as follows:

$$DoLP = \frac{\sqrt{S_1^2 + S_2^2}}{S_0} \ . \tag{2.5}$$

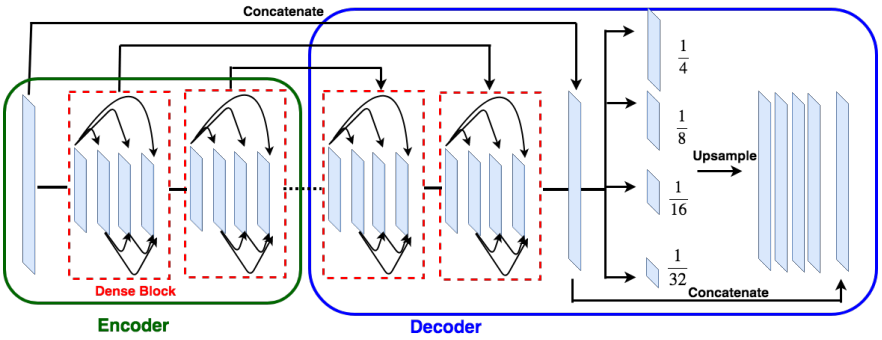


Figure 2.2: An overview of the pyramid densely connected network.

2.3 DenseNets

Traditional convolutional feed-forward networks such as VGG [120], connect the output of the l^{th} layer as the input to the next layer, which is equal to the following transition: $x_l = H_l(x_{l-1})$, where H_l is the convolutional mapping from $l-1$ to l. In Resnet [121], authors made a change in this transition information by adding a skip-connection which bypasses the non-linear transformation with an identity function:

$$x_l = H_l(x_{l-1}) + x_{l-1} \ . \tag{2.6}$$

A benefit of Resnet architecture is that through the identity function, the gradient of the cost function can progress directly from later layers to the earlier layers. However, the combination of the identity function and output of H_l might prevent the information flow in the network [122].

In order to improve the information flow between different layers, in Densenet [122] authors provided a different connectivity between different layers in which there is a direct connection between any layer and all the subsequent layers. Therefore the l^{th} layer receives the feature maps of all the previous layers, $x_0, x_1, ..., x_{l-1}$ as input:

$$x_l = H_l([x_0, x_1, ..., x_{l-1}]) \ , \tag{2.7}$$

where $[x_0, x_1, ..., x_{l-1}]$ represents the concatenation of the feature maps produced from the previous layers $0, ..., l-1$ [122] (see dense block in Fig. 2.2).

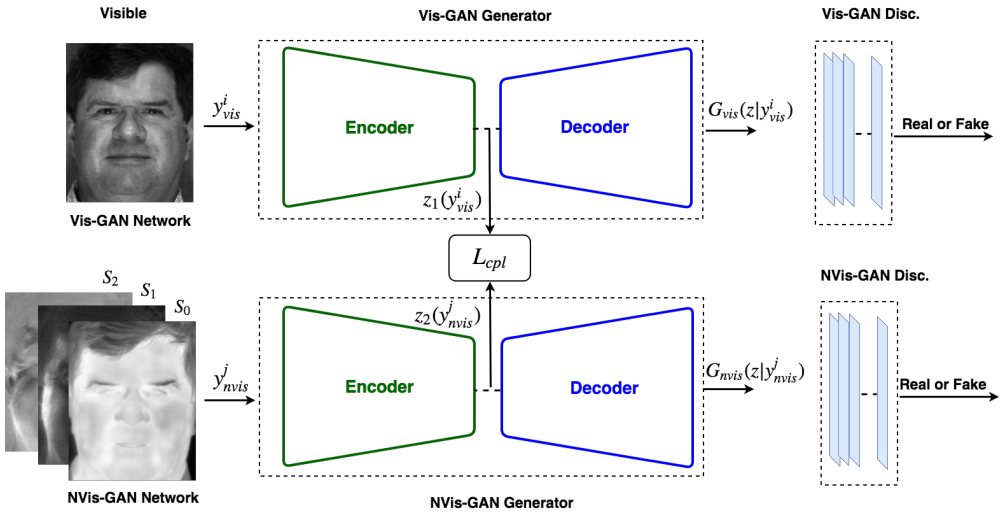


Figure 2.3: Proposed network using two GAN based sub-networks (Vis-GAN and NVis-GAN) coupled by contrastive loss function. Here, the input to NVis-GAN is polarimetric data (S_0, S_1, S_2). In the case of other non-visible modalities such as (NIR, MWIR, and LWIR) the framework remains the same and only the input to the NVIS-GAN is changed accordingly.

2.4 Generative adversarial networks

The generative adversarial network consists of two sub-networks, namely a generator and a discriminator which compete with each other in a minimax game. For the generator to learn the distribution p_g over the data x, the authors consider a prior on the input noise variables $p_z(z)$ [123]. Generator network G is a differentiable function with a parameter θ_g which performs a mapping to the data space $G(z;\theta_g)$. On the other hand, the discriminator network is also a differentiable function $D(.;\theta_d)$ which performs a binary classification between the real data x and the generated data $G(z)$. At the same time, network G tries to fool the discriminator by minimizing $log(1 - D(G(z)))$. In other words, D and G play a two-player minimax game which resembles minimizing the Jenson-Shannon divergence [123] as follows:

$$\min_G, \max_D E_{x\sim P_{data(x)}}[\log D(x)] + E_{z\sim P_z}[\log(1 - D(G(z)))].$$

2.5 Conditional generative adversarial networks

Conditional adversarial networks is an extension of generative adversarial networks in which both the generator and discriminator are conditioned on some auxiliary information y. The extra information y can be any kind of information such as class label or other modalities data. The objective of the conditional GAN

is the same as the classical GAN. The only exception is that in the conditional GAN both the discriminator and generator are conditioned on the auxiliary information as follows:

$$\min_G \max_D E_{x\sim P_{data(x)}}[\log D(x|y)] + E_{z\sim P_z}[\log(1 - D(G(z|y)))],$$

2.6 CpGAN method

The proposed CpGAN is illustrated in Fig. 2.3. The proposed approach consists of two generators and two discriminators which are coupled with each other [124]. In the following, we explain these modules in detail.

2.6.1 Pyramid Densely Connected Network.

This network is a densely connected encoder-decoder structure which utilizes the features from multi layers of a CNN [125]. In this framework a dense block [122] is used as the basic structure since it can maximize the information flow and has better convergence with connecting all the layers. The encoder part of the network consists of three dense blocks with their corresponding down-sampling operations which shrinks the feature map to 1/32 of the input size. The decoder part is responsible for reconstructing the original size image from the embedding subspace and it stacks five dense blocks with the refined up-sampling transition blocks [126, 127]. Moreover, the concatenations are performed on the feature maps with the same size. Inspired by the use of global context information in classification and segmentation, this network tries to capture more global information, using multi-level pyramid pooling blocks [128, 129]. This operation is done to make sure that features from different scales are embedded in the final result. Therefore, four different operations with pooling sizes of 1/32, 1/16, 1/8, and 1/4 is selected. All the four level features are up-sampled to the original size and are concatenated together. Fig. 2.2 illustrates the overview of the pyramid densely connected network.

2.7 Deep cross-modal face recognition

The final objective of the proposed model is identification of non-visible faces which we do not have access to them during the training phase. For this reason, we couple two pyramid densely connected networks

one dedicated to the visible spectrum (Vis-GAN) and the other one to the non-visible spectrum (NVis-GAN). Each network performs a non-linear transformation of the input space. The final objective of our proposed CpGAN is to find the global deep features representing the relationship between non-visible face images and their corresponding visible ones. In order to find a common latent embedding subspace between these two different domains we couple two pyramid densely connected networks (Vis-GAN and NVis-GAN) via a contrastive loss function [62]. The contrastive loss function (ℓ_{cont}) pulls the genuine pairs (i.e., a visible face image with its own corresponding non-visible face image) towards each other into a common latent feature subspace and push the impostor pairs (i.e., a visible face image of a subject with another subject's non-visible face image) apart from each other (see VisGAN and NVis-GAN networks at their bottlenecks in Fig. 2.3). Similar to [62], our contrastive loss is of the form:

$$\begin{aligned}\ell_{cont}(z_1(y^i_{vis}), z_2(y^j_{nvis}), y_{cont}) = \\ (1-y_{cont})L_{gen}(D(z_1(y^i_{vis}), z_2(y^j_{nvis})) + y_{cont}L_{imp}(D(z_1(y^i_{vis}), z_2(y^j_{nvis}))\ ,\end{aligned} \tag{2.8}$$

where y^i_{vis} is the input for the Vis-GAN (i.e., visible face image), and y^j_{nvis} is the input for the NVis-GAN (i.e., non-visible face images). y_{cont} is a binary label, L_{gen} and L_{imp} represent the partial loss functions for the genuine and impostor pairs, respectively, and $D(z_1(y^i_{vis}), z_2(y^j_{nvis}))$ indicates the Euclidean distance between the embedded data in the embedded common feature subspace. $z_1(.)$ and $z_2(.)$ are the deep convolutional neural network based embedding functions, which transform y^i_{vis} and y^j_{nvis} into a common latent embedding subspace, respectively. The binary label, y_{cont}, is assigned a value of 0 when both modalities, i.e., visible and non-visible, form a genuine pair, or, equivalently, the inputs are from the same class ($cl^i = cl^j$). On the contrary, when the inputs are from different classes, which means they form an impostor pair, y_{cont} is equal to 1. In addition, L_{gen} and L_{imp} are defined as follows:

$$\begin{aligned}L_{gen}(D(z_1(y^i_{vis}), z_2(y^j_{nvis}))) = &\frac{1}{2}||z_1(y^i_{vis}), z_2(y^j_{nvis})||^2_2 \\ &\text{for} \qquad cl^i = cl^j\ ,\end{aligned} \tag{2.9}$$

and

$$\begin{aligned}&L_{imp}(D(z_1(y^i_{vis}), z_2(y^j_{nvis}))) = \\ &\frac{1}{2}\,max(0, m - ||z_1(y^i_{vis}), z_2(y^j_{nvis})||^2_2) \quad \text{for} \quad cl^i \neq cl^j\ .\end{aligned} \tag{2.10}$$

where m is the contrastive margin. Therefore, the coupling loss function can be written as:

$$L_{cpl} = 1/N^2 \sum_{i=1}^{N} \sum_{j=1}^{N} \ell_{cont}(z_1(y_{vis}^i), z_2(y_{nvis}^j), y_{cont}) \ , \tag{2.11}$$

where N is the number of samples. It should be noted that the contrastive loss function (2.11) considers the subjects' labels implicitly. Therefore, it has the ability to find a discriminative embedding space by employing the data labels in contrast to some other metrics such as the Euclidean distance. This discriminative embedding space would be useful in identifying a non-visible probe photo against a gallery of visible photos.

2.7.1 Generative Adversarial Loss

Let G_{vis} and G_{nvis} denote the generators that synthesize a visible image from an input visible and a non-visible image, respectively. To synthesize the output and to make sure that the synthesized images generated by the two generators are indistinguishable from the corresponding ground truth visible image, we utilized the GAN loss function in [123]. As it is shown in Fig. 2.3, the first generator G_{vis} is responsible to generate a visible image when the network is conditioned on a visible image. On the other hand, the second generator G_{nvis} tries to generate the same visible image from the non-visible image which has a more challenging task compared to the first generator. Therefore, the total loss for the coupled GAN is as follows:

$$L_{GAN} = L_{vis} + L_{nvis}, \tag{2.12}$$

where the GAN loss function related to the Vis-GAN is given as:

$$L_{vis} = \underset{G_{vis}}{min}\ \underset{D_{vis}}{max} E_{x^i \sim P_{vis(x)}}[\log D(x^i|y_{vis}^i)] + E_{z \sim P_z}[\log(1 - D(G(z|y_{vis}^i)))],$$

where y_{vis}^i is the visible image used as condition for the Vis-GAN and x^i is the real data. It should be noted that for the Vis-GAN the real data x^i and the condition y_{vis}^i are the same. Similarly the loss for the NVis-GAN is given as:

$$L_{nvis} = \underset{G_{nvis}}{min}\ \underset{D_{nvis}}{max} E_{x^j \sim P_{vis(x)}}[\log D(x^j|y^j_{nvis})] + E_{z\sim P_z}[\log(1 - D(G(z|y^j_{nvis})))],$$

where y^j_{nvis} is the non-visible image used as condition for the NVis-GAN and x^j is the real data. It should be noted that x^i is the same as x^j if they refer to the same subject ($cl^i = cl^j$), otherwise they are not the same.

2.7.2 Overall Loss Function

The proposed approach contains the following loss function: the Euclidean $L_{E_{vis}}$ and $L_{E_{nvis}}$ losses which are enforced on the recovered visible images from the Vis-GAN and NVis-GAN networks, respectively, are defined as follows:

$$L_{E_{vis}} = ||G_{vis}(z|y^i_{vis}) - x^i||^2_2, \tag{2.13}$$

$$L_{E_{nvis}} = ||G_{nvis}(z|y^j_{nvis}) - x^j||^2_2, \tag{2.14}$$

$$L_E = L_{E_{vis}} + L_{E_{nvis}}. \tag{2.15}$$

The L_{GAN} (2.12) loss is also added to generate sharper images. In addition, based on the success of perceptual loss in low-level vision tasks [130, 131], the perceptual loss is added to the NVis-GAN to preserve more photo realistic details as follows:

$$L_{P_{nvis}} = \frac{1}{C_p W_p H_p} \sum_{c=1}^{C_p} \sum_{w=1}^{W_p} \sum_{h=1}^{H_p} ||V(G_{nvis}(z|y^j_{nvis}))^{c,w,h} - V(x^j)^{c,w,h}||,$$

where x^j is the ground truth visible image, $G_{nvis}(z|y^j_{nvis})$ is the output of NVis-GAN generator. V(.) represents a non-linear CNN transformation and C_p, W_p, H_p are the dimension of a particular layer in V. It should be noted that the perceptual loss is just used in the NVis-GAN.

Finally, the contrastive loss function (2.11) is added to train both networks Vis-GAN and NVis-GAN jointly to make the embedding space of the mentioned networks as close as possible and to preserve a more discriminative and distinguishable shared space. Therefore, the total loss function for the proposed CpGAN

Table 2.1: Summary of heterogeneous face recognition datasets used for comparing models.

Database	Source	Target	# subjects	Variations
WSRI	visible	MWIR	64	E
UND X1	visible	LWIR	241	E
NVESD	visible	MWIR & LWIR	50	E,D
Casia NIR-VIS 2.0	visible	NIR	725	P,E,G,D
Casia HFB	visible	NIR	202	P,E,G,D
Polarimetric thermal	visible	S_0, S_1, S_2	60	E,D

is as follows:

$$L_T = L_{cpl} + \lambda_1 L_E + \lambda_2 L_{GAN} + \lambda_3 L_{Pnvis}, \tag{2.16}$$

where L_{cpl} is the coupling loss (2.11) term which is the contrastive loss function, the second is the total $L2$ loss for the Vis-GAN and NVis-GAN. L_{GAN} and $L_{P_{nvis}}$ are the GAN, and perceptual loss functions for the Vis-GAN, respectively. λ_1, λ_2, and λ_3 are the hyper-parameters which weight the Euclidean, the adversarial, and the perceptual losses, respectively.

2.7.3 Testing Phase

During the testing phase, only the NVis-GAN is used. For a given test probe y^t_{nvis}, NVis-GAN is employed in the proposed CpGAN to synthesize the visible image $G_{nvis}(z|y^t_{nvis}) = \hat{x}^t_{vis}$. Eventually, the identification of face recognition is done, by calculating the minimum Euclidean distance between the synthesized image from the not-visible prob and visible gallery images as follows:

$$x^{t^*}_{vis} = \underset{x^t_{vis}}{argmin} \quad ||x^t_{vis}, \hat{x}^t_{vis}|| \,, \tag{2.17}$$

where $\hat{x}^t_{vis}$ is the synthesized probe face image and $x^{t^*}_{vis}$is the selected matching visible face image within the gallery of face images.

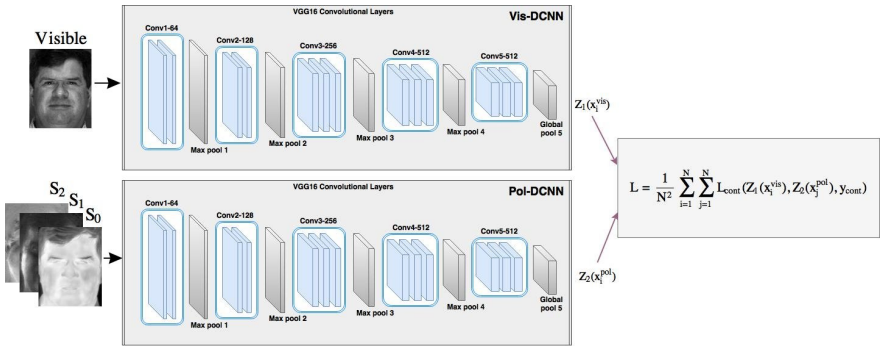


Figure 2.4: Proposed network using two convolutional networks (Vis-DCNN and Pol-DCNN) coupled by contrastive loss function.

2.8 CpDCNN

In this method, we used a VGG-16 like network [132] in our cross-spectrum recognition framework. The VGG-16 neural network comprised of five major convolutional components which are connected in series. The first two components, $Conv1-64$ and $Conv2-128$ consists of the following layers: a convolutional layer, a rectified linear unit layer, a second convolutional layer, a second rectified linear unit layer, and a max pooling layer. The remaining three components contain one additional convolutional layer and a rectifier linear unit layer. The only exception is in the last component, where global pooling was used instead of the max pooling to reduce the number of parameters.

The final objective of the proposed model is identification of the polarimetric thermal images of the probe faces while we do not have access to them during the training phase. For this reason, we coupled two VGG-16 like networks one dedicated to the visible spectrum (Vis-DCNN) and the other one to the polarimetric thermal (Pol-DCNN). Each DCNN performs a non-linear transformation of the input space. The ultimate goal of our proposed CpDCNN is to find the global deep features representing the relationship between polarimetric thermal face images and their corresponding visible ones. In other to find the common embedding space between these two different domains we coupled two VGG-16 structured networks (Vis-DCNN and Pol-DCNN) via a contrastive loss function [62] (see Fig. 2.4).

2.9 Experiments and results

2.9.1 Implementation details

The network is trained on a Nvidia Titan X GPU using the PyTorch framework. We choose $\lambda_4 = 0.5$ and $\lambda_{1,2,3} = 1$. For training, we used the Adam optimizer [133] with a first-order momentum of 0.5 and a learning rate of 0.0002 and a batch size of 4. The perceptual loss is assessed on relu3-1 layer of a pre-trained VGG [120] model for the Imagenet dataset [134].

2.9.2 Heterogeneous face recognition datasets

In order to evaluate the proposed CpGAN model, we utilize six different heterogeneous face recognition databases:

1) Wright State (WSRI) [135],

2) Notre Dame X1 (UND X1) [136],

3) Night Vision (NVESD) [137],

4) Casia NIR-VIS 2.0 [138],

5) Casia HFB [139],

6) Polarimetric thermal [14],

in order to test the NIR-to-visible, MWIR-to-visible, LWIR-to-visible and polarimetric thermal-to-visible face recognition applications. Table 2.1 provides an overview of the datasets. Each of the mentioned databases are described below:

WSRI dataset consists of 1,615 visible and 1,615 MWIR images from 64 different identities. There are 25 images per subject approximately with different facial expressions. For the visible modality, the original resolution of the images are 1004×1004, and for the MWIR modality and 512×640. After preprocessing, the images from both modalities are resized to 235×295 pixels. This database is split randomly into a set of 10 subjects for training set and remaining 54 subjects for testing set.

UND X1 dataset contains LWIR and visible images related to 241 subjects with different variations in lighting, expression and time lapse. The original resolutions of the images are 1600×1200 pixels for the visible modality and 320×240 pixels for the LWIR modality. Both modalities are resampled to 150×110 pixels after preprocessing.

The training set composed of 159 subjects captured in the visible and LWIR modalities with only one

image per subject. One the other hand, the test set contains the remaining 82 subjects with multiple images per subject. This database is challenging due to the low resolution and noise present in the LWIR imagery. This leads to significant difference between the two modalities in this dataset.

NVESD dataset is collected by the U.S Army CERDEC-NVESD in 2012 from 50 different subjects. The dataset composed of 450 images for each modality of visible, MWIR, and LWIR. The images were captured simultaneously from different identities with the original resolution of 640×480 pixels for all of the modalities. After preprocessing as in [24], the image resolution are resampled to 174×174 and dataset is split into training and testing sets.

CASIA NIR-VIS 2.0 dataset contains the visible and NIR images from 725 different identities. The images were not captured simultaneously. For each subject there are one to 22 visible images and five to 50 NIR images with different expressions, poses, glasses, and distance to camera/sensor. The original resolution of the images for both modalities are 640×480 pixels. After preprocessing, the cropped image sizes are 128×128. This database provides a part of data for the sake of parameter tuning, and 10 remaining parts for reporting the experimental results.

CASIA HFB dataset contains 202 subjects. Similar to the CASIA NIR-VIS 2.0 this dataset has two views where the first view is for parameter selection and View2 is for the sake of evaluation. This dataset contains about 1,000 visible images and 1,500 NIR images for training and similarly 1,000 visible and 1,500 NIR images for testing phase. The resolution of the images before and after preprocessing is the same as the NIR-VIS 2.0 dataset.

Polarimethric Thermal Face dataset [14] comprises polrimetric LWIR face images and their corresponding visible spectrum related to 60 subjects. Data was collected at three different distances: Range 1 (2.5 m), Range 2 (5 m), and Range 3 (7.5 m). At each range two different conditions, including baseline and expression are considered. In the baseline condition the subject is asked to keep a neutral expression looking at the polarimetric thermal sensor. On the other hand, in the expression condition the subject is asked to count out numerically from one upwards which results in different expressions in the mouth and to the eyes and consequently different variations in the facial imagery. Each subject has 16 images of visible and 16 polarimetric LWIR images in which four images are related to the baseline condition and the remaining 12 images are related to the expression condition.

2.9.3 WSRI and UND results

The network for the visible face images (Vis-GAN) and the network for the non-visible face images (NVis-GAN) have the same structure. These images are resized to 256×256 before passing to the network. To benefit from the pre-defined weights of the DenseNet [122], the first convolutional layer and the first three Dense-blocks have been taken up from a pre-trained DenseNet 121 as the encoder structure. At the end of the encoder part where the feature size is $1/32$ of the original input size, the two sub-networks (Vis-GAN and NVis-GAN) are coupled together via a contrastive loss function (see Fig. 2.3) to construct the CpGAN framework.

To increase the correlation between the two modalities of visible and thermal, each modality was pre-processed. We applied a band-pass filter so called difference of Gaussians (DoG), to emphasize the edges in addition to removing high and low frequency noise. The DoG filter which is the difference of two Gaussian kernels with different σ is defined as follows:

$$D(I, \sigma_0, \sigma_1) = [G(x, y, \sigma_0) - G(x, y, \sigma_1)] * I(x, y) \ , \tag{2.18}$$

where D is the DoG filtered image, $*$ is the convolution operator, and G is the Gaussian kernel which is defined in:

$$G(x, y, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2 + y^2}{2\sigma^2}} \ . \tag{2.19}$$

The training set is used to transform the visible and non-visible features to a shared latent embedding subspace. Also at the same time, the network tries to synthesize visible modality from the shared latent subspace in the GAN framework. To train the network, the genuine and impostor pairs are constructed. The genuine pair is constructed from the same subject images in two different modalities. For the impostor pair, a different subject is selected for each modality. In general, the number of the generated impostor pairs are significantly larger than the genuine pairs. For the sake of balancing the training set, we consider the same number of genuine and impostor pair. After training the network, during the testing phase, only the NVis-GAN sub-network is used for the evaluation. For a given probe, the network is used to synthesize the visible image. Afterwards, the Euclidean distance is used to match the synthesize image to its closest image

Table 2.2: Results of the proposed method and the baseline methods for WSRI and UND X1 datasets

Method	WSRI	UND X1
PLS	83.7%	41.0%
BCDL	93.1%	50.5%
K-BCDL	95.9%	52.0%
CpNN	97.2%	51.9%
CpGAN	**97.8%**	**76.4%**

from the gallery. The ratio of the number of correctly classified subjects and the entire number of subjects is computed as the identification rate.

The identification rate of our proposed approach for both WSRI and UND X1 datasets is reported in Table 2.2. In addition, we compare the performance of our method with some state-of-the-art methods in the literature such as CpNN [74], PLS [21], bilevel coupled dictionary learning (BCDL) [140], and kernel bilevel coupled dictionary learning (K-BCDL) [74]. The tabulated results show the improved performance of the proposed method and its effectiveness in synthesizing the visible modality from the non-visible modality.

2.9.4 NVESD results

We compare our proposed CpGAN with the reported results in the literature on the NVESD dataset. For the sake of comparison we perform the same split as in [74] on the dataset for the train and test set. Therefore, we train our proposed framework on training set with 10 subjects and report the rank-1 classification performance on the test set of 40 subjects. This database contain two different non-visible modalities, namely, MWIR and LWIR. Table 2.3 shows the reported results of our proposed method and as well as the other state-of-the-art models. As it is shown in Table 2.3, our proposed method performance surpasses the other methods in the literature for both MWIR-to-visible and LWIR-to-visble face recognition.

2.9.5 CASIA results

In this experiment, we compare our results with the results reported in [141]. For the sake of fair comparison, we perform the same set of experiments as in [141]. The dataset has two views in which View1 is used for the parameter tuning and View2 with 10 different setup is used for testing. Number of images in HFB dataset is about 1,000 visible images and 1,500 NIR images during the testing phase. The CASIA NIR-VIS 2.0 restricts algorithms to one gallery per subject during the testing phase. Therefore, there are only 358 gallery images for the comparison, while there are about 6,000 probe NIR images for testing.

Table 2.3: Results of the proposed method and the baseline methods for MWIR and LWIR on NVESD dataset

Method	MWIR	LWIR
PLS	82.4%	70.4%
BCDL	90.7%	90.6%
K-BCDL	93.3%	92.5%
CpNN	94.4%	89.1%
CpGAN	**96.1%**	**93.9%**

Table 2.4: Performance comparison to other baselines on View2 of CASIA NIR-VIS 2.0 dataset.

NIR-VIS 2.0	Rank 1	Std. Dev.	FAR=.001
CpNN	33.1%	6.6	76.35
C-CBFD [142]	81.8%	2.3	47.3
[143]	85.9%	0.9	78.0
[144]	86.2%	0.98	81.3
[145]	95.74%	0.52	91.03
[141]	92.6%	0.64	81.6
CpGAN	**96.63%**	0.56	87.05

In addition to the higher number of images in NIR-VIS 2.0, some of the images in this dataset has more challenging images with difficult poses, while the HFB images were taken in a more controlled environment. Moreover, the restriction of one image per gallery subject, makes the NIR-VIS 2.0 dataset more challenging. Table 2.4 and 2.5 shows the results of the proposed method compared to the other methods in the literature for the NIR-VIS 2.0 and HFB datasets, respectively. Following [141], the reported result is the average of 10 different experimental setups. The results show that our method performs very well compared to the other methods on the NIR-VIR 2.0 dataset which is more challenging. Moreover, since the other methods showed a good performance on the HFB dataset, even a little improvement is significant.

Table 2.5: Performance comparison to other baselines on View2 of CASIA HFB 2.0 dataset.

HFB	Rank 1	FAR=.01	FAR=.001
CpNN	39.8%	84.4	72.49
IDNet [119]	80.9%	70.4	36.2
P-RS [23]	87.8%	98.2	95.8
C-DFD [146]	92.2%	85.6	65.5
THFM [147]	99.28%	99.66	98.42
[144]	99.38%	-	92.25
[141]	99.52%	98.6	91.8
CpGAN	**99.64%**	98.4	89.7

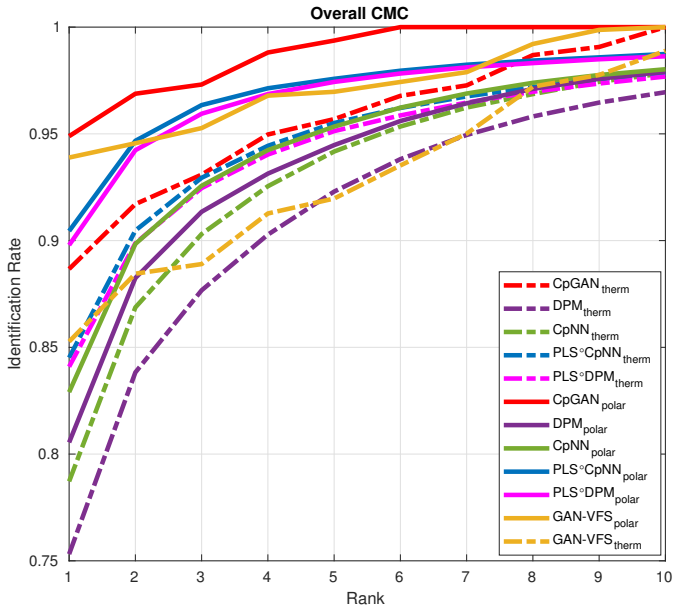


Figure 2.5: Overall CMC curves from testing PLS, DPM, CpNN, PLS∘DPM, PLS∘CpNN, GAN-VFS, and CpGAN using polarimetric and thermal probe samples, matching against a visible spectrum gallery.

2.9.6 Polarimetric thermal results

For the polarimetric thermal face dataset, we consider the same CpGAN architecture. We pass S_0, S_1, and S_2 to the NVis-GAN's three channels as the input as shown in Fig. 2.3.

In each experiment the dataset is partitioned randomly into the training and testing sets. The same set of training and testing data is used to evaluate PLS, DPM, CpNN, PLS∘DPM, PLS∘CpNN, GAN-VFS [78], and the proposed CpGAN network. Fig. 2.5 shows the overall cumulative matching characteristics (CMC) curves for our proposed method and the other state-of-the-art methods over all the three different data ranges as well as the expressions data at Range 1. For the sake of comparison, in addition to the polarimetric thermal-to-visible face recognition performance, Fig. 2.5 also shows the results for the conventional thermal-to-visible face recognition for some of the methods, namely PLS, PLS∘DPM, PLS∘CpNN, CpNN, and CpGAN. In the conventional thermal-to-visible face recognition, all the mentioned methods exactly follow the same procedure as before, with only using S_0 modality. Fig. 2.5 illustrates that by exploiting the polarization information of the thermal spectrum it enhances the cross-spectrum face recognition performance compared to the conventional one. Fig. 2.5 also shows the superior performance of our approach compared to the state-of-the-art methods. In addition, our method could achieve prefect accuracy of 1 at

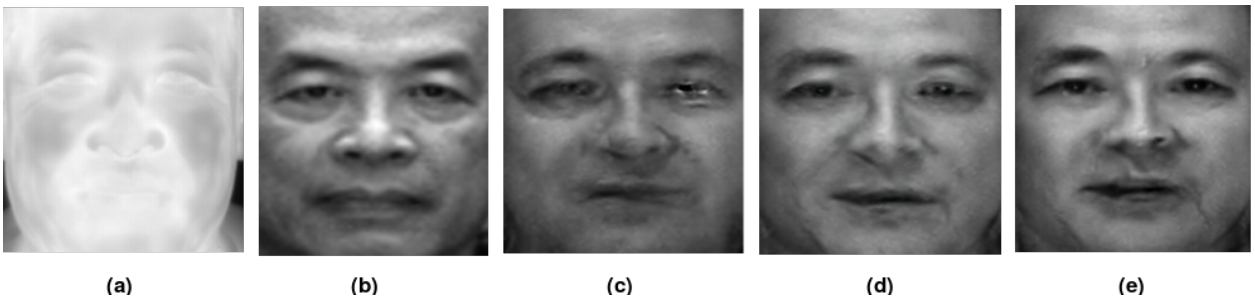


Figure 2.6: Comparison of visible face images synthesized with different experimental configurations. (a) Raw polarimetric image (S_0 is just shown in here). (b) Ground truth visible images. (c) Reconstructed images with $L_{cpl} + L_E$. (d) Reconstructed images with $L_{cpl} + L_E + L_{GAN}$. (e) Reconstructed images with CpGAN (Eq. 2.16).

Scenario	Rank-1 Identification Rate								
	Probe	PLS	DPM	CpNN	PLS∘DPM	PLS∘CpNN	GAN-VFS	CpDCNN	CpGAN
Overall	Polar	0.5867	0.8054	0.8290	0.8979	0.9045	0.9382	0.9408	**0.9549**
	Therm	0.5305	0.7531	0.7872	0.8409	0.8452	0.8561	0.8857	**0.8905**
Expressions	Polar	0.5658	0.8324	0.8597	0.9565	0.9559	0.9473	0.9637	**0.9684**
	Therm	0.6276	0.7887	0.8213	0.8898	0.8907	0.8934	0.9124	**0.9176**
Range 1 Baseline	Polar	0.7410	0.9092	0.9207	0.9646	0.9646	0.9653	0.9721	**0.9867**
	Therm	0.6211	0.8778	0.9102	0.9417	0.9388	0.9412	0.9534	**0.9637**
Range 2 Baseline	Polar	0.5570	0.8229	0.8489	0.9105	0.9187	0.9263	0.9317	**0.9659**
	Therm	0.5197	0.7532	0.7904	0.8578	0.8586	0.8701	0.8868	**0.8993**
Range 3 Baseline	Polar	0.3396	0.6033	0.6253	0.6445	0.6739	0.8491	0.8346	**0.8987**
	Therm	0.3448	0.5219	0.5588	0.5768	0.6014	0.7559	0.7754	**0.7912**

Table 2.6: Rank-1 identification rate for cross-spectrum face recognition using polarimetric thermal and conventional thermal (S_0) probe imagery.

Rank-5 and above.

Table 2.6 tabulates the Rank-1 identification rates for five different scenarios: overall (which corresponds to Fig. 2.5), Range 1 expressions, Range 1 baseline, Range 2 baseline, and Range 3 baseline. In our proposed approach, exploiting polarization information enhance the Rank-1 identification rate by 1.87%, 5.13%, 4.49%, and 5.92% for Range 1 baseline, Range 1 expression, Range 2 baseline, and Range 3 baseline compared to the conventional thermal-to-visible face recognition. This table reveals that using deep coupled generative adversarial network technique with the contrastive loss function to transform different modalities into a distinctive common embedding subspace is superior to the other embedding techniques such as PLS∘CpNN. It also shows the effectiveness of our method in exploiting polarization information to improve the cross-spectrum face recognition problem.

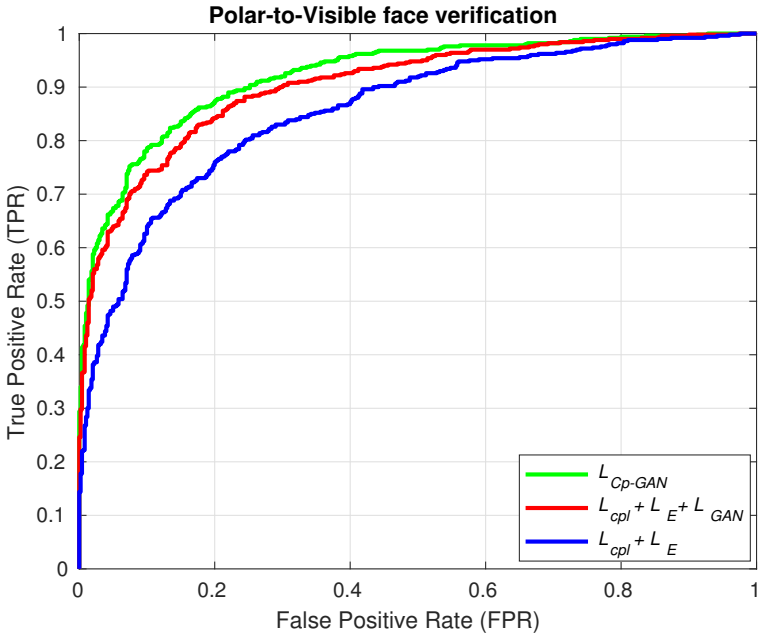


Figure 2.7: The ROC curves corresponding to the ablation study.

2.10 Ablation study

In order to illustrate the effect of adding different loss functions and their improvement in our proposed framework, we perform a study with the following evaluations using the polarimetic dataset: 1) Polar-to-visible using the coupled framework with using only $L_{cpl} + L_E$ loss, 2) Polar-to-visible using the proposed framework with $L_{cpl} + L_E + L_{GAN}$ loss functions, and 3) Polar-to-visible with all the loss functions in the proposed framework (2.16). Fig. 2.6 shows the reconstruction results for a random subject in this dataset. It can be conclude from Fig. 2.6 (c), that using $L_{cpl} + L_E$ loss it results in a blurry image and missing high frequency details. However, adding L_{GAN} loss function (2.12) to the framework leads to a sharper and more vivid images. Moreover, by adding the perceptual loss to the NVis-GAN sub-network, the results become more visually pleasing by removing some artifacts added by L_{GAN}.

For better understanding of different loss functions and their effect on the proposed framework results, we plot the receiver operation characteristic (ROC) curves corresponding to the mentioned three different settings of the framework. As it is shown in Fig. 2.7 the L_{GAN} has an important rule in the enhancement of our proposed approach. Also, adding a perceptual loss enhances the face recognition performance as well as generating visually more realistic images.

2.11 Conclusion

We have proposed two methods in this chapter. While CpDCNN uses the coupled deep network to bring the two modalities close to each other in the embedding domain the other (CpGAN) uses a coupled generative adversarial network to synthesize visible image from a non-visible image for the heterogeneous face recognition task. CpDCNN contains two VGG-based networks while CpGAN contains two GAN based sub-networks dedicated to visible and non-visible input images. CpGAN is capable of transforming the visible and non-visible modalities into a common discriminative embedding subspace and synthesizing the visible photos from that subspace. In order to efficiently synthesizing a realistic visible image from the non-visible modality a densely connected encoder-decoder structure is used as the generator in each sub-network. An ablation study was performed to demonstrate the enhancement obtained by different losses in the CpGAN method. Here, our main focus of accuracy improvement and we did not investigate about the timing of training or inference. The experiments on different HFR datasets with different range of electromagnetic spectrum showed the effectiveness of the CpGAN method compared to the other state-of-the-art methods. The results also revealed that both proposed frameworks could exploit polarimetric thermal information to enhance the thermal-to-visible face recognition performance. However, the proposed CpGAN showed a better performance compared to the CpDCNN.

Chapter 3

Face Recognition Assisted by Facial Attributes

3.1 Introduction

In recent years, there has been significant amount of research in Heterogeneous Face Recognition (HFR) [11]. The main issue in HFR is to match the visible face image to a face image that has been captured in another domain such as in infrared spectrum [11], or polarimetric [14] due to the significant phenomenological difference as well as the lack of training data. Infrared images are categorized into two major groups of reflection and emission. The reflection category, which contains near infrared (NIR) and shortwave infrared (SWIR) bands, is more informative about the facial details and it is very similar to the visible imagery. Due to this reflective phenomenology of the NIR and SWIR, there has been a significant performance on NIR-to-visible face recognition accuracy [12] and to some extent for SWIR-to-visible face recognition accuracy [13].

Multi-task learning (MTL) has been vastly applied in computer vision and biometrics problems. It basically attempts to solve correlated tasks concurrently with the help of knowledge sharing between tasks. [148] employed MTL technique to predict attributes such as age, gender, race, etc. Face photo can be viewed as having some positive or negative hidden relation with some of its soft facial biometric traits.

In this chapter, we propose two different methods to utilize the face attributes. The first approach is an attribute-assisted sketch recognition framework which uses relevant facial attributes, provided by a victim, to enhance the performance of our deep sketch recognition method. Our approach simultaneously learns a common embedding features of sketch and photo image by minimizing two supervisory loss functions,

namely the facial attributes identification and sketch-photo verification loss functions (tasks). Attribute identification task classifies photo and attribute assisted sketch images into a set of facial attributes, while verification task is to classify a pair of sketch-photo as belonging to the same person or not. The attribute identification loss is trying to pull the common features of photo and attribute assisted sketch closer in the shared latent subspace if they belong to the same set of attributes and push them apart if they belong to two different sets of attributes. Therefore, the learned features contain rich variations and can classify the photos and sketches to the classes containing the same sets of attributes in a latent feature subspace.

In the second approach, we propose an Attribute Guided Coupled Generative Adversarial Network (AGC-GAN), which considers the CNN weight sharing followed by the dedicated weights which are responsible for learning the representative features for each specific face attribute. The network also tries to find the common embedding space between the polarimetric and thermal utilizing coupling structure and adversarial training. Optimizing the coupled network by the guidance of the facial attributes leads to a more discriminative embedding space and can be utilized to enhance the performance of the main task which is face recognition.

In summary, the main contributions of this chapter include the following:

- A novel deep learning approach utilizing the facial attributes to improve sketch-photo recognition performance.
- A joint loss function which is based on an identification-verification framework in which the identification part is responsible for the facial attribute classification and the verification part is responsible for creating a common embedding subspace between the sketch and photo modalities. This loss function helps the proposed coupled deep architecture to produce a more discriminative embedding subspace which leads to a better sketch-photo recognition performance.
- Our method is able to fuse textural information of forensic sketches and complementary facial attributes such as skin color and hair color implicitly.
- A novel polarimetric thermal-to-visible face recognition system is proposed in which AGC-GAN is employed for synthesizing visible faces from the polarimetric thermal images using facial attributes.
- A multi-tasking framework is proposed to predict facial attributes from the polarimetric thermal faces. To the best of our knowledge, no such demonstration has been proposed in the literature.

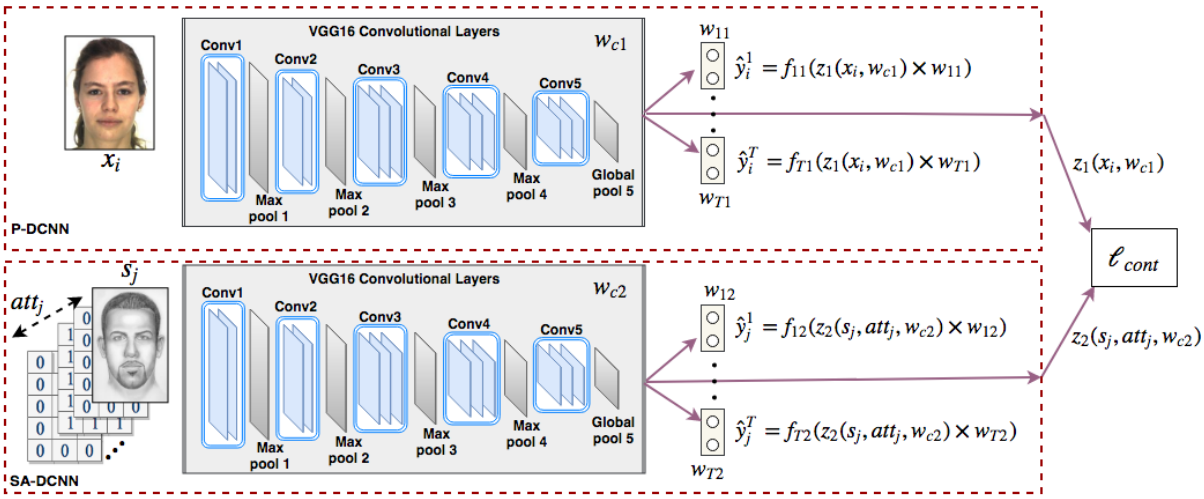


Figure 3.1: Attribute-assisted deep convolutional neural network. P-DCNN (upper network) and SA-DCNN (lower network) embed the photos and a pair of (sketch, attribute) into a common latent subspace.

- Extensive experiments are conducted on ARL polarimetric facial database [14] and the proposed method is compared to recent state-of-the-art methods.

3.2 First approach (Sketch-photo Recognition)

The network parameters in the proposed framework are learned by minimizing two supervisory loss functions namely the losses due to the sketch-photo verification and facial attribute identification tasks [149]. In the following we describe these two supervisory loss functions in details:

3.2.1 Sketch-Photo Verification Task:

Sketch-photo verification is the final objective of the proposed model which is identification of the suspect sketch in a gallery of mugshots. For this reason, we coupled two VGG-16 like networks one dedicated to the photo image domain (P-DCNN) and the other one to the sketch and complementary facial attributes modalities (SA-DCNN). Each DCNN performs a non-linear transformation on the input. The ultimate goal of our proposed attribute-assisted deep convolutional neural network, as shown in Figure 3.1, is to find the global deep features representing the relationship between sketches and their corresponding images. In order to find the common embedding space between these two different modalities we coupled two VGG-16 structured networks (P-DCNN and SA-DCNN) via a contrastive loss function [62]. This function (ℓ_{cont})

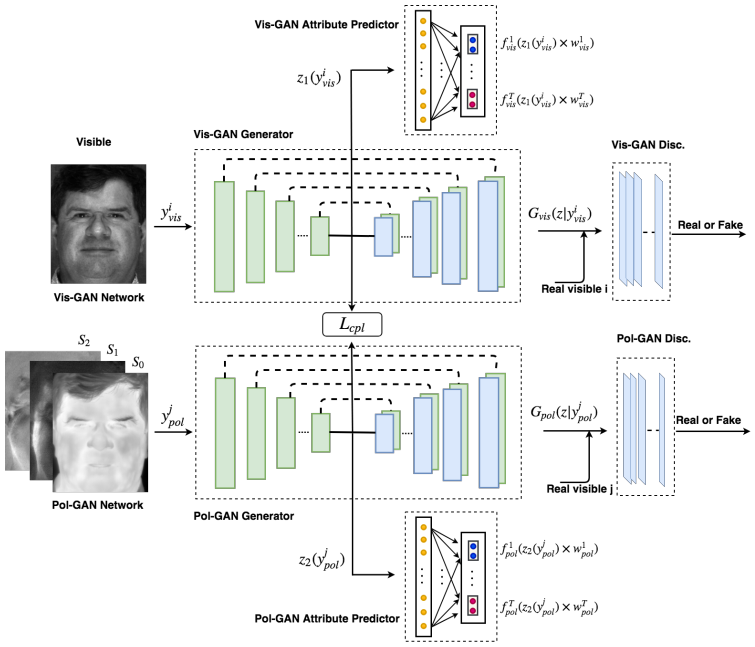


Figure 3.2: Proposed network using two GAN based sub-networks (Vis-GAN and Pol-GAN) coupled by contrastive loss function. The input to Pol-GAN is polarimetric data (S_0, S_1, S_2). The facial attributes are predicted from both sub-networks (Vis-GAN, Pol-GAN) in a multi-tasking paradigm.

pulls the genuine pairs (i.e., a face photo image with its own corresponding sketch image) towards each other into a common latent feature subspace and push the impostor pairs (i.e., a photo image with sketch image from another subject) apart from each other (see Fig. 3.3). Similar to [62], the contrastive loss is of the form:

$$\ell_{cont}(z_1(x_i), z_2(s_j, att_j), y_{cont}) = \tag{3.1}$$
$$(1 - y_{cont})L_{gen}(D(z_1(x_i), z_2(s_j, att_j)) + y_{cont}L_{imp}(D(z_1(x_i), z_2(s_j, att_j)) \ ,$$

where x_i is the input for the P-DCNN (i.e., a photo image), and (s_j, att_j) is the input for the SA-DCNN (i.e., an sketch image with its corresponding attributes provided by the eye witness). y_{cont} is a binary label, L_{gen} and L_{imp} represent the partial loss functions for the genuine and impostor pairs, respectively. z_1 and z_2 are the DNN-based embedding functions, which transform x_i and (s_j, att_j) into a common latent embedding subspace, respectively, and $D(z_1(x_i), z_2(s_j, att_j))$ indicates the Euclidean distance between the embedded data in the common feature subspace. The binary label, y_{cont}, is assigned a value of 0 when both modalities, i.e., photo and sketch, form a genuine pair, or, equivalently, the inputs are from the same subject. On the contrary, when the inputs are from different identities, which means they form an impostor pair, y_{cont} is

equal to 1. In addition, L_{gen} and L_{imp} are defined as follows:

$$L_{gen}(D(z_1(x_i), z_2(s_j, att_j))) = \frac{1}{2} D(z_1(x_i), z_2(s_j, att_j))^2 \quad \text{for} \quad y_i = y_j \ , \tag{3.2}$$

$$L_{imp}(D(z_1(x_i), z_2(s_j, att_j))) = \frac{1}{2}\, max(0, m - D(z_1(x_i), z_2(s_j, att_j)))^2 \quad \text{for} \quad y_i \neq y_j \ . \tag{3.3}$$

Therefore, the total loss function for the training dataset can be written as:

$$L_1 = 1/N^2 \sum_{i=1}^{N} \sum_{j=1}^{N} \ell_{cont}(z_1(x_i), z_2(s_j, att_j), y_{cont}) \ , \tag{3.4}$$

where N is the number of samples. It should be noted that the contrastive loss function [62] considers the subjects' labels inherently. Therefore, it has the ability to find a discriminative embedding space by employing the data labels in contrast to some other metrics such as the Euclidean distance. This discriminative embedding space would be useful in identifying an sketch probe against a gallery of mugshots. However, in our framework we incorporate the facial attribute identification task in addition to the contrastive function to make the embedding space more discriminative. The facial attributes identification task assigns each sketch or image domain to a set of attributes. The attributes are predicted using both the P-DCNN and SA-DCNN networks in a multi-tasking manner. In the following subsection, we describe the multi-tasking problem in the context of attribute prediction.

3.2.2 Multi-Attribute Prediction and Identification Task:

The objective of this model is to predict a set of attributes using a face photo or an sketch. Therefore, in this architecture a face photo (face sketch) is presented to the network as an input and a set of attributes are predicted. Suppose the input is an image $x_i \in X$, and its class label is $y_i \in Y$ for $i = 1, \ldots, N$ where N is the number of the training samples. Soft biometric traits, contain T different facial attributes or binary class labels provided by the eye witness. Therefore, in this framework we denote them as y^t for $t = 1, \ldots, T$. The loss function is defined as:

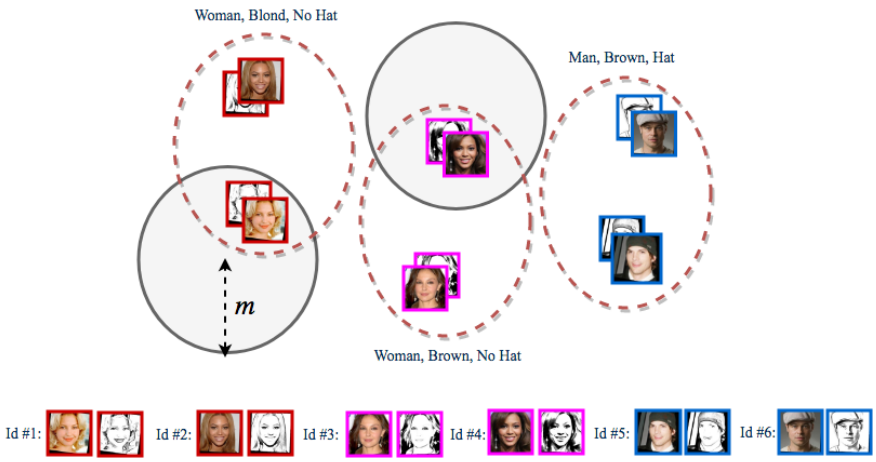


Figure 3.3: Visualization of the common latent subspace by leveraging facial attributes classification and verification loss functions simultaneously. Solid circles represent the contrastive margin in the embedding domain and the dashed circles depict the attributes classification. For the sake of clarity the contrastive margin is depicted for two Ids out of six Ids.

$$L_2 = 1/N \sum_{i=1}^{N} \sum_{t=1}^{T} \ell(f_{t1}(z_1(x_i)), y_i^t) \ , \tag{3.5}$$

where ℓ is a proper loss function (e.g., cross entropy) and $f_{t1}(z_1(x_i))$ is a binary classifier for the attribute t operated on the output of P-DCNN. Learning multiple CNNs separately is not optimal since different tasks may have some hidden relationships with each other and may share some common features. This is supported by [150] where they train a CNN features for the face recognition task and they used it directly for the face attribute estimation. Therefore, our network shares a big portion of its parameters among different tasks in order to enhance the performance of the recognition task. Thus, the loss function (3.5) can be reformulated as follows:

$$L_2 = 1/N \sum_{i=1}^{N} \sum_{t=1}^{T} \ell(f_{t1}(z_1(x_i, w_{c1}) \times w_{t1}), y_i^t) \ , \tag{3.6}$$

where ℓ is the cross entropy loss function. w_{c1} is the shared network parameters between all the tasks and w_{t1} represents the remaining parameters which are assigned separately for each facial attribute task.

The same procedure is performed in the other network (SA-DCNN) with an sketch as input. However, there are some attributes such as hair color and skin color which do not exist in the sketch modality while

they inherently exist in the RGB images. These are the soft biometric traits which is provided by the eye witness description. Therefore, these complementary soft biometric traits are given to the SA-DCNN network which is dedicated to sketch modality. The SA-DCNN network is also responsible to estimate a set of soft biometric attributes. It should be noted that although some of the attributes in the output are given to the network from the beginning, but this attributes are fused with sketch information through the network layers. Therefore, it is worth to estimate them accurately. Also, the set of attributes which are given to the network are not necessarily the same as the set of attributes predicted by the network.

Suppose the input is an sketch $s_j \in S$, and its class label $y_j \in Y$ for $j = 1, \ldots, N$ where N is the number of the training samples. The facial attributes provided by the eye witness are also given to the network as an input, denoted as att for the sake of clarity. The loss function will be defined as:

$$L_3 = 1/N \sum_{j=1}^{N} \sum_{t=1}^{T} \ell(f_{t2}(z_2(s_j, att_j)), y_j^t) \ , \tag{3.7}$$

where ℓ is the cross entropy loss function and $f_{t2}(z_2(s_j, att_j))$ is a binary classifier for the attribute t operated on the output of SA-DCNN. Here, as in P-DCNN network , we share a big portion of the network parameters among different tasks in order to enhance the performance of the recognition task. Therefore, the loss function (3.7) can be reformulated as follows:

$$L_3 = 1/N \sum_{j=1}^{N} \sum_{t=1}^{T} \ell(f_{t2}(z_2(s_j, att_j, w_{c2}) \times w_{t2}), y_j^t) \ , \tag{3.8}$$

where w_{c2} is the shared features between all the tasks. w_{t2} represents the remaining features which are assigned separately for each soft biometric prediction task.

3.2.3 Total Loss Function:

The total loss function L_T for the whole framework can be written as (See Fig. 3.1) :

$$
\begin{aligned}
L_T = L_1 + \lambda_1 L_2 + \lambda_2 L_3 = 1/N^2 \sum_{i=1}^{N}\sum_{j=1}^{N} \ell_{cont}(z_1(x_i), z_2(s_j, att_j), y_{cont}) \\
+ \lambda_1/N \sum_{i=1}^{N}\sum_{t=1}^{T} \ell(f_{t1}(z_1(x_i)), y_i^t) + \lambda_2/N \sum_{j=1}^{N}\sum_{t=1}^{T} \ell(f_{t2}(z_2(s_j, att_j)), y_j^t) \ ,
\end{aligned}
\tag{3.9}
$$

where the first term is the sketch-photo verification and the second and the third terms are the facial attributes classification loss for the P-DCNN network and SA-DCNN network, respectively. λ_1 and λ_2 are the hyper-parameters which weight facial identification cost functions of the P-DCNN network and SA-DCNN network, respectively. As it was mentioned earlier, the contrastive loss function has the ability to find a discriminative embedding space by employing the data labels. However, due to loss functions from the facial attributes classification term for photo (3.5) and for sketch (3.7), minimizing L_T will boost the discrimination in the common embedding domain. In another words, using just the contrastive loss it does not consider whether two subjects share similar facial attributes or not. Using the facial attribute classification, it enables the embedding space to be more discriminative from the attributes point of view.

Consider a subject sketch with $Id\#1$ (see Fig. 3.3). The contrastive loss function causes the corresponding photos from $Id\#1$ to move closer to $Id\#1's$ sketch and other Ids' photos to move farther away. Now, using the contrastive loss function in conjunction with the attribute classification makes $Id\#2$ to move closer to $Id\#1$ since they share the same set of attributes (see Fig. 3.3). In other words, it differentiates between different impostors of $Id\#1$. The same procedure is performed for the other identities during the training process. Figure 3.3 visualizes the overall concept of our joint loss function. As it is depicted, jointly training the model based on verification and facial identification will lead to a more discriminative embedding subspace which considers both the facial attributes and the geometrical relationship between the forensic sketches and photos.

During the testing phase, given a test probe with its facial attributes (s_t, att_t), the proposed AADCNN method transforms it to the common latent embedding domain, $z_2(s_t, att_t)$. In fact, after training our deep coupled network model, it has the ability to transform the photo and sketch images into a common discriminative embedding space. Therefore, the galley of the photo images is transformed to the mentioned embedding space. Eventually, the sketch image is identified, by calculating the minimum Euclidean distance between the transformed sketch prob and gallery of mugshots as follows:

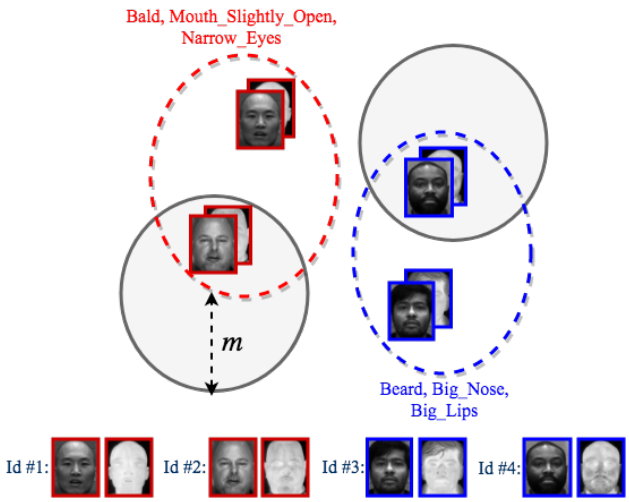


Figure 3.4: Visualization of the common latent subspace by leveraging facial attributes prediction loss function. Solid circles represent the contrastive margin and the dashed circles depict the attributes classification. For the sake of clarity the contrastive margin is depicted for two Ids out of four Ids.

$$x_i^* = \underset{x_i}{argmin} \quad D(z_1(x_i), z_2(s_t, att_t)) \qquad \text{for} \qquad i = 1, 2, .., M \ , \tag{3.10}$$

where (s_t, att_t) is an sketch probe with its facial attribute provided by the eye witness and x_i^* is the selected matching person within the gallery of mugshots of size M.

3.3 Second approach (Polarimetric-visible Recongition)

The proposed AGC-GAN [151] is illustrated in Fig. 3.2. The proposed approach consists of two coupled generators and two discriminators. Each generator is also responsible to predict facial attributes in a multi-tasking paradigm. In the following we explain these modules in detail.

3.3.1 Deep Coupled Framework

The final objective of the proposed model is identification of polarimetric faces which we do not have access to them during the training phase. For this reason, we couple two U-net networks [152] one is dedicated to the visible spectrum (Vis-GAN) and the other network is dedicated to the polarimetric spectrum (Pol-GAN). Each network performs a non-linear transformation of the input space. The final objective of our

proposed AGC-GAN is to find the global deep features representing the relationship between polarimetric face images and their corresponding visible ones. In order to find a common latent embedding subspace between these two different domains we couple two networks (Vis-GAN and Pol-GAN) via a contrastive loss function [62]. This loss function (ℓ_{cont}) pulls the genuine pairs (i.e., a visible face image with its own corresponding polarimetric face image) towards each other into a common latent feature subspace and push the impostor pairs (i.e., a visible face image of a subject with another subject's polarimetric face image) apart from each other (see Fig. 3.2). Similar to [62], our contrastive loss is of the form:

$$\begin{aligned}\ell_{cont}(z_1(y^i_{vis}), z_2(y^j_{pol}), y_{cont}) = \\ (1-y_{cont})L_{gen}(D(z_1(y^i_{vis}), z_2(y^j_{pol})) + y_{cont}L_{imp}(D(z_1(y^i_{vis}), z_2(y^j_{pol}))\ ,\end{aligned} \tag{3.11}$$

where y^i_{vis} is the input for the Vis-GAN (i.e., visible face image), and y^j_{pol} is the input for the pol-GAN (i.e., polarimetric face images). y_{cont} is a binary label, L_{gen} and L_{imp} represent the partial loss functions for the genuine and impostor pairs, respectively, and $D(z_1(y^i_{vis}), z_2(y^j_{pol}))$ indicates the Euclidean distance between the embedded data in the common feature subspace. $z_1(.)$ and $z_2(.)$ are the deep convolutional neural network based embedding functions, which transform y^i_{vis} and y^j_{pol} into a common latent embedding subspace, respectively. The binary label, y_{cont}, is assigned a value of 0 when both modalities, i.e., visible and polarimetric, form a genuine pair, or, equivalently, the inputs are from the same class ($cl^i = cl^j$). On the contrary, when the inputs are from different classes, which means they form an impostor pair, y_{cont} is equal to 1. In addition, L_{gen} and L_{imp} are defined as follows:

$$\begin{aligned}L_{gen}(D(z_1(y^i_{vis}), z_2(y^j_{pol}))) = &\frac{1}{2}||z_1(y^i_{vis}), z_2(y^j_{pol})||^2_2 \\ &\text{for} \qquad cl^i = cl^j\ .\end{aligned} \tag{3.12}$$

$$\begin{aligned}&L_{imp}(D(z_1(y^i_{vis}), z_2(y^j_{pol}))) = \\ &\frac{1}{2}\, max(0, m - ||z_1(y^i_{vis}), z_2(y^j_{pol})||^2_2) \quad \text{for} \quad cl^i \neq cl^j\ .\end{aligned} \tag{3.13}$$

Therefore, the coupling loss function can be written as:

$$L_{cpl} = 1/N^2 \sum_{i=1}^{N}\sum_{j=1}^{N} \ell_{cont}(z_1(y^i_{vis}), z_2(y^j_{pol}), y_{cont})\ , \tag{3.14}$$

where N is the number of samples. It should be noted that the contrastive loss function (3.14) considers the subjects' labels implicitly. Therefore, it has the ability to find a discriminative embedding space by employing the data labels in contrast to some other metrics such as the Euclidean distance. This discriminative embedding space would be useful in identifying a polarimetric probe photo against a gallery of visible photos.

3.3.2 Multi-Attribute Prediction and Identification Task:

The objective of this model is to predict a set of attributes using a visible or polarimetric face image. Therefore, in this architecture a visible face image (polarimetric face image) is presented to the network as an input and a set of attributes are predicted. Suppose the input is an image $y^i_{vis} \in Y$, and its class label is $cl_i \in CL$ for $i = 1, \ldots, N$ where N is the number of the training samples. Soft biometric traits, contain T different facial attributes or binary class labels. Therefore, in this framework we denote them as cl^t for $t = 1, \ldots, T$. Learning multiple CNNs separately is not optimal since different tasks may have some hidden relationships with each other and may share some common features. This is supported by [150] where they train a CNN features for the face recognition task and they used it directly for the face attribute estimation. Therefore, our network shares a big portion of its parameters among different tasks in order to enhance the performance of the recognition task. Thus, the loss function is as follows:

$$L_{avis} = 1/N \sum_{i=1}^{N} \sum_{t=1}^{T} \ell(f^t_{vis}(z_1(y^i_{vis}) \times w^t_{vis}), cl^{i,t}) \ , \tag{3.15}$$

where ℓ is a proper loss function (e.g., cross entropy) and $f^t_{vis}(.)$ is a binary classifier for the attribute t operated on the bottleneck of Vis-GAN (see Fig. 3.2). w^t_{vis} represents the remaining parameters which are assigned separately for each facial attribute task.

The same procedure is performed in the other network (Pol-GAN) with a polarimetric thermal image as input. The Pol-GAN network is also responsible to estimate a set of soft biometric attributes. Therefore, the loss function is:

$$L_{apol} = 1/N \sum_{j=1}^{N} \sum_{t=1}^{T} \ell(f^t_{pol}(z_2(y^j_{pol}) \times w^t_{pol}), cl^{j,t}) \ , \tag{3.16}$$

Table 3.1: Facial attributes used in this work.

Facial Attributes	Arched_Eyebrows, Big_Lips, Big_nose, Bushy_Eyebrows, Bald, Mustache, Narrow_Eyes, Beard, Mouth_Slightly_Open, Young

where ℓ is the cross entropy loss function and $f^t_{pol}(.)$ is a binary classifier for the attribute t operated on the bottleneck of Pol-GAN (see Fig. 3.2). w^t_{pol} represents the remaining features which are assigned separately for each facial attribute prediction task. The total attribute prediction loss function is:

$$L_a = L_{avis} + L_{apol}. \tag{3.17}$$

3.3.3 Generative adversarial loss

Let G_{vis} and G_{pol} denote the generators that synthesize the visible images from the visible and polarimetric images, respectively. To synthesize the output and to make sure that the synthesized images generated by the two generators are indistinguishable from the corresponding ground truth visible image, we utilized the GAN loss function [123]. As it is shown in Fig. 3.2, the first generator G_{vis} is responsible to generate a visible image when the network is conditioned on a visible image. On the other hand, the second generator G_{pol} tries to generate the visible image from the polarimetric image which is a more challenging task compared to the first generator. Therefore, the total loss for the coupled GAN is as follows:

$$L_{GAN} = L_{vis} + L_{pol}, \tag{3.18}$$

where the GAN loss function related to the Vis-GAN is given as:

$$L_{vis} = \underset{G_{vis}}{min}\ \underset{D_{vis}}{max} E_{x^i \sim P_{vis(x)}}[\log D(x^i|y^i_{vis})] + E_{z\sim P_z}[\log(1 - D(G(z|y^i_{vis})))],$$

where y^i_{vis} is the visible image used as condition for the Vis-GAN and x^i is the real data. It should be noted that for the Vis-GAN the real data x^i and the condition y^i_{vis} are the same. Similarly the loss for the Pol-GAN is given as:

$$L_{pol} = \min_{G_{pol}} \max_{D_{pol}} E_{x^j \sim P_{vis(x)}}[\log D(x^j|y^j_{pol})] + E_{z\sim P_z}[\log(1 - D(G(z|y^j_{pol})))],$$

where y^j_{pol} is the polarimetric image used as condition for the Pol-GAN and x^j is the real data. It should be noted that x^i is the same as x^j if they refer to the same person ($cl^i = cl^j$) and otherwise they are not the same.

3.3.4 Overall loss function

The proposed approach contains the following loss function: the Euclidean $L_{E_{vis}}$ and $L_{E_{pol}}$ losses which are enforced on the recovered visible images from the Vis-GAN and Pol-GAN networks, respectively, are defined as follows:

$$L_{E_{vis}} = ||G_{vis}(z|y^i_{vis}) - x^i||^2_2, \tag{3.19}$$

$$L_{E_{pol}} = ||G_{pol}(z|y^j_{pol}) - x^j||^2_2, \tag{3.20}$$

$$L_E = L_{E_{vis}} + L_{E_{pol}}. \tag{3.21}$$

The GAN loss is added to generate more sharp images with the use of adversarial loss. In addition, based on the usage of perceptual loss in low-level vision tasks [130], the perceptual loss is added to the Pol-GAN to preserve more photo realistic details as follows:

$$L_{P_{pol}} = \frac{1}{C_p W_p H_p} \sum_{c=1}^{C_p} \sum_{w=1}^{W_p} \sum_{h=1}^{H_p} ||V(G_{pol}(z|y^j_{pol}))^{c,w,h} - V(x^j)^{c,w,h}||,$$

where x^j is the ground truth visible image, $G_{pol}(z|y^j_{pol})$ is the output of Pol-GAN generator. V(.) represents a non-linear CNN transformation and C_p, W_p, H_p are the dimension of a particular layer in V. It should be noted that the perceptual loss is just used in the Pol-GAN. Similarly, we utilized a perceptual attribute loss which measures the difference between the facial attributes of the synthesized images and the real image. The pre-trained model needs to capture the facial attributes. Therefore, we fine-tune the pre-trained VGG-Face [153] on ten annotated facial attributes as tabulated in Table 3.1. Afterward, this network (attribute predictor) is utilized for perceptual attribute loss on Vis-GAN and Pol-GAN as follows:

$$L_{pavis} = ||A(G_{vis}(z|y^i_{vis})) - A(x^i)||^2_2, \tag{3.22}$$

$$L_{papol} = ||A(G_{pol}(z|y^j_{pol})) - A(x^j)||^2_2, \tag{3.23}$$

$$L_{pa} = L_{pavis} + L_{papol}. \tag{3.24}$$

where A is the fine-tuned VGG-Face network. L_{pa} is the perceptual attribute loss function which composed of the attribute loss function for Vis-GAN (L_{pavis}) and Pol-GAN (L_{papol}) sub-networks.

Finally, the coupling loss function (3.14) is added to train both networks Vis-GAN and Pol-GAN jointly to make the embedding space of the mentioned networks as close as possible and to preserve a more discriminative and distinguishable shared space. Therefore, the total loss function is as follows:

$$L_T = L_{cpl} + \lambda_1 L_E + \lambda_2 L_{GAN} + \lambda_3 L_a + \lambda_4 L_{Ppol} + \lambda_5 L_{pa},$$

where λ_1, λ_2, λ_3, λ_4, and λ_5 are the hyper-parameters which weight different loss functions in the total loss function.

3.3.5 Testing phase

During the testing phase, only the Pol-GAN is used. For a given test probe y^t_{pol}, Pol-GAN is employed in the proposed AGC-GAN to synthesize the visible image $G_{pol}(z|y^t_{pol}) = \hat{x}^t_{vis}$. Eventually, the identification of face recognition is done, by calculating the minimum Euclidean distance between the synthesized image from the polarimetric prob and visible gallery images as follows:

$$x^t_{vis} = \underset{x^i_{vis}}{argmin} \quad ||x^i_{vis}, \hat{x}^t_{vis}|| \ , \tag{3.25}$$

where $\hat{x}^t_{vis}$ is the synthesized probe face image and x^t_{vis} is the selected matching visible face image within the gallery of face images. In addition, the Pol-GAN sub-network can be employed to predict the facial attributes for the polarimetric face probe. The predicted facial attributes also can be used to find the person of interest in the visible databases.

3.4 Experiments (First approach: Sketch-photo recognition)

3.4.1 Implementation Details and Data Description

For the first approach, we used a VGG-16 like network [132] in our sketch-photo recognition framework. The VGG-16 neural network comprised of five major convolutional components which are connected in series. The first two components, $Conv1-64$ and $Conv2-128$ consists of the following layers: a convolutional layer, a rectified linear unit layer, a second convolutional layer, a second rectified linear unit layer, and a max pooling layer. The remaining three components contain one additional convolutional layer and a rectified linear unit layer. The only difference between our CNN network and the VGG-16 is in the last component, where the last three convolutional layers of VGG16 with the size of 512, are replaced with two convolutional layers of 256, and one convolutional layer of 64 respectively for the sake of parameter reduction. Also the network uses global pooling instead of the max pooling in the last component which results in a feature vector of size 64. The network which is dedicated to photo domain (P-DCNN) takes an RGB photo as an input and the other sketch-attribute network (SA-DCNN) gets an input consisting of multiple channels as shown in Fig. 3.1. The first channel is dedicated to a gray-scale sketch and the other channels are filled with 0 or 1 depending on the presence or absence of the attribute in the subject.

Experiment is performed using four main datasets, namely CUHK Face Sketch dataset (CUFS) [154] (containing 311 pairs), IIIT-D sketch dataset [155] (containing 238 viewed pairs, 140 semi-forensic pairs, and 190 forensic pairs), PRIP Viewed Software Generated Composite database (PRIP-VSGC) [19] (containing composite sketch and digital image pairs), extended-PRIP dataset (e-PRIP) [1], and unviewed Memory Gap Database (MGDB) [32] (containing 100 pairs). Since we are using the facial attribute classification in our proposed method, we utilized the CelebFaces Attributes dataset (CelebA) [156] (consisting of 200 k face images along with their attribute vectors of 40 attributes such as gender, face characteristics, skin color, hair color, etc.) to initialize the network. Since the CelebA dataset does not contain the sketch images we generated a synthetic sketch by employing xDOG [157] filter on each image. Twelve facial attributes namely bald, black hair, blond hair, brown hair, gray hair, male, Asian, Indian, White, Black, eye glasses and pale skin out of 40 attributes were selected. Since none of the sketch datasets used in this chapter have any facial attribute annotation, we utilized MOON [158], which is a well-known method in facial attributes recognition, in order to annotate them.

We pre-trained our deep coupled architecture using synthetic sketch-photo pair from the CelebA dataset.

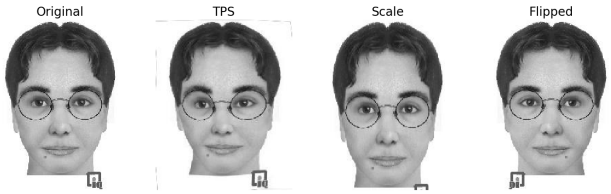


Figure 3.5: A sample of different augmentation techniques.

Table 3.2: Experimental Setup

Setup Name	Testing Dataset	Training Dataset	Train Size	Gallery Size	Prob Size
S1	e-PRIP	e-PRIP	48	75	75
S2	e-PRIP	e-PRIP	48	1500	75
S3	IIIT-D Semi-forensic	CUFS, IIIT-D Viewed, CUFSF, e-PRIP	1968	1500	135
	MGDB Unviewed				100

We used the final weights to initialize the network in all of our training scenarios. Since our coupled DNN has a large number of parameters and the size of the sketch datasets is relatively small it is prone to overfitting. In order to avoid the overfitting problem, we utilized multiple augmentation techniques namely deformation, scale and crop, and flipping. In the following we explain each method in details (see Fig. 3.5).

1- Deformation: Deforms the sketch and photos to compensate for the problem of geometrically mismatching between the sketch image and its corresponding photo. The deformation is performed by translating 25 preselected points with random magnitude and direction.

2- Scale and crop: One of the main mismatch problems between sketch images and their corresponding images is the ratio deformation. To address this problem, this method upscale the sketches and photos to several random sizes, and then cut a 250×200 crop from the center of the scaled image.

3- Flipping: In this method, the images are randomly flipped horizontally.

During the training phase, instead of picking the impostor pairs randomely, we considered an strategy to select them. For each genuine pair, we considered four impostor pairs. Two of the impostors were selected among the subjects which are sharing the same set of facial attributes and the other two were selected among the subjects which have different sets of attributes. This selection technique made the framework to see more variant impostors and also helped to avoid the overfitting problem.

3.4.2 Performance Evaluation:

Our proposed framework identify a person of interest in the galley of mugshots utilizing a sketch probe and a set of facial attributes provided. In this section, we compare our approach with several state-of-the-art

methods which are using both sketch and attributes and some other methods which are just using the sketch without using any attributes.

In order to evaluate performance of our method and compare with other methods, three different experiments are performed. For the sake of fair comparison, the first two experiment setups are adopted from [1]. In the first experimental setup which is the baseline (S1), the database is partitioned into two parts: training which is performed on 40% of the data and the remaining portion of the data is used for testing. e-PRIP dataset containing 123 subjects is used in this setup. Therefore, 48 identities are used in the training set and 75 subjects are considered for the testing phase. Only two out of the four different composite sketch datasets utilized in [1] are public at the time of writing this thesis. These two public datasets were created by Identi-Kit tool, and FACES tool and were used by Asian and Indian artists respectively. In the second experimental setup, called S2, the gallery is extended to 1500 subjects. In this chapter, the gallery is expanded utilizing WVU Multi-Modal [159], IIIT-D sketch, Multiple Encounter Dataset (MEDS) [160], and CUFS datasets. The facial attributes of the extended galley are obtained using MOON [158]. The training and probe datasets are the same as S1. The purpose of this experiment is to assess the robustness of the proposed method with a relatively large number of subject candidates. Finally, the method is evaluated on an unseen dataset. In this experimental setup (S3), we trained the network on IIIT-D Viewed, CUFS, and e-PRIP datasets and then tested it on IIIT-D Semi-forensic pairs and MGDB Unviewed. This setup represents the level of dependency of the network on the sketch styles in the training datasets. Table 3.2 shows different scenarios and the size of training set, prob and gallery for each scenario.

In the experiments, different values were selected for λ_1 and λ_2. We report our best results which belong to λ_1=λ_2=1. The evaluation performance is validated using ten fold random cross validation and the results are compared with the state-of-the-art approaches.

3.4.3 Results:

In [1], they propose an approach called attribute feedback to study the effect of facial attributes on their recognition system. They reported the rank 10 accuracy of 58.4% and 53.1% for the prob sketches generated by the Indian (Faces) and Asian (Identi-Kit) artists, respectively. Another approach called SGR-DA [2] utilizes the sketch modality without using the facial attribute information. They reported the rank 10 accuracy of 70% on the Identi-Kit dataset. On the other hand, our proposed approach accuracy was 76.4% and 72.3% on the Faces and Identi-Kit, respectively. We also consider a baseline version of our proposed

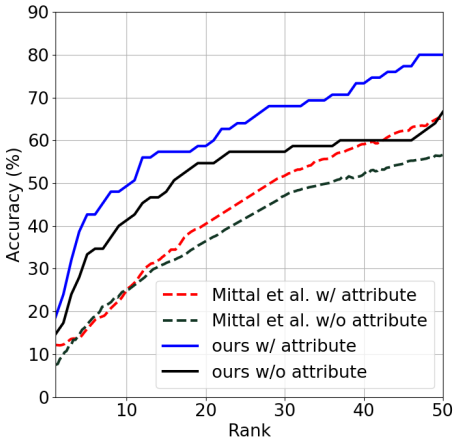


Figure 3.6: CMC curves of our proposed framework versus Mittal et al. algorithms [1] in the extended gallery experimental setup (S2) for the Indian dataset

method which is only based on the contrastive loss function and does not consider the facial attributes. This way, we could observe the benefit of utilizing the facial attributes in our framework. The baseline network has an accuracy of 69.1% and 67.6%, on Faces and Identi-Kit datasets, respectively. The results demonstrate that our method outperforms all the previous methods in the literature and also express the effectiveness of our framework in utilizing the facial attributes compare to the baseline. It should be noted that, the baseline framework outperform the state-of-the-art methods except SGR-DA [2] which support the superiority of deep models over the shallow models (see Table 3.3).

To evaluate the effectiveness of our proposed method by using a relatively large galley of mugshots, the same experiments were performed on the extended experimental setup (S2). Figure 3.6 shows the results of our method as well as the other methods for the extended galley of 1500 subjects. The results depicts that our approach outperforms the method in [1] by nearly 14% for rank 50 which shows the robustness of our algorithm utilizing the facial attributes. We compared our method with SGR-DA [2] for the Identi-Kit dataset, since [1] does not provide the results on this dataset. Figure 3.7 shows the superiority of our proposed method compared to SGR-DA. As shown in Fig. 3.7, although SGR-DA outperformed our baseline network in S1 scenario (see Table 3.3), its results were not as promising as our proposed method in the extended experimental setup (S2). Also, our attribute-assisted method outperformed our baseline method to support the effectiveness of utilizing the attributes in relatively large gallery of mugshots as well.

Eventually, we evaluated the robustness of our proposed method in S3 experimental setup in which

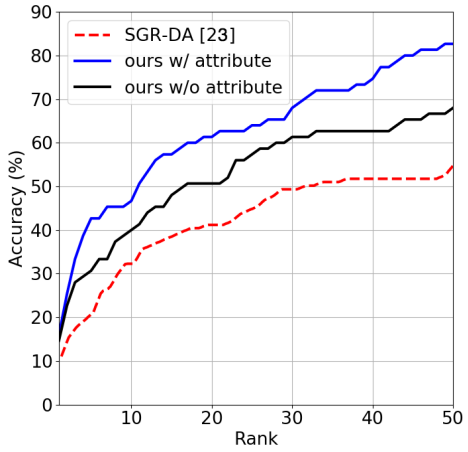


Figure 3.7: CMC curves of our proposed framework versus SGR-DA algorithm [2] in the extended gallery experimental setup (S2) for the Identi-Kit dataset

Table 3.3: Rank-10 identification accuracy (%) on the e-PRIP composite sketch database (S1 experimental setup).

Algorithm	**Faces (In)**	**IdentiKit (As)**
Mittal et al. [161]	53.3 ± 1.4	45.3 ± 1.5
Mittal et al. [88]	60.2 ± 2.9	52.0 ± 2.4
Mittal et al. [1]	58.4 ± 1.1	53.1 ± 1.0
SGR-DA [2]	-	70
Ours without attributes	69.1 ± 1.5	67.6 ± 1.9
Ours with attributes	**76.4 ± 1.2**	**72.3 ± 0.8**

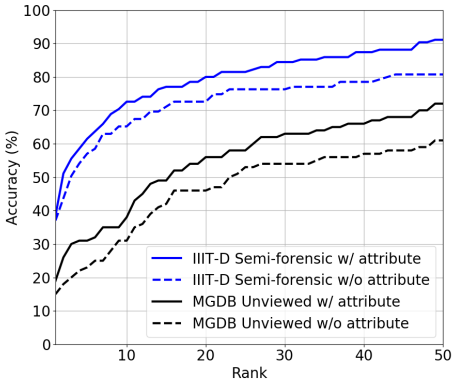


Figure 3.8: CMC curves of our proposed framework versus our baseline framework (without using attributes) for experimental setup (S3). The results support the robustness of our approach to different sketch styles.

the network is trained on more than 1900 sketch-photo pairs and is tested on two unseen datasets, namely MGDB Unviewed and IIIT-D Semi-forensic datasets. In this scenario the gallery of mugshots was also extended to 1500. We repeated this experimental scenario for our baseline method which is not utilizing the facial attributes. As shown in Fig. 3.8, the proposed method showed a better performance in this scenario on both datasets compared to the baseline method indicating the advantage of facial attributes in the proposed method on unseen datasets.

3.5 Experiments (Second approach: Polarimetric-visible recognition)

3.5.1 Implementation Details

A U-net structure [152] is employed as the network for the generator since it is able to address the vanishing gradient problem as well as better capturing large receptive field. Also, a patch-based discriminator [162] is used in the proposed method and it is trained iteratively with the generator. The entire network is trained in Pytorch. For the sake of training AGC-GAN, the hyper-parameters for all the loss functions considered as one except for the perceptual loss $L_{p_{pol}}$ and perceptual attribute loss L_{pa} which is equal to 0.5. For training we used Adam optimizer [133] with the first-order momentum of 0.5, the learning rate of 0.0002, and batch size of 4. For the generator the ReLU activation, and for the discriminator the Leaky ReLU activation with the slope of 0.2 is considered. The perceptual loss is assessed on relu3-1 layer in the pre-trained VGG [120] model. In order to fine-tune the attribute predictor network utilized for perceptual attribute loss, we manually annotate images with the attributes tabulated in Table 3.1.

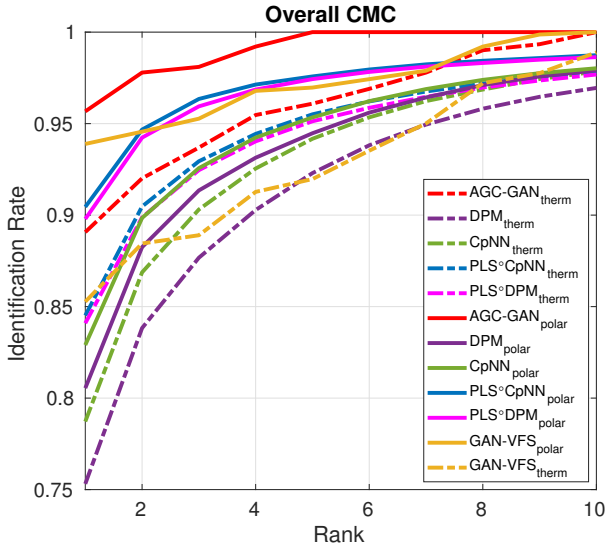


Figure 3.9: Overall CMC curves from testing DPM, CpNN, PLS∘DPM, PLS∘CpNN, GAN-VFS, and AGC-GAN using polarimetric and thermal probe samples, matching against a visible spectrum gallery.

3.5.2 Results

We evaluate the proposed face recognition method compared with several recent works [24, 75, 74, 14, 70, 78] on the ARL Multi-modal Face database [14].

Polarimetric Thermal Face dataset [14] comprises polarimetric LWIR face images and their corresponding visible spectrum related to 60 subjects. Data was collected at three different distances: Range 1 (2.5 m), Range 2 (5 m), and Range 3 (7.5 m). At each range two different conditions, including baseline and expression are considered. In the baseline condition the subject is asked to keep a neutral expression looking at the polarimetric thermal sensor. On the other hand, in the expression condition the subject is asked to count out numerically from one upwards which results in different expressions in the mouth and to the eyes and consequently different variations in the facial imagery.

To increase the correlation between the two modalities of visible and thermal, each modality was pre-processed. We applied a band-pass filter so called difference of Gaussians (DoG), to emphasize the edges in addition to removing high and low frequency noise.

We pass S_0, S_1, and S_2 to the Pol-GAN's three channels as the input as shown in Fig. 3.2. The training set is used to transform the visible and polarimetric features to a shared latent embedding subspace. Also at the same time, the network tries to synthesize visible modality from the shared latent subspace in the GAN framework. To train the network, the genuine and impostor pairs are constructed. The genuine pair

Table 3.4: Rank-1 identification rate for cross-spectrum face recognition using polarimetric thermal and conventional thermal (S_0) probe imagery.

Scenario	Rank-1 Identification Rate							
	Probe	PLS	DPM	CpNN	PLS∘DPM	PLS∘CpNN	GAN-VFS	AGC-GAN
Overall	Polar	0.5867	0.8054	0.8290	0.8979	0.9045	0.9382	**0.9654**
	Therm	0.5305	0.7531	0.7872	0.8409	0.8452	0.8561	**0.8925**
Expressions	Polar	0.5658	0.8324	0.8597	0.9565	0.9559	0.9473	**0.9733**
	Therm	0.6276	0.7887	0.8213	0.8898	0.8907	0.8934	**0.9217**
Range 1 Baseline	Polar	0.7410	0.9092	0.9207	0.9646	0.9646	0.9653	**0.9883**
	Therm	0.6211	0.8778	0.9102	0.9417	0.9388	0.9412	**0.9659**
Range 2 Baseline	Polar	0.5570	0.8229	0.8489	0.9105	0.9187	0.9263	**0.9643**
	Therm	0.5197	0.7532	0.7904	0.8578	0.8586	0.8701	**0.9178**
Range 3 Baseline	Polar	0.3396	0.6033	0.6253	0.6445	0.6739	0.8491	**0.9068**
	Therm	0.3448	0.5219	0.5588	0.5768	0.6014	0.7559	**0.8124**

is constructed from the same subject images in two different modalities. For the impostor pair, a different subject is selected for each modality. In general, the number of the generated impostor pairs are significantly larger than the genuine pairs. For the sake of balancing the training set, we consider the same number of genuine and impostor pair. After training the model, during the testing phase, only the polarimetric network is used for the evaluation. For a given probe, the network is used to synthesize the visible image. Afterwards, the Euclidean distance is used to match the synthesize image to its closest image from the gallery. The ratio of the number of correctly classified subjects and the entire number of subjects is computed as the identification rate.

In each experiment the dataset is partitioned to train and test randomly. The same set of train and test is used to evaluate PLS [24], DPM [75], CpNN [74], PLS∘DPM [14], PLS∘CpNN [70], GAN-VFS [78], and the proposed AGC-GAN network. Fig. 3.9 shows the overall cumulative matching characteristics (CMC) curves for our proposed method and the other state-of-the-art methods over all the three different ranges as well as the expressions data at Range 1. For the sake of comparison, in addition to the polarimetric thermal-to-visible face recognition performance, Fig. 3.9 also shows the results for the conventional thermal-to-visible face recognition for some of the methods, namely PLS∘DPM, PLS∘CpNN, CpNN, and AGC-GAN. In the conventional thermal-to-visible face recognition, all the mentioned methods exactly follow the same procedure as before, with only using S_0 modality. Fig. 3.9 illustrates that exploiting the polarization information of the thermal spectrum enhances cross-spectrum face recognition performance compared to the conventional one. Fig. 3.9 also shows the superior performance of our approach compared to the state-of-the-art methods. In addition, our method could achieve prefect accuracy of 1 at Rank-4 and above.

Table 3.5: Attribute prediction of the polarimetric face images using the proposed method and other frameworks and comparing it with the attribute prediction of visible faces.

Facial Attributes	Arched_Eyebrows	Big_Lips	Big_Nose	Bushy_Eyebrows	Bald	Mustache	Narrow_Eyes	Beard	Mouth_Slightly_Open	Young
Visible faces	96.7	98.4	99.1	95.9	99.3	99.4	95.7	98.9	97.7	96.9
Polar face	53.7	55.8	57.1	51.2	58.9	62.8	54.4	59.8	57.6	52.7
Polar faces (fine-tune)	78.2	83.9	85.3	80.7	88.4	89.3	79.3	88.9	81.9	76.5
Pol-GAN Network	89.6	95.3	96.6	90.4	96.2	95.2	91.9	94.8	93.7	91.1

Table 3.4 tabulates the Rank-1 identification rates for five different scenarios: overall (which corresponds to Fig. 3.9), Range 1 expressions, Range 1 baseline, Range 2 baseline, and Range 3 baseline. In our proposed approach, exploiting polarization information enhance the Rank-1 identification rate by 2.24%, 5.16%, 4.65%, and 9.44% for Range 1 baseline, Range 1 expression, Range 2 baseline, and Range 3 baseline compared to the conventional thermal-to-visible face recognition. This table reveals that using deep coupled generative adversarial network technique with the contrastive loss function as well as utilizing facial attributes to transform different modalities into a distinctive common embedding subspace is superior to the other embedding techniques such as PLS∘CpNN. It also shows the effectiveness of our method in exploiting polarization information to improve the cross-spectrum face recognition problem.

3.6 Ablation study

In order to illustrate the effect of adding different loss functions and their improvement in our proposed framework, we perform an study with the following evaluations using the polarimetic dataset: 1) Polar-to-visible using the coupled framework with using only $L_{cpl} + L_E$ loss, 2) Polar-to-visible using the proposed framework with $L_{cpl} + L_E + L_{GAN} + L_{p_{pol}} + L_{pa}$ loss functions, and 3) Polar-to-visible with all the loss functions in the proposed framework.

We plot the receiver operation characteristic (ROC) curves corresponding to the mentioned three different settings of the framework in the task of face verification. As it is shown in Fig. 3.10 the L_{GAN} has an important rule in the enhancement of our proposed approach by transforming the polarimetric modality to the visible one. Moreover, adding facial attribute prediction loss enhances the face recognition performance. The reason behind this is because using facial attributes loss in addition to contrastive loss function leads to a more discriminative embedding space and this leads to a better face recognition performance. Consider a polarimetric subject with $Id\#2$ (see Fig. 3.4). The contrastive loss function causes the corresponding visible images from $Id\#2$ to move closer to $Id\#2's$ polarimetric and other Ids' visible images to move

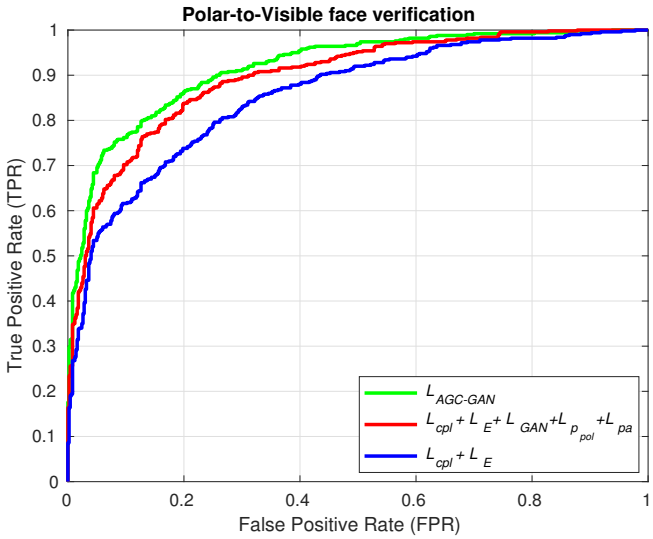


Figure 3.10: The ROC curves corresponding to the ablation study.

farther away. Now, using the contrastive loss function in conjunction with the attribute classification makes $Id\#1$ to move closer to $Id\#2$ since they share the same set of attributes (see Fig. 3.4). In other words, it differentiates between different impostors of $Id\#2$. The same procedure is performed for the other identities during the training process. Fig. 3.4 visualizes the overall concept of adding facial attributes prediction loss function. As it is depicted, addition of attribute prediction loss leads to a more discriminative embedding subspace. This leads to a better face recognition performance as it is shown in Fig. 3.10.

3.7 Attribute Prediction from Polarimetric thermal

One of the benefits of the proposed AGC-GAN is predicting facial attributes directly from polarimetric thermal modality. These attributes can be utilized directly or can be fused with other modalities to enhance recognition performance. In order to illustrate the effectiveness of the proposed method we performed attribute prediction in four different scenarios: 1) Attribute prediction of visible images with the attribute predictor. 2) Attribute prediction of polarimteric images with the attribute predictor. 3) Attribute prediction of polarimetric images with the fine-tuned attribute predictor. In this case, we fine-tuned the attribute predictor with the annotated polarimetric images and used it for the task of attribute prediction in the testing phase. 4) Attribute prediction of the polarimetric images using the Pol-GAN network from the proposed AGC-GAN. Table 3.5 shows the result of the prediction for the four mentioned frameworks. Although,

fine-tuning the attribute predictor increased the prediction performance (framework #3), but still its performance is less than our proposed framework. The proposed framework could outperform the other methods in polarimetric face recognition and it has a comparable performance to attribute prediction from the visible face images.

3.8 Conclusion

We have introduced a novel approach to exploit facial attributes information for the purpose of sketch-photo and polarimetric thermal-to-visible face recognition. In the first approach we proposed to use coupled deep neural network with facial attributes provided by eye witnesses. We simultaneously minimize the cost functions due to the facial attribute identification as well as the sketch-photo verification in order to increase inter-personal variations between different subjects with different sets of facial attributes and reducing intra-personal variations in the latent feature subspace. The combination of the two cost functions leads to a significantly more discriminative embedding subspace compared to the subspace that is created by either one of them. In the second approach, AGC-GAN contains two GAN based sub-networks dedicated to visible and polarimetric input images. This network is capable of transforming the visible and polarimetric thermal modalities into a common discriminative embedding subspace and synthesizing the visible photos from that subspace. It simultaneously minimizes the cost functions due to the facial attribute identification in addition to the other cost functions in order to increase inter-personal variations between different subjects with different sets of facial attributes in the latent feature subspace. This leads to a more discriminative embedding subspace. An ablation study was performed to demonstrate the enhancement obtained by different losses in the proposed method. Our main focus was on the accuracy performance and we did not consider the training or inference timing. We compared our method with state-of-the-art polarimetric thermal-to-visible face recognition methods and showed the superiority of our method over them.

Chapter 4

Soft Biometrics as a Privileged Data to Improve Deep Face Recognition

4.1 Introduction

In our approach, we propose a new framework where we learn features in a common latent subspace between the soft and hard biometric pairs of training data. We develop this framework based on a new training scheme, which empowers the network to outperform other networks with the same topology that are trained with no auxiliary data. More specifically, our proposed training scheme has the ability of transferring information from the auxiliary data and help the network to learn a more discriminative feature about the primary data. To this end, we utilize a coupled deep neural network (Cp-DNN) architecture which consists of a deep network (Net-I) dedicated for the primary image data coupled with another network (Net-S) dedicated for the auxiliary soft biometrics data, which together discover a common latent subspace between the two modalities during the training stage. This shared common feature subspace has a unique property; the Euclidean distance between the latent feature vectors from different modalities with the same identity are close in this latent subspace, while their Euclidean distance for different identities are far from each other.

In another framework, we investigate the concept of LUPI in the multi-task learning (MTL) paradigm. A face photo can be viewed as having some positive or negative hidden relation with some of its soft facial biometric traits. For example, a face with goatee beard has a male gender and this can help the classifier to refine the search space for the task of face recognition. Although, the soft biometric traits are correlated, they can be heterogeneous in data-type and scale [163] and semantic meaning [164]. While some soft

biometric traits such as height and age are ordinal, other soft biometric traits such as race and gender are nominal. Therefore, these two categories of face attributes belong to different heterogeneous data-types. In the MTL framework, we consider the early stage weight sharing which is mainly the CNN weights followed by the dedicated weights which are responsible for learning the representative features for each specific face attribute (privileged soft biometric trait). The first part of the network shares the same weights among different tasks in order to exploit a common information between different kinds of tasks by capturing the relationships between them. As a result, the trained embedding space is more discriminative and can be utilized to enhance the performance of the main task which is face recognition.

For the purpose of assessment, we study the problem of face recognition using paired image-attribute data, while the attributes (i.e., soft biometrics) are only available during the training phase. The performance of the proposed scheme is evaluated on four different datasets, namely CelebA [156], Morph [165], Biocop [159], and CMU-PIE [166] datasets. Our experimental results show that the proposed training scheme helps our architectures to transfer latent information from the auxiliary data to learn better discriminative features and to increase the overall recognition performance.

4.2 Related works

In computer vision and biometrics, it is a common trend to have more than one modality as the source of information. These modalities can vary from visual images to soft biometrics or a sequence of text data. Soft biometric traits can be applied to many useful applications. They can be directly used in a unimodal system for some applications [38] or can be fused with some primary biometric traits. This fusion is usually done for the sake of improvement in the recognition of a biometric trait. Jain et.al [167], considered a hybrid system which combined soft biometrics such as age, gender, and ethnicity with fingerprint in order to enhance the overall matching accuracy. There are some other notable research such as [168], [169], and [170] where soft biometrics are fused with the classical hard biometric traits. The task of using all modalities in a learning process, when they are available in both training and testing phases, has been studied under the context of multimodal or multiview learning. Different methodologies are proposed in the literature for this paradigm from feature stacking, in which features of different modalities are simply concatenated, to the fusion of modalities in early or late stages [42].

On the contrary, in this chapter we are addressing the problem of employing auxiliary information,

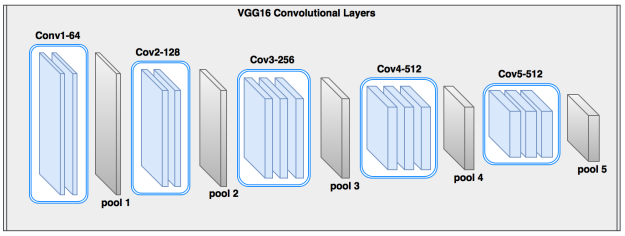


Figure 4.1: VGG-16 first 13 convolutional layers architecture.

which is only available during the training phase and not in the testing phase. Different settings of this problem have been investigated in the literature. In [171], a canonical correlation analysis based method was proposed for spectral clustering with paired data. The novelty of their method was that they used a distinct similarity measure for each modality. A shared feature space is found in [172] in which the Euclidean distance is meaningful for intra and inter modalities. [173] has proposed two visual annotator rationales for visual recognition. These rationales were then used as extra information to improve the absolute recognition accuracy.

The concept of LUPI was introduced in [41], for the first time, under the context of SVM+. In their methodology, the auxiliary data, or privileged information, was employed to predict the slack variables of an SVM classifier [174]. Since then, the idea of using privileged information has been vastly investigated in the literature and has been applied to different applications and contexts, e.g., object localization [175], facial feature detection [42], and metric learning [176]. In [42], the privileged information is used to distinguish between ”easy to classify” and ”hard to classify” examples. They showed that an extra source of information, such as text, bounding box, or attribute, can help a classifier to do a better classification in object recognition tasks. A framework was introduced in [92] to embed the main data into the latent feature subspace such that the mutual information between the embedded data and auxiliary data is maximized.

The concept of learning latent shared representation has also been studied in the area of transfer learning and domain adaptation. Authors in [177] have proposed a framework to learn the mapping from the available modalities to a new unseen modality using unlabeled data. An approach has been introduced in [178] to find a shared representation between RGB and depth modalities during the training phase. The authors used the extracted shared subspace from the training phase in order to help RGB modality during the testing phase. Hoffman et. al. [179] proposed an approach to enrich RGB information using depth data during the training phase as side information. They trained a hallucinating convolution neural network with the depth data in

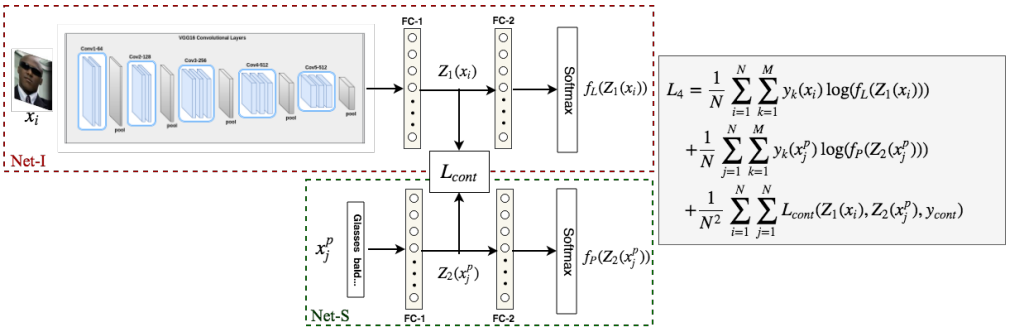


Figure 4.2: Proposed Cp-DNN model (Net-I+Net-S) using face images and soft biometric attributes for face classification during the training phase on Morph, Biocop and CelebA datasets. In the case of CMU multi-PIE dataset Net-S will be replaced with another Net-I.

order to utilize it jointly with the RGB network during the test time and boost accuracy in object detection tasks.

Multi-task learning has been vastly applied in computer vision and biometrics problems. It basically attempts to solve correlated tasks concurrently with the help of knowledge sharing between tasks. In [180] it was shown that MTL can boost the performance of different tasks. The early MTL approaches tried to improve the performance by finding the relationship among different tasks [181]. Afterwards, in [182] an MTL approach was utilized to find the common features between different tasks. In [183], the authors introduced an outlier matrix and introduced outliers that shared common features with other tasks and claimed that due to the high dimensionality of the data, the assumption of sharing features among different tasks is not reasonable. In other works [184, 185], the authors utilized techniques such as structure sparsity for feature selection in order to select meaningful common features. [148, 186] employed MTL technique to predict attributes such as age, gender, race, etc.

There is a growing interest on employing soft biometrics as complementary side information to low-level hard biometrics features. Authors in [187, 188] proposed fusion methods to combine these two different modalities. However, the soft biometric traits are considered to be available both during the training and testing phases. Our approach looks at this problem in a different way. We consider the case when soft biometric information does not exist during the testing phase, and we only have them during the training phase.

In this chapter, two face recognition approaches are proposed (Cp-DNN, and MTL-LUPI). Both approaches are developed for two main tasks, namely face classification and face verification. Our first approach employs Cp-DNN, in which soft biometric data is used only during the training phase, to improve

the accuracy of the trained network. The soft biometric attributes, which play the role of privileged data in our work, are not discriminative enough to perform a high accurate identification on their own, and they can change over time as well. However, we make use of them in the training phase to improve the performance of our deep convolutional neural network (CNN) dedicated for our primary data during the testing phase. In the second approach MTL-LUPI, we utilize the MTL technique, to predict the privileged soft biometric traits and adjust the network common features in a way to enhance the main goal which is the face recognition task.

4.3 Methodology

In traditional machine learning algorithms, a training set comprises of pairs of input and output. For example, in image classification, the input is an image $x_i \in X$, and the output is its class label $y_i \in Y$ for $i = 1, \ldots, N$ where N is the number of the training samples. The goal is to find a prediction function $f : X \rightarrow Y$. However, in our work, for each pair (x_i, y_i) we are also given an extra information x_i^p where in the literature it is called the privileged information. This extra information is only available during the training phase. Similar to the traditional algorithms, we aim to learn a prediction function $f_L : X \rightarrow Y$ but also exploiting the privileged data.

4.3.1 Traditional supervised algorithms

In general, when there is no privileged information, the parameters of the prediction function f, can be learned by minimizing an appropriate loss function as follows:

$$L_1 = 1/N \sum_{i=1}^{N} \ell(f(x_i), y_i) \ , \tag{4.1}$$

where ℓ is a desirable loss function, e.g., categorical cross entropy for multi-class classification. The prediction function f could be any of the popular classifiers in the literature, such as SVM [41], LVQ [189], or a multilayer perceptron (MLP). However, deep convolutional neural networks [120], which is a recent topic of interest in the literature, has successfully been applied to many computer vision problems. Hence, we selected a VGG-16 architecture (Fig. 4.1) as our basic prediction function. In the following section, the topology of our network is described in details.

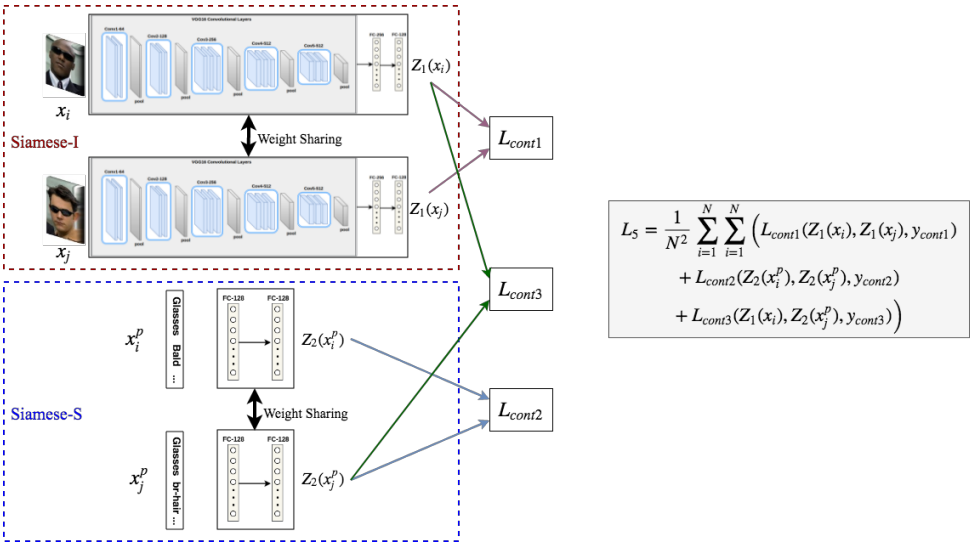


Figure 4.3: Proposed Cp-DNN framework using face images and soft biometric attributes for face verification during the training phase.

4.3.2 Learning using privileged information via coupled deep neural network

Several different techniques are proposed in the current literature on LUPI for making use of privileged information during the training phase. Authors in [92] added a mutual information term as a regularization factor to the cost function in Eq. 4.1. However, they pointed out a serious drawback that the labels of the samples are neglected in their loss function for LUPI, or, in other words, their LUPI technique is unsupervised. In contrast, we utilize a supervised loss function, which considers the sample labels. To that end, we use a contrastive loss function [62] (ℓ_{cont}) to pull the genuine pairs (i.e., a face image with its own corresponding soft biometric attributes), into a common feature subspace, toward each other and push the impostor pairs (i.e., a face image with another subject's soft biometric attributes) apart from each other (see Fig. 4.2). Similar to [62], the contrastive loss is of the form:

$$\ell_{cont}(z_1(x_i), z_2(x_j^p), y_{cont}) = \tag{4.2}$$
$$(1 - y_{cont})L_{gen}(D(z_1(x_i), z_2(x_j^p)) + y_{cont}L_{imp}(D(z_1(x_i), z_2(x_j^p)) \ ,$$

where x_i is the input for the main view (face image), and x_j^p is the input for the privileged view (soft biometric attributes). L_{gen} and L_{imp} represent the partial loss functions for the genuine and impostor pairs, respectively, and $D(z_1(x_i), z_2(x_j^p))$ indicates the Euclidean distance between the embedded data in the common feature subspace. y_{cont} is a binary label which is assigned a value of 0 when both modalities, i.e.,

main and privileged, form a genuine pair, or, equivalently, when the inputs are from the same class. On the contrary, when the inputs are from different classes, which means they form an impostor pair, y_{cont} is equal to 1. In addition, L_{gen} and L_{imp} are defined as follows:

$$L_{gen}(D(z_1(x_i), z_2(x_j^p))) = \frac{1}{2} D(z_1(x_i), z_2(x_j^p))^2 \quad \text{for} \quad y_i = y_j \ , \tag{4.3}$$

$$L_{imp}(D(z_1(x_i), z_2(x_j^p))) = \frac{1}{2}\, max(0, m - D(z_1(x_i), z_2(x_j^p)))^2 \quad \text{for} \quad y_i \neq y_j \ . \tag{4.4}$$

Therefore, the main loss function can be written as:

$$L_2 = 1/N \sum_{i=1}^{N} (\ell(f_L(z_1(x_i)), y_i) + 1/N^2 \sum_{i=1}^{N} \sum_{j=1}^{N} \ell_{cont}(z_1(x_i), z_2(x_j^p), y_{cont}) \ , \tag{4.5}$$

where the first term is the classification/verification term and the second term is the coupling term. z_1 and z_2 are the DNN-based embedding functions, which transform x_i and x_j^p into a common latent embedding subspace. As it was mentioned earlier, the contrastive loss function has the ability to find a discriminative embedding space by employing the data labels. Due to the classification/verification term in (4.5), minimizing L_2 will boost the discriminative common space of the main data, x_i. However, it does not enforce a discriminative ability on the common subspace for the privileged data. Therefore, we add a classification loss term $\ell(f_P(.))$ for the privileged data to the total loss function as follows (as shown in Fig. 4.2):

$$\begin{aligned} L_3 = 1/N \sum_{i=1}^{N} & (\ell(f_L(z_1(x_i)), y_i) \\ & + 1/N \sum_{j=1}^{N} \ell(f_P(z_2(x_j^p)), y_j) + 1/N^2 \sum_{i=1}^{N} \sum_{j=1}^{N} \ell_{cont}(z_1(x_i), z_2(x_j^p), y_{cont}) \ , \end{aligned} \tag{4.6}$$

where $f_L(.)$, and $f_P(.)$ are the classification functions which operate on z_1, and z_2, respectively (See Net-I and Net-S in Fig. 4.2). To visualize the common latent embedding subspace generated by the Net-I module (see Fig. 4.2), in the Cp-DNN framework for the Morph dataset trained according to (4.6), we reduced its dimensionality using principal component analysis (PCA) [190]. Afterwards, t-Distributed stochastic neighbor embedding (t-SNE) [191] is employed to project the transformed common features into two di-

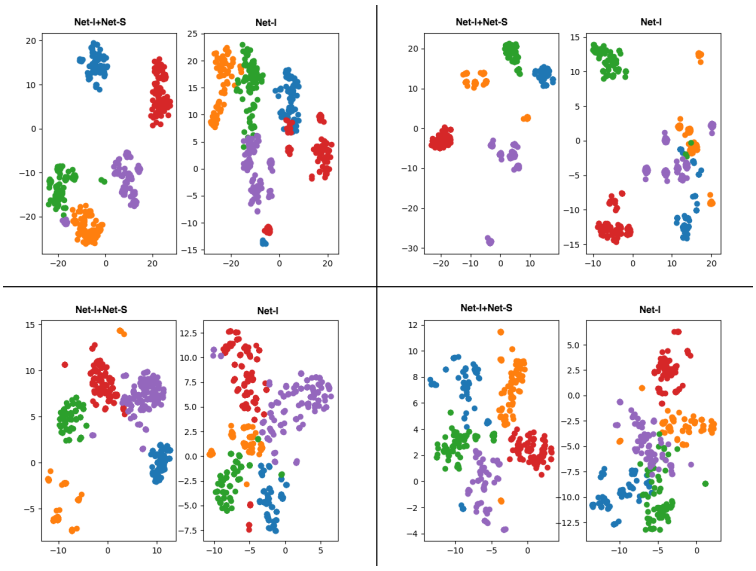


Figure 4.4: Common embedding subspace of Net-I training based on our proposed Cp-DNN framework (Net-I + Net-S) vs. training based on baseline network (Net-I) for four different groups of 5 classes (classification of Morph).

mensions. The final two dimensional embedding features are depicted in Fig. 4.4. The plots related to the Net-I embedding subspace which is trained using the (4.6) (Net-I + Net-S) show a more discriminative subspace compared to the base-model which is trained according to the Eq. 4.1 (Net-I). This emphasizes the effectiveness of our proposed network.

During the testing phase, for a given test sample x_i, only the prediction function f_L is used for the prediction (illustrated in Fig. 4.5), because its corresponding privileged information is not available. However, having f_P in the training phase will help to find a more discriminative common subspace which consequently leads to a better prediction performance in the testing phase.

We applied this paradigm for face classification and verification tasks. For classification task (see Fig. 4.2), $\ell(f_L(.))$ and $\ell(f_P(.))$ terms in (4.6) are the cross entropy loss functions, and for the verification task, $\ell(f_L(.))$ and $\ell(f_P(.))$ are the contrastive loss functions (see Fig. 4.3). Therefore, the total loss function for the classification is given by:

$$\begin{aligned} L_4 = 1/N \sum_{i=1}^{N}\sum_{k=1}^{M} y_k(x_i) log(f_L(z_1(x_i))) \\ + 1/N \sum_{j=1}^{N}\sum_{k=1}^{M} y_k(x_j^p) log(f_P(z_2(x_j^p))) + 1/N^2 \sum_{i=1}^{N}\sum_{j=1}^{N} \ell_{cont}(z_1(x_i), z_2(x_j^p), y_{cont}) \ , \end{aligned} \tag{4.7}$$

where $y_k, k = 1, ...M$ indicates the class labels. x_i and x_j^p refer to the main and privileged samples,

respectively, for $i, j = 1, ..., N$. $f_L(.)$, and $f_P(.)$ are the predicted class labels for x_i, and x_j, respectively. In the case of verification, the total loss function is defined as follows:

$$L_5 = 1/N^2 \sum_{i=1}^{N} \sum_{j=1}^{N} (\ell_{cont1}(z_1(x_i), z_1(x_j), y_{cont1}) + \ell_{cont2}(z_2(x_i^p), z_2(x_j^p), y_{cont2}) + \ell_{cont3}(z_1(x_i), z_2(x_j^p), y_{cont3})) \ , \tag{4.8}$$

where the first and second loss terms $\ell_{cont1}, \ell_{cont2}$ are the verification terms for the main samples (x_i, x_j) and the privileged samples (x_i^p, x_j^p), respectively and the third loss term ℓ_{cont3} is the coupling term (see Fig. 4.3). y_{cont_1}, y_{cont_2}, and y_{cont_3} are the binary labels which are assigned to Siamese-I, Siamese-S, and coupling Siamese-I and Siamese-S, respectively. While the whole approach remains the same for the classification and verification tasks, but their network structures are different. For verification task, we used a coupled Siamese network [62] which consists of a Siamese network dedicated for the verification task using the primary face dataset (Siamese-I) coupled with another Siamese network dedicated for their soft biometrics dataset (Siamese-S). In summary, the DNN architectures which are used for classification and verification during the training and testing stages are shown in Figs. 4.2, 4.3, 4.5, and 4.8.

4.3.3 Learning using privileged information via a multi-task learning architecture

Our goal is to simultaneously train the network layers using the face images (hard biometrics) and the face attributes (soft biometric traits). In this approach, instead of considering two different networks dedicated for the main data and the privileged information during the training phase, instead we consider to train them both simultaneously using only one network. However, considering the privileged information as an input to the network is not a good idea since the network does not have an access to such an input during the testing phase. In the proposed MTL architecture (see Fig. 4.6) the network gets the main data (face images) as an input and tries to estimate the privilege information in addition to the main task (i.e., classification/verification). Therefore, the network is jointly trained for both the face recognition task as well as the face attributes prediction tasks. Suppose the input is an image $x_i \in X$, and the output is its class label $y_i \in Y$ for $i = 1, \ldots, N$ where N is the number of training samples. Also, the privileged soft biometric traits, contain T different face attributes or class labels. Thus, in this framework we denote them as y^t for $t = 1, \ldots, T$. The loss function is defined as below:

$$L_6 = 1/N \sum_{i=1}^{N} (\ell(f_L(x_i), y_i) + 1/N \sum_{i=1}^{N} \sum_{t=1}^{T} (\ell(f^t(x_i), y_i^t) \ , \tag{4.9}$$

where $f^t(.)$ is the specific function for attribute prediction of task t. Note that the second term of the loss function is added because of the privileged soft biometric traits. However, learning multiple CNNs separately is not optimal since different tasks might have some hidden relationships with each other and may share some common features. This is supported by [150] where they train a CNN features for the face recognition task and they use it directly for the face attribute estimation. Therefore, our network also shares CNN features among different tasks in order to enhance the performance of the face recognition task. The loss function (4.9) can now be reformulated as follows:

$$L_7 = 1/N \sum_{i=1}^{N} (\ell(f_L(z(x_i, w_c) \times w_L), y_i) + 1/N \sum_{i=1}^{N} \sum_{t=1}^{T} (\ell(f^t(z(x_i, w_c) \times w^t), y_i^t) \ , \tag{4.10}$$

where w_c is the shared weights between all the tasks including the face recognition task. w_L and w^t for $t = 1, \ldots, T$ represent the remaining weights which are assigned separately for the main task and the privileged soft biometric tasks, respectively (see Fig. 4.6). Our MTL-LUPI framework is applied for the tasks of classification and verification.

While the whole approach remains the same for the classification and verification tasks, but their network structures are different. In the classification task, $\ell(f_L(.))$ would be replaced by a cross entropy loss function. However, in the verification task, we use a Siamese network [62] for the primary face dataset (Siamese-I). Each attribute predictor of the Siamese-I is trained for a particular privileged task as shown in Fig. 4.7. Therefore, the total loss for MTL-LUPI in the case of verification would be:

$$\begin{aligned} L_8 = 1/N^2 \sum_{i=1}^{N} \sum_{j=1}^{N} (\ell_{cont}(z(x_i, w_c), z(x_j, w_c), y_{cont}) \\ + \sum_{t=1}^{T} (\ell(f_i^t(z(x_i, w_c) \times w^t), y_i^t) + \sum_{t=1}^{T} (\ell(f_j^t(z(x_j, w_c) \times w^t), y_j^t)) \ , \end{aligned} \tag{4.11}$$

where (x_i, x_j) is a pair of face images and ℓ_{cont} is a loss function for the verification task with y_{cont} label.

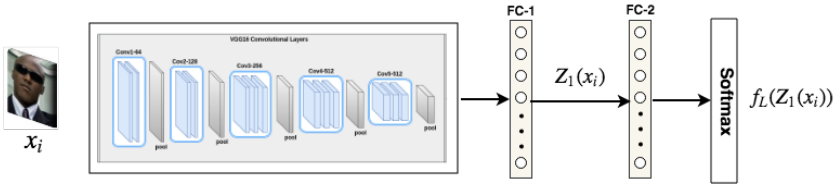


Figure 4.5: Classification framework during the testing phase.

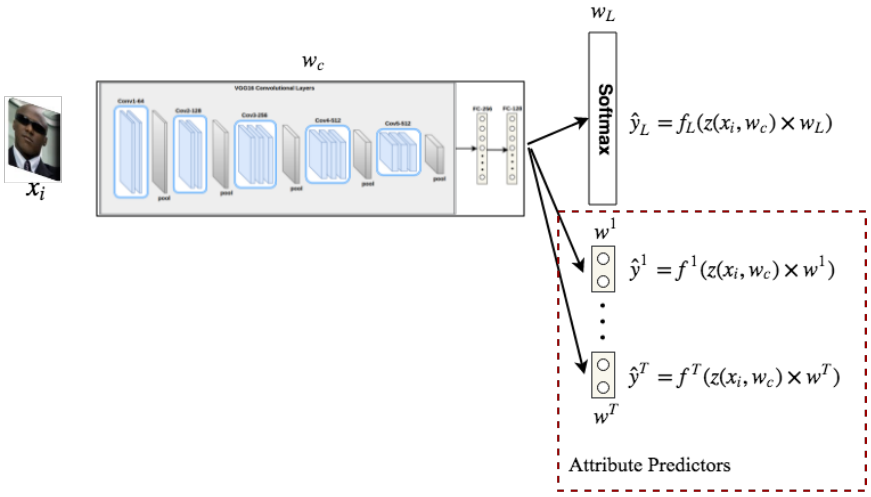


Figure 4.6: Classification using face images (primary data) in MTL-LUPI framework during the training phase.

$t = 1, \ldots, T$ indicate the different tasks which are exploited from the privileged soft biometric traits. For simplicity we consider the same loss function for all the attribute tasks but this can be generalized for tasks to use different loss functions. y_i^t, y_j^t are the real class labels (privileged information) and $f_i^t(.), f_j^t(.)$ are the predicted class labels of the (x_i, x_j) pair for the task t. w_c is the common weights for the verification and attribute tasks. w_i^t and w_j^t represent the remaining weights which are assigned separately for the privileged soft biometric task t dedicated to the input face images x_i and x_j, respectively (see Fig. 4.7).

4.4 Experiments

We divided the experiments into two parts, namely classification and verification tasks. For the classification task, we compared our results with two state-of-the-art LUPI approaches, i.e., SVM+ and LMIBPI [92], on four different benchmark datasets, namely Morph [165], Biocop [159], CelebA [156], and CMU Multi-PIE [166]. We consider two different approaches to utilize the privileged information (soft biometrcis) for the classification task, namely Cp-DNN (Net-I+Net-S) and MTL-LUPI. Similarly, for the verification task,

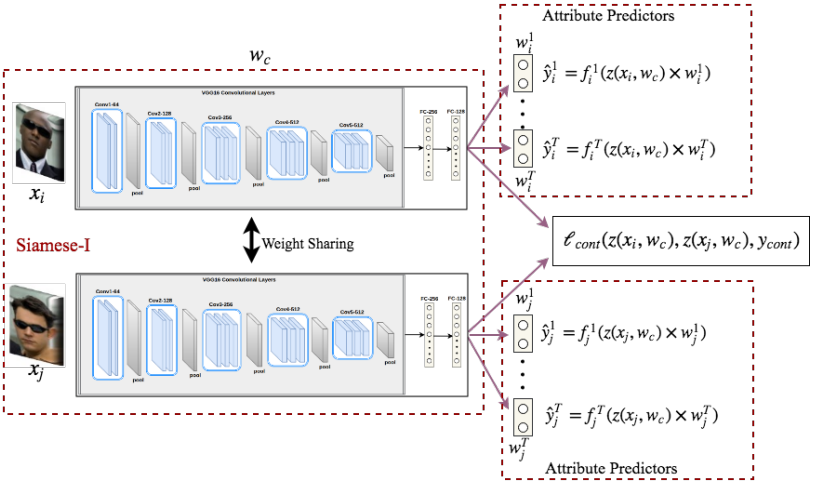


Figure 4.7: Verification using face images in MTL-LUPI framework during the training phase.

we evaluate the effect of using the privileged soft biometrics in two different LUPI frameworks namely Cp-DNN (Siamese-I+Siamese-S), and MTL-LUPI. We explain each framework separately and eventually we compare the results with each other and the base-model. The experiments are conducted using NVidia Titan GPU (12 GB).

CelebA- The images of this dataset come from the Celeb-Face dataset [65] along with their attribute vectors of 40 attributes such as gender, face characteristics, skin color, hair color, etc. CelebA dataset contains $200K$ images of $10K$ different identities. Since in the CelebA dataset the identities of the images are not provided, we consider the gender attribute, i.e., male or female, as the image label and the remaining 39 features as the soft biometric data. We use this dataset for gender verification and classification.

Morph dataset [192] consists of two albums I, II. Album I contains more than $1K$ face images and album II has more than $55K$ images, respectively. Each face image has three soft biometrics traits namely: age, gender, and race. The age in the dataset is distributed in the range of 15-68. In this chapter, we consider 8 different classes to cover the age range since the main goal is not age estimation. In total the dataset images are related to more than $10K$ subjects. There are 103 subjects with more than 15 samples, which enables us to use this dataset for classification and verification tasks.

CMU Multi-PIE dataset [166] consists of 337 identities, from 15 different view points and 19 illumination conditions. This dataset contains $750K$ images collected in four different sessions. We use frontal view point of the face images as the main data and 45 degree point view as the auxiliary data. This dataset is utilized only for the sake of classification task.

Dataset	fc1	fc2	softmax
Morph	256	128	103
Biocop	512	512	400
CelebA	256	128	2
CMU Multi-PIE	256	256	129

Table 4.1: Output sizes of the fc layers in Net-I for the Cp-DNN framework

Biocop dataset [159] is collected under four disjoint years; 2008, 2009, 2012 and 2013. Each label consists of different biometric modalities for each subject; face, iris, fingerprint, palm print, hand geometry, voice and soft biometric traits including age, gender, ethnicity, hair color, eye color, beard, mustache, weight, and height.

Under each label, the biometrics are acquired during either one or two separate sessions. The 2012 and 2013 databases contain $2K$ subjects [159]. For the task of classification and verification we used the face images of 2012 and 2013 subjects with their corresponding soft biometric traits. We randomly select 400 subjects from the dataset and perform the classification task on them. In the verification task, we construct genuine and impostor pairs from 600 subjects and test the framework on the remaining subjects.

4.4.1 Classification task using coupling framework (Cp-DNN: Net-I+Net-S)

Our coupled deep architecture is applied to all four datasets. We use different architectures for the main data (face images) and the privileged data (attributes). The network for face images (Net-I) composed of 13 convolutional layers of VGG-16 [132] (see Fig. 4.1), pre-trained on the CMU Multi-PIE dataset, followed by three fully connected layers with different output sizes. All the convolutional and fully connected layers, except the last fully connected layer, are equipped with the *Relu* activation function. For the sake of classification, *softmax* activation is used for the last fully connected layer. Table 4.1 shows the size of the fully connected layers as well as the softmax layer for different datasets. The network for the soft biometric data (Net-S) consists of three fully connected layers. The size for each fully connected layer in Net-S is shown in Table 4.2. The first fully connected layers of the Net-I and Net-S are coupled together via the contrastive loss. The network structure is depicted in Fig. 4.2. However, for the CMU Multi-PIE dataset, the second network dedicated to the auxiliary data has the same structure as the first one since the privileged data are the 45 degree view face images.

Implementation details. We randomly divided all the datasets into three parts, i.e., training (%70), validation (%15) and testing (%15). For training, the number of genuine and impostor pairs are very high,

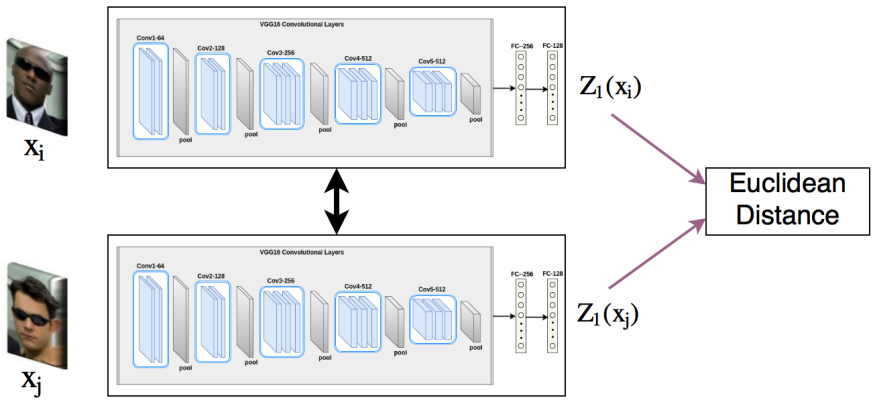


Figure 4.8: Verification framework during the testing phase.

therefore, we randomly selected $20K$ genuine pairs and $20K$ impostor pairs. Each pair contains five entries, $[x_i, x_j^p, y_i, y_j^p, y_{cont}]$ where x_i is the input of the Net-I (cropped face) with the class label y_i, x_j^p is the input of the Net-S (attributes) with the class label y_j^p, and y_{cont} is the binary label which is 0 when x_i and x_j^p belong to the same class, and it is equal to 1 otherwise. For the purpose of validation, we randomly selected $7.5K$ impostor and $7.5K$ genuine pairs. Similarly, we randomly selected $7.5K$ impostor and $7.5K$ genuine pairs for the testing phase. During the validation/testing of a probe the soft biometric data is not available, therefore, we only use Net-I as illustrated in Fig. 4.5. Therefore, each probe sample in validation and testing only contains two entries, $[x_i, y_i]$, in which y_i is used for evaluating the classification accuracy. We also consider a base-model when there is no soft biometric data available during the training phase. In this case, we only have Net-I in the training, validation, and testing phases. By comparing the results of our model with the base-model, we can evaluate whether the privileged soft biometric data (only available during the training) would improve the classification accuracy or not and if so what is the percentage in improvement. In addition to the base-model, we also compared our results with two state-of-the-art LUPI approaches, namely SVM+ and LMIBPI. Since these approaches are binary classifiers, we compared the results of the gender classification task on CelebA with them. The SVM+ and LMIBPI are trained using the features extracted from the first fully connected layer of the Net-I (in the base-model case) as the main input, and the soft biometric vectors as the auxiliary input. For the Morph, Biocop, and CMU Multi-PIE dataset, we constructed a multi-class SVM+, and a multi-class LMIBPI by training 103, 400, and 129 one-versus-all binary SVM+, and LMIBPI classifiers, respectively. We also compare our results with the classical SVM which could also be considered as a base-classifier.

Dataset	fc1	fc2	softmax
Morph	256	128	103
Biocop	512	512	400
CelebA	256	128	2

Table 4.2: Output sizes of the fc layers in Net-S for the Cp-DNN framework

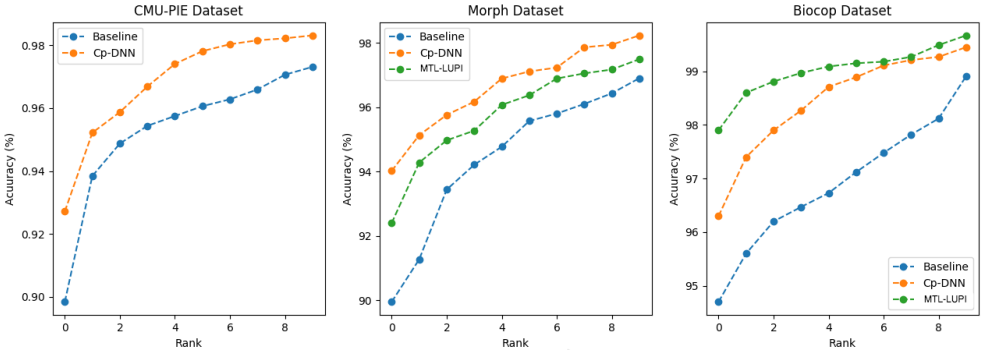


Figure 4.9: Overall cumulative matching characteristics (CMC) curves for CMU Multi-PIE, Morph, and Biocop datasets.

4.4.2 Classification task using MTL framework (MTL-LUPI)

Our MTL-LUPI Framework composed of 13 convolutional layers of VGG-16 (see Fig. 4.1), pre-trained on the CMU Multi-PIE dataset, followed by two fully connected layers shared by all the tasks and finally a separate fully connected layer for each individual task (see Fig. 4.6). The size of the last fully connected layer for each task depends on the number of classes within each attribute. For example, in the binary attribute classification for gender the size of the fully connected layer is two, but in other tasks such as race or hair color the size would be more than two. All the convolutional and fully connected layers, except the last fully connected layer, are equipped with $Relu$ activation function. For the last fully connected, $softmax$ was considered for the classification of each individual task. Table 4.3, shows the size of the shared fully connected layers and the last fully connected layer for each task on different datasets. Depending on the dataset, the types of the available soft biometrics are variant. Also, the number of classes for the same attribute can be different depending on the dataset. For example, in the Biocop dataset, hair color has six classes while it has only three different classes in the CelebA dataset.

Dataset	fc1	fc2	main task	Age	Gender	Race	Eye color	Hair color	Beard	Mustache	Eye glasses
Morph	256	128	103	8	2	2	✗	✗	✗	✗	✗
Biocop	512	512	400	7	2	7	6	6	2	2	✗
CelebA	256	128	2	✗	2	✗	✗	3	2	2	2

Table 4.3: size of shared fully connected layers and the last fully connected layer for different tasks (MTL-LUPI framework)

	Base-classifier		LUPI			
Dataset	SVM	Net-I	SVM+	LMIBPI	Net-I + Net-S	MTL-LUPI
Morph	73.45	89.95	77.14	75.69	**94.02**	92.4
Biocop	89.65	94.7	92.14	92.78	96.3	**97.9**
CelebA	95.82	99.2	95.61	95.54	99.6	**99.7**
CMU Multi-PIE	82.31	89.6	84.68	86.43	**92.7**	-

Table 4.4: Classification results of our proposed methods (Cp-DNN:Net-I+Net-S and MTL-LUPI), our baseline method (Net-I), LMIBPI, SVM, and SVM+ on Morph, Biocop, CelebA, and CMU Multi-PIE datasets.

Implementation details. We randomly divided all the datasets into three parts, i.e., training (%70), validation (%15) and testing (%15). Each input sample to the network contains three entries, $[x_i, y_i, y^t]$ where x_i is the input cropped face with the class label y_i, and y^t is the class label for the privileged task t. During the validation, and testing phases of a probe, the soft biometric data is not available. Moreover, in the testing phase, the network would be able to do the attribute prediction as well as the classification task. For the sake of comparison, we also consider a base-model when there is no MTL framework. It should be noted that the base-model for this framework is the same as the base-model for the coupling architecture (Cp-DNN) and we only have Net-I during the training, validation, and testing phases.

4.4.3 Classification results

Table 4.4, lists the classification accuracy results. For all the four datasets, both LUPI-based Cp-DNN and MTL-LUPI show better performance compared to SVM+ and LMIBPI. In addition, they outperform the base-model (Net-I), highlighting the usefulness of the privileged information in the proposed frameworks. For the CelebA dataset, Net-I already shows a very high classification accuracy. Therefore, even a little improvement is worthy. In the Cp-DNN framwork, privileged data helped to improve the classification accuracy by 4.5%, 1.7%, and 5.8% for Morph, Biocop, and CMU Multi-PIE datasets, respectively. When compared to Net-I. MTL-LUPI enhanced the classification accuracy by 2.7% and 3.4% in the Morph and Biocop datasets, respectively, by utilizing the privileged information. It should be noted that MTL-LUPI cannot be applied to the CMU Multi-PIE dataset since this dataset does not have the soft biometric traits and the privilege information is just the face images at 45 degree. Furthermore, Net-I has better performance compared to SVM+ and LMIBPI, highlighting the superiority of our deep model over the shallow models.

The evaluation performance is validated using ten fold random cross validation. Fig. 4.9 shows the overall cumulative matching characteristics (CMC) curves of our proposed methods and the baseline model for the Morph, Biocop, and CMU Multi-PIE datasets.

	Base-model	LUPI	
Dataset	Siamese-I	Siamese-I + Siamese-S	MTL-LUPI
Morph	87.37	90.66	**91.53**
Biocop	87.6	88.27	**90.07**
CelebA	96.1	96.8	**97.3**

Table 4.5: Verification results for the proposed methods (Cp-DNN: Siamese-I+Siamese-S and MTL-LUPI) and baseline model on Morph, Biocop, and CelebA datasets.

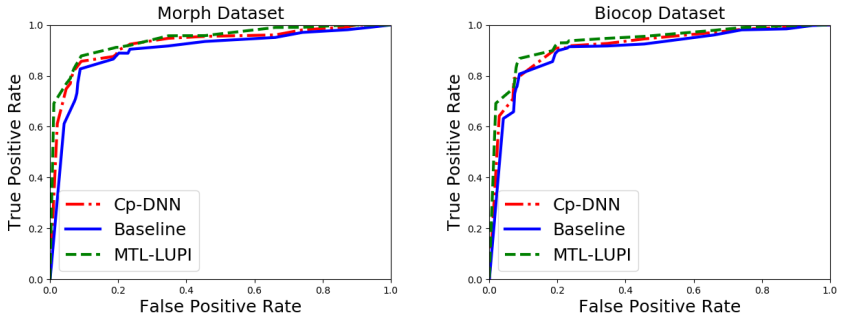


Figure 4.10: ROC curves for verification results for our proposed methods (Cp-DNN:Siamese-I+Siamese-S and MTL-LUPI), and baseline model (Siamese-I) on Morph, and Biocop datasets.

4.4.4 Verification task using coupling framework (Cp-DNN: Siamese-I+Siamese-S)

For the verification task, we train two Siamese DNNs that are coupled together. We utilize a Siamese structure, which means its underlying networks share their weights with each other. Each Siamese network consists of two identical VGG-16 nets when the first 13 convolutional layers of the VGG-16 nets are pre-trained on the CMU Multi-PIE dataset, followed by 2 fully connected layers, which have 256 and 128 nodes, respectively. The whole structure is displayed in Fig 4.3. The output of Siamese-I network acts as the embedding function $z_1(x_i, x_j)$. Since the input of this Siamese network are the cropped face images, we refer to it as Siamese-I. For $z_2(x_i^p, x_j^p)$, we use an MLP-based Siamese structure in which each MLP contains two fully connected layers, each of which has 128 nodes. Since the input of this Siamese network is the soft biometrics data, it is denoted by Siamese-S in this chapter. The output of the upper branch in Siamese-I is coupled with the output of the lower branch in Siamese-S by the contrastive loss, L_{cont3}, as shown in Fig. 4.3. Since the weights in Siamese-I and Siamese-S are shared, it does not matter which branches in Siamese networks are coupled.

Implementation details. The same network architecture is used for all the datasets (Morph, Bio-Cup, and CelebA). We randomly divided the datasets into three parts, i.e., training (%70), validation (%15)

and testing (%15). For the training part, the number of genuine and impostor pairs are very high, therefore, we randomly selected $20K$ genuine and $20K$ impostor pairs. Each pair contains seven entries, $[x_i, x_j, x_i^p, x_j^p, y_{cont1}, y_{cont2}, y_{cont3}]$ where x_i and x_j are the inputs to the Siamese-I, x_i^p and x_j^p are the inputs to the Siamese-S, and y_{cont1}, y_{cont2}, and y_{cont3} are binary labels which are 0 when the inputs in (x_i, x_j), (x_i^p, x_j^p), and (x_i, x_j^p) come from the same class, respectively, and are equal to 1 otherwise. For validation phase, due to the high number of pairs we randomly selected $7.5K$ impostor and $7.5K$ genuine pairs. Similar to the validation phase, we randomly selected $7.5K$ impostor and $7.5K$ genuine pairs for the testing phase. During the validation and testing phases the soft biometric data is not available. Therefore, we only use Siamese-I in these phases (see Fig. 4.8). Thus, each pair in the validation and testing phases contains three entries, $[x_i, x_j, y_{cont1}]$. We also consider the case when there is no soft biometric data in the training. In this case, we only have Siamese-I in all the training, validation, and testing phases. Similar to the classification experiments, the results of our model is compared with the results of the base-model to evaluate the effectiveness of the privileged soft biometric data on the verification performance improvement.

4.4.5 Verification task using MTL framework (MTL-LUPI)

For the verification task in MTL-LUPI framework, we train a Siamese DNN consisting of two identical VGG-16 nets when the first 13 convolutional layers of the VGG-16 nets are pre-trained on the CMU Multi-PIE dataset, followed by two fully connected layers shared between all the tasks and finally at the last layer we have a bank of fully connected layers each dedicated for a different task. The size of the two shared fully connected layers for all the datasets are 256 and 128 respectively. The whole structure is displayed in Fig 4.7. The size of the last fully connected depends on number of tasks' classes. All the common weights of different tasks are shared among two VGG-16 nets in the Siamese network. The last fully connected layers which are dedicated for each task are not shared between the two VGG-16 nets in order to help the network to estimate each task separately, by utilizing the shallow subnetworks at the end. All the convolutional and fully connected layers, except the last fully connected layer, are equipped with $Relu$ activation function. The last fully connected $softmax$ was considered for the classification of each individual task.

Implementation details. The same network architecture is used in all the datasets (Morph, Bio-Cup, and CelebA). We randomly divided the datasets into three parts, i.e., training (%70), validation (%15) and testing (%15). For the training part, we randomly selected $20K$ genuine and $20K$ impostor pairs. Each pair contains five entries, $[x_i, x_j, y_{cont}, y_i^t, y_j^t]$ where x_i and x_j are the inputs to the network, and y_{cont} is a binary

label which is 0 when the inputs x_i and x_j come from the same class and is equal to 1 otherwise. y_i^t, and y_j^t are the class labels for the privileged task t related to x_i, and x_j, respectively.

For validation phase, we randomly selected $7.5K$ impostor and $7.5K$ genuine pairs. Similar to validation, we randomly selected $7.5K$ impostor and $7.5K$ genuine pairs for the testing phase. During the validation and testing phases the soft biometric data is not available. Therefore, we only use the Siamese network to do the verification task in these phases although it can be used to predict the attributes if needed. Thus, each pair in the validation and testing phases contains three entries, $[x_i, x_j, y_{cont}]$. We compared the result of this architecture with the Cp-DNN framework and the base-model.

4.4.6 Verification results

The verification results for the Morph, Biocop, and CelebA datasets are listed in Table 4.5. The results show that both proposed methods exploit the privileged soft biometrics to improve the performance of verification task on all the datasets. Cp-DNN framework (Siamise-I+Siamese-S) could enhance the Morph, Biocop, and CelebA datasets by 3.8%, 0.7%, and 0.7%, respectively. On the other hand, the MTL-LUPI framework generally shows a better performance compared to the Cp-DNN framework and enhanced the verification results of the base-model by 4.7%, 2.8%, and 1.2% on the Morph, Biocop, and CelebA datasets, respectively. The ROC curves of verification results on the Morph and Biocop datasets are shown in Fig 5.4. The same as the classification experiments, the evaluation performance is validated using ten fold random cross validation. The results indicate that the Cp-DNN framework did not significantly improve the performance of the baseline network on the Biocop and CelebA datasets while the MTL-LUPI had slightly better performance in all the datasets.

4.5 Conclusion

We have introduced a new approach to exploit additional information in the form of soft biometrics during the training phase for the purpose of face recognition performance enhancement. In this work, two different architectures Cp-DNN and MTL-LUPI were employed in the training phase. In Cp-DNN a coupled DNN architecture is employed in the training phase, which discovers a discriminative latent common subspace utilizing soft biometrics and RGB images. On the other hand, the MTL-LUPI approach considers the privileged information as extra tasks which the network tries to estimate simultaneously with the main

task using the image data. The proposed approaches have been evaluated on two different tasks of face classification and verification. We have compared our method with two state-of-the-art LUPI methods in the literature for the task of face classification and showed the superiority of our method over them. Moreover, to the best of our knowledge, the privileged information has not been used in the LUPI framework for the verification task. We also compared our two LUPI-based algorithms with each other and with the base-model where there is no privileged information. During the inference time, the architecture degenerates to the baseline for both frameworks. Therefore the inference time for LUPI-based algorithm is the same as the baseline. While both methods showed relatively the same amount of improvement in the face classification task performance depending on the datasets, the MTL-LUPI algorithm showed a better results in the verification task compared to the Cp-DNN which is one direction for further studies. Also, the results have revealed that the soft biometrics information could be more helpful for the task of classification compared to the verification which still needs to be investigated more in the future works.

Chapter 5

Attribute Adaptive Margin Softmax Loss using Privileged Information

5.1 Introduction

A multimodal recognition system with multiple modalities, such as the face, fingerprint, and iris, is expected to be more reliable and accurate due to the utilization of different sources of information. However, acquisition of this information is costly and a tedious task which can affect the popularity and ease of using multimodal recognition systems. On the other hand, there are some informative traits, such as age, gender, ethnicity, race, and hair color, which are not distinctive enough for the sake of recognition, but still can act as complementary information to other primary information, such as face and fingerprint. These traits, which are known as soft biometrics, can improve recognition algorithms performance.

The design of a recognition system comprises two major phases, namely training and testing. However, in some cases, there are extra information which is only available during the training phase and is missing during the testing phase. In other words, the training data is augmented with some extra auxiliary information. For example, in object recognition, the labeled images may be annotated with texts which can provide semantic information about the object, or any other extra knowledge, such as the boundary information of an object which determines the exact location of a specific object [193]. This extra information can be regarded as an auxiliary to the primary modality of the data. Unlike the domain adaptation and transfer learning problems in which the data is similar in both the source and target domains but statistically different [194, 195], here, the available data in the source domain has an extra modality which is not available in the target domain.

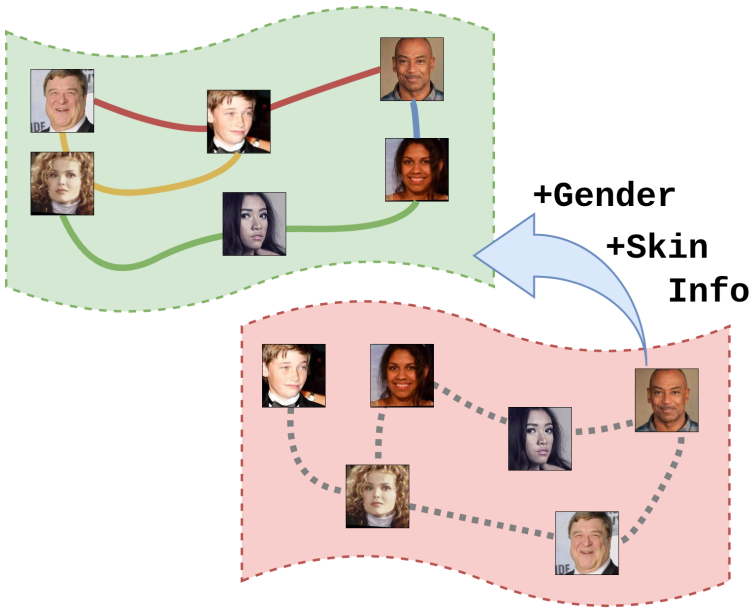


Figure 5.1: The proposed framework employs attributes information to improve the semantic correlation of the faces in the embedding space. Top and bottom figures illustrate the deep embedding of faces using the proposed and conventional methods, respectively.

The concept of learning using privileged information (LUPI) was introduced in [196], for the first time, under the context of SVM+. In their methodology, the auxiliary data, or privileged information, was employed to predict the slack variables of an SVM classifier [197]. Since then, the idea of using privileged information has been vastly investigated in the literature and has been applied to different applications and contexts, e.g., object localization [198], facial feature detection [199], and metric learning [200]. A framework was introduced in [193] to embed the main data into the latent feature subspace such that the mutual information between the embedded data and auxiliary data is maximized.

Softmax loss is widely used in training CNN features [201], which is specified as a combination of the last fully connected layer, a softmax function and a crossentropy loss [202]. However, features through softmax loss are learned with limited discriminative power. To address the limitation, various supervision objectives have been proposed to enhance the discriminativeness of the learned features, such as contrastive loss [203], triplet loss [204], center loss [205]. In contrast to most of the other loss functions which use Euclidean margin, [206, 207, 208] showed the effectiveness of angular margin to squeeze each class. However, these methods have an implicit hypothesis that all the classes have sufficient samples to describe their distributions, so that a constant margin is enough to equally squeeze each intra-class variations. However this is not the case in many public unbalanced datasets.

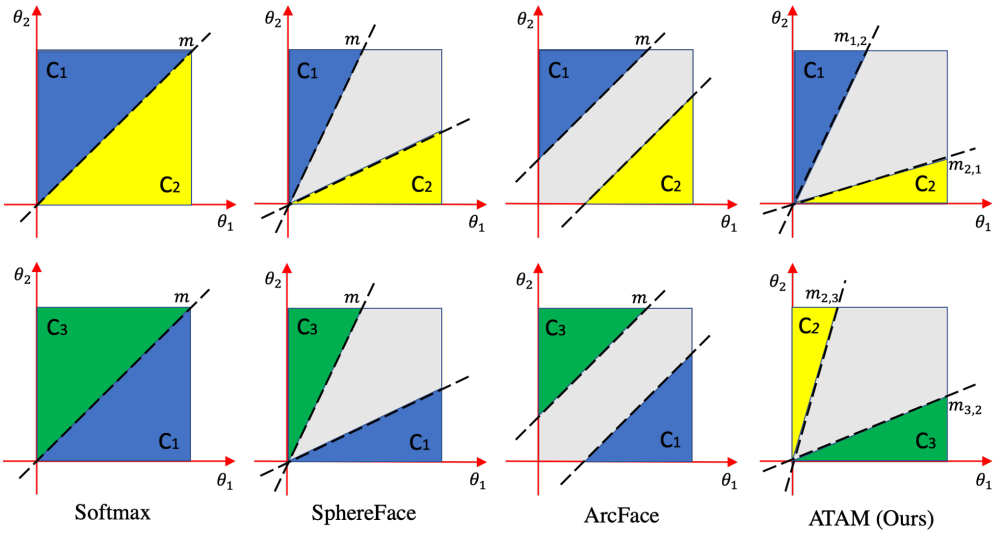


Figure 5.2: Decision margins of different loss functions for three different classes C1, C2, and C3 (in blue, yellow, and green, respectively). The dashed line represents the decision boundary, and the grey areas are the decision margins.

Our proposed training scheme has the ability to transfer information from the auxiliary data and help the network to learn a more discriminative features regarding the primary data (see Fig. 5.1 (b)). To this end, we propose a novel loss function, Attribute Adaptive Margin Softmax Loss (ATAM), to adaptively find the appropriate margins utilizing attributes during training phase. Specifically, we make the margin m particular and learnable and directly train the network to find the adaptive margins. We show its important applications to face recognition and person re-identification. Note such tasks can be assessed under either closed- or open-set protocol. The open-set protocol is harder since the testing classes may be unseen from the training classes. It usually requires discriminative feature representations with built-in large margins, which are embodied in our approach.

Our method is motivated by the observation that the minority classes often encompasses very few samples with high degree of visual variability. The scarcity and high variability makes the neighborhood of these samples easy to be invaded by other imposter nearest neighbors. To this end, we propose to learn an embedding utilizing attributes along with a novel loss function to ameliorate the invasion. For the purpose of assessment, we study the problem of face recognition and person re-identification using paired image-attribute data, while the attributes (i.e., soft biometrics) are only available during the training phase (see Fig. 5.3).

The major contributions of this chapter are: i) rewriting the angular softmax loss using a set of inter-class

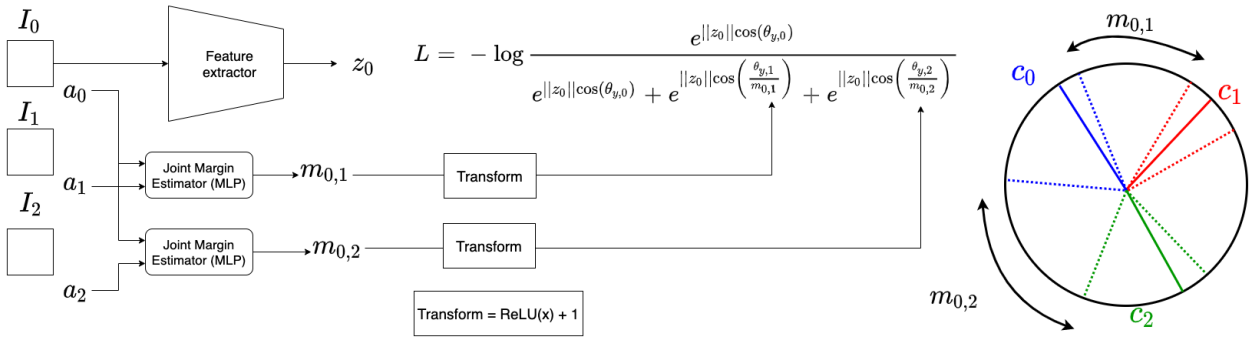


Figure 5.3: Proposed ATAM loss utilizing privileged attributes.

margins, ii) proposing a framework enhanced by auxiliary information to learn these margins simultaneously during the training, iii) providing an example for the auxiliary information by means of the discrepancy between the attributes. The performance of the proposed scheme is evaluated on five different datasets, namely MegaFace [209], YTF [210], LFW [3], Market-1501 [211], and DukeMTMC-reID [212] datasets. On both of the mentioned tasks, we demonstrate the superiority of ATAM loss with performance on par with the state of the art.

5.2 Methodology

In this section we detail our methodology [213]. First we revisit the A-Softmax [206] which maps faces on the hyperspace manifold. Then we present our proposed attribute adaptive margin Softmax loss which exploits the attributes to learn more discriminative feature space.

5.2.1 A-Softmax

Lets begin with the most widely used Softmax loss. The Softmax loss maximizes the posterior probability of each class to separate features of different classes. Its formulation is provided as follows:

$$L_s = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{W_{y_i}^T z_i + b_{y_i}}}{\sum_{j=1}^{M} e^{W_j^T z_i + b_j}}, \tag{5.1}$$

where $W_j \in \mathbb{R}^d$ is the weights of last layer of class j with d dimension and $b_j \in \mathbb{R}$ is the bias term. $z_i \in \mathbb{R}^d$ is the learned feature of sample i, and y_i is the ground truth class label. N and M are the number of samples and classes, respectively. The inherent angular distribution of learned deep features by Softmax loss

suggests using cosine distance as the metric instead of using Euclidean distance [206]. Modified Softmax loss normalizes the weights $||W_i||= 1$ and zero the biases as follows:

$$L_m = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{||z_i||\cos(\theta_{y_i,i})}}{\sum_{j=1}^{M}e^{||z_i||\cos(\theta_{j,i})}}. \tag{5.2}$$

Given a query point, the model compares its angle with weights of different classes and select the one with minimum angle. Although the features learned using modified Softmax have angular boundary, they are not necessarily discriminative enough. SphereFace [206] proposed a natural way to produce the angular margins through an A-Softmax loss. The angle between a query point and target class is multiplied by the margin parameter m:

$$L_{a-s} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{||z_i||\psi(\theta_{y_i,i})}}{e^{||z_i||\psi(\theta_{y_i,i})}+\sum_{j\neq y_i}e^{||z_i||\cos(\theta_{j,i})}}\ , \tag{5.3}$$

where $\psi(\theta_{y_i}, i)$ is a monotonically decreasing angle function and defined as $(-1)^k\cos(m\theta_{y_i}, i) - 2k$, and $\theta_{y_i}, i \in [\frac{k\pi}{m}, \frac{(k+1)\pi}{m}]$, $k \in [0, m-1]$ to compensate for the limitation of θ_{y_i}, i in $\cos(m\theta_{y_i}, i)$. While A-Softmax loss manually tune m to squeeze the intra-class variation and increase the angular distance between different classes, it fails to consider the feature distribution in the hyperspace manifold. Other works such as CosFace [207] and ArcFace [208] also have similar assumption and consider same distributions for different classes. Thus, they fail to exploit the holistic feature space. [214, 215] tried to address this limitation by designing a new loss function and constructing a weighted combination with A-Softmax loss function.

Here, we propose a natural way to learn and tune m in an end-to-end fashion in order to make the the holistic feature space more discriminative and consider the inter-class space between different classes via their attributes.

5.2.2 Attribute Adaptive Margin Softmax Loss

Given an Image I and its specific attribute set a, we propose to learn adaptive margins which discriminate the holistic feature space and reflects the characteristics of corresponding attribute in the image. If there are k attributes $\frac{k\times(k-1)}{2}$ margins can be learned in the feature space. The proposed model architecture is composed of a feature extraction branch (CNN) combined with a small network from attributes. The attribute network is a MLP network which is responsible for learning margins. Concretely, attribute sets a_j, a_{y_i} (each of which has dimensionality of k) related to the specific samples from classes j, y_i, respec-

tively, are fused with simple concatenation and further fed into three subsequent FC layers (MLP network) to obtain the specific margin (m_{j,y_i} or $m_{y_i,j}$).

The margin m is usually set manually in SphereFace [206], and also kept constant during the training phase. In order to address the problem mentioned above and have a better holistic feature space, we propose to have a learnable attribute-specific parameter m. The Eq. 5.3 can be modified as follows:

$$L_{ATAM} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{||z_i||\cos(\theta_{y_i,i})}}{e^{||z_i||\cos(\theta_{y_i,i})} + \sum_{j\neq y_i} e^{||z_i||\cos(\frac{\theta_{j,i}}{m_{j,y_i}})}}, \tag{5.4}$$

where m_{j,y_i} is the score that is provided by the attribute network (MLP network) with the input $[a_j, a_{y_i}]$ (where $[,]$ is the concatenation operation). Note that since we want the margins to be greater or equal to 1 ($m_{y_i,j} \geq 1$), we normalize the output of MLP network using $ReLU$ function. Afterwards, this positive score is added to 1 (see Fig 5.2). This transformation ($ReLU(m_{j,y_i}) + 1$) ensures that $m_{j,y_i} = 1$ in the worst case situation. Both of the CNN and MLP branches are trained jointly using Eq. 5.4 in an end-to-end fashion.

5.2.3 Discussion

In this subsection we compare our proposed ATAM loss with Softmax, A-Softmax [206], and ArcFace [208] as illustrated in Fig. 5.1 (a). For simplicity of analysis, we consider three classes of C1, C2, and C3.

Softmax decision boundary depends on both magnitudes of weight vectors and cosine of angles, which results in an overlapping decision area (margin ¡ 0) in the cosine space. The popular A-Softmax reduces intra-class variations and allocates equal space for each class, without considering its sample distribution. In contrast to A-Softmax which has a nonlinear angular margin, ArcFace has a constant linear angular margin throughout the whole interval. However, it still fails to capture the true distribution of each class in the holistic feature space leading to not accurate margins between classes.

We introduce learnable attribute-specific margins ($m_{1,2}$, $m_{1,3}$, and $m_{2,3}$ in Fig. 5.1 (a)) by utilizing the privileged attribute information during the training phase. This loss intrinsically takes attributes into account in a unified loss function leading to the more discriminative holistic feature space. The attribute-specific margins discriminate embedding domain into different clusters and each cluster may have one or

Method	Protocol	MF1 Rank1	MF1 Veri.
Beijing FaceAll Norm 1600	Large	64.80	67.11
Google - FaceNet v8 [204]	Large	70.49	86.47
NTechLAB - facenx large	Large	73.30	85.08
SIATMMLAB TencentVision	Large	74.20	87.27
DeepSense V2	Large	81.29	95.99
YouTu Lab	Large	83.29	91.34
Vocord - deepVo V3	Large	91.76	94.96
CosFace [207]	Large	82.72	96.65
UniformFace [214]	Large	79.98	95.36
AdaptiveFace [215]	Large	95.02	95.61
Softmax	Large	71.37	73.05
SphereFace [206]	Large	92.24	93.42
CosFace [207]	Large	93.94	94.11
ArcFace [208]	Large	94.64	94.85
ATAM	Large	96.51	97.14

Table 5.1: Face identification and verification evaluation on MF1. "Rank 1" refers to rank-1 face identification accuracy and "Veri." refers to face verification TAR under 10^{-6} FAR.

more classes. This also embraces the multimodality of class distribution: it preserves not only locality across the same-class with different attributes but also discrimination between classes. Hence, it is capable of preserving discrimination in any local neighborhood, and forming local class boundaries with the most discriminative samples with regards to the attributes (More distant classes contain more different attributes: class C2 (yellow) is placed between class C1 (blue) and class C3 (green) due to the more discrepancy between the attributes of C1 and C3 compared to C1 and C2 in Fig. 5.1 (a)).

UniformFace [214] and AdaptiveFace [215] also try to make the embedding space more efficient and discriminative. UniformFace address this by adding another loss function to A-Softmax loss with another hyperparameter in a similar fashion to old joint loss functions (i.e. different combination of contrastive loss or center loss with softmax). It disperses classes in the embedding domain uniformly which does not reflect the true underlying distributions of classes. Adaptive loss also tackles this issue by adjusting margin m adaptively. However, it also introduce extra hyperparameter λ to jointly train additional loss with the modified A-Softmax. The additional loss function is taking the average of all classes' margins leading to overlooking specific margin for each class.

The proposed loss utilizes the privileged information to learn the margins that are usually set manually in other loss functions such as cosFace loss. Our method enhances the training of the feature extractor network in two ways. First, instead of defining a global margin that is constant for all the classes, it defines k(k-1)/2 margins for the pairs of classes with different attribute information. Second, we incorporate the attribute

information to adaptively control the inter-class margins and regularize the distribution of the features in the embeddings. Hence, in contrast to the previous approaches that impose a global prior for the separability of the class distributions, ATAM enables the training framework to carefully exploit the auxiliary information to learn the distribution of the features based on a set of adaptive priors constructed using the local properties of inter-class relationships, i.e., relative discrepancy of attributes. After the training, we solely use the feature extractor network for the face recognition/ re-id.

5.3 Experiments

In this section we conduct extensive experiments on five datasets to demonstrate the effectiveness of our proposed method. First, we describe the implementation setup of the proposed model in subsection 5.3.1. Eventually, we compare the proposed ATAM to other baselines for two tasks of face recognition and person re-identification in subsections 5.3.2 and 5.3.3, respectively.

5.3.1 Implementation Details

We adopt ResNet [216] as the base network of our proposed model. We performed standard preprocessing on faces. MTCNN [217] is used to detect and align each face via five landmarks (two eyes, two mouth corners and nose) from train and test sets. Afterwards, we cropped the image into 112×112. We also normalized each pixel in RGB images by subtracting 127.5 and then dividing by 128.

All CNN models in the experiments use the same architecture in this work, which is a 50-layer residual network [216]. It contains four residual blocks and finally gets a 512 dimensional feature by average pooling. The networks are trained using Stochastic Gradient Descent (SGD) on TITANX GPUs and the batch size is set to fill all the GPU memory. The initial value for the learning rate is set to 0.1 and multiplied by 0.9 in intervals of five epochs until its value is less than or equal to 10^{-6}. All models are trained for $600K$ iterations.

Training: We trained our model on the refined MSCeleb-1M [218] dataset. MS-Celeb-1M originally contained about 10M images from 100K identities. We removed the images which were far away from the class centers to enhance the quality of the training data and cleared the identities with less than 3 images to diminish the long-tail distribution [208, 219]. The refined MS-Celeb-1M dataset contained 85K identities with 3.84M images. The face images are horizontally flipped for data augmentation.

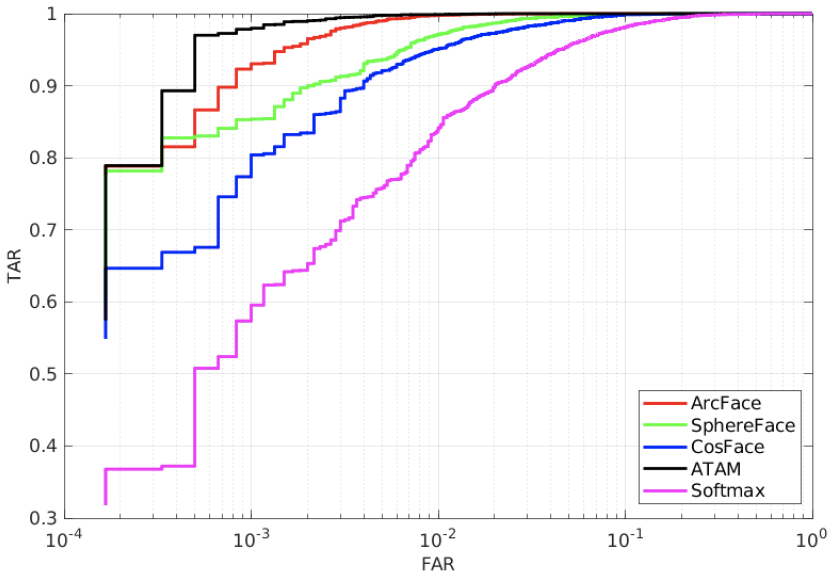


Figure 5.4: ROC curves for matching face images for different methods on LFW [3].

Evaluation Setup: For each image, we extract features only from the original image as the final representation. We didn't extract features from both the original image and the flipped one and concatenate them as the final representation. Therefore, the dimension of the final representation is 512 for each image. The score is measured by the cosine distance of two features. Eventually, verification and identification are conducted by thresholding and ranking the scores.

5.3.2 Face Recognition: Overall Benchmark Comparisons

Experiments on MegaFace.

MegaFace [209] is one of the most challenging testing benchmark for large-scale face identification and verification, which intends to assess the performance of face recognition models at the million scale of distractors. The gallery set of MegaFace is a subset of Flickr photos, contains more than one million face images. The probe sets are two existing databases: FaceScrub [220] and FGNet. The FaceScrub dataset contains 106,863 face images of 530 celebrities. The FGNet dataset is mainly used for testing age invariant face recognition, with 1002 face images from 82 persons.

We evaluated the proposed ATAM loss on FaceScrub of MegaFace Challenge 1, including both face identification and verification tasks. We followed the protocol of large training set as the training dataset contains more than 0.5M images, where the identities appearing in FaceScrub were removed from the train-

Method	Training size	#Models	LFW	YTF
Deep Face	4M	3	97.4	91.4
FaceNet	200M	1	99.7	95.1
DeepFR	2.6M	1	98.9	97.3
DeepID2+	300K	25	99.5	93.2
Center Face	0.7M	1	99.3	94.9
Baidu	1.3M	1	99.1	-
SphereFace	0.5M	1	99.4	95.0
CosFace	5M*	1	99.7	97.6
UniformFace	6.1M	1	99.8	97.7
AdaptiveFace	5M	1	99.6	-
Softmax	5M	1	98.8	95.7
SphereFace	5M	1	99.6	96.6
CosFace	5M	1	99.5	96.2
ArcFace	5M	1	99.6	96.8
ATAM	5M	1	99.7	97.9

Table 5.2: Face verification (%) on the LFW and YTF datasets. "*" indicates although the dataset of CosFace contains 5M images, it is composed of several public datasets and a private face dataset, containing about more than 90K identities.

ing set. In addition, there are some noisy images from FaceScrub and MegaFace, hence we used the noises list proposed by [208] to clean it.

We employed an attribute predictor to predict the attributes for the MegaFace training set. In [221], an ontology of 40 facial attributes are defined. We utilized the predicted attributes as an input in our proposed ATAM loss. For fair comparison, we implemented the Softmax, A-Softmax, CosFace, ArcFace, and our ATAM loss with the same architecture. Table 5.1 shows the results of our models trained on the large protocol of MegaFace. The proposed model obtains the best performance on both identification and verification tasks, compared with related methods including SphereFace, ArcFace, and AdaptiveFace. This shows the effectiveness of the proposed ATAM loss through the final recognition rates on the MegaFace dataset.

Experiments on YTF and LFW.

We evaluated our proposed model on the widely-used YTF [210] and LFW [3] datasets. YTF contains 3,425 videos of 1,595 different persons downloaded from YouTube, with different variations of pose, illumination and expression, which is a popular dataset for unconstrained face recognition. In YTF, there are about 2.15 videos available for each person and a video clip has 181.3 frames on average. LFW [3] is a famous image dataset for face recognition, which contains 13,233 images from 5,749 different identities. The images are captured from the web in wild conditions, varying in pose, illumination, expression, age and background, leading to large intra-class variations.

We followed the standard protocol of unrestricted with labeled outside data [222] to evaluate our model,

and reported the result on the 5,000 and 6,000 pair testing images from YTF and LFW, respectively. From the table 5.2, we observed that the usage of ATAM loss boosts the performance of 1.1% on YTF compared to ArcFace with the same training data. It should be noted that the attributes which are passed to the proposed model during the training, are noisy due the error in the attribute prediction model. However, our proposed ATAM could exploit them to make the holistic feature space more discriminative and boost face recognition performance.

5.3.3 Person re-identification:

Person re-identification (reID) goal is to spot the appearance of a same person in different scene. We evaluated our proposed method on two popular datasets, i.e., Market-1501 [211] and DukeMTMC-reID [212]. Market-1501 contains 1,501 identities, 12,936 training images and 19,732 gallery images captured with 6 cameras. We utilized 27 attributes such as gender, hair length, carrying backpack, etc., annotated by [223]. The DukeMTMC-reID dataset for re-ID has 1,812 identities from eight cameras. There are 1,404 identities appearing in more than two cameras and 408 identities (distractor ID) who appear in only one camera. We randomly picked 702 IDs as the training set and the remaining 702 IDs as the testing set. In the testing set, we select one query image for each ID in each camera and put the remaining images in the gallery. As a result, we get 16,522 training images with 702 identities, 2,228 query images of the other 702 identities, and 17,661 gallery images. We used 23 attributes such as gender, wearing hat, wearing boots, carrying backpack, etc., annotated by [223].

We adopted two network architectures, i.e. a global feature learning model backboned on ResNet-50 and a partfeature model named MGN [224]. We used MGN due to its competitive performance and relatively concise structure. The original MGN uses a Softmax loss on each part feature branch for training. MGN [224] is one of the state-of-the-art method which can learn multi-granularity part-level features. It utilizes both Softmax loss and triplet loss to facilitate a joint optimization. Following [225], we only used a single loss function in implementation of "MGN (ResNet-50)+ ATAM loss" for simplicity. In addition, all the part features were concatenated into a single feature vector. We evaluate ATAM loss on re-ID task in Table 5.3.

We make following observations from Table 5.3. First, comparing ATAM loss against state-of-the-art, we find that ATAM loss achieves competitive re-ID accuracy, with a single loss function and without using auxiliary loss functions. Second, using privileged attributes, here also improves the accuracy of our proposed

Method	Market-1501		DukeMTMC-reID	
	R-1	mAP	R-1	mAP
PCB [226] (Softmax)	93.8	81.6	83.3	69.2
MGN [224] (Softmax+Triplet)	95.7	86.9	88.7	78.4
JDGL [227]	94.8	86.0	86.6	74.8
APR [223]	84.3	64.7	73.9	55.6
AANet-50 [228]	93.9	82.5	86.4	72.6
ResNet50 + AMSoftmax [229]	92.4	83.7	83.9	68.5
ResNet50 + CircleLoss [225]	94.2	84.9	-	-
ResNet50 + ATAM	95.1	86.5	87.8	77.3
MGN + AMSoftmax	95.3	86.6	85.7	72.3
MGN + CircleLoss	96.1	87.4	-	-
MGN + ATAM	97.1	88.3	89.6	79.1

Table 5.3: Evaluation of ATAM loss on re-ID task. We report R-1 accuracy (%) and mAP (%).

model which is consistent with the experimental results on face recognition task. Third, comparing ATAM loss with APR [223] and AANet-50 [228] methods which utilize attributes both in the training and inference time, we observe the superiority of ATAM loss.

5.4 Conclusion

Most existing methods aim to learn discriminative deep features, encouraging large inter-class distances and small intra-class variations. However, they ignore the distribution of different classes in the holistic feature space, which may lead to severe locality and unbalance. Recent deep representation learning methods typically adopt class re-sampling or cost-sensitive learning schemes. In this chapter, we proposed a novel approach which introduces the adaptive margins utilizing attributes to adaptively minimize intra-class variances. On two major deep feature learning tasks, i.e., face recognition and person re-identification, our feature learning achieves performance on par with the state-of-the-art. We believe that our approach could be very helpful for unbalanced data training in practice.

Chapter 6

Robust Facial Landmark Detection via Aggregation on Geometrically Manipulated Faces

6.1 Introduction

Facial landmark detection goal is to identify the location of predefined facial landmarks (*i.e.*, tip of the nose, corner of the eyes, and eyebrows). Reliable landmark estimation is part of the procedure for more complicated vision tasks. It can be applied to the variant tasks such as 3D face reconstruction [230], head pose estimation [231], facial reenactment [232], and face recognition [233]. However, it remains challenging due to the necessity of handling non-rigid shape deformations, occlusions, and appearance variations. For example, facial landmark detection must handle not only coarse variations such as illumination and head pose but also finer variations including skin tones and expressions.

Many approaches solve the face alignment problem with multi-tasking approaches. However, the task of face alignment might not be in parallel with the other tasks. For example, in the classification task, the output needs to be invariant to small deformations such as translation. However, in tasks such as landmark localization or image segmentation both the global integration of information as well as maintaining the local information and pixel-level detail is necessary. The goal of precise landmark localization has led to evolving new architectures such as dilated convolutions [234], recombinator-networks [235], stacked what where auto-encoders [236], and hyper-columns [237] where each of them attempts to preserve pixel-level information.

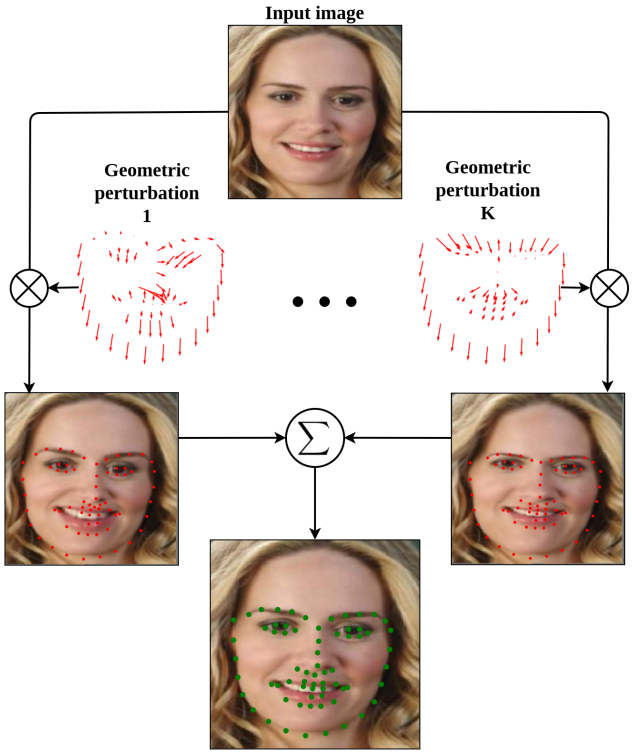


Figure 6.1: An input face image is manipulated utilizing geometric perturbations that target important locations of face images for the task of landmark detection. K different manipulated faces are generated where each of them contains the important displacements from the input image. The aggregation on these manipulated images leads to robust landmark detection.

In this chapter, we propose a geometry aggregated network (GEAN) for face alignment which can comfortably deal with rich expressions and arbitrary shape variations. We design a novel aggregation framework which optimizes the landmark locations directly using only one image without requiring any extra prior which leads to robust alignment given arbitrary face deformations. We provide three different approaches to produce deformed images using only one image and aggregate them in a weighted manner according to their amount of displacement to estimate the final locations of the landmarks. Extensive empirical results indicate the superiority of the proposed method compared to existing methods on challenging datasets with large shape and appearance variations, *i.e.*, 300-W [238] and ALFW [239].

6.2 Proposed Method

Given a face image $I \in \mathbb{R}^{w \times h}$ with spatial size $W \times H$, the facial landmark detection algorithm aims to find a prediction function $\Phi : \mathbb{R}^{W \times H} \rightarrow \mathbb{R}^{2 \times L}$ which estimates the 2D locations of L landmarks. We seek to

find a robust and accurate version of Φ by training a deep function through the aggregation of geometrically manipulated faces. The proposed method [240] consists of different parts which will be described in detail.

6.2.1 Aggregated Landmark Detector.

The proposed approach attempts to provide a robust landmark detection algorithm to compensate for the lack of a specific mechanism to handle arbitrary shape variations in the literature of landmark detection. The method builds upon aggregating set of manipulated images to capture robust landmark representation. Given a face image I, a set of manipulated images are constructed such that $\hat{I}_k = M(I, \theta_k)$ is the k-th manipulated face image and θ_k is its related parameters for the manipulating function M. Considering the set of manipulated images, we seek a proper choice of M such that aggregating landmark information in the set $\{\Phi(\hat{I}) : k = 1 \dots K\}$ provides a more accurate and robust landmark features compared to $\Phi(I)$ which solely uses the original image I. Therefore, one important key in the aggregated method is answering the question of "how" to manipulate images. Face images typically have a semantic structure which have a similar global structure but the local and relative characteristics of facial regions differ between individuals. Hence, a straightforward and comprehensive choice of the manipulation function M should incorporate the prior information provided by the global consistency of semantic regions and uniqueness of relative features which can be interpreted as the ID information. Hence, we build our work based on a choice of M which incorporates geometric transformations to manipulate relative characteristics of inputs samples while preserving the semantic and global structure of input faces.

To incorporate ID information, we consider a pretrained face recognizer $f : \mathbb{R}^{W \times H} \rightarrow \mathbb{R}^{n_z}$ mapping an input face image to an ID representation $z \in \mathbb{R}^{n_z}$, where cardinality of the embedding subspace is n_z (typically set to be 128 [204]). Having f makes it possible to compare IDs of two samples by simply measuring the ℓ_2-norm of their representation in the embedding space. Hence, we geometrically manipulate the input face image to change its ID. It should be noted that since f is trained on face images, the corresponding embedding space of IDs captures a meaningful representation of faces. Therefore, the manipulated faces contain rich information with regards to face IDs.

To manipulate the face image I based on landmark coordinates, we consider coarse landmark locations $P = \{(x_0, y_0), \dots, (x_{L-1}, y_{L-1})\}$ and define the displacement field d to manipulate the landmark locations. Given the i-th source landmark (x_i, y_i), we compute its manipulated version using the displacement vector

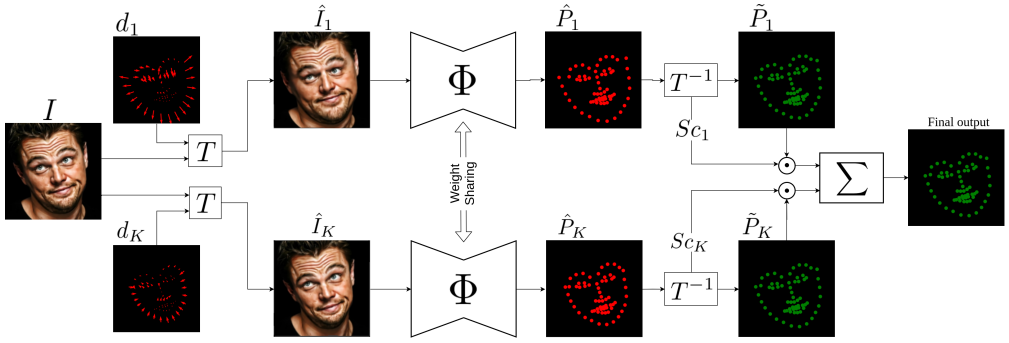


Figure 6.2: Overview of the proposed aggregated framework (GEAN). It consists of four steps: 1) K different manipulated faces are generated; 2) Each manipulated face is given to the shared landmark detector Φ to extract its landmarks; 3) The inverse of transformation matrix is applied to the extracted landmarks to compensate for the displacement of step 1; 4) The normalization score values for each landmark of each branch is calculated and the aggregation is performed to extract the final landmark locations.

$d_i = (\Delta x_i, \Delta y_i)$. The manipulated landmark $p_i + d_i$ is as follows:

$$p_i + d_i = (x_i + \Delta x_i, y_i + \Delta y_i) \ . \tag{6.1}$$

We present three different approaches to find a proper displacement (d) for manipulating face images.

6.2.2 Manipulation by Adversarial Attack.

In the first approach we use adversarial attacks [241] to manipulate facial landmarks to fool a face recognizer. Xiao et al. [242], proposed stAdv attack to generate adversarial examples using spatially transforming benign images. They utilize a displacement field for all the pixels in the input image. Afterward, they computed the corresponding location of pixels in the adversarial image using the displacement field d. However, optimizing a displacement field for all the pixels in the image is a highly non-convex function. Therefore, they used the L-BFGS [243], with a linear backtrack search to find the optimal displacement field which is computationally expensive. Here, our approach considers the fact that the facial landmarks provide highly discriminative information for face recognition tasks [244]. In fact, face recognition tasks are highly linear around the original coordinates of the facial landmarks as it is shown in [245].

In contrast to [242] which computes the displacement field for all the pixels, our proposed method is inspired by [245] and estimates the d only for L landmarks and it does not suffer from the computational complexity. In addition, it is possible to apply the conventional spatial transformation to transform image. Therefore, the adversarial (manipulated) image using the transformation T is as follows:

$$\hat{I} = T(P, P + d, I) \ , \tag{6.2}$$

where T is the thin plate spline (TPS) [246] transformation mapping from the source landmarks (control points) P to the target ones $P + d$. In order to make the whole framework differentiable with respect to the landmark locations, we select a differentiable interpolation function (*i.e.*, differentiable bilinear interpolation) [247] so that the prediction of the face recognizer is differentiable with respect to the landmark locations.

In this approach, we employ the gradient of the prediction in a face recognition model to update the displacement field d and geometrically manipulate the input face image. We extend Dabouei et al. [245] work in a way to generate K different adversarial faces where each face represents a different ID (K different IDs will be generated). Considering an input image I, a face recognizer f , and a set of $k-1$ manipulated images $S_I = \{\hat{I}_1, ..., \hat{I}_{k-1}\}$ the cost is defined as follows for the k-th adversarial face:

$$\mathcal{L} = \sum_{I' \in S_I} ||f(T(P, P + d, I)) - f(I')||_2 \ . \tag{6.3}$$

Inspired by FGSM [241], we employ the direction of the gradients of the prediction to update the adversarial landmark locations $P+d$, in an iterative manner. Considering $P+d$ as P^{adv}, using FGSM [241], the t-th step of optimization is as follows:

$$P_t^{adv} = P_{t-1}^{adv} + \epsilon \ sign(\nabla_{P_{t-1}^{adv}} \mathcal{L}) \ . \tag{6.4}$$

In addition, we consider the clipping technique to constrain the displacement field in order to prevent the model from generating distorted face images. The algorithm continues the optimization for the k-th landmark locations until $\underset{I' \in S_I}{min}\{||f(\hat{I}) - f(I')||_2\} < \tau$ is failed, where τ is simply the distance threshold in the embedding space. In this way, we make sure that the k-th manipulated face has a minimum distance of τ to the other manipulated images in the face embedding subspace. Algorithm 1 shows the proposed procedure for generating K different manipulated faces.

Algorithm 1 Adversarial Face Generation

Input: Image I, number of branches K, face recognizer f, distance threshold τ, clipping threshold δ.
Output: Set of adversarial faces $S = \{\hat{I}_1, ..., \hat{I}_K\}$.
Initialize $\hat{I} \leftarrow I$ and $S = \{I\}$.
for $k = 1$ **to** K **do**
  $\hat{I}_{t=0,k} \leftarrow I$;
  while $\min_{I' \in S_I} \{||f(\hat{I}_{t,k}) - f(I')||_2\} < \tau$ **do**
    $\mathcal{L} = \sum_{I' \in S_I} ||f(T(P, P_t^{adv}, I)) - f(I')||_2$;
    $P_{t+1}^{adv} = P_t^{adv} + \epsilon \; sign(\nabla_{P_t^{adv}} \mathcal{L})$;
    $P_{t+1}^{adv} = clip\,(P_{t+1}^{adv}, \delta)$;
    $\hat{I}_{t+1,k} = T(P, P_{t+1}^{adv}, I)$;
  $S \leftarrow \{S, \hat{I}_k\}$;
return $S - \{I\}$;

6.2.3 Manipulation of Semantic Groups of Landmarks using Adversarial Attacks.

In the first approach, we consider a fast and efficient approach to generate different faces based on the given face image. However, the first approach does not directly consider the fact that different landmarks semantically placed in different groups (*i.e.*, landmarks related to lip, left eye, right eye, etc.). This might lead to generating severely distorted adversarial images.

We added the clipping constraint to mitigate this issue in the first approach. Here, we perform semantic landmarks grouping [245]. We categorize the landmarks into n semantic groups $P_i, i \in \{1, \dots n\}$, where $p_{i,j}$ denotes the j-th landmark in the group i which contains c_i landmarks. These groups are formed based on different semantic regions which construct the face shape (*i.e.*, lip, left eye, right eye, etc.). Semantic landmark grouping considers a scale and translation for each semantic group, instead of independently displacing each landmark. This consideration allows us to increase the total amount of displacement while preserving the global structure of the face.

Let P_i represents the i-th landmark group (*e.g.*, group of landmarks related to the lip). The adversarial landmark locations which are semantically grouped can be obtained as following:

$$P_i^{adv} = \alpha_i (P_i - \bar{p_i}) + \beta_i \;, \tag{6.5}$$

where $\bar{p_i} = \dfrac{1}{c_i} \sum_{j=1}^{c_i} p_{i,j}$ is the average location of all the landmarks in group P_i, and α_i and β_i for each group can be computed using the closed-from solution in [245]. It should be noted that the value of displacement d for each branch used for computing semantic scales and translations is obtained using Algorithm 1.

The only difference is that we add semantic perturbations constructed by Eq. 6.5 instead of the random perturbation in line 8 of Algorithm 1. Therefore, the scale and translation of each semantic part of the face is different from other manipulated images in the set S.

6.2.4 Manipulation of Semantic Group of Landmarks with Known Transformation.

In this approach, we semantically group the landmark locations in the same manner as the previous approach. Afterward, we uniformly sample ranges $[0.9, 1.1]^2$ and $[-0.05 \times W, 0.05 \times W]^2$ for the scale and translation of each semantic group, respectively. It may be noted that in the post-processing stage, we make sure that the semantic inter-group structure is preserved, i.e., eyes region does not interfere with the eyebrows region and they are symmetric according to each other. Therefore, the heuristic post-processing limits the above ranges based on the properties of each group. For instance, eyebrows could achieve higher vertical displacement compared to eyes since there is no semantic part above them to impose a constraint.

6.2.5 Landmark Detector.

Next, the hourglass network proposed in [49] is employed to estimate the facial landmarks location. Hourglass is designed based on residual blocks [248]. It is a symmetric top-down and bottom-up fully convolutional network. The residual modules are able to capture high-level features based on the convolutional operation, while they can maintain the original information with the skip connections. The original information is branched out before downsampling and concatenated together before each up-sampling to retain the resolution information. Therefore, hourglass is an appropriate topology to capture and consolidate information from different resolutions and scales.

After manipulating the face images (employing either of the three aforementioned approaches), we employ the hourglass network, Φ, to extract the landmarks from the manipulated images. The network Φ is shared among all the branches of the framework as it is shown in Fig. 6.2. Each landmark has a corresponding detector, which convolutionally extracts a response map. Taking r_i as a i-th response map, we use the weighted mean coordinate as the location of the i-th landmark as follows:

$$\hat{p}_i = (x_i, y_i) = \frac{1}{\zeta_i} \sum_{u=1}^{H} \sum_{v=1}^{W} (u, v).r_i(u, v) \ , \tag{6.6}$$

where H and W are the height and width of the response map which are the same as the spatial size of the

input image, and $\zeta_i = \sum_{u,v} r_i(u, v)$.

6.2.6 Aggregation.

After extracting the facial landmarks using the shared landmark detector Φ for each of the manipulated face images, we aim to move the predicted landmarks $\hat{P}$ toward their original locations. Let T be a transformation that is used to convert the original faces to the manipulated ones (*i.e.*, via adversarial attack approaches or the known transformation approach). We employ the inverse of the transformation matrix on the predicted landmarks to compensate for the displacement of them and denote the new landmark locations as $\tilde{P}$.

The proposed approach contains a set of landmarks from K branches, *i.e.*, $\tilde{P} = \{\tilde{p}_{i,k}\}$ in which $i \in \{1, \ldots, L\}$, and $k \in \{1, \ldots, K\}$ is the i-th landmark location in k-th branch of the framework. Each branch considers a score value which normalizes the displacement of landmarks caused by the manipulation approach (*i.e.*, via adversarial attacks or known transformations) in each branch of the aggregated network as follows:

$$Sc_{i,k} = \frac{\sqrt{\Delta x_{i,k}^2 + \Delta y_{i,k}^2}}{\sum_{k=1}^{K} \sqrt{\Delta x_{i,k}^2 + \Delta y_{i,k}^2}} , \tag{6.7}$$

where $Sc_{i,k}$ represents the displacement value for the i-landmark in the k-th branch. This score is utilized as a weight to cast appropriate loss punishment in different branches during the optimization of the proposed aggregated landmark detection as follows:

$$\mathcal{L}_T = \frac{1}{LK} \sum_{i=1}^{L} \sum_{k=1}^{K} Sc_{i,k} ||p_i^* - \tilde{p}_{i,k}||_2 , \tag{6.8}$$

where $\tilde{p}_{i,k}$ represents the i-th estimated landmark at k-th branch and p_i^* indicates the ground truth for i-th landmark location.

Given a test image, we extract the rough estimation of the landmark coordinates employing the trained landmark detector Φ in the aggregated approach and consider them as the coarse landmarks P. Afterward, we perform the manipulation approach on the extracted landmarks P and generate manipulated images. The extracted landmarks and manipulated images are used in the aggregated framework to produce the final landmarks. The final landmarks are calculated as follows:

Figure 6.3: The representative results for three face images from the 300-W dataset. For each face, the first row represents displacement fields for the aggregated network with $K = 3$ (the arrows are exaggerated for the sake of illustration). The second row shows manipulated images using the corresponding displacement field, and the third row represents the extracted landmarks given the corresponding manipulated images to the landmark detector,$\Phi(\hat{I})$. The fourth row represents landmarks' locations on the input image I from the base detector (in blue), ground-truth (in green), and GEAN landmark detector (in magenta), respectively.

Methods	ERT [249]	LBF [250]	CFSS [5]	CCL [44]	Two-St. [46]	SAN [251]	ODN [4]	LRef. [10]	GEAN_{adv}	GEAN_{GK}	GEAN_{Gadv}
AFLW-Full	4.35	4.25	3.92	2.72	2.17	1.91	1.63	1.63	1.69	1.64	**1.59**
AFLW-Front	2.75	2.74	2.68	2.17	-	1.85	1.38	1.46	1.44	1.38	**1.34**

Table 6.1: Comparison of different methods based on normalized mean errors (NME) on AFLW dataset.

$$p_i^f = \sum_{k=1}^{K} Sc_{i,k}.(\tilde{p}_{i,k}) \ , \tag{6.9}$$

where p_i^f is the coordinate of the i-th landmark employing the proposed aggregated network such that $\Phi(I) = P^f$ for L landmark locations.

As it is mentioned in the manuscript, during the training phase the manipulation is performed on the coarse landmarks' locations and we have access to them. However, given a test image, we extract the landmarks using the trained landmark detector Φ and then use them as the coarse landmarks' locations in the aggregated framework to predict the final landmark locations.

One question that comes to mind is: what if the predicted landmark locations using the trained landmark detector Φ are not an accurate representation for the original coarse landmarks? To compensate this issue and make the conditions equal for the training and testing phases, we add random noise to the ground truth landmarks such that $P = P^* + \eta$ where P^* is the ground truth for landmark coordinates and η is random noise. Afterward, we employ these landmarks as the coarse landmarks P in the aggregated framework during the training phase.

6.3 Experiments

In the following section, we consider three variations of our GEAN approach. $GEAN_{adv}$ (6.2.2) represents the case when the manipulated faces are generated using the adversarial attack approach. $GEAN_{Gadv}$ (6.2.3) and $GEAN_{GK}$ (6.2.4) represent the cases when the manipulated faces are generated using the semantically grouped adversarially attack and known transformations approach, respectively. In order to show the effectiveness of GEAN we evaluate its performance on three following datasets:

300-W [238]: The dataset annotates five existing datasets with 68 landmarks: LFPW [252], AFW [253], HELEN [254], iBug, and XM2VTS. Following the common setting in [251, 46], we consider 3,148 training images from LFPW, HELEN, and the full set of AFW. The testing dataset is split into three categories of common, challenging, and full groups. The common group contains 554 testing images from LFPW and HELEN datasets, and the challenging test set contains 135 images from the IBUG dataset. Combining these two subsets form the full testing set.

AFLW [239]: This dataset contains 21,997 real-world images with 25,993 faces in total with a large variety in appearance (*e.g.*, pose, expression, ethnicity, and age) and environmental conditions. This dataset provides at most 21 landmarks for each face. Having faces with different pose, expression, and occlusion makes this dataset challenging to train a robust detector. Following the same setting as in [46, 251], we do not consider the landmark of two ears. This dataset has two different categories of AFLW-Full and AFLW-Frontal [44]. AFLW-Full contains 20,000 training samples and 4,386 testing samples. AFLW-Front uses the same set of training samples as in AFLW-Full, but only contains 1,165 samples with the frontal face for the testing set.

COFW [7]: This dataset contains 1,345 images for training and 507 images for test. Originally this dataset annotated with 21 landmarks for each face. However, there is a new version of annotation for this dataset with 68 landmarks for each face [255]. We used a new version of annotation to evaluate proposed method and comparison with the other methods.

Evaluation: Normalized mean error (NME) and and Cumulative Error Distribution (CED) curve are usually used as metric to evaluate performance of different methods [44, 46]. Following [250], we use the inter-ocular distance to normalize mean error on 300-W dataset. For the AFLW dataset we employ the face size to normalize mean error as there are many faces with inter-ocular distance closing to zero in this dataset [46].

Method	Common	Challenging	Full Set
LBF [250]	4.95	11.98	6.32
CFSS [5]	4.73	9.98	5.76
MDM [258]	4.83	10.14	5.88
TCDCN [259]	4.80	8.60	5.54
Two-Stage [46]	4.36	7.42	4.96
RDR [260]	5.03	8.95	5.80
Pose-Invariant [261]	5.43	9.88	6.30
SAN [251]	3.34	6.60	3.98
ODN [4]	3.56	6.67	4.17
LRefNets [10]	2.71	4.78	3.12
GEAN	**2.68**	**4.71**	**3.05**

Table 6.2: Normalized mean errors (NME) on 300-W dataset.

Implementation Details: We employ the face recognition model developed by Schroff et al. [204] which obtain the state-of-the-art accuracy on the Labeled Faces in the Wild (LFW) [3] dataset as the face recognizer. We train this model on more than 3.3M training images and the average of 360 images per ID (subject) from VGGFace2 dataset [256] to recognize 9,101 celebrities. The landmarks are divided to five different categories based on facial regions as: 1) P_1 : right eye and eyebrow, 2) P_2 : left eye and eyebrow, 3) P_3 : nose, 4) P_4 : mouth, and 5) P_5 : jaw. The number of landmarks in each group is as: $\{n_1 = 11$, $n_2 = 11$, $n_3 = 9$, $n_4 = 20$, $n_5 = 17\}$. We set $\tau = 0.6$, δ to 5% of the width of the bounding box of each face.

The landmarks' coordinates are scaled to lie inside the range $[-1, 1]^2$ where $(-1, -1)$ is the top left corner and $(1, 1)$ is the bottom right corner of the face image. All the coordinates are assumed to be continuous values since TPS has no restriction on the continuity of the coordinates because of the differentiable bilinear interpolation [247]. The face images are cropped and resized to (256×256). We follow the same setting in [257] and use four stacks of hourglass network for the landmark detection network. We train our model with the batch size of 8, weight decay of 5×10^{-4}, and the starting learning rate of 5×10^{-5} on two GPUs. The face bounding boxes are expanded by the ratio of 0.2 and random cropping is performed as data augmentation.

6.3.1 Comparison with State-of-the-arts Methods:

Results on 300-W. Table 6.2 shows the performance of different facial landmark detection methods on 300-W dataset. We compare our method to the most recent state-of-the-art approaches in the literature [251, 260, 261, 10]. The number of branches in training and testing phases is set to $K = 5$. Among the three proposed approaches, we consider $GEAN_{Gadv}$ as the final proposed method to compare with the state-of-

	Common test set				**Challenging test set**				**Full test set**						
Train	1	3	5	7	1	3	5	7	1*	1	3	5	7	10	14
1	4.40	3.77	3.48	3.43	5.44	5.35	5.30	5.28	4.80	4.80	4.49	4.07	4.02	4.00	4.03
3	3.67	3.25	3.03	2.98	5.33	5.27	5.18	5.10	4.68	4.46	4.01	3.77	3.74	3.72	3.79
5	3.35	2.99	2.68	2.66	5.26	4.97	4.71	4.67	4.63	4.04	3.64	3.05	3.01	2.99	3.05
7	3.32	2.93	2.65	2.63	5.22	4.90	4.65	4.60	4.59	3.96	3.56	3.00	2.97	2.96	3.00

Table 6.3: Comparison of NME on three test sets of 300-W with different numbers of branches for the training and testing. The column with asterisk demonstrates the results for evaluating the performance of our model without aggregation.

the-art methods. The results show its superiority compared to the other methods for both types of bounding boxes. The superiority of the proposed method shows the effect of manipulated images which target the important locations in the input face image. Aggregation of these images improves the facial landmark detection by giving more attention to the keypoint locations of the face images.

Results on AFLW. We conduct our experiments on the training/testing splits and the bounding box provided from [44, 5]. Table 6.1 shows the effectiveness of proposed GEAN. AFLW dataset provides a comprehensive set of unconstrained images. This dataset contains challenging images with rich facial expression and poses up to $\pm 120^{\circ}$ for yaw and $\pm 90^{\circ}$ for pitch and roll. Evaluation of proposed method on this challenging dataset shows its robustness to large pose variations. Indeed, the weighted aggregation of predictions obtained on the set of deformed faces reduces the sensitivity of GEAN to large pose variations.

Results on COFW. Figure 6.4 shows the evaluation of our proposed method in a cross-dataset scenario. We conduct evaluation using the models trained on the 300-W dataset and test them on re-annotated COFW dataset with 68 landmarks [255]. The comparison is performed using the CED curves as plotted in Figure 6.4. The best performance belongs to our method (GEAN) with 4.24% mean error compared to the previous best [10] with 4.40% mean error. This shows the robustness of our method compared to other state-of-the-art methods in detecting facial landmarks.

Timing of the proposed approach directly depends on the number of branches and also the approach that we take to generate the manipulated faces. It is shown in [245] that semantically grouping the landmark locations increases the time of manipulated faces generation. However, it can overcome the problem of face distortion due to considering the semantic grouping. Therefore, there is a trade-off between the speed and accuracy of the proposed framework. However, in the case of aggregating with five branches and employing $GEAN_{Gadv}$ for generating manipulated faces, the framework runs in 17 FPS with NVIDIA TITAN X GPU.

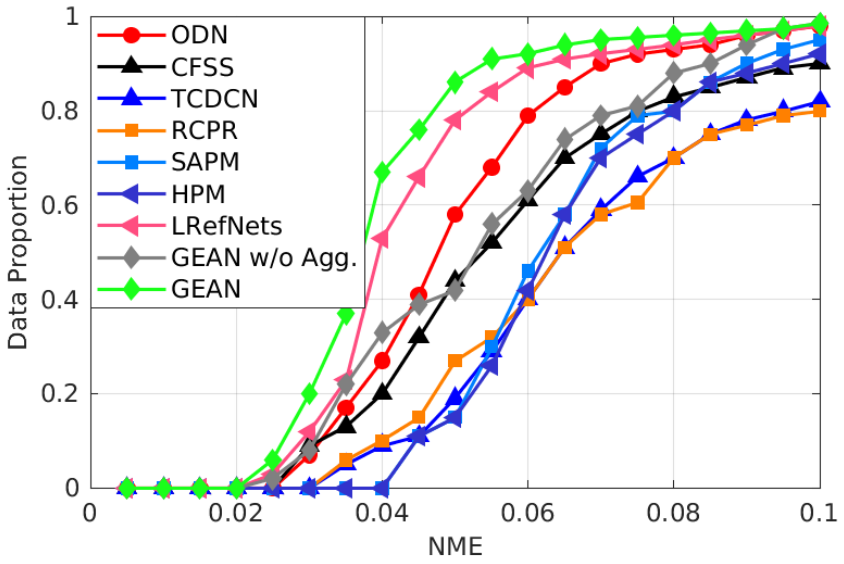


Figure 6.4: Comparison results of different methods (ODN [4], CFSS [5], TCDCN [6], RCPR [7], SAPM [8], HPM [9], LRefNets [10], and GEAN) on COFW dataset.

6.3.2 Ablation Studies

Number of branches: In this section, we observe the effect of adding branches on the performance of the aggregated framework. We start with $k = 1$ in which there is no aggregation and one manipulated image is generated. We increase the number of branches from one to seven and measure the performance of aggregated network on the common, challenging, and full split of the 300-W dataset.

In addition, the number of branches in the training and testing phases is not necessarily the same. For example, the number of branches in the aggregated framework can be three while the number of branches in the testing phase is equal to 10. This is essentially important due to the time complexity of the framework during the training and testing phases. In addition, one can train the network on two or three branches while test it on more branches to get more accurate results. Table 6.3 shows the evaluation results of 16 training and testing combinations, *i.e.*, four different training architectures ($K = 1, 3, 5, 7$) multiply four different testing architectures on 300-W common, challenging, respectively. We increased the number of branches to 14 during the inference time on full test set.

As we can observe, the performance will be increased if the number of branches is increased during the training phase. However, we observe that adding more than five branches to the framework does not significantly improve the results with the cost of more computational complexity. The same behavior is observed for the testing framework. By increasing the number of branches in the testing phase, the accuracy is

increased. This is useful when we want to reduce the computational complexity in training and maintaining the performance in the testing phase to some extent. Considering both accuracy and speed, we choose the framework with the number of training and testing branches equal to five for the sake of comparison with state-of-the-art (6.3.1). We also increased the number of branches to 14 on full test set. As it is shown in Table 6.3, by increasing the numbers to 10 and more the model gains slight improvement in accuracy with the cost of more computational and complexity. However as the numbers increased to 14, the accuracy drops and the model with 14 branches has less accuracy with more time complexity (the model has less accuracy compare to its counterpart with 5 branches).

We also conduct another experiment to demystify the effect of aggregation part in the proposed GEAN. In this case, GEAN with just one branch is trained on all the deformed and manipulated faces without the aggregation part. Table 6.3 and Figure 6.4 show the performance of $GEAN\ w/o\ Agg.$ compared to the proposed GEAN. For the sake of fair comparison, we trained the network on the same number of manipulated face images for both methods. By comparing column (1^*) with column (1) of 300-W full test set, it is shown that the proposed GEAN which is trained with the exact same faces is superior to its counterpart without aggregation. Figure 6.4 also confirms the effectiveness of aggregation part and illustrates the fact that proposed GEAN performs beyond a careful augmentation.

A Comparison between Three Different Variations of GEAN: Three different approaches of $GEAN_{adv}$, $GEAN_{Gadv}$, and $GEAN_{GK}$ have been introduced in this work. Through this section, we evaluate the performance of three different variations of our GEAN method on AFLW dataset. As Table 6.1 shows, both $GEAN_{Gadv}$ and $GEAN_{GK}$ outperform $GEAN_{adv}$ approach. We attribute this to the fact that $GEAN_{adv}$ does not consider grouping different landmarks semantically. This causes inconsistent displacements for the landmarks of one region (*e.g.*, left eye) and generate distorted images. In addition, the amount of displacement of landmarks in manipulated images might be greater than the manipulated images with the other two methods. However, this displacement might not be beneficial as it does not consider the general shape of each face region. Utilizing clipping constraint can mitigate this issue to some extent. However, this approach still suffers from not considering the semantic groups.

$GEAN_{Gadv}$ works the best among all three proposed approaches. Several reasons can explain this superiority. This approach considers the semantic relationship among the landmarks of same regions of the face. In addition, the manipulated images in this approach have different face IDs from the original face image. Therefore, in the framework with K branches, the aggregation is performed in K different

face IDs. This makes this approach to preserve a reasonable relative distance among different groups of landmarks since it could fool the recognizer to misclassify it. However, this is not necessarily the case for the $GEAN_{GK}$ approach. This makes the $GEAN_{Gadv}$ to capture more important landmark displacement for the image manipulation which is beneficial for the aggregation. The advantages of the other two approaches (*i.e.*, adversarially attack technique in $GEAN_{adv}$ and semantic grouping of landmarks in $GEAN_{GK}$) is unified in $GEAN_{Gadv}$ which leads to a better landmark detection performance.

6.4 Conclusion

In this chapter, we introduce a novel approach for facial landmark detection. The proposed method is an aggregated framework in which each branch of the framework contains a manipulated face. Three different approaches are employed to generate the manipulated faces and two of them perform the manipulation via the adversarial attacks to fool a face recognizer. This step can decouple from our framework and potentially used to enhance other landmark detectors [251, 46, 257]. Aggregation of the manipulated faces in different branches of GEAN leads to robust landmark detection. An ablation study is performed on the number of branches in training and testing phases and also on the effect different approaches of face image manipulation on the facial landmark detection. The results on the AFLW, 300-W, and COFW datasets show the superiority of our method compared to the state-of-the-art algorithms.

Chapter 7

HGAN: Hybrid Generative Adversarial Network

In this chapter, we propose a simple effective GAN architecture and a training strategy with the goal of adversarially distilling the explicit information of the data distribution provided by the autoregressive model in addition to mimicking the real data which leads to generating samples with a distribution very close to the actual data distribution and helps to avoid possible mode collapse. To resolve the issue of sharp good-looking samples but poor likelihood estimation in the case of adversarial learning (and vice versa in the case of maximum likelihood estimation), our proposed hybrid model bridges implicit and explicit learning models by augmenting the adversarial learning with an additional autoregressive model. Our approach combines the implicit and explicit density function estimation into a unified objective function. In our model, the HGAN generator is guided by exploiting the explicit data probability density from the knowledge provided by the autoregressive model while it is also responsible to learn the data distribution via the adversarial learning. HGAN model exploits the complementary statistical properties of data obtained from an autoregressive model by utilizing a GAN to effectively diversify the estimated density function and capturing different modes of the data distribution as well as avoiding possible mode collapse.

In short, our main contributions are: (i) a novel adversarial model to train a generator in a GAN framework in order to stabilize the training process; (ii) the proposed model is able to estimate the data density by mimicking an autoregressive model and simultaneously combining it with the adversarial learning process; (iii) a comprehensive performance evaluation of our proposed method on real-world large-scale datasets of diverse natural scenes as well as mitigating adversarial examples in a defense scenario.

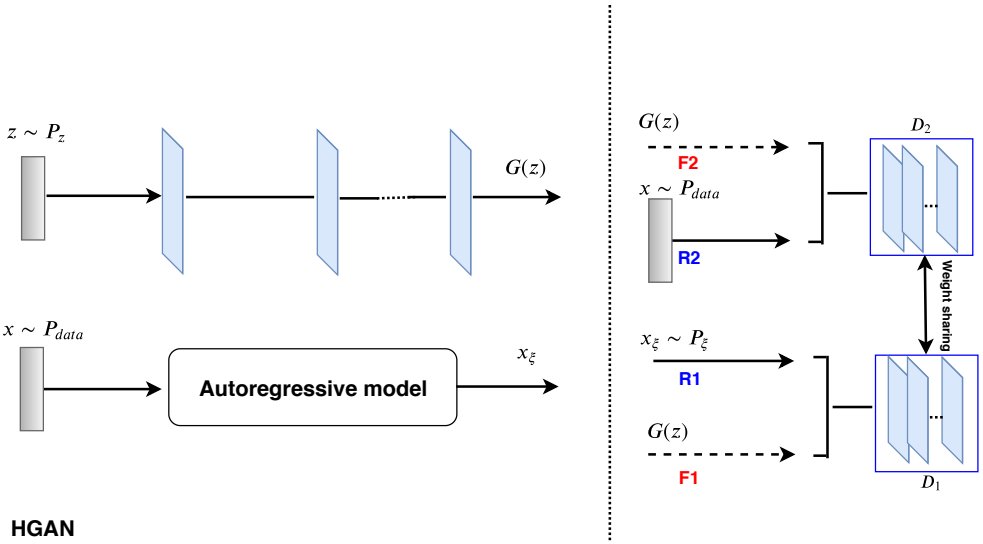


Figure 7.1: Proposed HGAN framework with an autoregressive model, a generator, and a discriminator is trained by using two types of real data.

7.1 Background

7.1.1 Generative Adversarial Nets:

GAN is a min-max game between a generator G and a discriminator D, both parameterized via neural networks [262]. Training a GAN can be formulated as the following objective function:

$$\min_G \max_D E_{x \sim P_{data(x)}}[\log D(x)] + E_{z \sim P_z}[\log(1 - D(G(z)))],$$

where x is from a real data distribution P_{data} and z is a sample from a prior distribution P_z. The generator is a mapping function from z which approximates P_{model}. GAN alternatively optimizes D and G in a minimax game using stochastic gradient-based algorithm. Generator is prone to map every z to a single x that is most probable to be recognized as a true data, and this leads to a mode collapse. Another issue with GAN is that at optimal point of D, minimizing the generator is equal to minimizing the JSD between the true data distribution and model distribution which is empirically shown to cause the mode collapse by generating few modes and ignoring other modes [59].

Figure 7.2: Samples generated by the proposed HGAN compared with the samples generated from DCGAN and AutoGAN on CIFAR-10.

7.1.2 Autoregressive Models:

Autoregressive models can be designed by using recurrent networks (PixelRNNs) or a CNN (PixelCNNs) [263]. These models learn the join distribution of pixels of an image x as a product of conditional distributions $p(x_i|x_1, ..., x_{i-1})$, where x_i is a single pixel. The ordering of pixel dependencies is row by row and in each row, pixel by pixel. Therefore, every pixel (x_i) depends on all the pixels above and left of it ($x_1, ..., x_{i-1}$).

7.1.3 Knowledge distillation:

Knowledge distillation is mostly used in image classification problem where the output of neural network is a probability distribution over categories. The probability is calculated by applying a softmax function over *logits* which are the output of the last fully connected layer. Hinton et al. [264] used logits to transfer the embedded information in a teacher network to student network. In order to train a student network F to generate student logits $F(x_i)$, a parameter called temperature T is introduced. Afterwards, the generalized softmax layer converts logits vector $t_i = (t_i^1, ..., t_i^C)$ to a probability distribution q_i,

$$M_T(t_i) = q_i, \quad where \quad q_i^j = \frac{exp(t_i^j/T)}{\sum_k exp(t_i^k/T)}. \tag{7.1}$$

where higher temperature T produces softer probability over categories.

Hinton et al. [264] proposed to minimize the KL divergence between teacher and student output as

follows:

$$L_{KD}(F,T) = \frac{1}{N}\sum_{i=1}^{N} KL(M_T(t_i)||M_T(F(x_i))). \tag{7.2}$$

In [265] instead of forcing the student to exactly mimic the teacher by minimizing KL-divergence in Equation (7.2), the knowledge is transferred from teacher to student via discriminator in a GAN-based approach.

7.2 Proposed Hybrid GAN:

We now present our novel hybrid approach to tackle the problem of mode collapse in GANs [266, 267]. In general, GANs can generate good-looking samples but have intractable likelihoods. On the other hand, autoregressive models are likelihood-based generative models which can return explicit probability densities. The idea is to utilize a mixture of these two models rather than a single model as in a typical GAN.

In our proposed hybrid model, the generator's first task is to learn the data distribution without any explicit model just like a regular GAN such that $G(z) \simeq x \sim P_{data}$. On the other hand, its second task is to perform sampling where it samples a random vector $z \sim P_z$ and maps it to an autoregressive model P_ξ such that $G(z) \simeq x_\xi \sim P_\xi$. This forces our hybrid model to learn the probability density of the autoregressive model using the adversarial training method. These two tasks together provide a hybrid model which gives more attention to the likelihood of the data for estimating P_{model} in the data space.

A natural question to ask is why one should use adversarial learning when autoregressive model can return a tractable likelihood. The reason is that the synthesis from these autoregressive models are difficult to parallelize and usually inefficient on parallel hardware [268]. Therefore synthesizing image using them is much slower than a generator. Moreover, it is not practical to perform accurate data manipulation since the hidden layers of autoregressive models have unknown marginal distributions [112]. However, GAN models are fast in synthesizing and also can have useful latent space for downstream tasks especially in ones which have an encoder such as AGE or ALI [269, 270]. Fig. 7.1 illustrates the architecture of our proposed Hybrid GAN (HGAN) model.

In naive GAN, the odds that the two distributions p_g and p_{data} share support in high-dimensional space, especially early in training, are very small. If p_g and p_{data} have non-overlapping support the Jensen-Shannon

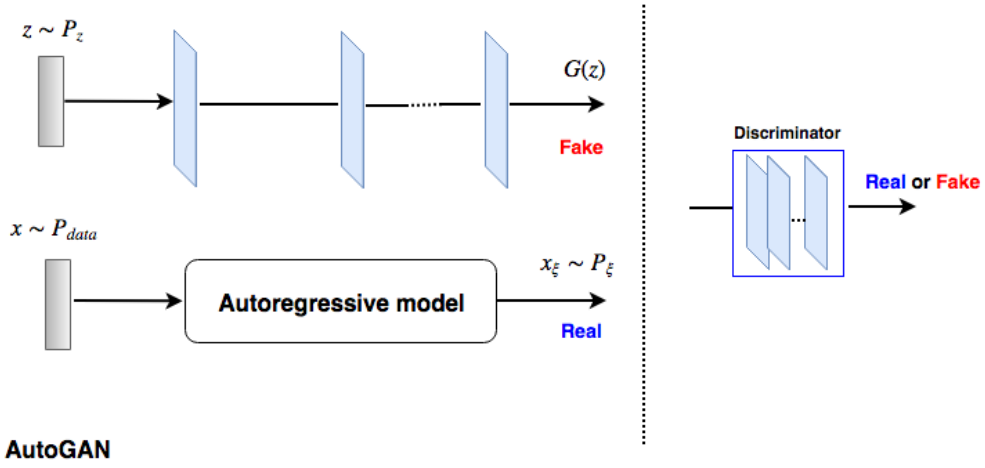
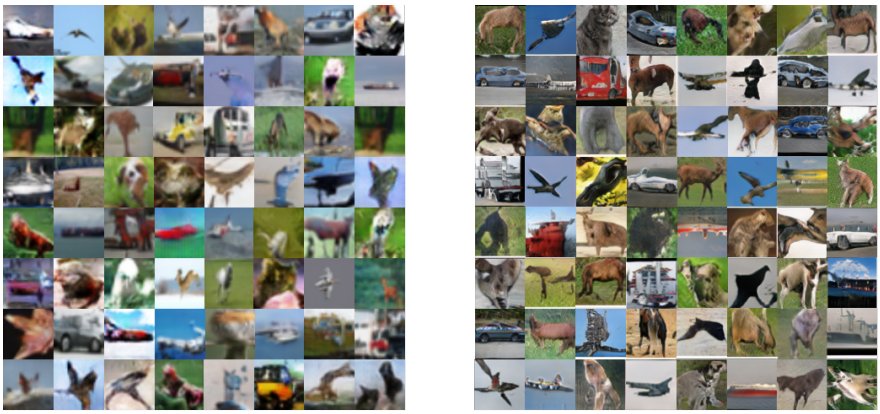


Figure 7.3: Training $G(z)$ to mimic the autoregressive model's output with an adversarial learning process. In this network which we denote as AutoGAN, the real data is obtained from the autoregressive model's output and fake data is the generated output from $G(z)$.

divergence is saturated as is locally constant in θ. Also, there might be a large set of near-optimal discriminators whose logistic regression loss is very close to optimum, but each of these possibly provides very different gradients to the generator. Therefore, training the discriminator might find a different near-optimal solution each time depending on initialization, even for a fixed g_θ and p_{data}. We instead employ autoregressive model to augment the gradient information obtained by ordinary back-propagation. In fact, we are interested in manipulating the feature space of a discriminator, using the autoregressive model as a tool to tell us *how* to perform that manipulation.

In our Hybrid GAN, the discriminator observes two types of real inputs: the real data x and the output of the autoregressive model x_ξ. The fake input, $G(z)$, is mimicking the output of the autoregressive model $x_\xi \sim P_\xi$ in addition to the real data $x \sim P_{data}$. We consider two terms for the discriminator D, namely D_1 and D_2. D_1 is the first discriminator which is related to the first task, where $G(z)$ is fake and x_ξ is real. D_2 is related to the second task, where $G(z)$ is fake and x is real. However, all the parameters are shared between discriminators D_1 and D_2 and in fact there is only one discriminator D.

In the first fake input of discriminator, the generator attempts to generate data that is as close as possible to the autoregressive model's output. Therefore, the generator's task is to make $G(z) \simeq x_\xi \sim P_\xi$. However, for the second round of fake input, the generator tries to fool the discriminator in a way that its generated data is as close as possible to the real data. Thus, it is responsible to make $G(z) \simeq x \sim P_{data}$. While G acts similar to a typical generator in a regular GAN, our hybrid method tries to maximize the likelihood of



(a) $CIFAR-10$ (b) $STL-10$

Figure 7.4: Images generated by our proposed HGAN trained on natural image datasets.

a mixture model by adversarially distilling the properties of autoregressive model.

7.3 Experiments

We show the effectiveness of our proposed approach in different experiments with real-world datasets. For the fair evaluation, we use the same experimental settings that are identical to prior works [56, 271, 272]. Therefore, we use the results from the latest state-of-the-art GAN-based models to compare with ours.

We used Pytorch [273] to implement our framework. The generator and discriminator architecture is adopted from DCGAN [274]. In addition, pixelCNN++ [110] architecture is chosen for the autoregressive model. For training we used Adam optimizer [275] with the first-order momentum of 0.5, the learning rate of 0.0002, and batch size of 64. For the generator the ReLU activation [276], and for the discriminator the Leaky ReLU activation with the slope of 0.2 is considered. Weights are initialized from an isotropic Gaussian: $\mathcal{N}(0,\, 0.01)$ and zero biases. To show the effectiveness of the proposed framework, we perform two types of experiments on MNIST dataset and compare our methods to the other well-know GANs, namely WGAN [58], MIX+WGAN [113], DFM [117], Improved-GAN [113], ALI [269], BEGAN [277], MAD-GAN [271], GMAN [278], DCGAN [274], MGAN [272], SNGAN [279], and SAGAN [280]. It should be noted that our method cannot be compared directly with BigGAN [281] and StyleGAN [282] since the mentioned models are based on larger models and different settings (i.e., BigGAN is using class conditional setting or StyleGAN purpose is for having more control over the latent space for high resolution image generation). Following [271], we reuse the KL-divergence [283] and the number of captured modes [284] as

Table 7.1: Experiment on MNIST dataset containing 10 different modes.

GAN Variants	Chi-square ($\times 10^5$)	KL Div
WGAN	1.32	0.614
MIX+WGAN	1.25	0.567
DFM	1.46	0.623
Improved-GAN	1.13	0.436
ALI	2.34	0.875
BEGAN	1.06	0.944
MAD-GAN	0.24	0.145
GMAN	1.86	1.345
DCGAN	0.90	0.322
MGAN	0.32	0.211
SAGAN	0.29	0.148
SNGAN	0.25	0.146
HGAN	**0.23**	**0.141**

Table 7.2: Results for the Inception scores on CIFAR-10 dataset.

Objective	Inception Score
DCGAN	6.40
AutoGAN	6.17
HGAN	**7.46**

the criteria for the comparison to illustrate the superiority of our method compared to others. Moreover, we perform the quantitative experiments on more complicated real-world datasets namely the CIFAR-10 [285] and STL-10 [286] datasets.

7.3.1 MNIST

The data distribution of the MNIST dataset can be approximated with ten dominant modes. Here, following [284] we define the term 'mode' as a connected component in the data manifold.

For the sake of evaluation, we train a four-layer CNN classifier on the MNIST digits and then apply it to compute the mode scores in the generated samples from the proposed method. We repeat the procedure and apply the trained classifier to discover the mode scores on different baseline GAN methods. We also have the ground truth by measuring the performance of classifier on the MNIST test set. The number of generated samples from each method is equal to the number of test set which is 10,000. Afterwards, we use

Table 7.3: Results for the test MODE scores on the MNIST dataset.

Objective	MODE Score
DCGAN	9.28
AutoGAN	9.32
HGAN	**9.51**

Table 7.4: Stacked-MNIST experiment. There are 1,000 modes in the dataset.

GAN Variants	KL Div	# Mode Covered
WGAN	1.02	868
MIX+WGAN	0.98	874
DFM	1.13	843
Improved-GAN	1.45	847
ALI	2.03	802
BEGAN	1.89	819
MAD-GAN	0.91	890
GMAN	2.17	756
DCGAN	2.15	712
MGAN	0.94	**896**
SAGAN	0.97	886
SNGAN	0.91	889
HGAN	**0.88**	891

Chi-square distance and the KL-divergence to compute distance between the two histograms (ground truth vs. each GAN model). Table 7.1 shows the performance of our proposed HGAN compared to the other methods. From Table 7.1, it is evident that our proposed method could outperform the other methods in capturing all the modes in the MNIST dataset.

7.3.2 Stacked and Compositional MNIST

In this experiment, the goal is to explore the performance of our proposed HGAN in a more challenging scenario. In order to illustrate and compare HGAN with other baselines, we utilized similar setup as in [271]. Authors in [114] created a Stacked MNIST with 25,600 samples where each sample has three channels stacked together with a random digit from MNIST in each of them. Therefore, the Stacked MNIST contains 1,000 distinct modes in the data distribution. In [287], a similar process is applied to MNIST dataset. They created the Compositional MNIST whereby they took three random MNIST digits and placed them at three quadrants of a 64×64 dimensional image. This also resulted in a data distribution with 1,000 modes. Distribution of the generated samples was estimated with a pre-trained MNIST classifier which classifies the digits in each channel or quadrants, and consequently decides which of the 1,000 modes is generated by the particular GAN method's generator.

Table 7.4 and 7.5 show the performance of the proposed method as well as other GAN methods in terms of the KL divergence and the number of modes recovered for the Stacked and Compositional MNIST datasets. As shown in Table 7.4, our method outperformed all the other GAN methods in terms of the KL divergence and the number of captured modes. MGAN surpasses ours in only the number of captured modes. It is evident from Table 7.5 that our proposed HGAN outperforms all the other baselines in terms of

Table 7.5: Compositional-MNIST experiment. There are 1,000 modes in the dataset.

GAN Variants	KL Div	# Mode Covered
WGAN	0.25	1000
MIX+WGAN	0.21	1000
DFM	0.23	965
Improved-GAN	0.67	934
ALI	1.23	967
BEGAN	0.19	999
MAD-GAN	0.074	1000
GMAN	0.57	929
DCGAN	0.18	980
MGAN	0.12	1000
SAGAN	0.095	1000
SNGAN	0.083	1000
HGAN	**0.078**	**1000**

Table 7.6: Inception scores on the CIFAR-10 and STL-10 datasets.

Model	CIFAR-10	STL-10
Real data	11.24 ± 0.16	26.08 ± 0.26
WGAN	3.82 ± 0.06	–
MIX+WGAN	4.04 ± 0.07	–
DFM	7.72 ± 0.13	8.51 ± 0.13
Improved-GAN	4.36 ± 0.04	–
ALI	5.34 ± 0.05	–
BEGAN	5.62	–
MAD-GAN	7.34	–
GMAN	6.00 ± 0.19	–
DCGAN	6.40 ± 0.05	7.54
MGAN	$\mathbf{8.33 \pm 0.10}$	$\mathbf{9.22 \pm 0.11}$
SAGAN	7.51 ± 0.15	8.61 ± 0.11
SNGAN	7.58 ± 0.12	8.79 ± 0.14
HGAN	7.46 ± 0.11	8.94 ± 0.13

Table 7.7: FIDs on CIFAR-10 and STL-10 (lower is better).

Model	DCGAN	DCGAN+TTUR [288]	WGAN-GP [289]	GAN-GP	MGAN	SAGAN	SNGAN	HGAN
CIFAR-10	37.7	36.9	40.2	37.7	26.7	26.3	**25.5**	26.1
STL-10	-	-	55.1	-	-	43.6	43.2	**42.1**

Table 7.8: Classification accuracies of using Defense-GAN and Defense-HGAN strategies on the MNIST dataset with L = 200 and R = 10.

Attack	No Attack (Defense-GAN)	Defense-GAN	No Attack (Defense-HGAN)	Defense-HGAN
FGSM ($\epsilon = 0.3$)	0.989	0.961	0.991	0.974
PGD	0.989	0.956	0.991	0.969
CW (l_2 norm)	0.989	0.945	0.991	0.965

Table 7.9: Classification accuracies of using Defense-GAN and Defense-HGAN strategies on the CIFAR-10 dataset with L = 200 and R = 10.

Attack	No Attack (Defense-GAN)	Defense-GAN	No Attack (Defense-HGAN)	Defense-HGAN
FGSM ($\epsilon = 0.3$)	0.763	0.684	0.794	0.741
PGD	0.763	0.671	0.794	0.738
CW (l_2 norm)	0.763	0.646	0.794	0.731

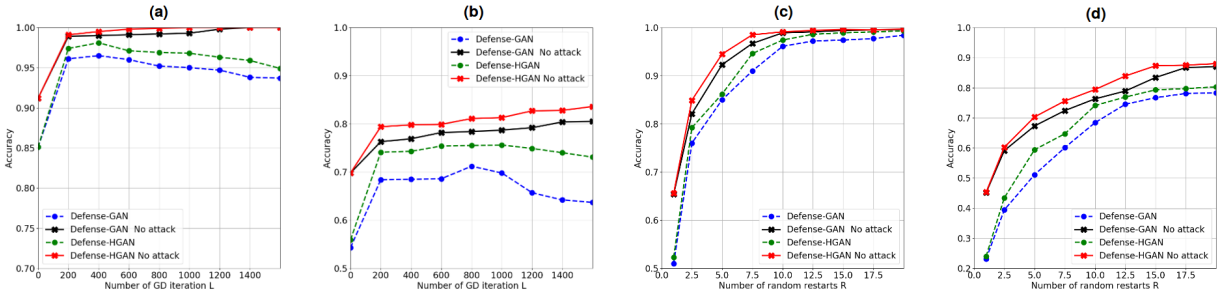


Figure 7.5: Classification accuracy of Defense-GAN and Defense-HGAN on the MNIST and CIFAR-10 datasets in the case of no attack and also under FGSM white-box attack with $\epsilon = 0.3$. (a) MNIST classification accuracy varying L (with R = 10). (b) CIFAR-10 classification accuracy varying L (with R = 10). (c) MNIST classification accuracy varying R (with L = 100). (d) CIFAR-10 classification accuracy varying R (with L = 100).

the KL divergence and it is the closest to the true data distribution. Also, in terms of the number of captured modes, our method as well as MGAN, MAD-GAN, WGAN, SNGAN and MIX+WGAN capture all the 1,000 modes in the Compositional MNIST experiment.

7.3.3 Real-world Datasets

In this section, the proposed HGAN framework is applied on more complicated real-world datasets to evaluate its effectiveness on more challenging large-scale image data.

Datasets.

We use two widely-adopted datasets, namely CIFAR-10 [285] and STL-10 [290]. CIFAR-10 dataset contains 50,000 training images with the resolution of 32×32 for 10 different classes: airplane, automobile, bird, cat, deer, dog, frog, horse, ship, and truck. STL-10 dataset subsampled from the ImageNet [291] and is more diverse database compared to CIFAR-10. This dataset composed of 100,000 images with the resolution of 96×96. For the sake of fair comparison with the baselines in [117], we follow the same procedure as in [292] to resize the STL-10 down to 48×48.

Evaluation Protocols.

For quantitative evaluation, we consider the Inception score which was introduced in [113]. This metric computes $exp(\mathbb{E}_x[D_{KL}(p(y|x)||p(y))])$, where $p(y|x)$ is the conditional label distribution for image x estimated by the reference Inception model. The metric rewards good and varied samples and is found to be well-correlated to human judgment. The code provided in [113], is used to compute the Inception score

for 10 partitions of 50,000 generated samples. For qualitative evaluation of the quality of images generated by our proposed HGAN framework, we show the samples generated by HGAN which are drawn randomly rather than cherry-picked.

Inception Results.

Table 7.6 shows the Inception scores obtained by our proposed HGAN method as well as the baselines. For the fair comparison, only models which are trained completely in an unsupervised manner without the label information are included in Table 7.6. Also, the reported results on STL-10 for DCGAN and D2GAN are based on the models trained on 32×32 resolution. Table 7.6 shows the superiority of our proposed HGAN compared to the other methods in the literature for both the STL-10 and CIFAR-10 datasets.

Image Generation.

For the qualitative assessment, we present samples which are randomly selected from the images generated by the proposed HGAN. It can be seen from Fig. 7.4 that the images generated by HGAN are visually recognizable images of cars, ships, trucks, birds, airplanes, dogs, and horses in the CIFAR-10 database. Moreover, in the case of the STL-10 dataset, HGAN is able to produce images including car, trucks, ships, airplanes, and different kinds of animals including horses, cats, monkeys, deers, and dogs with wider range of background such as sky, cloudy sky, sea, and forest. These visually appealing images confirms the diversity of the generated samples by HGAN.

7.3.4 Frechet Inception Distance results.

The main disadvantage of the inception score is that it does not compare the statistics of the synthetic samples and the real world ones. Therefore, we evaluate HGAN using the Frechet Inception Distance (FID) proposed in [288]. Table 7.7 compares the FIDs obtained by HGAN with baselines collected in [272, 279]. It should be noted that some methods in the literature use the Resnet [293] architecture. Here, for the fair comparison we show the results of different methods when using DCGAN architecture.

7.3.5 Ablation Study

In the previous sections, we examined the mode coverage of the proposed framework compared to the other baselines in three separate experiments. In order to show the effectiveness of our HGAN framework,

we perform another experiment with two different datasets, namely the MNIST and CIFAR-10 datasets. In this setup, we consider $G(z)$ within two complete separate training approaches. In the first approach, $G(z)$ is trained as a regular GAN such as a DCGAN, and in the second approach $G(z)$ is trained to mimic the autoregressive model's output with an adversarial training. We denote the first approach as DCGAN and the second approach as AutoGAN. Fig. 7.3 depicts the framework of AutoGAN. We compare the performance of these two networks with the proposed HGAN in terms of sample quality.

Table 7.3 and 7.2 show the highest Inception/MODE scores [113] of DCGAN, AutoGAN, and HGAN monitored during the training phase. The samples generated by each of the mentioned methods on the CIFAR-10 dataset is also shown in Fig. 7.2.

As it is illustrated in Table 7.3 and 7.2, HGAN outperforms both DCGAN and AutoGAN in terms of sample quality. One possible reason behind this is in HGAN, the addition of adversarially distillation of the data information from the autoregressive model (pixelCNN++) in the $G(z)$ objective function can stabilize its optimization, thus avoiding the mode collapse issue. Finally, the hybrid nature of the proposed method leads to a better performance for both datasets.

7.3.6 Comparison with WGAN in Defense Framework

Despite a very rich research work leading to very interesting GAN algorithms, it is still challenging to assess which algorithm performs better compared to others. In this experiment we evaluate the effectiveness of HGAN compared to WGAN in a defense scenario. We believe this could be another way of assessment for GAN frameworks.

Adversarial examples [294] are neural network inputs which are designed to force misclassification. These inputs often appear normal to humans while cause the neural network to make inaccurate predictions. Various defenses have been proposed to mitigate the effect of adversarial attacks [295, 296, 297]. In this experiment we use our proposed HGAN as a defense mechanism against three different white-box attacks: Fast Gradient-Sign Method (FGSM) [298], Carlini-Wagner (CW) attack (with l_2 norm) [299], and Projected Gradient Descent (PGD) [300]. For the fair comparison, we adopt the same set of experiment as Defense-GAN [295]. Instead of using WGAN, we use our proposed HGAN in Defense-GAN framework which we denote as Defense-HGAN. We also compare Defense-HGAN with Defense-GAN in the case of no attack. Table 7.8 and 7.9 show the classification performance of our method compared to Defense-GAN on the MNIST and CIFAR-10 datasets, respectively. It should be noted that the classification accuracy results on the

MNIST and CIFAR-10 is 0.994 and 0.886, respectively. We note that Defense-HGAN outperforms Defense-GAN which shows the superiority of our HGAN comparing to WGAN in Defense-GAN framework.

We also compared the effect of different numbers of iterations L and random restarts R for Defense-GAN and Defense-HGAN on the MNIST and CIFAR-10 datasets. Both methods need to look for an appropriate datapoint in the latent space which leads to generating an image closer to the input image. As it is shown in Fig. 7.5 classification performance of HGAN is better than Defense-GAN which means that HGAN could do a better job in capturing the data distribution compared to WGAN on the MNIST and CIFAR-10 datasets.

7.4 Conclusion

We have proposed a novel approach to address the mode collapse issue in GANs. Our idea is to design a hybrid model which tries to learn the distribution of data via a mixture of density estimating models utilizing an autoregressive model and an adversarial learning. For this purpose, we introduce a minimax game between a generator, an autoregressive model, and a discriminator to optimize the problem of minimizing the JSD between P_{data} and P_{model}. In our proposed HGAN, the generator is responsible to learn the autoregressive model output in addition to modeling the real data just like a regular GAN. Distillation of autoregressive model is beneficial for the HGAN since it also models the distribution of the same data but in an explicit way. It makes the generator to give more attention to the likelihood of the data and stabilize its optimization. This helps the proposed model to capture more data modes which leads to generating a more diversified set of images. Comprehensive study on MNIST and also more challenging real-world datasets show the effectiveness of our HGAN in covering data modes and avoiding mode collapse as well as generating diverse and visually appealing images.

References

[1] P. Mittal, A. Jain, G. Goswami, M. Vatsa, and R. Singh, "Composite sketch recognition using saliency and attribute feedback," *Information Fusion*, vol. 33, pp. 86–99, 2017.

[2] C. Peng, X. Gao, N. Wang, and J. Li, "Sparse graphical representation based discriminant analysis for heterogeneous face recognition," *arXiv preprint arXiv:1607.00137*, 2016.

[3] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database forstudying face recognition in unconstrained environments," 2008.

[4] M. Zhu, D. Shi, M. Zheng, and M. Sadiq, "Robust facial landmark detection via occlusion-adaptive deep networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3486–3496.

[5] and, C. C. Loy, and X. Tang, "Face alignment by coarse-to-fine shape searching," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015, pp. 4998–5006.

[6] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, "Learning deep representation for face alignment with auxiliary attributes," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 5, pp. 918–930, 2015.

[7] X. P. Burgos-Artizzu, P. Perona, and P. Dollár, "Robust face landmark estimation under occlusion," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 1513–1520.

[8] G. Ghiasi, C. C. Fowlkes, and C. Irvine, "Using segmentation to predict the absence of occluded parts." in *BMVC*, 2015, pp. 22–1.

[9] G. Ghiasi and C. C. Fowlkes, "Occlusion coherence: Localizing occluded faces with a hierarchical deformable part model," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 2385–2392.

[10] J. Su, Z. Wang, C. Liao, and H. Ling, "Efficient and accurate face alignment by global regression and cascaded local refinement," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 0–0.

[11] S. Ouyang, T. Hospedales, Y.-Z. Song, and X. Li, "A survey on heterogeneous face recognition: Sketch, infra-red, 3d and low-resolution," *arXiv preprint arXiv:1409.5114*, 2014.

[12] B. Klare and A. K. Jain, "Heterogeneous face recognition: Matching NIR to visible light images," in *Pattern Recognition (ICPR), 2010 20th International Conference on*. IEEE, 2010, pp. 1513–1516.

[13] F. Nicolo and N. A. Schmid, "Long range cross-spectral face recognition: matching SWIR against visible light images," *IEEE Transactions on Information Forensics and Security*, vol. 7, no. 6, pp. 1717–1726, 2012.

[14] S. Hu, N. J. Short, B. S. Riggan, C. Gordon, K. P. Gurton, M. Thielke, P. Gurram, and A. L. Chan, "A polarimetric thermal database for face recognition research," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2016, pp. 119–126.

[15] E. Gonzalez-Sosa, R. Vera-Rodriguez, J. Fierrez, and V. M. Patel, "Millimetre wave person recognition: Hand-crafted vs learned features," in *Identity, Security and Behavior Analysis (ISBA), 2017 IEEE International Conference on*. IEEE, 2017, pp. 1–7.

[16] D. Yi, R. Liu, R. Chu, Z. Lei, and S. Z. Li, "Face matching between near infrared and visible light images," in *International Conference on Biometrics*. Springer, 2007, pp. 523–530.

[17] T. Bourlai, N. Kalka, A. Ross, B. Cukic, and L. Hornak, "Cross-spectral face verification in the short wave infrared (SWIR) band," in *Pattern Recognition (ICPR), 2010 20th International Conference on*. IEEE, 2010, pp. 1343–1347.

[18] X. Wang and X. Tang, "Face photo-sketch synthesis and recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 11, pp. 1955–1967, 2009.

[19] H. Han, B. F. Klare, K. Bonnen, and A. K. Jain, "Matching composite sketches to face photos: A component-based approach," *IEEE Transactions on Information Forensics and Security*, vol. 8, no. 1, pp. 191–204, 2013.

[20] H. Kazemi, S. Soleymani, A. Dabouei, M. Iranmanesh, and N. M. Nasrabadi, "Attribute-centered loss for soft-biometrics guided face sketch-photo recognition," pp. 499–507, 2018.

[21] J. Choi, S. Hu, S. S. Young, and L. S. Davis, "Thermal to visible face recognition," MARYLAND UNIV COLLEGE PARK, Tech. Rep., 2012.

[22] T. Bourlai, A. Ross, C. Chen, and L. Hornak, "A study on using mid-wave infrared images for face recognition," in *Proc. SPIE*, vol. 8371, 2012, p. 83711K.

[23] B. F. Klare and A. K. Jain, "Heterogeneous face recognition using kernel prototype similarities," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 6, pp. 1410–1422, 2013.

[24] S. Hu, J. Choi, A. L. Chan, and W. R. Schwartz, "Thermal-to-visible face recognition using partial least squares," *JOSA A*, vol. 32, no. 3, pp. 431–442, 2015.

[25] D. Xu, W. Ouyang, E. Ricci, X. Wang, and N. Sebe, "Learning cross-modal deep representations for robust pedestrian detection," *arXiv preprint arXiv:1704.02431*, 2017.

[26] K. P. Gurton, A. J. Yuffa, and G. W. Videen, "Enhanced facial recognition for thermal imagery using polarimetric imaging," *Optics letters*, vol. 39, no. 13, pp. 3857–3859, 2014.

[27] N. Short, S. Hu, P. Gurram, K. Gurton, and A. Chan, "Improving cross-modal face recognition using polarimetric imaging," *Optics letters*, vol. 40, no. 6, pp. 882–885, 2015.

[28] N. Short, S. Hu, P. Gurram, and K. Gurton, "Exploiting polarization-state information for cross-spectrum face recognition," in *Biometrics Theory, Applications and Systems (BTAS), 2015 IEEE 7th International Conference on*. IEEE, 2015, pp. 1–6.

[29] C. Peng, N. Wang, X. Gao, and J. Li, "Face recognition from multiple stylistic sketches: Scenarios, datasets, and evaluation," in *European Conference on Computer Vision*. Springer, 2016, pp. 3–18.

[30] L. Best-Rowden, H. Han, C. Otto, B. F. Klare, and A. K. Jain, "Unconstrained face recognition: Identifying a person of interest from a media collection," *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 12, pp. 2144–2157, 2014.

[31] L. Gibson, *Forensic art essentials: a manual for law enforcement artists*. Academic Press, 2010.

[32] S. Ouyang, T. M. Hospedales, Y.-Z. Song, and X. Li, "Forgetmenot: Memory-aware forensic facial sketch matching," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 5571–5579.

[33] Y. Wang, L. Zhang, Z. Liu, G. Hua, Z. Wen, Z. Zhang, and D. Samaras, "Face relighting from a single image under arbitrary unknown lighting conditions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 11, pp. 1968–1984, 2009.

[34] B. Klare and A. K. Jain, "Sketch-to-photo matching: a feature-based approach," in *Biometric Technology for Human Identification VII*, vol. 7667. International Society for Optics and Photonics, 2010, p. 766702.

[35] H. S. Bhatt, S. Bharadwaj, R. Singh, and M. Vatsa, "Memetically optimized mcwld for matching sketches with digital face images," *IEEE Transactions on Information Forensics and Security*, vol. 7, no. 5, pp. 1522–1535, 2012.

[36] C. Galea and R. A. Farrugia, "Face photo-sketch recognition using local and global texture descriptors," in *Signal Processing Conference (EUSIPCO), 2016 24th European*. IEEE, 2016, pp. 2240–2244.

[37] D.-A. Huang and Y.-C. F. Wang, "Coupled dictionary and feature space learning with applications to cross-domain image synthesis and recognition," in *Computer Vision (ICCV), 2013 IEEE International Conference on*. IEEE, 2013, pp. 2496–2503.

[38] A. Dantcheva, P. Elia, and A. Ross, "What else does your biometric data reveal? a survey on soft biometrics," vol. 11, no. 3. IEEE, 2016, pp. 441–467.

[39] A. Jain, R. Bolle, and S. Pankanti, "Biometrics: Personal identification in networked security," *Bolle, and S. Pankanti, Eds.: Kluwer Academic Publishers*, 1999.

[40] L. Hong, A. K. Jain, and S. Pankanti, "Can multibiometrics improve performance?" *Proceedings AutoID*, vol. 99, pp. 59–64, 1999.

[41] V. Vapnik and A. Vashist, "A new learning paradigm: Learning using privileged information," *Neural networks*, vol. 22, no. 5, pp. 544–557, 2009.

[42] V. Sharmanska, N. Quadrianto, and C. H. Lampert, "Learning to rank using privileged information," *IEEE International Conference on Computer Vision (ICCV)*, pp. 825–832, 2013.

[43] Y. Sun, X. Wang, and X. Tang, "Deep convolutional network cascade for facial point detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 3476–3483.

[44] S. Zhu, C. Li, C. C. Loy, and X. Tang, "Unconstrained face alignment via cascaded compositional learning," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016, pp. 3409–3417.

[45] X. Yu, F. Zhou, and M. Chandraker, "Deep deformation network for object landmark localization," in *Computer Vision – ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham: Springer International Publishing, 2016, pp. 52–70.

[46] J. Lv, X. Shao, J. Xing, C. Cheng, and X. Zhou, "A deep regression architecture with two-stage re-initialization for high performance facial landmark detection," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017, pp. 3691–3700.

[47] S. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016, pp. 4724–4732.

[48] A. Bulat and G. Tzimiropoulos, "How far are we from solving the 2d amp; 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks)," in *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct 2017, pp. 1021–1030.

[49] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *Computer Vision – ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham: Springer International Publishing, 2016, pp. 483–499.

[50] S. Reed, K. Sohn, Y. Zhang, and H. Lee, "Learning to disentangle factors of variation with manifold interaction," in *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ser. ICML'14. JMLR.org, 2014, pp. II–1431–II–1439. [Online]. Available: http://dl.acm.org/citation.cfm?id=3044805.3045052

[51] S. E. Reed, Y. Zhang, Y. Zhang, and H. Lee, "Deep visual analogy-making," in *Advances in Neural Information Processing Systems 28*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds. Curran Associates, Inc., 2015, pp. 1252–1260. [Online]. Available: http://papers.nips.cc/paper/5845-deep-visual-analogy-making.pdf

[52] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, Oct 2016.

[53] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct 2017, pp. 2980–2988.

[54] L. Theis, A. v. d. Oord, and M. Bethge, "A note on the evaluation of generative models," *arXiv preprint arXiv:1511.01844*, 2015.

[55] S. Nowozin, B. Cseke, and R. Tomioka, "f-gan: Training generative neural samplers using variational divergence minimization," in *Advances in Neural Information Processing Systems*, 2016, pp. 271–279.

[56] T. Nguyen, T. Le, H. Vu, and D. Phung, "Dual discriminator generative adversarial nets," in *Advances in Neural Information Processing Systems*, 2017, pp. 2670–2680.

[57] D. Bang and H. Shim, "Mggan: Solving mode collapse using manifold guided training," *arXiv preprint arXiv:1804.04391*, 2018.

[58] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein gan," *arXiv preprint arXiv:1701.07875*, 2017.

[59] F. Huszár, "How (not) to train your generative model: Scheduled sampling, likelihood, adversary?" *arXiv preprint arXiv:1511.05101*, 2015.

[60] Y. He, S. Xiang, C. Kang, J. Wang, and C. Pan, "Cross-modal retrieval via deep and bidirectional representation learning," *IEEE Transactions on Multimedia*, vol. 18, no. 7, pp. 1363–1377, 2016.

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

[62] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1. IEEE, 2005, pp. 539–546.

[63] Y. Sun, Y. Chen, X. Wang, and X. Tang, "Deep learning face representation by joint identification-verification," in *Advances in neural information processing systems*, 2014, pp. 1988–1996.

[64] Y. Sun, D. Liang, X. Wang, and X. Tang, "Deepid3: Face recognition with very deep neural networks," *arXiv preprint arXiv:1502.00873*, 2015.

[65] Y. Sun, X. Wang, and X. Tang, "Deep learning face representation from predicting 10,000 classes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1891–1898.

[66] ——, "Deeply learned face representations are sparse, selective, and robust," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 2892–2900.

[67] D. Chen, X. Cao, L. Wang, F. Wen, and J. Sun, "Bayesian face revisited: A joint formulation," *Computer Vision–ECCV 2012*, pp. 566–579, 2012.

[68] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 815–823.

[69] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," Technical Report 07-49, University of Massachusetts, Amherst, Tech. Rep., 2007.

[70] B. S. Riggan, N. J. Short, and S. Hu, "Optimal feature learning and discriminative framework for polarimetric thermal to visible face recognition," in *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*. IEEE, 2016, pp. 1–7.

[71] A. J. Yuffa, K. P. Gurton, and G. Videen, "Three-dimensional facial recognition using passive long-wavelength infrared polarimetric imaging," *Applied optics*, vol. 53, no. 36, pp. 8514–8521, 2014.

[72] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1. IEEE, 2005, pp. 886–893.

[73] H. Wold, "Estimation of principal components and related models by iterative least squares," *Multivariate analysis*, pp. 391–420, 1966.

[74] B. S. Riggan, C. Reale, and N. M. Nasrabadi, "Coupled auto-associative neural networks for heterogeneous face recognition," *IEEE Access*, vol. 3, pp. 1620–1632, 2015.

[75] M. S. Sarfraz and R. Stiefelhagen, "Deep perceptual mapping for thermal to visible face recognition," *arXiv preprint arXiv:1507.02879*, 2015.

[76] B. S. Riggan, N. J. Short, S. Hu, and H. Kwon, "Estimation of visible spectrum faces from polarimetric thermal faces," in *Biometrics Theory, Applications and Systems (BTAS), 2016 IEEE 8th International Conference on*. IEEE, 2016, pp. 1–7.

[77] B. S. Riggan, N. J. Short, and S. Hu, "Thermal to visible synthesis of face images using multiple regions," *arXiv preprint arXiv:1803.07599*, 2018.

[78] H. Zhang, V. M. Patel, B. S. Riggan, and S. Hu, "Generative adversarial network-based synthesis of visible faces from polarimetrie thermal faces," in *Biometrics (IJCB), 2017 IEEE International Joint Conference on*. IEEE, 2017, pp. 100–107.

[79] S. Motiian, Q. Jones, S. Iranmanesh, and G. Doretto, "Few-shot adversarial domain adaptation," in *Advances in Neural Information Processing Systems*, 2017, pp. 6673–6683.

[80] A. Dabouei, H. Kazemi, S. M. Iranmanesh, J. Dawson, and N. M. Nasrabadi, "Fingerprint distortion rectification using deep convolutional neural networks," pp. 1–8, 2018.

[81] S. Soleymani, A. Dabouei, H. Kazemi, J. Dawson, and N. M. Nasrabadi, "Multi-level feature abstraction from convolutional neural networks for multimodal biometric identification," in *24th International Conference on Pattern Recognition (ICPR)*, 2018.

[82] S. Soleymani, A. Torfi, J. Dawson, and N. M. Nasrabadi, "Generalized bilinear deep convolutional neural networks for multimodal biometric identification," in *IEEE International Conference on Image Processing (ICIP)*, 2018.

[83] S. M. Iranmanesh, A. Dabouei, H. Kazemi, and N. M. Nasrabadi, "Deep cross polarimetric thermal-to-visible face recognition," in *2018 International Conference on Biometrics (ICB)*. IEEE, 2018, pp. 166–173.

[84] A. Torfi, S. M. Iranmanesh, N. Nasrabadi, and J. Dawson, "3d convolutional neural networks for cross audio-visual matching recognition," *IEEE Access*, vol. 5, pp. 22 081–22 091, 2017.

[85] A. Broumand, M. S. Esfahani, B.-J. Yoon, and E. R. Dougherty, "Discrete optimal bayesian classification with error-conditioned sequential sampling," *Pattern Recognition*, vol. 48, no. 11, pp. 3766–3782, 2015.

[86] D. Alhelal, K. A. Aboalayon, M. Daneshzand, and M. Faezipour, "Fpga-based denoising and beat detection of the ecg signal," in *Systems, Applications and Technology Conference (LISAT), 2015 IEEE Long Island*. IEEE, 2015, pp. 1–5.

[87] C. Galea and R. A. Farrugia, "Forensic face photo-sketch recognition using a deep learning-based architecture," *IEEE Signal Processing Letters*, vol. 24, no. 11, pp. 1586–1590, 2017.

[88] P. Mittal, M. Vatsa, and R. Singh, "Composite sketch recognition via deep network-a transfer learning approach," in *Biometrics (ICB), 2015 International Conference on*. IEEE, 2015, pp. 251–256.

[89] B. F. Klare, S. Klum, J. C. Klontz, E. Taborsky, T. Akgul, and A. K. Jain, "Suspect identification based on descriptive facial attributes," in *Biometrics (IJCB), 2014 IEEE International Joint Conference on*. IEEE, 2014, pp. 1–8.

[90] B. Klare, Z. Li, and A. K. Jain, "Matching forensic sketches to mug shot photos," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 3, pp. 639–646, 2011.

[91] S. Ouyang, T. Hospedales, Y.-Z. Song, and X. Li, "Cross-modal face matching: beyond viewed sketches," in *Asian Conference on Computer Vision*. Springer, 2014, pp. 210–225.

[92] S. Motiian, M. Piccirilli, D. A. Adjeroh, and G. Doretto, "Information bottleneck learning using privileged information for visual recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[93] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan, "A theory of learning from different domains," *Machine learning*, vol. 79, no. 1-2, pp. 151–175, 2010.

[94] K. Saenko, B. Kulis, M. Fritz, and T. Darrell, "Adapting visual category models to new domains," *European conference on computer vision*, pp. 213–226, 2010.

[95] J. D. Farquhar, D. R. Hardoon, H. Meng, J. Shawe-Taylor, and S. Szedmak, "Two view learning: SVM-2K, theory and practice," in *Advances in Neural Information Processing Systems (NIPS)*, vol. 3, no. 5, 2005, p. 6.

[96] P. Gehler and S. Nowozin, "On feature combination for multiclass object classification," *IEEE International Conference on Computer Vision (ICCV)*, pp. 221–228, 2009.

[97] C. Xu, D. Tao, and C. Xu, "Large-margin multi-view information bottleneck," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 8, pp. 1559–1572, 2014.

[98] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, "Active shape models-their training and application," *Computer vision and image understanding*, vol. 61, no. 1, pp. 38–59, 1995.

[99] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, no. 6, pp. 681–685, 2001.

[100] J. Zhang, S. Shan, M. Kan, and X. Chen, "Coarse-to-fine auto-encoder networks (cfan) for real-time face alignment," in *European Conference on Computer Vision*. Springer, 2014, pp. 1–16.

[101] X. Dong, J. Huang, Y. Yang, and S. Yan, "More is less: A more complicated network with less inference complexity," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5840–5848.

[102] A. Toshev and C. Szegedy, "Deeppose: Human pose estimation via deep neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1653–1660.

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

[104] P. Dollár, P. Welinder, and P. Perona, "Cascaded pose regression," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, 2010, pp. 1078–1085.

[105] X. Peng, R. S. Feris, X. Wang, and D. N. Metaxas, "A recurrent encoder-decoder network for sequential face alignment," in *European conference on computer vision*. Springer, 2016, pp. 38–56.

[106] S. Ren, X. Cao, Y. Wei, and J. Sun, "Face alignment at 3000 fps via regressing local binary features," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1685–1692.

[107] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[108] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," *arXiv preprint arXiv:1312.6199*, 2013.

[109] A. v. d. Oord, N. Kalchbrenner, and K. Kavukcuoglu, "Pixel recurrent neural networks," *arXiv preprint arXiv:1601.06759*, 2016.

[110] T. Salimans, A. Karpathy, X. Chen, and D. P. Kingma, "Pixelcnn++: Improving the pixelcnn with discretized logistic mixture likelihood and other modifications," *arXiv preprint arXiv:1701.05517*, 2017.

[111] G. Ostrovski, M. G. Bellemare, A. v. d. Oord, and R. Munos, "Count-based exploration with neural density models," *arXiv preprint arXiv:1703.01310*, 2017.

[112] A. Grover, M. Dhar, and S. Ermon, "Flow-gan: Combining maximum likelihood and adversarial learning in generative models," *arXiv preprint arXiv:1705.08868*, 2017.

[113] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," in *Advances in Neural Information Processing Systems*, 2016, pp. 2234–2242.

[114] L. Metz, B. Poole, D. Pfau, and J. Sohl-Dickstein, "Unrolled generative adversarial networks," *arXiv preprint arXiv:1611.02163*, 2016.

[115] Q. Hoang, T. D. Nguyen, T. Le, and D. Phung, "Multi-generator gernerative adversarial nets," *arXiv preprint arXiv:1708.02556*, 2017.

[116] R. Wang, A. Cully, H. J. Chang, and Y. Demiris, "Magan: Margin adaptation for generative adversarial networks," *arXiv preprint arXiv:1704.03817*, 2017.

[117] D. Warde-Farley and Y. Bengio, "Improving generative adversarial networks with denoising feature matching," 2016.

[118] Y. Yazici, K.-H. Yap, and S. Winkler, "Autoregressive generative adversarial networks," 2018.

[119] C. Reale, N. M. Nasrabadi, H. Kwon, and R. Chellappa, "Seeing the forest from the trees: A holistic approach to near-infrared heterogeneous face recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2016, pp. 54–62.

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

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

[122] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks." in *CVPR*, vol. 1, no. 2, 2017, p. 3.

[123] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[124] S. M. Iranmanesh, B. Riggan, S. Hu, and N. M. Nasrabadi, "Coupled generative adversarial network for heterogeneous face recognition," *Image and Vision Computing*, vol. 94, p. 103861, 2020.

[125] H. Zhang and V. M. Patel, "Densely connected pyramid dehazing network," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[126] S. Jégou, M. Drozdzal, D. Vazquez, A. Romero, and Y. Bengio, "The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on*. IEEE, 2017, pp. 1175–1183.

[127] Y. Zhu and S. Newsam, "Densenet for dense flow," in *Image Processing (ICIP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 790–794.

[128] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2881–2890.

[129] H. Zhang, K. Dana, J. Shi, Z. Zhang, X. Wang, A. Tyagi, and A. Agrawal, "Context encoding for semantic segmentation," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[130] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European Conference on Computer Vision*. Springer, 2016, pp. 694–711.

[131] H. Zhang and V. M. Patel, "Density-aware single image de-raining using a multi-stream dense network," *arXiv preprint arXiv:1802.07412*, 2018.

[132] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *ICLR*, 2014.

[133] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[134] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. Ieee, 2009, pp. 248–255.

[135] "https://wsri.wright.edu/."

[136] X. Chen, P. J. Flynn, and K. W. Bowyer, "Visible-light and infrared face recognition," *Proceedings of the Workshop on Multimodal User Authentication*, pp. 48–55, 2003.

[137] K. A. Byrd, "Preview of the newly acquired nvesd-arl multimodal face database," in *Proc. SPIE*, vol. 8734, 2013, p. 34.

[138] S. Li, D. Yi, Z. Lei, and S. Liao, "The casia nir-vis 2.0 face database," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2013, pp. 348–353.

[139] S. Z. Li, Z. Lei, and M. Ao, "The hfb face database for heterogeneous face biometrics research," in *Computer Vision and Pattern Recognition Workshops, 2009. CVPR Workshops 2009. IEEE Computer Society Conference on*. IEEE, 2009, pp. 1–8.

[140] C. Reale, N. M. Nasrabadi, and R. Chellappa, "Coupled dictionaries for thermal to visible face recognition." in *ICIP*, 2014, pp. 328–332.

[141] C. Reale, H. Lee, and H. Kwon, "Deep heterogeneous face recognition networks based on cross-modal distillation and an equitable distance metric," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 32–38.

[142] J. Lu, V. E. Liong, X. Zhou, and J. Zhou, "Learning compact binary face descriptor for face recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 10, pp. 2041–2056, 2015.

[143] S. Saxena and J. Verbeek, "Heterogeneous face recognition with cnns," in *European Conference on Computer Vision*. Springer, 2016, pp. 483–491.

[144] D. Yi, Z. Lei, and S. Z. Li, "Shared representation learning for heterogenous face recognition," in *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, vol. 1. IEEE, 2015, pp. 1–7.

[145] X. Liu, L. Song, X. Wu, and T. Tan, "Transferring deep representation for nir-vis heterogeneous face recognition," in *Biometrics (ICB), 2016 International Conference on*. IEEE, 2016, pp. 1–8.

[146] Z. Lei, M. Pietikäinen, and S. Z. Li, "Learning discriminant face descriptor," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 2, pp. 289–302, 2014.

[147] J.-Y. Zhu, W.-S. Zheng, J.-H. Lai, and S. Z. Li, "Matching NIR face to VIS face using transduction," *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 3, pp. 501–514, 2014.

[148] A. H. Abdulnabi, G. Wang, J. Lu, and K. Jia, "Multi-task CNN model for attribute prediction," *IEEE Transactions on Multimedia*, vol. 17, no. 11, pp. 1949–1959, 2015.

[149] S. M. Iranmanesh, H. Kazemi, S. Soleymani, A. Dabouei, and N. M. Nasrabadi, "Deep sketch-photo face recognition assisted by facial attributes," in *2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS)*. IEEE, 2018, pp. 1–10.

[150] Y. Zhong, J. Sullivan, and H. Li, "Face attribute prediction using off-the-shelf CNN features," in *Biometrics (ICB), 2016 International Conference on*. IEEE, 2016, pp. 1–7.

[151] S. M. Iranmanesh and N. M. Nasrabadi, "Attribute-guided deep polarimetric thermal-to-visible face recognition," in *2019 International Conference on Biometrics (ICB)*. IEEE, 2019, pp. 1–8.

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

[153] O. M. Parkhi, A. Vedaldi, A. Zisserman *et al.*, "Deep face recognition." in *BMVC*, vol. 1, no. 3, 2015, p. 6.

[154] X. Tang and X. Wang, "Face sketch synthesis and recognition," in *Computer vision, 2003. proceedings. ninth ieee international conference on.* IEEE, 2003, pp. 687–694.

[155] H. S. Bhatt, S. Bharadwaj, R. Singh, and M. Vatsa, "Memetic approach for matching sketches with digital face images," Tech. Rep., 2012.

[156] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 3730–3738.

[157] H. Winnemöller, J. E. Kyprianidis, and S. C. Olsen, "Xdog: an extended difference-of-gaussians compendium including advanced image stylization," *Computers & Graphics*, vol. 36, no. 6, pp. 740–753, 2012.

[158] E. M. Rudd, M. Günther, and T. E. Boult, "Moon: A mixed objective optimization network for the recognition of facial attributes," in *European Conference on Computer Vision.* Springer, 2016, pp. 19–35.

[159] "Biometrics and identification innovation center, wvu multi- modal dataset. available at http://biic.wvu.edu/,."

[160] A. P. Founds, N. Orlans, W. Genevieve, and C. I. Watson, "Nist special databse 32-multiple encounter dataset ii (meds-ii)," Tech. Rep., 2011.

[161] P. Mittal, A. Jain, G. Goswami, R. Singh, and M. Vatsa, "Recognizing composite sketches with digital face images via ssd dictionary," in *Biometrics (IJCB), 2014 IEEE International Joint Conference on.* IEEE, 2014, pp. 1–6.

[162] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," *CVPR*, 2017.

[163] A. K. Jain and R. C. Dubes, *Algorithms for clustering data.* Prentice-Hall, Inc., 1988.

[164] S. Samangooei, B. Guo, and M. S. Nixon, "The use of semantic human description as a soft biometric," in *Biometrics: Theory, Applications and Systems, 2008. BTAS 2008. 2nd IEEE International Conference on.* IEEE, 2008, pp. 1–7.

[165] K. Ricanek and T. Tesafaye, "Morph: A longitudinal image database of normal adult age-progression," in *Automatic Face and Gesture Recognition, 2006. FGR 2006. 7th International Conference on.* IEEE, 2006, pp. 341–345.

[166] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-pie," *Image and Vision Computing*, vol. 28, no. 5, pp. 807–813, 2010.

[167] A. K. Jain, S. C. Dass, and K. Nandakumar, "Soft biometric traits for personal recognition systems," in *Biometric Authentication.* Springer, 2004, pp. 731–738.

[168] W. J. Scheirer, N. Kumar, K. Ricanek, P. N. Belhumeur, and T. E. Boult, "Fusing with context: a bayesian approach to combining descriptive attributes," in *International Joint Conference on Biometrics (IJCB).* IEEE, 2011, pp. 1–8.

[169] M. C. D. C. Abreu and M. Fairhurst, "Enhancing identity prediction using a novel approach to combining hard-and soft-biometric information," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 41, no. 5, pp. 599–607, 2011.

[170] U. Park and A. K. Jain, "Face matching and retrieval using soft biometrics," *IEEE Transactions on Information Forensics and Security*, vol. 5, no. 3, pp. 406–415, 2010.

[171] M. B. Blaschko and C. H. Lampert, "Correlational spectral clustering," in *2008 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2008, pp. 1–8.

[172] N. Quadrianto and C. H. Lampert, "Learning multi-view neighborhood preserving projections," in *Proceedings of the 28th international conference on machine learning (ICML)*, 2011, pp. 425–432.

[173] J. Donahue and K. Grauman, "Annotator rationales for visual recognition," *IEEE International Conference on Computer Vision (ICCV)*, pp. 1395–1402, 2011.

[174] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, and B. Scholkopf, "Support vector machines," *IEEE Intelligent Systems and their applications*, vol. 13, no. 4, pp. 18–28, 1998.

[175] J. Feyereisl, S. Kwak, J. Son, and B. Han, "Object localization based on structural SVM using privileged information," *Advances in Neural Information Processing Systems (NIPS)*, pp. 208–216, 2014.

[176] S. Fouad, P. Tino, S. Raychaudhury, and P. Schneider, "Incorporating privileged information through metric learning," *IEEE transactions on neural networks and learning systems*, vol. 24, no. 7, pp. 1086–1098, 2013.

[177] C. M. Christoudias, R. Urtasun, M. Salzmann, and T. Darrell, "Learning to recognize objects from unseen modalities," *European Conference on Computer Vision (ECCV)*, pp. 677–691, 2010.

[178] L. Chen, W. Li, and D. Xu, "Recognizing RGB images by learning from RGB-D data," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1418–1425, 2014.

[179] J. Hoffman, S. Gupta, and T. Darrell, "Learning with side information through modality hallucination," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 826–834, 2016.

[180] R. Caruana, "Multitask learning," in *Learning to learn*. Springer, 1998, pp. 95–133.

[181] B. Hariharan, L. Zelnik-Manor, M. Varma, and S. Vishwanathan, "Large scale max-margin multi-label classification with priors," in *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*. Citeseer, 2010, pp. 423–430.

[182] P. Rai and H. Daume III, "Infinite predictor subspace models for multitask learning," in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, 2010, pp. 613–620.

[183] P. Gong, J. Ye, and C. Zhang, "Robust multi-task feature learning," in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2012, pp. 895–903.

[184] A. Kumar and H. Daume III, "Learning task grouping and overlap in multi-task learning," *International Conference on Machine Learning (ICML)*, 2012.

[185] Q. Zhou, G. Wang, K. Jia, and Q. Zhao, "Learning to share latent tasks for action recognition," in *Computer Vision (ICCV), 2013 IEEE International Conference on*. IEEE, 2013, pp. 2264–2271.

[186] H. Han, A. K. Jain, S. Shan, and X. Chen, "Heterogeneous face attribute estimation: A deep multi-task learning approach," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2018.

[187] J. Zhu, S. Liao, D. Yi, Z. Lei, and S. Z. Li, "Multi-label CNN based pedestrian attribute learning for soft biometrics," in *International Conference on Biometrics (ICB)*. IEEE, 2015, pp. 535–540.

[188] L. An, X. Chen, M. Kafai, S. Yang, and B. Bhanu, "Improving person re-identification by soft biometrics based reranking," in *Seventh International Conference on Distributed Smart Cameras (ICDSC)*. IEEE, 2013, pp. 1–6.

[189] T. Kohonen, "Learning vector quantization," in *Self-Organizing Maps*. Springer, 1995, pp. 175–189.

[190] I. Jolliffe, *Principal component analysis*. Wiley Online Library, 2002.

[191] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[192] X. Geng, Z.-H. Zhou, and K. Smith-Miles, "Automatic age estimation based on facial aging patterns," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 29, no. 12, pp. 2234–2240, 2007.

[193] S. Motiian, M. Piccirilli, D. A. Adjeroh, and G. Doretto, "Information bottleneck learning using privileged information for visual recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[194] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan, "A theory of learning from different domains," *Machine learning*, vol. 79, no. 1-2, pp. 151–175, 2010.

[195] K. Saenko, B. Kulis, M. Fritz, and T. Darrell, "Adapting visual category models to new domains," *European conference on computer vision*, pp. 213–226, 2010.

[196] V. Vapnik and A. Vashist, "A new learning paradigm: Learning using privileged information," *Neural networks*, vol. 22, no. 5, pp. 544–557, 2009.

[197] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, and B. Scholkopf, "Support vector machines," *IEEE Intelligent Systems and their applications*, vol. 13, no. 4, pp. 18–28, 1998.

[198] J. Feyereisl, S. Kwak, J. Son, and B. Han, "Object localization based on structural SVM using privileged information," *Advances in Neural Information Processing Systems (NIPS)*, pp. 208–216, 2014.

[199] V. Sharmanska, N. Quadrianto, and C. H. Lampert, "Learning to rank using privileged information," *IEEE International Conference on Computer Vision (ICCV)*, pp. 825–832, 2013.

[200] S. Fouad, P. Tino, S. Raychaudhury, and P. Schneider, "Incorporating privileged information through metric learning," *IEEE transactions on neural networks and learning systems*, vol. 24, no. 7, pp. 1086–1098, 2013.

[201] Y. Sun, X. Wang, and X. Tang, "Deep learning face representation from predicting 10,000 classes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 1891–1898.

[202] W. Liu, Y. Wen, Z. Yu, and M. Yang, "Large-margin softmax loss for convolutional neural networks." in *ICML*, vol. 2, no. 3, 2016, p. 7.

[203] Y. Sun, Y. Chen, X. Wang, and X. Tang, "Deep learning face representation by joint identification-verification," in *Advances in neural information processing systems*, 2014, pp. 1988–1996.

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

[205] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *European conference on computer vision*. Springer, 2016, pp. 499–515.

[206] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "Sphereface: Deep hypersphere embedding for face recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 212–220.

[207] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu, "Cosface: Large margin cosine loss for deep face recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5265–5274.

[208] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4690–4699.

[209] I. Kemelmacher-Shlizerman, S. M. Seitz, D. Miller, and E. Brossard, "The megaface benchmark: 1 million faces for recognition at scale," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4873–4882.

[210] L. Wolf, T. Hassner, and I. Maoz, "Face recognition in unconstrained videos with matched background similarity," in *CVPR 2011*. IEEE, 2011, pp. 529–534.

[211] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1116–1124.

[212] Z. Zheng, L. Zheng, and Y. Yang, "Unlabeled samples generated by gan improve the person re-identification baseline in vitro," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 3754–3762.

[213] S. M. Iranmanesh, A. Dabouei, and N. M. Nasrabadi, "Attribute adaptive margin softmax loss using privileged information," *arXiv preprint arXiv:2009.01972*, 2020.

[214] Y. Duan, J. Lu, and J. Zhou, "Uniformface: Learning deep equidistributed representation for face recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3415–3424.

[215] H. Liu, X. Zhu, Z. Lei, and S. Z. Li, "Adaptiveface: Adaptive margin and sampling for face recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 11 947–11 956.

[216] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

[217] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016.

[218] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, "Ms-celeb-1m: A dataset and benchmark for large-scale face recognition," in *European conference on computer vision*. Springer, 2016, pp. 87–102.

[219] J. Deng, Y. Zhou, and S. Zafeiriou, "Marginal loss for deep face recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 60–68.

[220] H.-W. Ng and S. Winkler, "A data-driven approach to cleaning large face datasets," in *2014 IEEE international conference on image processing (ICIP)*. IEEE, 2014, pp. 343–347.

[221] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," *IEEE International Conference on Computer Vision (ICCV)*, pp. 3730–3738, 2015.

[222] G. B. Huang and E. Learned-Miller, "Labeled faces in the wild: Updates and new reporting procedures," *Dept. Comput. Sci., Univ. Massachusetts Amherst, Amherst, MA, USA, Tech. Rep*, pp. 14–003, 2014.

[223] Y. Lin, L. Zheng, Z. Zheng, Y. Wu, Z. Hu, C. Yan, and Y. Yang, "Improving person re-identification by attribute and identity learning," *Pattern Recognition*, vol. 95, pp. 151–161, 2019.

[224] G. Wang, Y. Yuan, X. Chen, J. Li, and X. Zhou, "Learning discriminative features with multiple granularities for person re-identification," in *Proceedings of the 26th ACM international conference on Multimedia*, 2018, pp. 274–282.

[225] Y. Sun, C. Cheng, Y. Zhang, C. Zhang, L. Zheng, Z. Wang, and Y. Wei, "Circle loss: A unified perspective of pair similarity optimization," *arXiv preprint arXiv:2002.10857*, 2020.

[226] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, "Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 480–496.

[227] Z. Zheng, X. Yang, Z. Yu, L. Zheng, Y. Yang, and J. Kautz, "Joint discriminative and generative learning for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2138–2147.

[228] C.-P. Tay, S. Roy, and K.-H. Yap, "Aanet: Attribute attention network for person re-identifications," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7134–7143.

[229] F. Wang, J. Cheng, W. Liu, and H. Liu, "Additive margin softmax for face verification," *IEEE Signal Processing Letters*, vol. 25, no. 7, pp. 926–930, 2018.

[230] F. Liu, D. Zeng, Q. Zhao, and X. Liu, "Joint face alignment and 3d face reconstruction," in *European Conference on Computer Vision*. Springer, 2016, pp. 545–560.

[231] Y. Wu, C. Gou, and Q. Ji, "Simultaneous facial landmark detection, pose and deformation estimation under facial occlusion," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3471–3480.

[232] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, and M. Nießner, "Face2face: Real-time face capture and reenactment of rgb videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2387–2395.

[233] X. Zhu, Z. Lei, J. Yan, D. Yi, and S. Z. Li, "High-fidelity pose and expression normalization for face recognition in the wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 787–796.

[234] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," *arXiv preprint arXiv:1511.07122*, 2015.

[235] S. Honari, J. Yosinski, P. Vincent, and C. Pal, "Recombinator networks: Learning coarse-to-fine feature aggregation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 5743–5752.

[236] J. Zhao, M. Mathieu, R. Goroshin, and Y. Lecun, "Stacked what-where auto-encoders," *arXiv preprint arXiv:1506.02351*, 2015.

[237] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik, "Hypercolumns for object segmentation and fine-grained localization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 447–456.

[238] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "300 faces in-the-wild challenge: The first facial landmark localization challenge," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2013, pp. 397–403.

[239] M. Köstinger, P. Wohlhart, P. M. Roth, and H. Bischof, "Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization," in *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, Nov 2011, pp. 2144–2151.

[240] S. M. Iranmanesh, A. Dabouei, S. Soleymani, H. Kazemi, and N. Nasrabadi, "Robust facial landmark detection via aggregation on geometrically manipulated faces," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2020, pp. 330–340.

[241] I. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *International Conference on Learning Representations*, 2015. [Online]. Available: http://arxiv.org/abs/1412.6572

[242] C. Xiao, J.-Y. Zhu, B. Li, W. He, M. Liu, and D. X. Song, "Spatially transformed adversarial examples," *CoRR*, vol. abs/1801.02612, 2018.

[243] D. C. Liu and J. Nocedal, "On the limited memory bfgs method for large scale optimization," *Mathematical Programming*, vol. 45, no. 1, pp. 503–528, Aug 1989.

[244] M. O. Irfanoglu, B. Gokberk, and L. Akarun, "3d shape-based face recognition using automatically registered facial surfaces," in *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, vol. 4, Aug 2004, pp. 183–186 Vol.4.

[245] A. Dabouei, S. Soleymani, J. M. Dawson, and N. M. Nasrabadi, "Fast geometrically-perturbed adversarial faces," *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1979–1988, 2019.

[246] F. L. Bookstein, "Principal warps: thin-plate splines and the decomposition of deformations," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11, no. 6, pp. 567–585, June 1989.

[247] M. Jaderberg, K. Simonyan, A. Zisserman, and k. kavukcuoglu, "Spatial transformer networks," in *Advances in Neural Information Processing Systems 28*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds. Curran Associates, Inc., 2015, pp. 2017–2025. [Online]. Available: http://papers.nips.cc/paper/5854-spatial-transformer-networks.pdf

[248] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

[249] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, June 2014, pp. 1867–1874.

[250] S. Ren, X. Cao, Y. Wei, and J. Sun, "Face alignment via regressing local binary features," *IEEE Transactions on Image Processing*, vol. 25, no. 3, pp. 1233–1245, 2016.

[251] X. Dong, Y. Yan, W. Ouyang, and Y. Yang, "Style aggregated network for facial landmark detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 379–388.

[252] P. N. Belhumeur, D. W. Jacobs, D. J. Kriegman, and N. Kumar, "Localizing parts of faces using a consensus of exemplars," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 12, pp. 2930–2940, 2013.

[253] X. Zhu and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, June 2012, pp. 2879–2886.

[254] V. Le, J. Brandt, Z. Lin, L. Bourdev, and T. S. Huang, "Interactive facial feature localization," in *European conference on computer vision*. Springer, 2012, pp. 679–692.

[255] G. Ghiasi and C. C. Fowlkes, "Occlusion coherence: Detecting and localizing occluded faces," *arXiv preprint arXiv:1506.08347*, 2015.

[256] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "Vggface2: A dataset for recognising faces across pose and age," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 2018, pp. 67–74.

[257] J. Yang, Q. Liu, and K. Zhang, "Stacked hourglass network for robust facial landmark localisation," in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, July 2017, pp. 2025–2033.

[258] G. Trigeorgis, P. Snape, M. A. Nicolaou, E. Antonakos, and S. Zafeiriou, "Mnemonic descent method: A recurrent process applied for end-to-end face alignment," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4177–4187, 2016.

[259] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, "Facial landmark detection by deep multi-task learning," in *Computer Vision – ECCV 2014*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Cham: Springer International Publishing, 2014, pp. 94–108.

[260] S. Xiao, J. Feng, L. Liu, X. Nie, W. Wang, S. Yan, and A. Kassim, "Recurrent 3d-2d dual learning for large-pose facial landmark detection," in *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct 2017, pp. 1642–1651.

[261] A. Jourabloo, M. Ye, X. Liu, and L. Ren, "Pose-invariant face alignment with a single cnn," *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 3219–3228, 2017.

[262] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[263] A. van den Oord, N. Kalchbrenner, L. Espeholt, O. Vinyals, A. Graves *et al.*, "Conditional image generation with pixelCNN decoders," in *Advances in Neural Information Processing Systems*, 2016, pp. 4790–4798.

[264] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[265] Z. Xu, Y.-C. Hsu, and J. Huang, "Learning loss for knowledge distillation with conditional adversarial networks," *arXiv preprint arXiv:1709.00513*, 2017.

[266] S. M. Iranmanesh and N. M. Nasrabadi, "Hgan: Hybrid generative adversarial network," *Journal of Intelligent & Fuzzy Systems*, no. Preprint, pp. 1–12, 2021.

[267] H. Kazemi, S. M. Iranmanesh, and N. Nasrabadi, "Style and content disentanglement in generative adversarial networks," in *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2019, pp. 848–856.

[268] D. P. Kingma and P. Dhariwal, "Glow: Generative flow with invertible 1x1 convolutions," *arXiv preprint arXiv:1807.03039*, 2018.

[269] V. Dumoulin, I. Belghazi, B. Poole, O. Mastropietro, A. Lamb, M. Arjovsky, and A. Courville, "Adversarially learned inference," *arXiv preprint arXiv:1606.00704*, 2016.

[270] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Adversarial generator-encoder networks," *CoRR, abs/1704.02304*, 2017.

[271] A. Ghosh, V. Kulharia, V. Namboodiri, P. H. Torr, and P. K. Dokania, "Multi-agent diverse generative adversarial networks," *CoRR, abs/1704.02906*, vol. 6, p. 7, 2017.

[272] Q. Hoang, T. D. Nguyen, T. Le, and D. Phung, "Mgan: Training generative adversarial nets with multiple generators," 2018.

[273] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," 2017.

[274] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015.

[275] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[276] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th international conference on machine learning (ICML-10)*, 2010, pp. 807–814.

[277] D. Berthelot, T. Schumm, and L. Metz, "Began: boundary equilibrium generative adversarial networks," *arXiv preprint arXiv:1703.10717*, 2017.

[278] I. Durugkar, I. Gemp, and S. Mahadevan, "Generative multi-adversarial networks," *arXiv preprint arXiv:1611.01673*, 2016.

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

[280] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," in *International Conference on Machine Learning*. PMLR, 2019, pp. 7354–7363.

[281] A. Brock, J. Donahue, and K. Simonyan, "Large scale gan training for high fidelity natural image synthesis," *arXiv preprint arXiv:1809.11096*, 2018.

[282] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2019, pp. 4401–4410.

[283] S. Kullback and R. A. Leibler, "On information and sufficiency," *The annals of mathematical statistics*, vol. 22, no. 1, pp. 79–86, 1951.

[284] T. Che, Y. Li, A. P. Jacob, Y. Bengio, and W. Li, "Mode regularized generative adversarial networks," *arXiv preprint arXiv:1612.02136*, 2016.

[285] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," Citeseer, Tech. Rep., 2009.

[286] A. Coates, A. Ng, and H. Lee, "An analysis of single-layer networks in unsupervised feature learning," in *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, 2011, pp. 215–223.

[287] T. Che, Y. Li, A. P. Jacob, Y. Bengio, and W. Li, "Mode regularized generative adversarial networks," *arXiv preprint arXiv:1612.02136*, 2016.

[288] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," in *Advances in Neural Information Processing Systems*, 2017, pp. 6626–6637.

[289] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of wasserstein gans," in *Advances in Neural Information Processing Systems*, 2017, pp. 5767–5777.

[290] A. Coates, A. Ng, and H. Lee, "An analysis of single-layer networks in unsupervised feature learning," in *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, 2011, pp. 215–223.

[291] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

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

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

[294] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis," *arXiv preprint arXiv:1605.05396*, 2016.

[295] P. Samangouei, M. Kabkab, and R. Chellappa, "Defense-gan: Protecting classifiers against adversarial attacks using generative models," *arXiv preprint arXiv:1805.06605*, 2018.

[296] N. Papernot, P. McDaniel, X. Wu, S. Jha, and A. Swami, "Distillation as a defense to adversarial perturbations against deep neural networks," in *2016 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2016, pp. 582–597.

[297] D. Meng and H. Chen, "Magnet: a two-pronged defense against adversarial examples," in *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*. ACM, 2017, pp. 135–147.

[298] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples (2014)," *arXiv preprint arXiv:1412.6572*.

[299] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *2017 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2017, pp. 39–57.

[300] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," *arXiv preprint arXiv:1607.02533*, 2016.