WestVirginiaUniversity
THE RESEARCH REPOSITORY @ WVU

Graduate Theses, Dissertations, and Problem Reports

2020

Deep Learning Based Face Detection and Recognition in MWIR and Visible Bands

Suha Reddy Mokalla
sumokalla@mix.wvu.edu

Recommended Citation

Mokalla, Suha Reddy, "Deep Learning Based Face Detection and Recognition in MWIR and Visible Bands" (2020). *Graduate Theses, Dissertations, and Problem Reports*. 7632.
https://researchrepository.wvu.edu/etd/7632

DEEP LEARNING BASED FACE DETECCTION AND RECOGNITION IN MWIR AND VISIBLE BANDS

Suha Reddy Mokalla

Thesis 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

Master of Science in
Electrical Engineering

Thirimachos Bourlai, Ph.D., Chair
Jeremy Dawson, Ph.D.
Yuxin Liu, M.S.
Lane Department of Computer Science and Electrical Engineering

Morgantown, West Virginia
2020

Keywords: Face Detection, Recognition, Multi-spectral, Mid-Wave infrared, GAN, Data Augmentation

ABSTRACT

DEEP LEARNING BASED FACE DETECCTION AND RECOGNITION IN MWIR AND VISIBLE BANDS

Suha Reddy Mokalla

In non-favorable conditions for visible imaging like extreme illumination or nighttime, there is a need to collect images in other spectra, specifically infrared. Mid-Wave infrared images can be collected without giving away the location of the sensor in varying illumination conditions. There are many algorithms for face detection, face alignment, face recognition etc. proposed in visible band till date, while the research using MWIR images is highly limited. Face detection is an important pre-processing step for face recognition, which in turn is an important biometric modality. This thesis works towards bridging the gap between MWIR and visible spectrum through three contributions. First, a dual band based deep face detection model that works well in visible and MWIR spectrum is proposed using transfer learning. Different models are trained and tested extensively using visible and MWIR images and the one model that works well for this data is determined. For this model, experiments are conducted to learn the speed/accuracy trade-off. Following this, the available MWIR dataset is extended through augmentation using traditional methods and generative adversarial networks (GANs). Traditional methods used to augment the data are brightness adjustment, contrast enhancement, applying noise to and de-noising the images. A deep learning based GAN architecture is developed and is used to generate new face identities. The generated images are added to the original dataset and the face detection model developed earlier is once again trained and tested. The third contribution is the proposal of another GAN that converts given thermal face images into their visible counterparts. A pre-trained model is used as discriminator for this purpose and is trained to classify the images as real and fake and an identity network is used to provide further feedback to the generator. The generated visible images are used as probe images and the original visible images are used as gallery images to perform face recognition experiments using a state-of-the-art visible-to-visible face recognition algorithm.

I dedicate this thesis to my family...

Acknowledgments

I would like to take this opportunity to express my deep gratitude and appreciation towards my advisor and committee chair Dr. Thirimachos Bourlai for all the support and guidance throughout my Masters' journey. This would not have been possible without his constant efforts to guide me in the right direction and his faith in me. I consider it a privilege having him as my mentor.

I would like to thank my committee members, Dr. Jeremy Dawson and Dr. Yuxin Liu for their suggestions and guidance.

I would like to extend my gratitude towards my family and friends who continuously encouraged me and believed in me.

I would like to thank my colleagues Neeru and Michael for guiding me in the right direction in my first steps towards my goal. I would also like to thank Jake, Ananya, Greg, Dylan and Kelsey for their co-operation and support.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Biometrics is the measurement and statistical analysis of people's unique physical and behavioral characteristics or traits. The technology is mainly used for identification and access control, or for identifying individuals under surveillance. The basic premise of biometric authentication is that every individual can be accurately identified by his/her intrinsic physical and behavioral traits. There are several types of biometric modalities, including, but not limited to, fingerprint and retinal scanning, facial recognition and voice analysis. Face recognition holds high importance since it is non-intrusive, understandable, and can be collected using a non-contact sensor in a covert manner at various stand-off distances. It has a wide variety of applications including, but not limited to access control to computer and other devices, buildings, auto-screening at airports, for secure banking. It can be used either independently or can be combined with other facial biometric traits like age, gender estimation and ethnicity recognition to improve the recognition performance.

Traditionally, FR system comprises of five modules - image acquisition, face detection, face alignment, feature extraction and matching as shown in Figure 1.1 .Image acquisition refers to capturing images using visible, MWIR (Mid-Wave Infrared) cameras etc. Face detection finds the

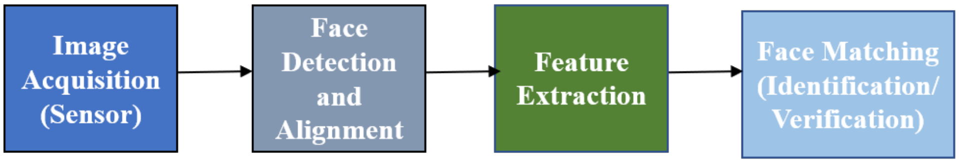


Image Acquisition (Sensor)
Face Detection and Alignment
Feature Extraction
Face Matching (Identification/ Verification)

Figure 1.1: Basic Building Blocks of a Traditional Face Recognition System

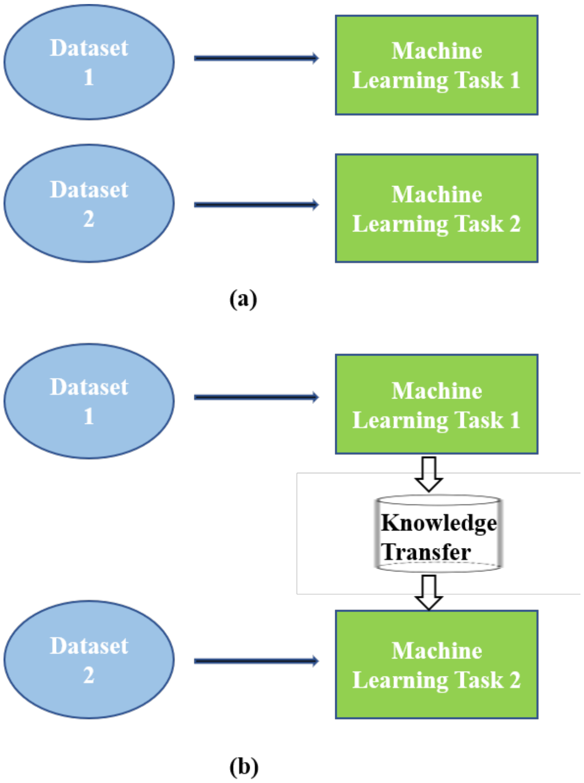


Figure 1.2: Machine Learning Approaches - (a) Traditional and (b) Transfer Learning

location of the face in a given image, while face alignment is normalizing a face in an image, so that it is in the center and the line joining the centers of eyes is parallel to the horizontal axis. Face alignment helps improve face recognition accuracy. Feature extraction can be defined as a process that reduces the dimensionality of an image by transforming the raw pixels in an image to a refined and useful format, like matrices that are straight forward for matching. These include hand-crafted features like LBP (Local Binary Patterns), HOG (Histogram of Oriented Gradients) and deep learning features referred to as embeddings. Face recognition is then performed by matching the generated features and it is divided into two categories — face identification and face verification. Face identification is the process of searching for a face in a database consisting many faces, which can be referred as 1-to-n matching where 'n' is the number of faces in the database. Face verification is comparing one face to another, and can be referred as 1-to-1 matching. Accuracy of a face recognition system is determined using match scores (degree of similarity) between a gallery image (image in the database) and a probe image (query or input).

The many robust algorithms proposed using traditional and deep learning based methods work well with images collected in well illuminated environments and are not very efficient in face recognition (FR) in low-light to no light conditions. The ability to perform face recognition outside the visible spectrum (wavelength - 400-700nm) is of prime importance in many surveillance, law enforcement, and military organizations as it is useful in challenging, in-the-wild scenarios, especially when operating in no-light and low-light conditions. One such spectrum is IR (Infrared) which is classified into active and passive IR. Active IR comprises of NIR(Near IR) and SWIR (Short Wave IR) at wavelengths 0.75-1 μm and 1-2.5 μm respectively and passive IR comprises of MWIR (Mid-Wave Infrared) and LWIR (Long-Wave IR) at 3-5 μm and 8-12 μm respectively. This work is focused on MWIR spectrum, in which IR radiation emitted from the subject's face in the form of heat is detected by the camera sensor whenever data is acquired. MWIR sensors provide a significant capability of acquiring human biometric signatures under obscure environments. In addition, when operating in the MWIR band, the location of the sensor cannot be detected and the images obtained are not affected by the extreme illumination conditions.

While face recognition is a very important biometric authentication technique, face detection is an important pre-processing step for face recognition. There are many face detection algorithms proposed in visible spectrum, however the face detection knowledge available in thermal spectrum is highly limited. The reason for this is the limited availability of face datasets in the thermal spectrum. The high cost of thermal cameras and other challenges related to Institutional Review Boards (IRB) or Export Control Issues contribute to the limited availability of data. Developing an original deep face detection model requires a deep learning training phase using millions of images representing each band, as well as tuning of numerous parameters. This is a difficult task owing to the aforementioned reasons and the time and computational cost required to train it. Transfer learning can be used as an acceptable alternative to solve these problems as it requires images in the order of a few hundreds to thousands and fine-tuning a small set of parameters and requires less time to train than it is required to train an original model. Transfer learning can be defined as re-utilizing the knowledge gained from one problem to solve another related problem. This is addressed as the future of machine learning by many researchers, and is shown in Figure 1.2. Using this, one or more of the deep learning models developed for face detection or object detection in general can be used to develop another deep face detection model in MWIR band.

Another important and interesting development in the field of deep learning is the introduction of GANs (Generative Adversarial Networks) —that consists a generative model *(G)* that generates new images that are similar to the original images and a discriminative model *(D)* that discriminates between original and generated images. The end goal of GANs is to generate images that look indistinguishable from the original ones. Goodfellow et al. [1] refers to GANs as a two-player minimax game played by D and G where G is a culprit trying to generate counterfeit money and D is a cop

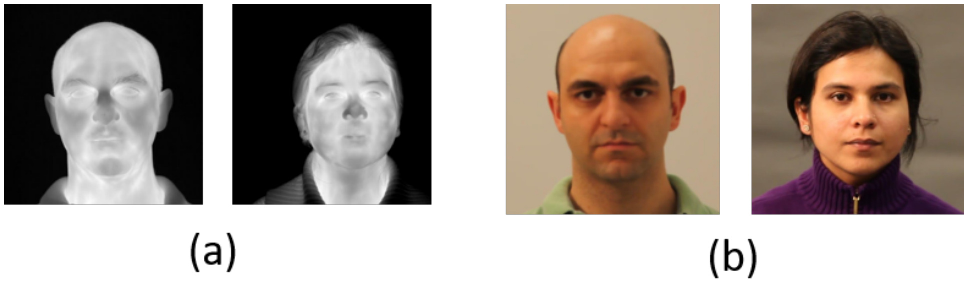


Figure 1.3: Modality gap between Visible and Thermal images

trying to distinguish between the real and fake currency.

The idea of GANs can be used to improve the limited thermal data sets by generating new identities from the existing face images. This solves the problem of not having enough data to train a CNN (Convolutional Neural Network) and avoids the expenses of sensors and data collections. Though, the face detection model developed using transfer learning yields good results, generating new identities further improves the accuracy. Many variations of GANs are proposed since the introduction and one of them is the DCGAN (Deep Convolutional GAN) [2]. In DCGAN, the original GANs are modified by proposing suitable architectural topology and several constraints to train a successful GAN and develop a discriminator that not only discriminates between real and fake images, but also performs unsupervised learning tasks such as clustering. We use this as our base network to augment thermal images and generate new identities. Then we use the generated images to improve the face detection models.

After DCGANs, another significant development to GANs is conditional GANs. In conditional GANs, D and G are conditioned on some extra information y, where y could be any auxiliary information, such as class labels or data from other modalities. The FR algorithms available for visible band cannot be directly applied to match thermal images to their counterparts due to the modality gap between the two as shown in Figure 1.3. Conditional GANs are an excellent way to generate visible images from their MWIR counterparts and these can be used for visible-to-visible FR using any of the widely available FR algorithms.

1.2 Problem Statement

MWIR imaging became an area of growing interest among researchers and government institutions equally. However, many modalities including face detection, normalization, FR needs to be developed to be able to be used in practical applications. The challenges here are: 1. Limited availability of datasets to train and test an original deep face model (detection and recognition) because of the cost of the sensors, 2. The models are expensive to train, even after acquiring the required datasets, both computationally and economically and 3. Modality gap between the visible and thermal bands making it difficult to apply the already developed visible band FR models in MWIR and cross-spectral matching.

1.3 Contributions of Thesis

In this work, a dual band deep face detection model and a visible-to-thermal FR strategy are proposed. The contributions of this thesis are three-fold.

First, a deep learning based face detection model that works equally well in MWIR and Visible bands is proposed. To overcome the problem of limited availability of data to train an original deep CNN, transfer learning is used which requires images in the order of a few hundreds or a thousands and the pre-trained models SSD (Single-shot Multibox Detector) [3], R-FCN (Region based Fully Convolutional Network) [4] and Faster R-CNN (Region based CNN) [5] are fine-tuned, optimized and re-trained to achieve high accuracies in thermal and visible band face detection, and trade-off between speed and accuracy of deep face detection in thermal band is analyzed. The efficiency of these face detection model is further demonstrated by performing same-spectral cross-scenario FR.

Second, to further solve the problem of data availability, GANs are used to develop new face identities in thermal band increasing the size of dataset to a great extent. These generated images are added to the original dataset and the models are re-trained. This addition increased the face detection accuracy and increased the speed to a great extent. This is achieved by implementing DCGANs, explained in Section 1.1.

Third, a GAN based approach is proposed to overcome the problem of modality gap between visible and thermal bands by generating visible images from their MWIR counterparts. Prior to this, transfer learning is used to re-train Facenet [6] FR model using original visible and thermal images and the results obtained re-confirm the modality gap between the bands. Facenet is used to extract features form original visible images and these features are used to condition G and D of the GAN. Pix2pix [7], an application of conditional GAN that performs image-to-image translation is used for this purpose with numerous changes in the architectures of G and D networks. After the visible images are generated, the Facenet model is used to perform visible-to-visible FR without any training.

1.4 Organization of Thesis

The rest of the work is organized as follows:

- Chapter 2 describes existing work in detection in visible and thermal bands using traditional and deep learning approaches, GAN based approaches for data augmentation and same spectral and cross-spectral FR algorithms.

- Chapter 3 describes the development of a dual band based deep face detection model by fine-tuning and optimizing the existing pre-trained models and trade-off between speed and accuracy of deep face detection model in thermal band that could detect faces indoor and outdoor at 5m and 10m images.

- Chapter 4 describes the GAN based image augmentation to generate new face identities in MWIR band to improve the datasets, thereby improving the efficiency of the face detection model.

- Chapter 5 describes the GAN based approach to synthesize visible images from their MWIR counter parts, in order to be able to use the existing visible-to-visible FR approaches without further training.

- Chapter 6 presents the results and evaluation of the proposed approach.

- Chapter 7 concludes the work and explains the scope for future work.

Chapter 2

Related Work

2.1 Face Detection

2.1.1 Face Detection in the Thermal Band

Most of the traditional and deep learning based face detection algorithms are proposed in the visible spectrum. Some of them are proposed in thermal spectrum, most of which follow traditional approaches — extracting features from images using feature extractors including HOG, facial landmark points etc. Adaboost classifier with local features such as Haar-like, MB-LBP (Multi-Block Local Binary Pattern) and HOG (Histogram of Oriented Gradients) is an example [8] for hand-crafted features. Zheng [9] proposes face and eye glass detection in thermal band that uses region growth algorithm to segment face and image de-noising and normalization. Murata et al. [10] propose a face detection algorithm that automatically extracts the target facial region from the thermal image by focusing on the temperature distribution of the facial thermal images as well as examines the automation of the evaluation. Kopaczka et al. [11] propose an LWIR (Long Wave Infrared) face tracking method based on an active appearance model (AAM) to address the problem of in-plane rotation and occlusion for detection facial landmarks.

One of the very few deep learning based thermal face detection models is proposed by Kwasnieska et al. [12] that detects faces in low resolution thermal images. The authors localize objects by restoring the spatial information about features distribution from the classification network. Kwasniewska et al. [13] propose a deep learning approach for face detection and tracking in thermography using transfer learning technique to re-train the Inception *v3* using thermal images and modifying the last layer of model to improve its localization ability. They demonstrate that this approach can be used with low resolution thermal images collected using thermal cameras that can be embedded into everyday devices like cell phones or an indoor remote monitoring solution.

2.1.2 Face Detection in the Visible Band

There are many face detection algorithms in visible spectrum that use deep learning approaches. Zhang et al. [14] propose MT-CNN (Multitask CNN) where, the image is first re-sized to form an image pyramid followed by three stages. The first stage is Proposal net that obtains several candidate windows and their corresponding bounding box regression vectors, the second stage is a Refine net that rejects a large number of false candidates and the third stage is Output net that describes face in more detail and outputs five facial landmarks. Each of these stages are followed by NMS (Non-Maximum Suppression). Farfade et al. [15] propose a multi-view face detection algorithm which they call DDFD (Deep Dense Face Detector) that does not require pose and landmark annotation and is able to detect faces in a wide range of orientations. They extracted training samples from AFLW (Annotated Facial Landmarks in the Wild) dataset [16] and used different data augmentation strategies and increased the size of the dataset ten times that of the original. Alex-Net [17] is fine-tuned and the sliding window approach is selected to decrease the system complexity.

Ranjan et al. [18] propose HyperFace, a single CNN that performs face detection, landmark localization, pose estimation and gender classification in three modules: the first one generates class independent region proposals and scales them, second is a CNN which takes in the re-sized candidate regions and classifies them as face or non-face and provides landmarks for face regions. The third module is a post-processing step which involves iterative region proposals and landmarks-based NMS. Yang et al. [19] propose Faceness-Net, which detects faces in different poses and occlusions in two stages — In the first stage, a full image is input to a CNN that generates partness maps of each face part (hair, eye, nose, mouth and beard) using an attribute aware network for each part and generates face proposals. In the second stage, face proposals generated in the first stage are refined by training a multi-task CNN, where face classification and bounding box regression are jointly optimized. Wu et al. [20] propose a CNN for facial landmark regression by examining the intermediate features of a standard CNN trained for landmark detection and show that features extracted from later, more specialized layers capture rough landmark localizations. They propose a tweaked CNN by providing a natural means of applying differential treatment midway throughout the network.

Sun et al. [21] propose an improved Faster R-CNN for face detection which is a light head based two-stage framework which they call FDNet1.0. Their findings suggest that some modifications including multi-scale training, multi-scale testing, keeping small proposals at training and test stages, directly selecting top ranked proposals without NMS in the RPN stage for R-CNN, feature concatenation, hard negative mining, and model pre-training improved the final face detection performance and implemented them. Cheng et al. [22] propose a two-layer CNN to learn the high-level features which performs face identification via sparse representation. Sparse Representation Classifier (SRC) represents the image by a subset of the training data. They improve its performance via a precisely selected feature extractor. Sun et al. [23] propose a method for estimation of the positions of facial keypoints with three-level convolutional networks and the outputs at each level are fused for robust and accurate estimation and global high-level features are extracted in the initial stages which help to locate accurate keypoints. Zhang et al. [24] propose a multi-task deep CNN to address the problems of occlusion and pose variation. They investigate the possibility of improving detection robustness through multi-task learning and to optimize facial landmark together with heterogeneous but subtly correlated tasks like headpose estimation and facial attribute inference and developed a tasks-constrained deep model with task-wise early stopping to facilitate learning convergence.

2.2 GANs for Image Augmentation

Neural networks often require large amounts of data for effective functionality. There are many standard techniques to generate more images from a limited number of images like rotation, flip etc. however, the data generated by these techniques is again limited. GANs (Generative Adversarial Networks) are introduced by Goodfellow et al. [1], in which two networks, namely, generator and discriminator compete with each other, the former trying to generate data that looks as similar as possible to the original data and the latter trying to discriminate between the real and fake distributions. Therefore, adversarial training leads to the generation of new data that looks similar to the original data. Following this lead many algorithms are developed to generate broader datasets using GANs.

In [25], Antoniou et al. propose DAGAN (Data Augmentation GAN) in which an encoder takes an input image and projects it to a lower dimension manifold (bottleneck) and a transformed random vector is concatenated to the bottleneck which is then passed to the generator. A combination of UNet [26] and ResNet [27] is used for generator and DenseNet [28] is used for discriminator. Zhu et al. [29] propose a data augmentation strategy to create a balanced dataset from an unbalanced one using cycleGANs [30] for emotion classification. They use reference and target images and feed them to a cycleGAN, which is used as generator to generate images with rare emotions such as disgust from sad or happy expressions and they develop a CNN based emotion classifier. Out of many variants of GANs, DCGAN [2] became popular which focuses on unsupervised learning of GANs and implements a set of changes on top of the convolutional GANs to improve stability. All pooling layers are replaced with strided convolutions for discriminator and fractional-strided convolutions for generator. Batch Normalization that stabilizes learning by normalizing the input to each unit to have zero mean and unit variance is used in both the networks. A ReLU activation layer is used in all layers of the generator except the output, and a leaky ReLU is used in all the layers of the discriminator. Zhao et al. [31] propose EBGAN (Energy-Based GAN) that views discriminator as an energy function that attributes low energies near the data manifold and higher energies to the other regions. This facilitates the use of a variety of architectures and loss functions in addition to the usual binary classifier (real or fake) and found that this is more stable and that a single-scale architecture can be used to generate high-resolution images.

In [32], Frogner et al. propose BEGAN (Boundary Equilibrium GAN) in which they use auto-encoder as a discriminator and match auto-encoder loss distributions using a loss derived from the Wasserstein distance [33]. They achieve this by adding an equilibrium term to the GAN objective to balance the discriminator and generator. Shin et al. [34] introduce a GAN algorithm to generate synthetic abnormal MRI (Magnetic Resonance Imaging) with brain tumors by adopting image-to-image translation conditional GAN to translate label-to-MRI (synthetic image generation) and MRI-to-label (image segmentation) and obtained comparable results when trained on synthetic data versus when trained on the original data.

2.3 Face Recognition

2.3.1 Same Spectral Matching

There are numerous face recognition algorithms available in the visible spectrum. ArcFace (Additive Angilar Margin Loss) [36] is one of the most recent visible face recognition model, which uses arc-cosine function (dot product) between the CNN feature and the last fully connected layer is equal to the cosine distance after feature and weight normalization. Additive angular weight is added to the target angle and all the logits are re-scaled by a fixed feature norm, and the subsequent steps are exactly the same as in the softmax loss. FaceNet [6] is one of the most popular face recognition algorithms that defined a term called "triplet loss", which uses triplets, containing a pair of objects from the same class and another one that is not. FaceNet model learns a mapping from face images to a compact Euclidean space where distances directly correspond to a measure of face similarity.

Huang et al. [37] propose CLMLE (Cluster based Large Margin Local Embedding) from all the examples in clusters, rather than only hard examples in clusters by enforcing margins between hard-mined clusters in the local neighborhood from same and different classes to address the problem of class imbalanced data in deep learning experiments. This margins introduce a tight constraint to generate more balanced class boundaries locally. CosFace [38] is proposed by reformulating the softmax loss as a cosine loss, LMCL (Large Margin Cosine Loss) by normalizing both the features and weight vectors to remove radial variations. The introduced cosine margin term maximizes the decision margin in the angular space. Minimum intra-class variance and maximum inter-class variance are achieved by normalization and cosine decision margin maximization. Masi et al. [39] propose a face recognition approach that considers and handles pose variability by learning PAMs (Pose Aware Models) for frontal, half-profile and full profile faces. In addition, they propose multiple ideal co-ordinates for out-of-plane face alignment and co-training, which addresses the problem of training CNN models for extreme poses where relatively few example faces are available for training.

2.3.2 Multi-Spectral Matching

Though there are many face recognition algorithms available in the visible spectrum, they cannot be directly used for IR (Infrared) images, due to the limited availability of IR data. To overcome this problem, Lezama et al. [40] propose a NIR-VIS face recognition approach that has two core components - cross spectral hallucination and low-rank embedding to optimize input and output of a visible deep model for cross-spectral face recognition. Cross hallucination CNN is trained on pairs of corresponding NIR-VIS patches that are mined from CASIA NIR-VIS 2.0 [41]. Low-rank embedding is a geometrically motivated transformation that is learned to restore a within-class low rank structure while introducing a maximally separated inter-class structure.

Narang et al. [42] propose a SWIR based face recognition system in which they develop an SSMW (Single Sensor Multi-Wavelength) imaging system that operates in SWIR band and captures in three different wavelengths of SWIR spectrum. They also develop an automated image quality-based score-level fusion scheme for the automated classification of multi-wavelength face images to individual wavelengths and an image quality weighted-based, score level fusion scheme developed for frontal vs. non-frontal classification. Finally, they determined which SWIR wavelength provides good quality face images and high recognition rates. Osia et al. [43] propose a fully automated direct matching based FR approach that operates when images from either visible or passive IR bands are used. Input images are first geometrically normalized using a pre-processing pipeline, after which face-based features including wrinkles, veins as well as edges of facial characteristics are detected and extracted for each operational band. Finally, global and local face-based matching is applied, before fusion is applied at score level.

To make GANs conditional [44], both the generator and discriminator networks are conditioned on some extra information such as labels, images from other domains etc. This information is utilized to use images from one domain and condition the GAN to generate images in other domains including using thermal images to generate visible images that facilitates cross-spectral matching. To this end, some algorithms are proposed to match thermal images to visible. TV-GAN (Thermal-to-Visible GAN) [45] introduces a network with a generator that transforms thermal image into its visible counterpart that still carries sufficient identity information for the face recognition task. An identity loss function is introduced in the discriminator, so that the discriminator does not only provide the fake or real discrimination result, but also provides closed-set face recognition.

Di et al. [46] propose AP-GAN (Attribute Preserving GAN) to generate visible images from thermal images using the attributes generated using a pre-trained VGG-Face network. AP-GAN uses U-Net as a building block for the generator and a patch-based discriminator. The adversarial loss for discriminator has unconditional loss to discriminate between real and fake images and conditional loss to match real images to reconstructed images. The generator, in addition to the adversarial loss, has perceptual loss, identity loss, attribute loss and L1-Norm loss between the target and reconstructed image. Semantic-Guided GAN [47] is another model that performs thermal to visible matching, using semantic labels extracted by a face parsing network during training. These semantic labels denote high level facial component information associated with each pixel. Identity loss and perceptual loss are used in addition to the adversarial loss.

Wang et al. [48] propose a network that combines a generator network with a detector. The generator is based on CycleGAN and learns bi-directional translation between visible and thermal images. The detector network is designed following [49] and extracts face shape features, which are constituted by important landmarks of visible faces. These face shape features help the generator synthesize visible images of better visual quality and with more realistic identity preserving features.

Chapter 3

MWIR and Visible Based DeepFace Detection Model

3.1 Methodology overview

The methodological approach of this work is as follows: first step is to manually annotate all the images to generate bounding box co-ordinates. Second step is to train and validate the deep learning models, fine-tune, and determine the model that yields the best results for the data and the last step is to optimize the selected model further to improve the accuracy and speed and to find an accuracy/speed trade-off. Figure 3.1 shows the proposed methodological approach for the system.

3.2 Meta Architectures

The deep learning models used to train the face detection are SSD, R-FCN, Faster R-CNN with VGG-16 [50], ResNet-101 [27], Inception *v2* [51], Inception *v3* [52], Inception ResNet (v2) [53] and MobileNet [54] as feature extractors. All the above-mentioned models are trained and tested using the MWIR and visible data separately and the results obtained from all the combinations of models and feature extractors are compared with each other. All the machine learning algorithms have one objective — to minimize the loss. The loss function in [55] is provided below:

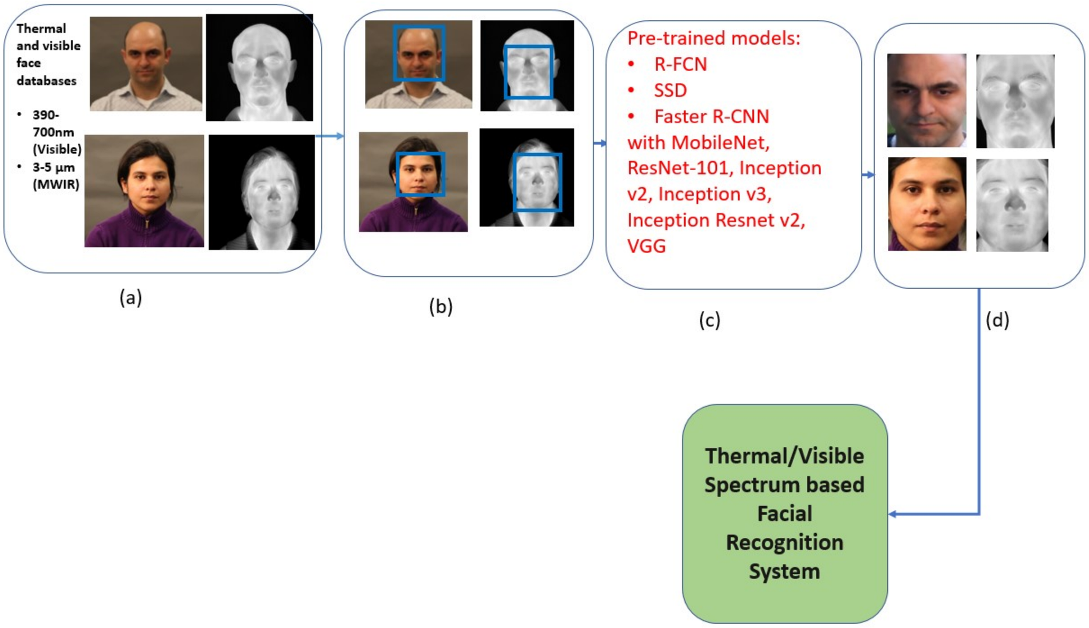

Thermal and visible face databases
• 390-700nm (Visible)
• 3-5 µm (MWIR)
(a)
(b)
Pre-trained models:
• R-FCN
• SSD
• Faster R-CNN with MobileNet, ResNet-101, Inception v2, Inception v3, Inception Resnet v2, VGG
(c)
(d)
Thermal/Visible Spectrum based Facial Recognition System

Figure 3.1: Overview of the proposed approach - (a) Images collected in Visible and Thermal bands, (b) Manual annotation of images to generate bounding boxes, (c) Pre-trained models trained and tested, (d) Output (Cropped images that are used for face recognition)

$$L(a, I; \theta) = \alpha * 1[a\ is\ positive] * l_{loc}(\phi(b_a; a) - f_{loc}(I; a, \theta)) + \beta * \\ l_{cls}(y_a, f_{loc}(I; a, \theta)) \qquad (3.1)$$

All the models used in this work follow this loss function where α and β are weights of the network balancing localization and classification losses, a is the anchor (anchors are predicted boxes overlaid on the image I at different spatial locations, scales and aspect ratios), b_a is the best ground truth match for a (if exists), y_a is class label if anchor is positive i.e. object is present in a (only face in our case; 1 if positive, 0 if negative), $\phi(b_a; a)$ is the vector encoding of b_a with respect to a, f_{cls} is discrete class prediction for each anchor and f_{loc} is continuous prediction of offset by which the anchor needs to be shifted to fit the ground truth bounding box. The meta architectures of the models used are described below. They are tuned, optimized and tested to determine the most efficient model in terms of the speed and detection accuracy.

3.2.1 Deep face detection models

Faster R-CNN

Faster R-CNN [5] is a relatively fast version of Fast R-CNN [56]. It has two components. First is an RPN (Region Proposal Network) that takes image as an input and outputs a set of rectangular object proposals, each with score of the presence of object class in the box. To generate proposals, a small network is slid over the convolutional feature map output by the last shared convolutional layer and each sliding window is mapped into a lower-dimensional vector which is fed into two sibling fully-connected layers — a box regression layer and a box classification layer. The second component is the Fast R-CNN. Fast R-CNN takes an entire image and a set of box proposals as input. The network processes the whole image with several convolutional and max pooling layers to produce a convolutional feature map. Then, a region of interest (RoI) pooling layer extracts a fixed length feature vector from the feature map for each of the object proposals. Each vector is then fed into a sequence of fully connected layers that branch into two output layers: one that produces softmax probability estimates and another layer that outputs four real valued numbers that encode refined bounding-box positions.

R-FCN

R-FCN [4] follows R-CNN in adopting the two-stage object detection strategy explained above. Candidate regions are extracted by RPN, after which R-FCN classifies the ROIs (Regions of Interest) into categories and background. The last layer of R-FCN is a position-sensitive ROI pooling layer that aggregates the outputs of the last convolutional layer. The backbone architecture of R-FCN is ResNet-101 which is explained later in this section. The significant change made in the ResNet-101 architecture is reducing the effective stride from 32 pixels to 16 pixels, increasing the score map resolution.

SSD

SSD [3] is a feed-forward CNN that produces a set of bounding boxes with confidence scores for the presence of object in those boxes. This is followed by a non-maximum suppression step that detects the object. The early networks are standard CNN layers that are used for classification in high quality images called base network and are truncated before any classification layers. Convolutional feature layers are added to the truncated base network that decrease in size and generates detections at multiple scales. Each added feature layer produces a fixed set of predictions using a set of convolutional networks which are different for each feature layer.

3.2.2 Feature Extractors

MobileNet

MobileNet [54] is basically designed for Mobile Vision applications and is a relatively fast convolutional network. It is based on depth-wise separable convolutionals which is a form of factorized convolutions which factorize a standard convolution into a depth-wise convolution and a 1×1 convolution called a point-wise convolution. The depth-wise convolution applies a single filter to each input channel and the point-wise convolution applies a 1×1 convolution to combine the outputs of the former.

Inception *v2*

Inception [57], popularly known as GoogleNet stands the base architecture for Inception *v2* and Inception *v3*. It follows the basic idea to operate filters with multiple sizes on the same level. This model has 9 inception modules stacked linearly in 22 layers and uses global average pooling at the end of the last inception module. To address the problem of vanishing gradient [51], which is common in any very deep classifier, two auxiliary layers are introduced. Inception *v2* is similar to GoogleNet with the following changes. Representational bottleneck is reduced as reducing the dimensions drastically may cause loss of information. To achieve this, filter banks in the module are expanded. 5×5 convolutions are factorized to two 3×3 convolutions and any n×n convolutions are factorized to n×1 and 1×n convolutions. This reduces the computational cost, as large convolutions are extremely expensive compared to the smaller ones.

Inception *v3*

Inception *v3* [52] includes all the upgrades mentioned above for Inception *v2*. Additionally, the following details are added to the architecture: RMSProp optimizer, factorized 5×5 convolutions, batch normalization is applied to the auxiliary layers and label smoothing to prevent over-fitting.

VGG-16

VGG-16 [50] replaces the large kernel-sized filters (mostly 11×11 and 7×7) with multiple 3×3 filters. Blocks with the same filter size are applied multiple times to extract more complex and representative features. VGG convolutional layers are followed by 3 fully connected layers. Width of the network starts at a small value of 64 and increases by a factor of 2 after every sub-sampling pooling layer.

ResNet-101

ResNet-101 [27] uses referenced mapping instead of un-referenced mapping i.e., the input from one layer is directly connected to the next layer along with the output from the previous layer. The intuition behind this approach is that it is easier to use a referenced mapping than it is to optimize a non-referenced mapping.

Inception ResNet *v2*

In Inception ResNet *v2* [53], residual connections are introduced that add the output of the convolutional operation of the inception module to the input. To match the depth sizes for residual connections, 1×1 convolutions are used after tthe original convolutions. Pooling operation is replaced to favor residual connections. Networks with residual units deeper in the architecture made the network unstable and this problem is addressed by scaling the residual activations to 0.1 to 0.3.

3.3 Parameter tuning

Deep learning models explained in the above section are combined with all the aforementioned feature extractors explained and all the choices for training are made closely following the original work [55] such as convolutional layers to be used for generating region proposals for region-based networks, number of region proposals, usage of multiple feature maps, output stride of ResNet etc. ArgMax matching is used for matching anchors (generated bounding boxes) with ground truth instances in which anchors are discarded if the overlap between anchor and ground truth is lower than a threshold (0.5 IoU). Smooth L1 loss [58] is used for all the experiments following the original work. Learning rate is varied from 0.002 to 2e-6 for each of the combination of model and feature extractor. Different optmizers are trained and tested, which include AdaGrad, Momentum, RMSProp and Adam optimizers.

3.3.1 Momentum Optimizer

Momentum optimizer helps accelerate SGD (Stochastic Gradient Descent) in the relevant direction and dampens oscillations. It is achieved by adding a fraction γ (called momentum) of the update vector of the past time step to the curernt update vector. The momentum term increases for dimensions whose gradients point in the same directions and reduces updates for dimensions whose gradients change directions. The default value of γ is 0.9.

3.3.2 AdaGrad Optimizer

Adagrad optimizer adapts the learning rate to the parameters, performing smaller updates (larger learning rate) for parameters associated with frequently occurring terms and larger updates (larger learning rate) for infrequent terms. The default learning rate value for AdaGrad optimizer is 0.01 and ϵ (smoothing term that avoids division by 0) is 1e-8.

3.3.3 RMSProp Optimizer

RMSProp (Root Mean Square Propagation) optimizer is an extension of AdaGrad optimizer that seeks to reduce its aggressive and monotonically decreasing learning rate. It divides the learning rate by an exponentially decaying average of squared gradients. The default momentum value is 0.9.

3.3.4 ADAM Optimizer

Adam optimizer computes adaptive learning rates for each parameter. It stores exponentially decaying average of past squared gradients (similar to RMSProp) and exponentially decaying average of past gradients (similar to momentum) as the first moment (the mean) and the second moment (the variance) respectively. Default values for the first and second estimates are 0.9 and 0.999 respectively and that of ϵ is 1e-8.

3.4 Faster R-CNN with ResNet-101 (Proposed Model)

Results obtained from the experiments show that the Faster R-CNN model with ResNet-101 performed better than the other models for our data. This section includes a brief description of the training parameters for this particular model. The main difference in the implementation of the Faster R-CNN is that the Adam optimizer [59] is used instead of the momentum optimizer. This optimizer is found to be the most effective in this case than other optimizers after training and testing the models with Momentum, AdaGrad and RMSProp optimizers. The Adam optimizer is used with the default values of 0.9 and 0.999 for the first and second moment estimates respectively, and Epsilon is set to 1e-8. This optimizer learns the learning rates itself, i.e. it adapts the learning rate as the loss decreases. Note also that the mini-batch size for RPN training is set to be 64 while for box classifier is set to be 1. Though the network performs well, the detection time is higher than the other networks. To address this problem, further experiments are performed by reducing the number of proposals in the Fast R-CNN, which reduces the time required for the RPN to propose regions, thereby reducing the detection time. However, the speed is increased at the expense of accuracy of the system. This is further explained in Chapter 6. Basic blocks of Faster R-CNN is shown in Figure 3.2 and difference between a regular and a residual connection is shown in Figure 3.3.

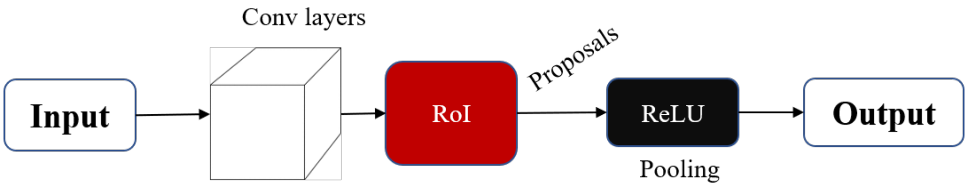
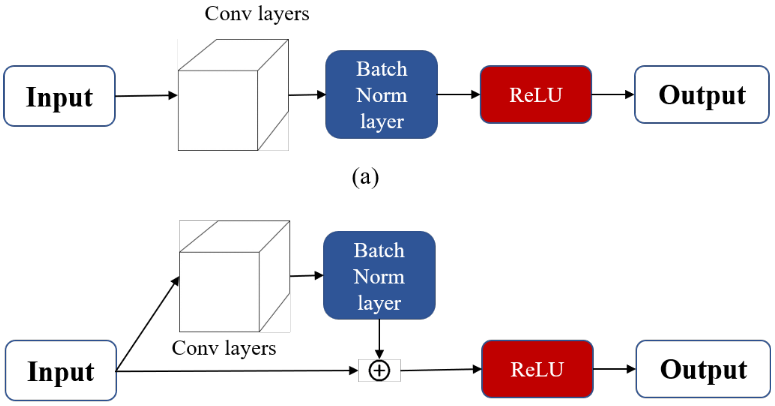


Figure 3.2: Overview of Faster R-CNN model



(a)

(b)

Figure 3.3: Network (a)without and (b) with a residual connection

3.5 Same Spectral face Recognition

Same spectral cross-scenario (indoor vs outdoor) face recognition experiments are performed to further demonstrate the efficiency and benefits of the proposed automated face detection framework. The results are compared against the face recognition experiment results obtained using manually cropped faces. FaceNet [6] face recognition model is used for visible face recognition. Thermal face recognition is performed using Histogram of Oriented Gradients (HOG) and Linear Binary Pattern

(LBP) hand-crafted features. These are fused using SS (Simple Sum) score-level fusion, thereby forming a single distance metric.

Chapter 4

Improving MWIR DeepFace Detection Accuracy Using Data Augmentation - Traditional and GAN Approach

Fig 4.1 shows the proposed methodology followed in this work to increase the size of the database and perform various experiments to validate the impact of different image augmentation techniques. The Faster R-CNN with ResNet101 model for face detection used in all the experiments in this work is the model that is proposed and developed in Chapter 3. The first step is to augment the data using different traditional image augmentation techniques like brightness, contrast adjustment, and noising and de-noising algorithms. The second step is to use the generated data to train and validate the face detection model to study the positive and negative effects of using different image processing techniques. The learning rate is changed each time generated images are added to the training and test datasets. In each experiment, 90% of the data is used for training and the remaining 10% is used for testing. Care is taken that equal percentage of data from each domain (original and each augmentation technique) is used for training and testing. Then, GANs are used to develop new identities for addition to the original dataset used to train and test the Faster R-CNN model.

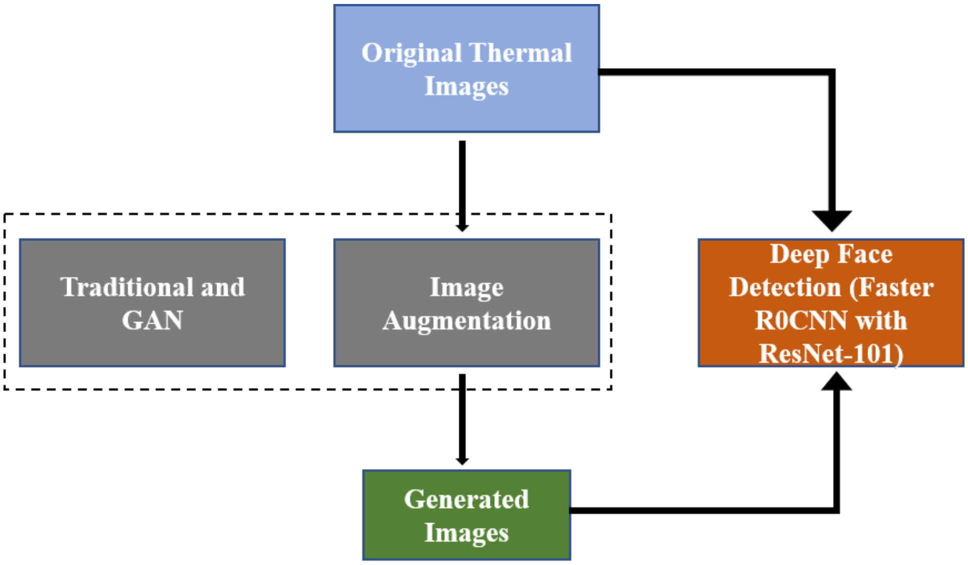


Figure 4.1: Overview of the proposed approach - Data Augmentation

4.1 Traditional Image Augmentation Techniques

Image enhancement techniques are used to augment the data and to find the positive and negative impacts of using such techniques on the MWIR images. The image enhancement techniques used are brightness, contrast and filtering. Brightness is varied to add darker and brighter images to the dataset. When brightness of an image is increased, all the pixels' values in the image are increased, making the entire image lighter and when brightness of an image is decreased, pixel values are decreased, making the entire image darker.

The second image augmentation technique used is contrast enhancement. Three types of contrast enhancement techniques are used in this study. One is the regular contrast enhancement which saturates the top and bottom 1% of the pixel values. The second is histogram equalization where the most frequent intensity values are spread out. The third contrast enhancement method used is CLAHE (Contrast Limited Adaptive Histogram Equalization), which is similar to histogram equalization. CLAHE differs from histogram equalization in that it uses tiles of an image at a time instead of using the entire image at once.

The third augmentation technique used is applying noise and then de-noising all the images. Gaussian and Salt & Pepper noise is applied to images as these types of noise are common in images. Additionally they do not alter the appearance of objects in an image drastically compared to Poisson or Speckle noise [60]. Then Wiener and Median filters are applied to remove Gaussian and Salt & Pepper noise respectively, since these are proved to work for these particular kinds of noise.

4.2 GAN Augmentation

GANs are used here to generate new face identities in thermal spectrum. The generator and discriminator architectures are explained below. The guidelines laid out by Radford et al. in [2] are followed to develop the GAN architecture.

4.2.1 Generator Architecture

- All the fully connected layers in the vanilla GAN are replaced with deep convolutional layers. Also, spatial pooling layers are replaced with fractional-strided convolutions for generator. This allows the generator network to learn its own upsampling.
- Generator is built using four fractionally strided convolutional layers.
- No fully connected layers are used in any of the hidden units.
- Each of the deconvolutional layer is followed by a Batch Normalization layer except the last one.
- ReLU (Rectified Linear Unit) is used as activation for all the hidden units.
- Input to the generator is a noise vector which can be either uniform or normal. We use uniform distribution in this case as we need diversity in the generated images and since uniform noise has high entropy, it facilitates the generation of highly diverse output images.
- The output layer of the generator is a Tangent Hyperbolic activation function. Generator architecture is shown in Fig 4.2.

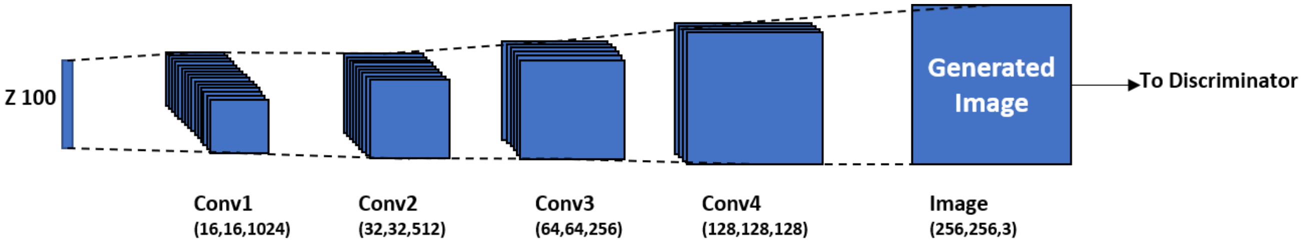


Figure 4.2: Generator Architecture

4.2.2 Discriminator Architecture

- In the discriminator, all the spatial pooling layers are replaced with strided convolutional layers. This helps the discriminator learn its own downsampling.
- Discriminator comprises of three fully convolutional layers, followed by a fully connected layer that outputs a single logit. Each convolutional layer reduces the size of the input by half.
- Each of the convolutional layers, except the first, is followed by a Batch Normalization layer.
- All the hidden units use LeakyReLU activation function.
- The output layer of the discriminator is a sigmoid activation function. The discriminator architecture is shown in Fig 4.3.

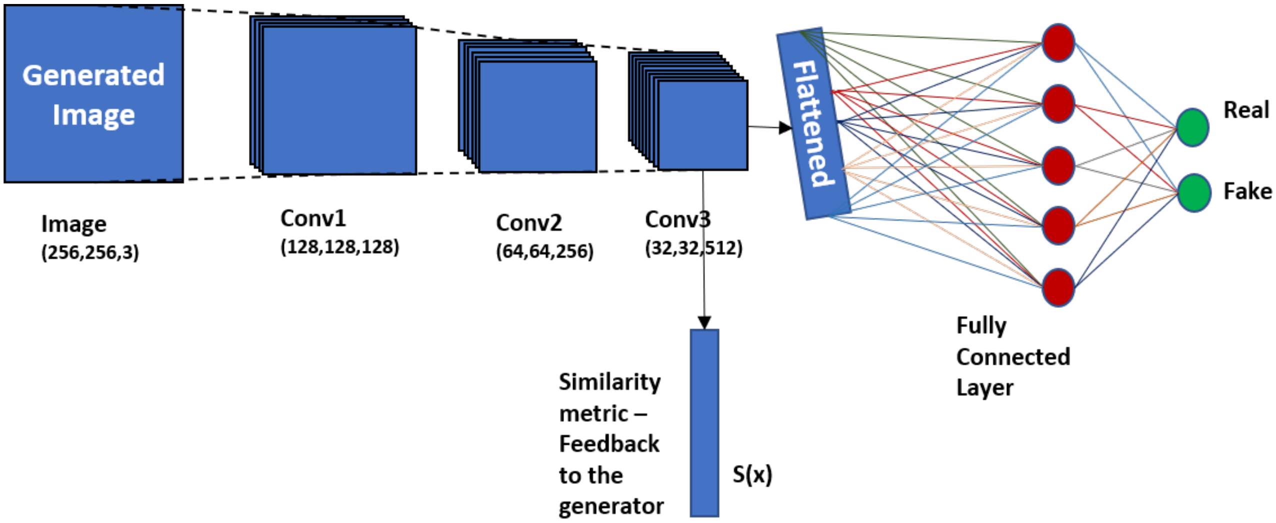


Figure 4.3: Discriminator Architecture

For generator and discriminator, number of feature maps depends on the size of the convolution window, 256 in our network.

4.2.3 Mode Collapse

Mode collapse is a problem that occurs while training GANs, when G is trained extensively without updates to D. The problem can be defined as follows: the generator finds an optimal output image that successfully fools the discriminator into thinking that it is a real image and tries to generate more instances of that one image regardless of the input provided, collapsing the mode to one point (image in this case). Instead of generating new thermal face images of different identities, the generator keeps generating more images of one or two identities. To solve this problem, an additional layer is included towards the end of the discriminator network before the sigmoid activation — a minibatch discrimination layer [61].

To achieve minibatch discrimination, we send output from the generated (fake images) and real images to the discriminator in separate mini batches. The discriminator now has an additional task to discriminate between each image in a mini batch of fake images to the rest of the images in the same mini batch. Similarity 's' between an image (x_i) in the mini batch and all the other images in the mini batch is calculated by discriminator. If the mode starts to collapse, the similarity between the generated images increases and this in turn penalizes the generator. To allow minibatch discrimination, a layer to calculate similarity is added to the discriminator before the sigmoid activation. To keep this addition simple and cost effective, HOG features are extracted from the generated images and are used to measure the similarity using L1-Norm.

4.2.4 GAN Loss Functions

The loss function proposed by Goodfellow et al. in [1] is used with the addition of loss from the HOG features for minibatch discrimination, and can be defined as below:

$$L_{GAN} = E_{x,y}[logD(x,y)] + E_{x,z}[log(1 - D(x, G(x,z))] - L_{HOG}, \tag{4.1}$$

where L_{HOG} is the loss from minibatch discrimination layer calculated using L1-Norm, and is obtained as follows:

$$L_{HOG} = \sum_{i=1}^{N} L_{HOG_i} \tag{4.2}$$

where

$$L_{HOG_i} = \sum_{j=1}^{N-1} \left\| x_i - x_j \right\|_1 \tag{4.3}$$

where x_i is the HOG of the input image, x_j are the HOG of the rest of the images from the minibatch and N is the number of images in the minibatch.

Chapter 5

Thermal-to-Visible Face Recognition using GANs

5.1 Methodological Approach

The methodology of the proposed thermal-to-visible GAN face matcher, is as follows: Faces are detected and cropped from all the visible and thermal images. These cropped images are then saved and geometrically normalized by locating the eye centers manually. The geometrically normalized visible faces are used to generate features (embeddings) using the Facenet model, which are used to condition the generator and the discriminator networks. Once the features are generated, the conditional GAN is trained and the generated visible images are used as probe images against the original visible gallery and face matching is performed using Facenet face recognition. The entire methodology is presented in Figure 5.1.

5.2 CNN Architecture

The CNN architecture used has the following building blocks — 1. Face Detection network to detect faces from visible and thermal face images 2. A feature extractor to extract features from visible images, 3. A generator network, 4. A discriminator network 5. An identity network.

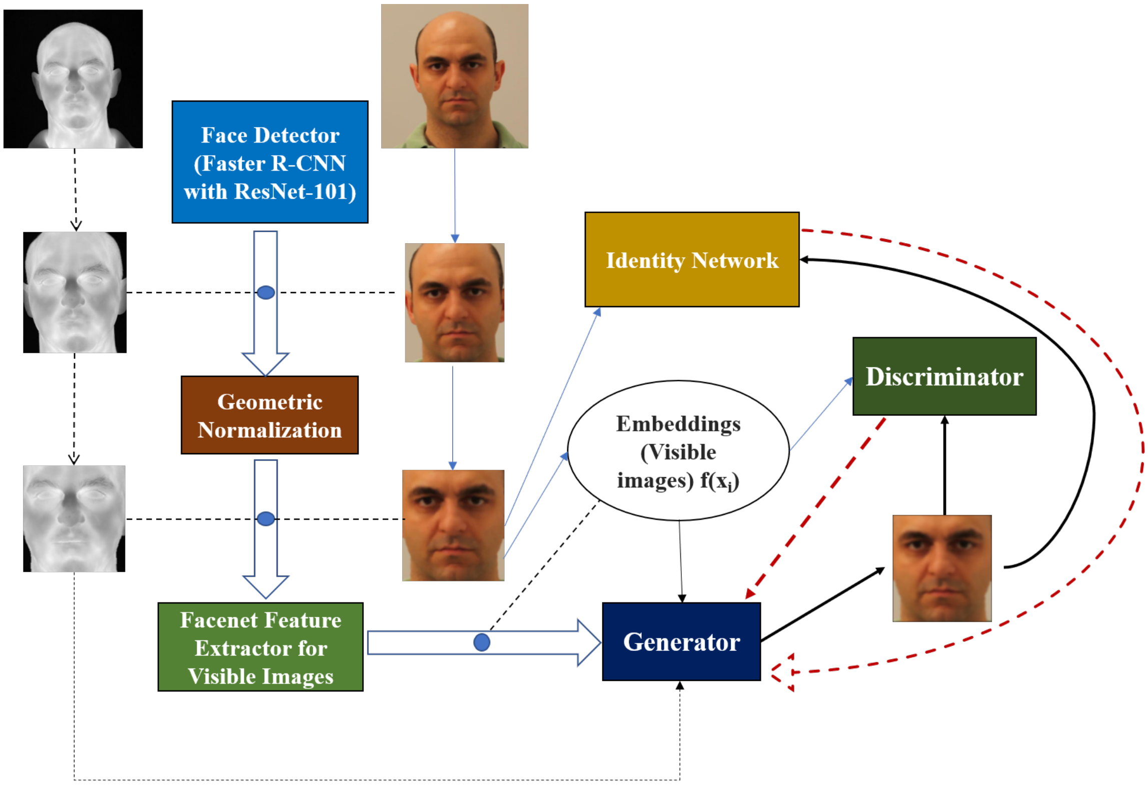

Face Detector (Faster R-CNN with ResNet-101)
Geometric Normalization
Facenet Feature Extractor for Visible Images
Identity Network
Discriminator
Embeddings (Visible images) f(x_i)
Generator

Figure 5.1: Overview of the proposed approach to generate visible images from thermal - thin solid line shows the flow of thermal images, dotted line-visible images, thick solid line-generated visible images and dashed line-feedback

5.2.1 Face Detection

Visible and thermal faces are detected using the face detection models developed in Chapter 3. These faces are geometrically normalized by manually locating the eye centers from each image and rotating the image so that the eye centers are on a horizontal line and the image is cropped. This generated images of size 256×256.

5.2.2 Feature Extractor

Facenet [6] face recognition model is proven to be a highly reliable and robust face matcher. Therefore, this is used without any training to extract features from visible face images which are used to condition generator and discriminator networks. The Facenet model used here is obtained by training an Inception Resnet model using 200 million face thumbnails and the face recognition accuracy over LFW Dataset was 99.63%. The network generates an embedding *f(x)* from an image x into a feature space, such that the squared distance between images of the same identity is smaller than that of different identities. A 224×224 window size waas used due to the fact that it is the closest in size to our image, 256×256. The 256×256 window size achieved the highest accuracy among the others (160×160, 96×96).

5.2.3 Generator Network

Pix2pix [7] serves as a base network for our GAN architecture. Following this work, we use a U-Net encoder based generator network. The inputs for this network are the geometrically normalized thermal images and the embeddings generated by the feature extractor network work as condition for the generator. Vanilla GANs use noise as input to the generator, however in this case, thermal images are used as input. If noise is completely omitted, the generator produces deterministic results and if noise is provided as input alongside images, the generator simply learns to ignore the noise as the training progresses. To avoid this, noise is applied only as dropout on several layers of generator. The generator has the following architecture:

CL64-CL128-CL256-CL512-CL512-CL512-CL512 — CDR512-CDR1024-CDR1024-CDR1024-CDR512-CR256-CR128-CT

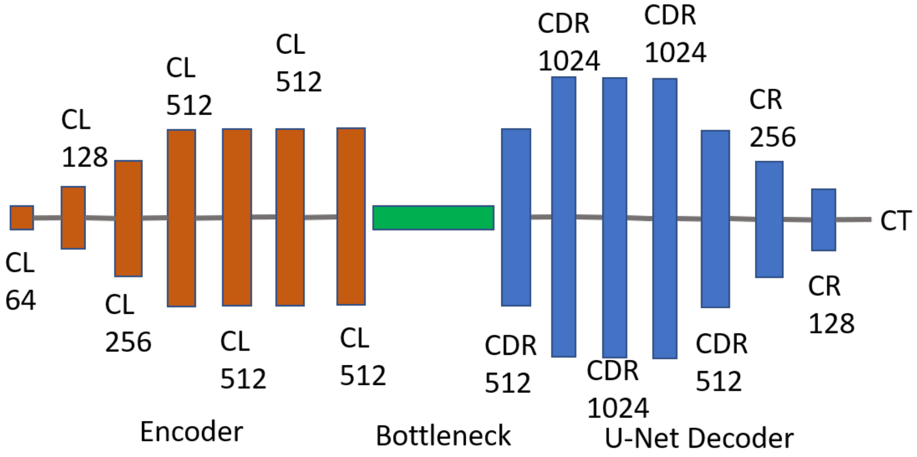


Figure 5.2: Generator Architecture using U-Net Decoder — CL:Convolution-BatchNorm-LeakyReLU, CDR:Convolution-BatchNorm-Dropout-ReLU, T: Tangent Hyperbolic Function

where CLk denotes Convolution-BathNorm-LeakyReLU layer with k filters, with the exception of the first layer of the encoder where BatchNorm is not applied and CDRk denotes Convolution-BatchNorm-Dropout-ReLU layer with a dropout rate of 50%. The architecture is represented in Figure 5.2. All the convolutions are 4×4 spatial filters applied with stride 2 and convolutions in the encoder down-sample by a factor of 2 and in the decoder up-sample by a factor of 2. The CT layer is the convolution applied to map to the number of output channels followed by a Tanh activation.

All the leaky ReLUs have a slope of 0.2. As mentioned earlier, noise is applied as dropout, in this work it is Gaussian with '0' mean and a standard deviation of 0.01.

5.2.4 Discriminator Network

The patch-based discriminator in pix2pix is replaced by a pre-trained Inception ResNet *v1* model in this work. The pre-trained Inception ResNet model classifies images into various categories from the ILSVRC [62] dataset. This model is re-trained to classify the images as real or fake. The last layer of this network receives 1792 dimensional vector as input, followed by average pooling, dropout and softmax layers. In general, if the generator and discriminator networks are trained at the same rate, discriminator starts classifying all the images generator generates as fake, which makes it difficult for the generator to converge. To overcome this, discriminator has trained at a slower rate than generator, usually by dividing the objective function by 2 when training the discriminator. In this case, since we fine tune and use a robust classifier as discriminator, it is necessary to train the discriminator at an even slower rate. This is achieved by dividing the objective function when training the discriminator by 4.

5.2.5 Identity Network

An identity network is introduced to provide G with more feedback about the identity of the image being generated. This network is again a pre-trained Facenet model, re-trained and fine tuned along side D, to provide feedback to the generator about the identity of the generated images. The Facenet model is re-trained using triplet loss method [6].

As the name suggests, the triplet loss method uses triplets of data to train the network. Each triplet consists of an anchor image, a positive image and a negative image, where the anchor and the positive image belong to the same identity and the negative image is of any other different identity. The aim of the training is to ensure that, in all the triplets, the positive image is closer to the anchor than the negative image in the embedding space, which live on the d-dimensional Euclidean hypersphere as shown in Figure 5.3. Thus, the relation between the three images in a triplet can be defined as:

$$\left\|x_i^a - x_i^p\right\|_2^2 + \alpha < \left\|x_i^a - x_i^n\right\|_2^2 \ \forall \ (x_i^a, x_i^p, x_i^n) \in \tau \tag{5.1}$$

where $\mathrm{x_i}^\mathrm{a}$ is the anchor image, $\mathrm{x_i}^\mathrm{p}$ is the positive image and $\mathrm{x_i}^\mathrm{n}$ is the negative image in the triplet and τ is the set of all possible triplets in the data set. α is a margin enforced between positive

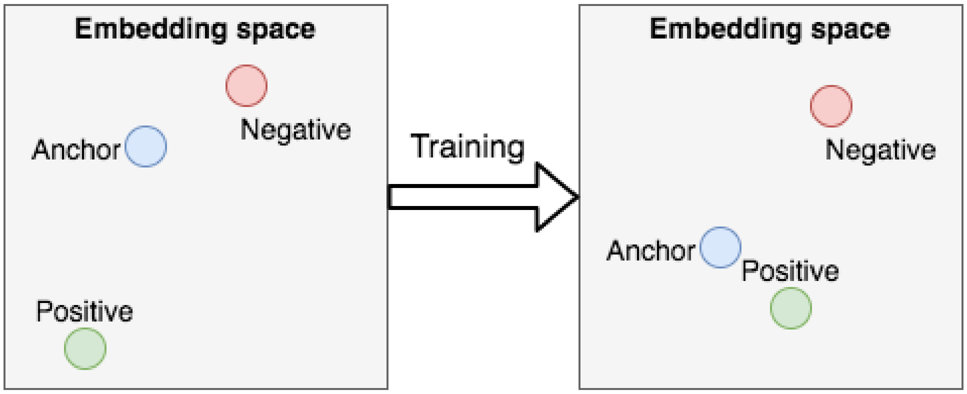


Figure 5.3: Illustration of Triplet Training

and negative pairs. If all the possible triplets are used for training, it becomes impossible for the network to converge and the the triplets that already satisfy equation 5.1 would not contribute to the training. To ensure fast convergence, it is important to select triplets that do not satisfy the above equation, which means, for any x_i^a, x_i^p is selected such that the distance between x_i^a and x_i^p is maximum $(argmax_{x_i^p} \|f(x_i^a) - f(x_i^p)\|_2^2)$ and x_i^n such that the distance between x_i^a and x_i^n is minimum $(argmin_{x_i^n} \|f(x_i^a) - f(x_i^n)\|_2^2)$. The data set for training original facenet model consists about 200 million images, so they used online hard mining to select hard triplets from with in a mini-batch. Since our data has only around 456 images, we do not use any mining technique to select triplets, rather compute all the possible argmin and argmax and use the triplets from these distances. After training, this network provides feedback to generator in the form of the identity number.

The basic building blocks of the inception resnet *v1* are shown in Figure 5.4. The Reduction network is less expensive while also being an efficient way to reduce the dimensionality.

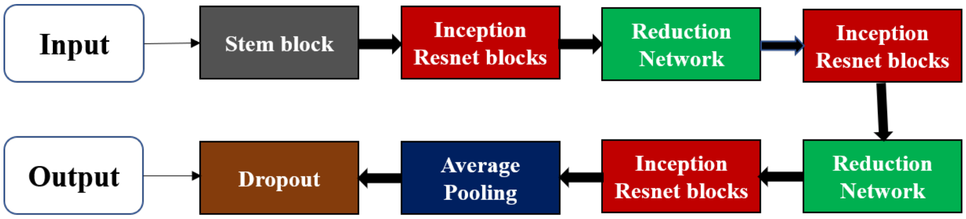


Figure 5.4: Basic building blocks of Inception Resnet *v1*

5.3 Objective Function

Our thermal-to-visible GAN is optimized by minimizing the following loss function:

$$L_{FGAN} = L_{GAN} + \lambda_1 L_1 + \lambda_t L_t \tag{5.2}$$

where L_{GAN} is the adversarial loss for GAN, L_1 is the L1 loss and L_t is the triplet loss. λ_1 and λ_t are weights for L1 and triplet losses respectively.

5.3.1 Adversarial Loss

The GAN loss is:

$$L_{GAN} = E_{x,y}[logD(x,y)] + E_{x,z}[log(1 - D(x, G(x,z))], \tag{5.3}$$

where X is the input thermal image, Y is the embedding generated by the feature extractor and z is the dropout noise. G tries to minimize this loss, while D tries to maximize it.

5.3.2 L1 loss

Generator is able to fool the discriminator with the above loss functions, however using L1 loss between real and generated images proved to provide the generator enough feedback to be able to generate near ground truth images.

$$L_1 = \|y - G(x,z)\|_1 \tag{5.4}$$

5.3.3 Triplet Loss

Triplet loss function, used in the identity network described above is defined as:

$$L_t = \sum_{i=1}^{N}[\left\|f(x_i^a) - f(x_i^p)\right\|_2^2 - \left\|f(x_i^a) - f(x_i^n)\right\|_2^2 + \alpha] \tag{5.5}$$

where $f(x_i^a), f(x_i^p) and f(x_i^n)$ are the embeddings of anchor, positive and negative images respectively.

After the images are generated, Wiener filter with window size [5 5] is used to de-blur the images

Chapter 6

Experimental Evaluation and Results

All the experiments are performed using Ubuntu 16.04/18.04 LTS 64-bit on a Titanx GPU in TensorFlow.

6.1 Face Detection in Thermal and Visible Spectrum

The steps include manually annotating all the images to generate bounding boxes for ground truth labels, training and testing several models, fine-tuning.

6.1.1 Datasets

Images in the thermal and visible bands are collected at two different stand-off distances — 5m and 10m and in indoor and outdoor environments. Therefore, a total of four categories for each of thermal and visible bands are used in this study from a total of 57 subjects and around 6000 images in total (3000 from each band). No image augmentation or image pre-processing techniques are used. The scenarios used are 5m—indoor, 5m—outdoor, 10m—indoor, 10m—outdoor.

6.1.2 Training and Parameter Tuning

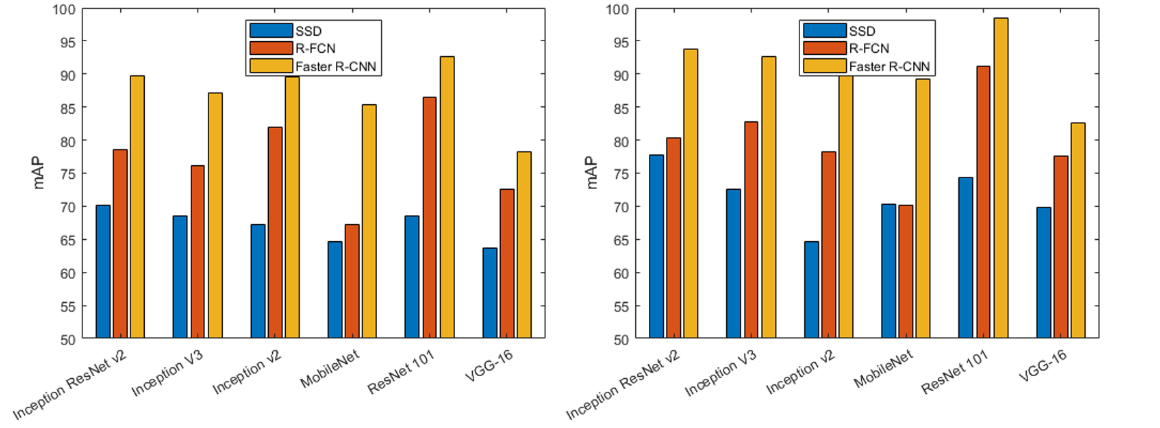
Learning rate is varied from 0.002 to 2e-6 for each of the combination of model and feature extractor using AdaGrad, Momentum, Adam and RMSProp optimizers which lead to a total of 720 experiments, summarized in Table 6.1. These experiments are repeated for visible and thermal data separately with 90% of the data for training and the remaining 10% for validation.

For all the combinations of models and feature extractors, for learning rates below 1e-4, the

Table 6.1: Accuracy of detector using different models assessed

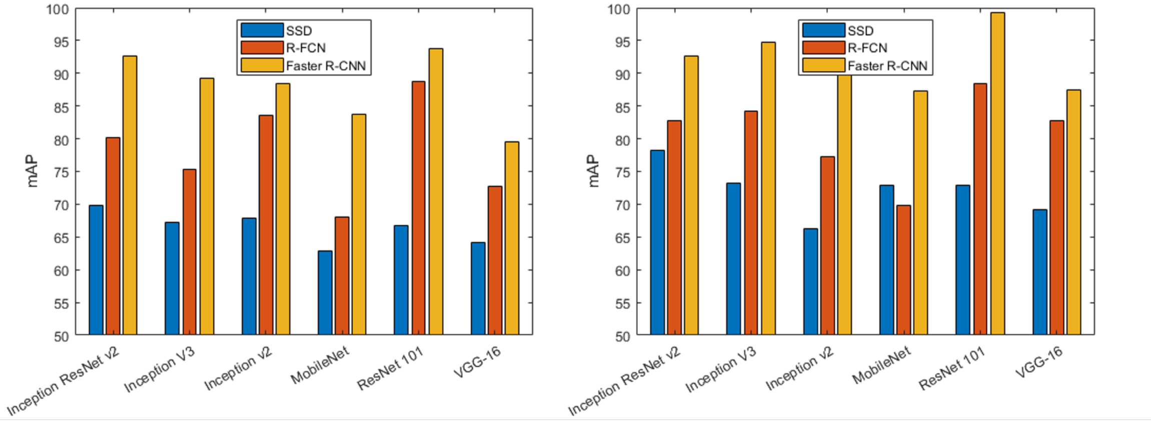
Model	Learning rate	mAP-Thermal (%)	mAP-Visible (%)
SSD with VGG 16	1e-4	63.7	64.2
SSD with ResNet 101		68.6	66.7
SSD with Inception *v*2		67.3	67.8
SSD with Inception *v*3		68.6	67.2
SSD with Inception ResNet *v2*		70.2	69.8
SSD with MobileNet		64.6	62.8
R-FCN with VGG 16		72.6	72.8
R-FCN with ResNet 101		86.5	88.7
R-FCN with Inception *v*2		81.9	83.6
R-FCN with Inception *v*3		76.2	75.4
R-FCN with Inception ResNet *v2*		78.6	80.2
R-FCN with MobileNet		67.2	68.1
Faster R-CNN with VGG 16		78.3	79.6
Faster R-CNN with Resnet 101		92.6	93.7
Faster R-CNN with Inception *v*2		89.6	88.4
Faster R-CNN with Inception *v*3		87.2	89.3
Faster R-CNN with Inception ResNet *v2*		89.8	92.7
Faster R-CNN with MobileNet		85.3	83.8
SSD with VGG 16	3e-5	69.8	69.2
SSD with ResNet 101		74.3	72.9
SSD with Inception *v*2		64.6	66.2
SSD with Inception *v*3		72.6	73.2
SSD with Inception ResNet *v2*		77.8	78.2
SSD with MobileNet		70.3	72.9
R-FCN with VGG 16		77.6	82.7
R-FCN with ResNet 101		91.2	88.4
R-FCN with Inception *v*2		78.3	77.2
R-FCN with Inception *v*3		82.7	84.2
R-FCN with Inception ResNet *v2*		80.4	82.8
R-FCN with MobileNet		70.2	69.8
Faster R-CNN with VGG 16		82.6	87.5
Faster R-CNN with Resnet 101		**98.4**	**99.2**
Faster R-CNN with Inception *v*2		93.8	93.6
Faster R-CNN with Inception *v*3		92.6	94.8
Faster R-CNN with Inception ResNet *v2*		93.7	92.6
Faster R-CNN with MobileNet		89.2	87.3

models did not converge. After extensive training and validation experiments, 3e-5 as learning rate is found to be optimal for all the models. Therefore, the results obtained using learning rate values of 1e-4 and 3e-5 are presented in Table 6.1 using Adam optimizer. The accuracy of all our models is evaluated using mAP (mean Average Precision) metric, which is the average of all the precision values over the recall range of 0 to 1.



(a) (b)

Figure 6.1: Thermal face detection accuracy for(a) learning rate of 1e-4, (b) learning rate of 3e-5



(a) (b)

Figure 6.2: Visible face detection accuracy for(a) learning rate of 1e-4, (b) learning rate of 3e-5

SSD model with MobileNet resulted in 74.3% and 72.9% mAP (mean Average Precision) over the thermal and visible sets respectively and is the fastest among all with 1ms for single detection. SSD with ResNet-101 yielded an mAP of 74.3.5% and 72.9% over thermal and visible sets respectively. Faster R-CNN with ResNet-101 yielded the highest mAP over thermal and visible data sets with 98.4% and 99.2% with 60ms for each detection task.

6.1.3 Experiments by reducing the number of proposals in the RPN for Faster R-CNN and R-FCN

In all the above experiments, the number of proposals generated by RPN during detection time is 300. Speed can be reduced by decreasing the number of these proposals. Therefore, another set of experiments were performed by reducing the number of proposals in R-FCN and Faster R-CNN models with ResNet-101. The number of proposals used in the above set of experiments is 300 and this number is reduced by 50 each time. For R-FCN, when number of proposals were decreased, time taken for single detection remained the same (2ms), however accuracy decreased by a large amount. For 150 proposals, time taken for a single detection for Faster R-CNN decreased to 20ms, while the mAP remained at 95.28%, after which speed increased only a little and mAP decreased more.

Table 6.2: Accuracy of Faster R-CNN for different number of region proposals

Number of proposals	Thermal (%)	Visible (%)	time (ms)
300	98.40	99.20	60
250	98.17	96.54	54
200	96.87	96.19	45
150	**95.28**	**95.72**	**20**
100	84.88	86.54	17
50	76.26	72.55	12

Table 6.3: Accuracy of R-FCN for different number of region proposals

Number of proposals	Thermal (%)	Visible (%)	time (ms)
300	91.20	88.40	60
250	88.29	86.67	54
200	85.94	82.53	45
150	**84.28**	**81.90**	**20**
100	72.78	70.53	17
50	65.37	62.48	12

Speed-Accuracy trade-off curves are shown in Figure 6.3.

Figure 6.4 shows two examples of images where the detection is successful and two examples where it is not. A large percentage of images used in the training and test sets have the subject looking at the camera. The algorithm failed in the images where this is not the case, i.e., it failed when the subject is looking in other direction.

6.1.4 Face Recognition Experiments

For thermal face recognition, using conventional FR approach (SS score-level fusion of HOG and LBP), the Rank-1 face identification accuracy values yielded are 95.26% and 85.78% for the 5m and 10m distances respectively. For visible, these are 99% and 97.4% for the 5m and 10m distances respectively. In contrast, when the same FR models were used but the face images were manually cropped, the yielded accuracy is 97.67% (5m MWIR) and 91.20% (10m MWIR) and 100% in both 5m and 10m visible band face data sets. Face recognition experiments and the results are presented

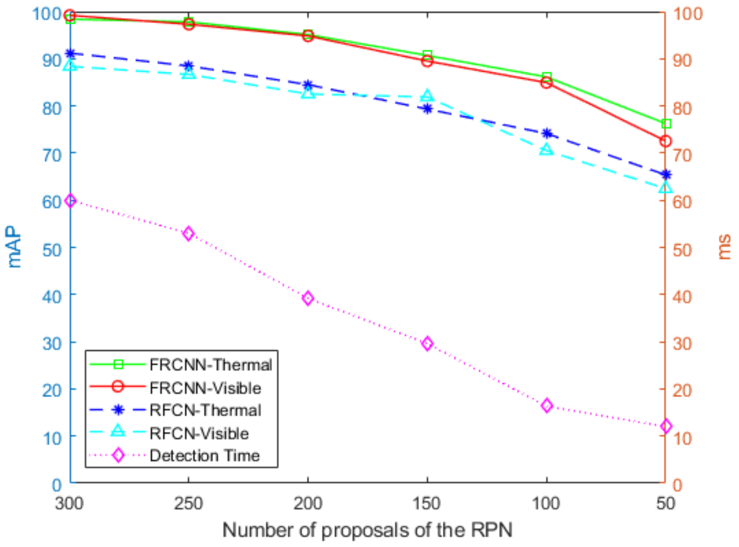


Figure 6.3: Speed-Accuracy Trade-off curves for Faster R-CNN and R-FCN with change in number of proposals of RPN

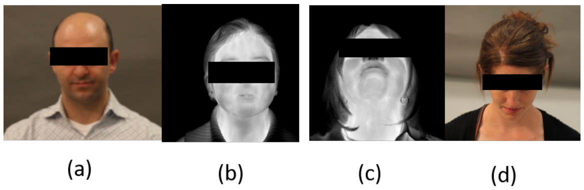

[redacted]
[redacted]
[redacted]
[redacted]

Figure 6.4: (a,b)-Images where face detection worked, (c,d)-Images where face detection failed

in Table 6.4.

Table 6.4: Face Recognition Results - I:Indoor, O:Outdoor

Datasets	Gallery	Probe	Manual (%)	Automated (%)
Visible	5m Indoor	5m Outdoor	100	99
	10m Indoor	10m Outdoor	100	97.4
Thermal	5m Indoor	5m Outdoor	97.67	96.19
	10m Indoor	10m Outdoor	91.20	86.54

6.2 Data Augmentation to Improve Detection Performance in MWIR Spectrum

6.2.1 Datasets

Data sets used are thermal face images, collected at two different distances (5m and 10m) and two different scenarios (indoor and outdoor). This contributes to a total of four scenarios i.e., 5m, indoor; 5m, outdoor: 10m, indoor and 10m outdoor. Data from 56 different individuals is used in this study constituting a total of 3000 thermal face images.

6.2.2 Traditional Image Enhancement

For all the enhancement techniques presented in this section, images generated using one method and one parameter are added to the original images and the face detection model is trained and tested. At the end, all the images generated using one technique changing all the parameters are included to train and test the face detector and this increased the size of the dataset by a large amount. Instead of changing the learning rate for each of the experiments, batch size is fixed to be the same. The learning rate is only changed when all images from an augmentation technique are used for training and testing to 3e-6.

Brightness

The first set of experiments were performed by varying brightness of images. Brightness is changed in increments of 10 until the brightness of each pixel is increased to 50 and the images of each brightness level are added to the the original data and are used to train and validate the Faster R-CNN model. At the end, all the images are used at once to train and test the model. Images with increased and decreased brightness levels are shown in Figures 6.5 and 6.6 and the mAP values for proposals of the RPN from 300 to 50 are presented in Table 6.5 and Fig 6.7.

Table 6.5: Accuracy of FRCNN for different brightness levels

Proposals	+10	+20	+30	+40	-10	-20	-30	-40	-50	All Data
300	98.64	97.95	98.90	98.82	98.76	98.85	98.82	99.01	98.63	99.64
250	98.28	98.66	98.59	98.67	98.23	98.26	98.77	98.38	97.26	99.18
200	96.89	97.23	96.54	97.63	96.98	96.90	97.26	97.38	98.23	97.43
150	**95.78**	**96.89**	**95.27**	**96.54**	**95.57**	**95.60**	**96.29**	**96.24**	**95.98**	**96.88**
100	87.66	86.47	87.28	86.26	85.03	85.38	86.24	87.23	86.98	87.03
50	78.29	77.98	76.20	77.49	77.28	76.46	77.23	78.91	77.29	77.95

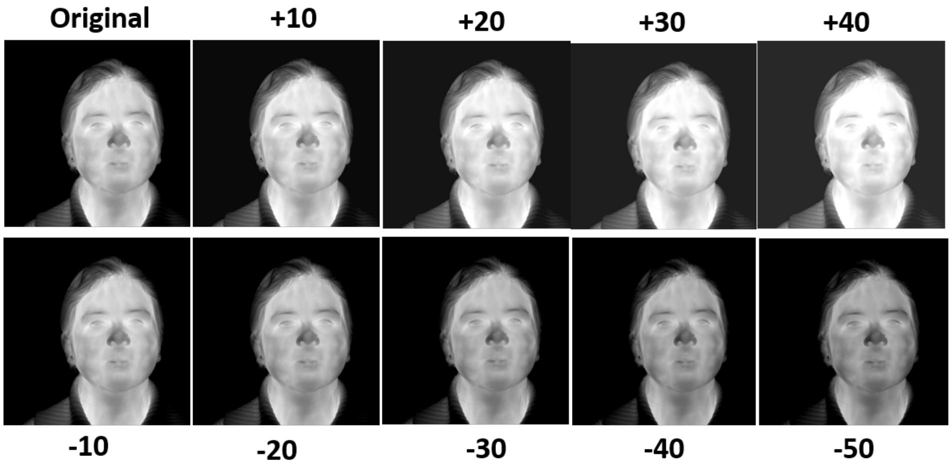


Figure 6.5: Images generated by varying brightness - Subject1

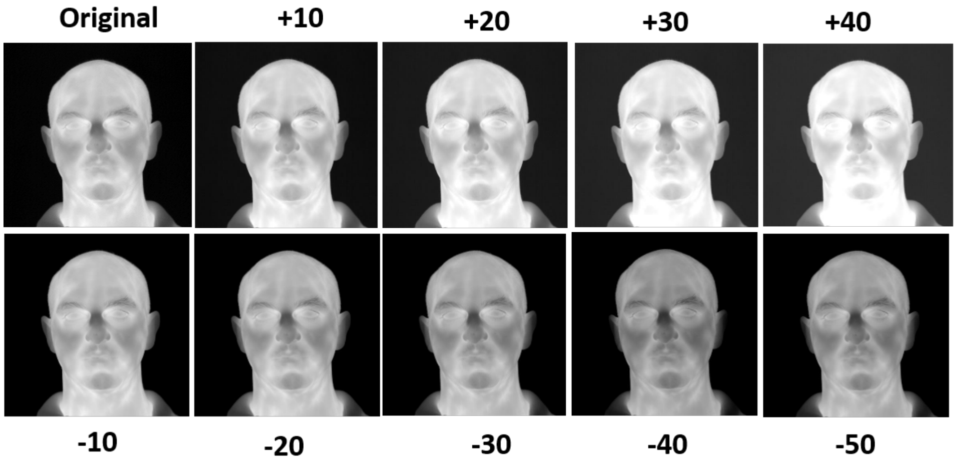


Figure 6.6: Images generated by varying brightness - Subject2

Contrast Enhancement

Three types of contrast enhancement techniques are used for this purpose. Images generated are shown in Fig 6.8 and results are presented in Table 6.6 and 6.9.

Table 6.6: Accuracy of FRCNN for different contrast enhancement methods

Proposals	imadjust	histeq	CLAHE	All Data
300	98.67	99.03	98.98	99.23
250	99.23	99.05	98.34	99.35
200	96.89	96.43	96.98	97.80
150	**96.75**	**97.20**	**97.03**	**97.64**
100	85.19	86.28	86.48	87.27
50	76.46	77.43	77.96	78.02

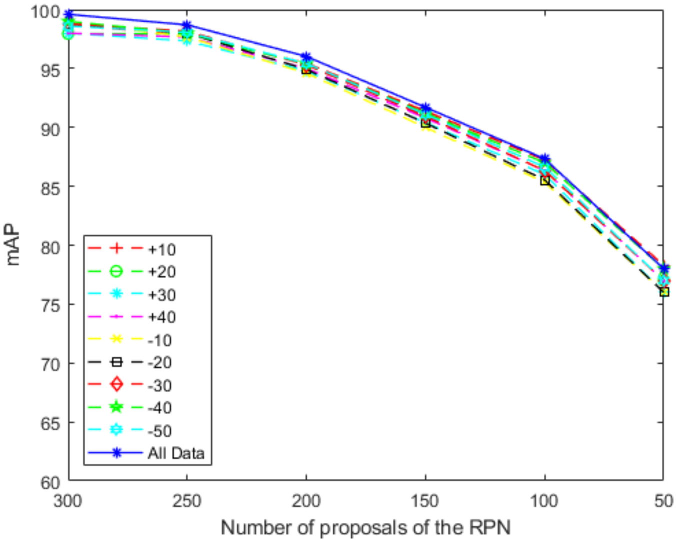


Figure 6.7: mAP vs number of proposals of RPN for varying brightness values

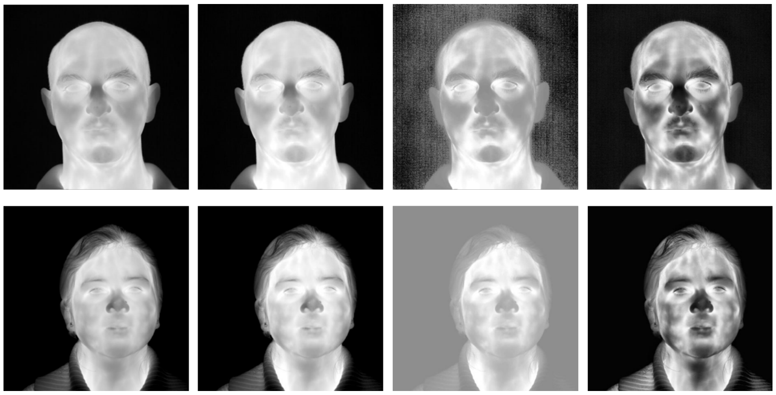


Figure 6.8: Images generated through contrast enhancement — left to right — original image, regular contrast enhancement, histogram equalization, CLAHE

Applying Noise and De-noising the Images

The next image processing technique used is applying noise and filtering out the noise. Gaussian noise is applied to the images with a mean of '0' and a standard deviation of 0.025. This applied Gaussian noise is then filtered out using Wiener filter with 5×5 as the size of the neighborhood window used to estimate the local mean and variance. Salt & Pepper noise is applied to the original images with a density of 0.02 i.e., 2% of the pixels are affected. Then, median filter is used to de-noise the images. Fig shows original, noisy and filtered images. Images generated by applying Gaussian noise and Wiener filter are shown in Figure 6.10 and by applying Salt & Pepper noise and median filter are shown in Figure 6.11. The accuracy of face detector for different number of proposals of RPN is shown in Table 6.7 and presented in Fig 6.12.

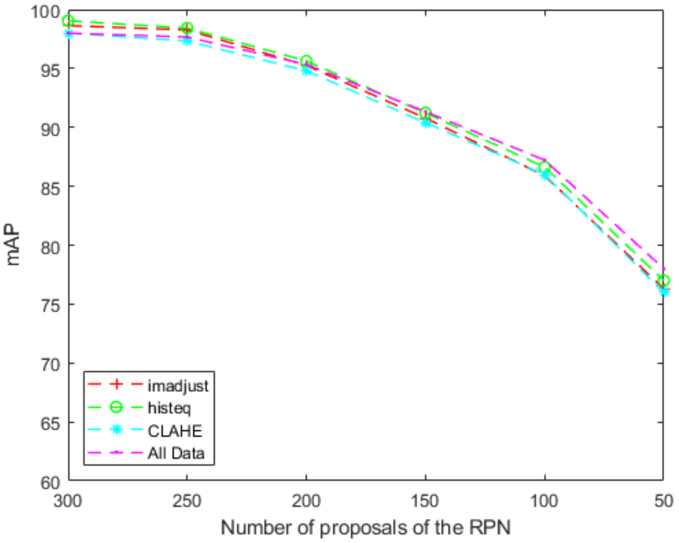


Figure 6.9: mAP vs number of proposals of RPN for different contrast enhancement techniques

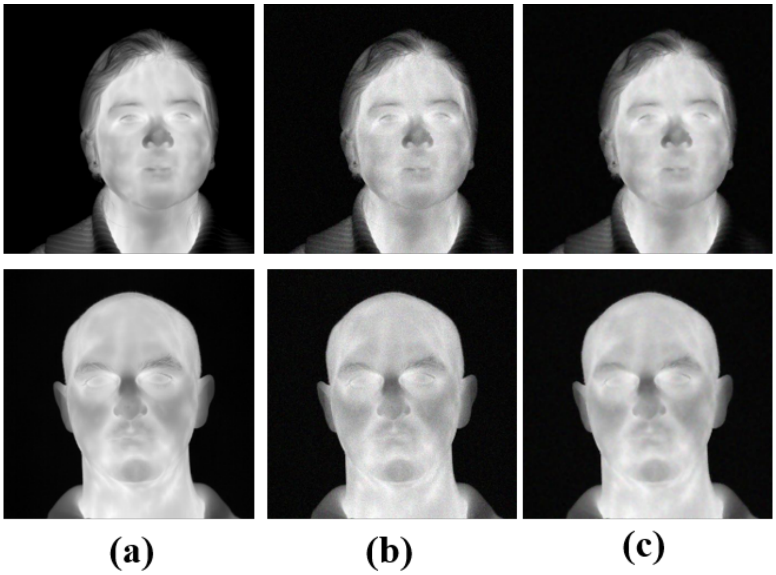


Figure 6.10: Images generated by noising and de-noising the images, Left—to—right, Original image, Gaussian noise applied, Wiener filter applied

From Tables 6.5, 6.6 and 6.7, it can be observed that the accuracy increased by a small amount for any of the augmentation techniques while the time required to train the models increased by a large amount for each of these augmentation techniques. By augmenting the data using traditional augmentation techniques, a robust model is trained that works well with noisy images, bright images, dark images etc.

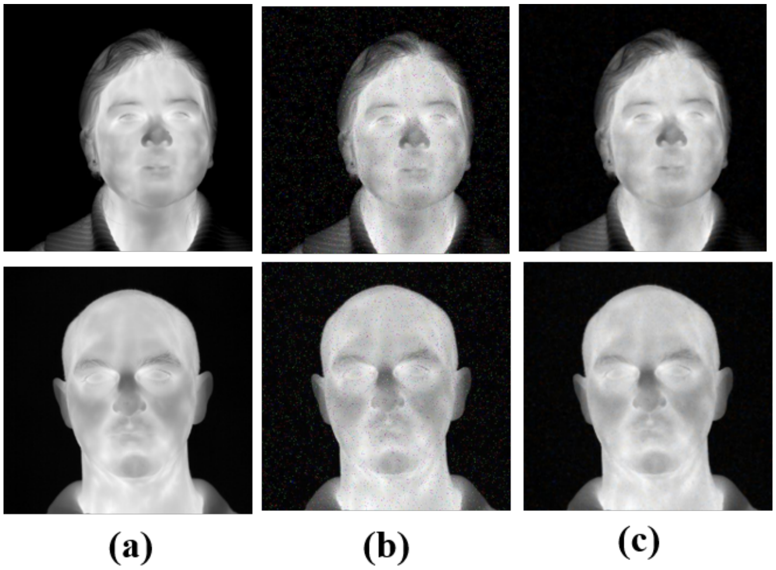


Figure 6.11: Images generated by noising and de-noising the images, Left—to—right, Original image, Salt and Pepper noise applied, Median filter applied

Table 6.7: Accuracy of FRCNN for different noising and de-noising algorithms

Proposals	Gaussian noise	wiener filter	Salt and Pepper noise	Median filter	All images
300	98.23	98.72	98.76	98.58	99.67
250	97.99	98.20	98.27	98.45	99.40
200	97.16	97.91	97.77	98.23	99.15
150	**95.47**	**96.38**	**96.44**	**96.47**	**97.42**
100	85.26	86.43	87.21	87.13	89.25
50	77.10	78.00	77.85	78.24	79.56

6.2.3 GAN Data Augmentation

Original thermal face dataset used for face detection in Chapter 3 is used here to train the GAN. This includes 6000 images from thermal spectrum.

Pre-processing

Face detection is applied to the original images and all the images are cropped to include only the face area. These images are resized to 256×256, This is the window size for our GAN architecture and are scaled to the range of [-1,1] for tanh activation.

Adversarial Training

The model is trained with mini-batch SGD (Stochastic Gradient Descent) with a mini-batch size of 32. All the weights are initialized with a zero-centered mean with standard deviation of 0.02. In

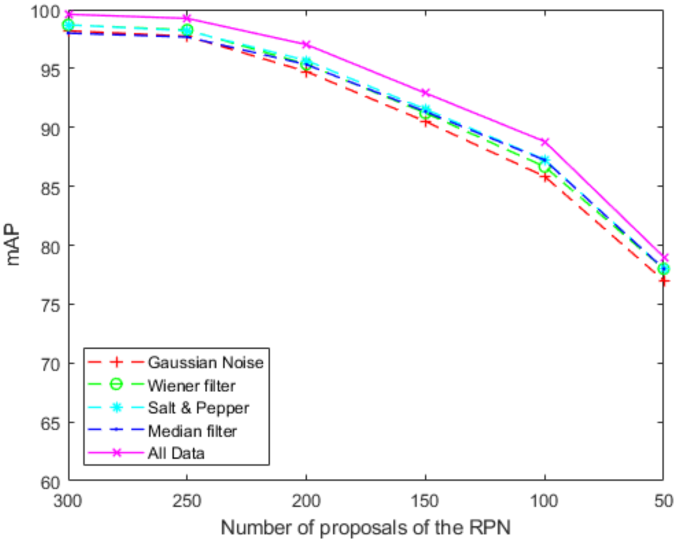


Figure 6.12: mAP vs number of proposals of RPN for different types of noise and filters

the leaky ReLU, the slope of the leak used is 0.2 in all the layers. The training process is accelerated using an ADAM optimizer with tuned hyper-parameters. The learning rate used is 2e-5. Default values for the first and second estimates of the ADAM optimizer are 0.9 and 0.999 respectively and that of ϵ is 1e-8. However, the value of 0.9 for the first estimate resulted in training oscillation and instability. This is solved by reducing its value to 0.5. The model converged after 150 epochs and this took 200 hours using a Titanx GPU. After the images are generated, similarity score using HOG is calculated once again between each image generated and rest of the generated images.

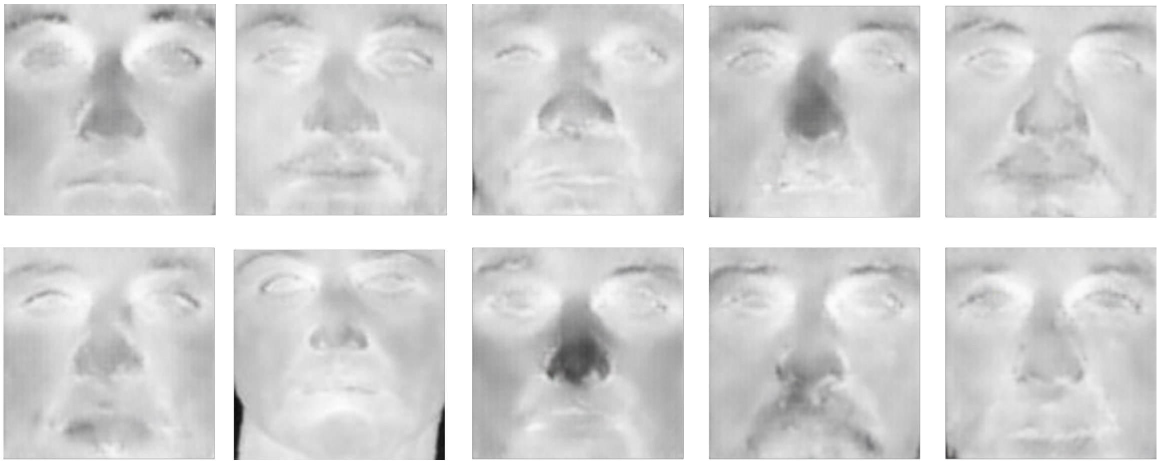


Figure 6.13: Instances of generated faces using GANs

The images that are too similar are rejected and this resulted in the generation of 62 new faces in thermal spectrum, few instances of generated faces are shown in Fig 6.13 The generated images are used to train the Faster R-CNN face detector alone and are then used with the original images. The learning rate used when training with only the generated images with original and generated images together is 3e-5. mAP is calculated at 300 to 50 proposals. Results are tabulated in 6.8 and are presented in Fig 6.14.

Table 6.8: Accuracy of FRCNN for original, generated and original+generated images

Proposals	Generated	Original+Generated
300	97.26	100
250	98.25	99.96
200	97.13	99.88
150	**95.73**	**99.83**
100	85.76	92.67
50	75.92	80.68

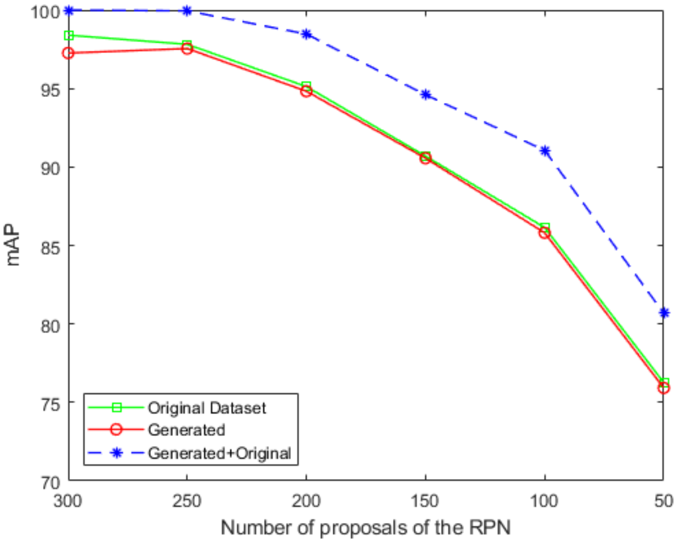


Figure 6.14: mAP of the Faster R-CNN when trained with original, generated and original+generated images vs number of proposals of RPN

6.3 Thermal-to-Visible Face Recognition

6.3.1 Datasets

MWIR and visible images collected at a 5m and 10m distances, indoor and outdoor are used for all the thermal-to-visible experiments. The data set includes images from 51 subjects, 16 in visible and 16 in thermal for each subject, resulting in a total of 3264 images.

6.3.2 preliminary Experiments

Two preliminary experiments are conducted to understand the efficiency of FaceNet model for matching original visible images to their thermal counterparts. These are explained below:

- *Experiment 1* — Facenet model is used to match original visible-to-thermal images without re-training. The Rank-1 accuracy yielded here is 8.24%

- *Experiment 2* — Facenet model is re-trained with a learning rate of 0.00002 using ADAM optimizer (0.5, 0.999) for 100 epochs with a batch size of 1. The accuracy yielded here is 11.76%. The train-test split is 90-10.

6.3.3 GAN Implementation

The original Inception style Facenet model is trained on 200 million face thumbnails of visible images. It is used without any training to extract features (embeddings) from visible face images in our dataset. The GAN network is trained with default values for ADAM optimizer (0.5, 0.999), with a batch size of 1 for a total of 200 epochs. The initial learning is fixed at 0.0002 and is decreased after 125 epochs by a factor of 1/10 after every 5 epochs. The value for λ_1 is 10 (following [46]) and λ_t is 0.5. The identity network is trained using the same initial learning rate of 0.0002 for the first 100 epochs and is then decreased by a factor of 1/100 for every 10 epochs. The parameter and hyper parameter values assigned for the final experiments are decided after running experiments with different combinations of weights. This network is evaluated over 51 subjects out of which 85% (42 subjects) are used for training with 32 images for each subject — 16 visible and 16 thermal. Since the size of our dataset is too small, all the networks are trained using 5-fold cross validation.

6.3.4 GAN Experiments

A number of experiments are conducted with GAN, adding different networks and weights to generate visible images from thermal images. This is followed by performing visible-to-visible face matching experiments using Facenet.

- *Experiment 3* — First, the pix2pix [7] is trained without any changes to the network using the original U-Net encoder and patch based discriminator with a patch size of 70×70. We also trained and tested pixelGAN that matches an image to another through pixel by pixel matching and image GAN of size 256×256, however the results using 70×70 patch are visually pleasing and yielded better recognition accuracy. However, the accuracy is very low at 13.23%

Table 6.9: Rank-1 Face Recognition Accuracy - (1) is pre-trained feature extractor, (2) pre-trained discriminator, (3) identity network

Network	Rank-1 Recognition Accuracy (%)
Facenet model	7.26
Facenet model after re-training	11.76
pix2pix	13.23
pix2pix with (1)	58.66
pix2pix with (1) and (2)	87.98
pix2pix with (1), (2) and (3)	92.83
pix2pix with (1), (2) and (3) with de-blurring	**97.63**

- *Experiment 4* — For the next experiments, the patch based discriminator is replaced by a pre-trained Inception ResNet model [53], and is fine tuned and optimized to be able to classify images into fake and real categories. The pre-trained weights derived from training the model with COCO dataset are used for transfer learning here. The accuracy increased significantly to 58.66% by using a pre-trained model for discriminator.

- *Experiment 5* — For the next set, a pre-trained Facenet model is used to extract features from the visible images and the embeddings generated are used to condition the GAN. The pre-trained Facenet model weights obtained by training Inception ResNet *v1* on ILSVRC [] dataset are used directly to extract the features without any re-training. The addition of this pre-trained feature extractor increased the face recognition accuracy to 87.98%.

- *Experiment 6* — Last addition is the identity network, which is again a pre-trained Facenet model with weights from ILSVRC. This model is retrained using original visible and generated visible images. This yielded a Rank-1 face recognition accuracy of 92.83%. Images are generated with distinct features of individuals after 100 epochs of training, however the generated images are noisy. After 200 epochs, images started losing quality and therefore, the training is stopped at 200 epochs and that is where the model converged.

- *Experiment 7* — Finally, synthetic visible images are de-noised using Wiener filter. After this step, the recognition accuracy increased to 97.63%. The images generated using the final model with all the components and de-noising are shown in Figure 6.15 along with the original MWIR and visible images.

All the experiments and results are summarized in Table 6.9 and the CMC (Cumulative Match Characteristic) curves for different networks are presented in Fig 6.16.

Since the test set for all the above experiments is very small (9 subjects), the final model, GAN

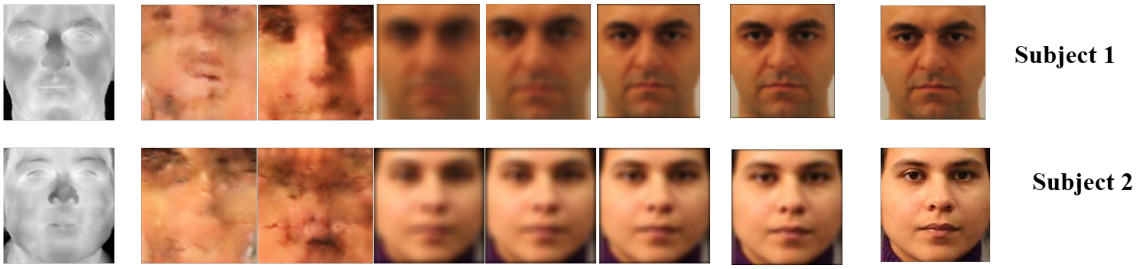


Figure 6.15: Output from GAN — left to right — original MWIR image, images generated after 25, 50, 100, 150 and 200 epochs, de-blurred image and original visible image

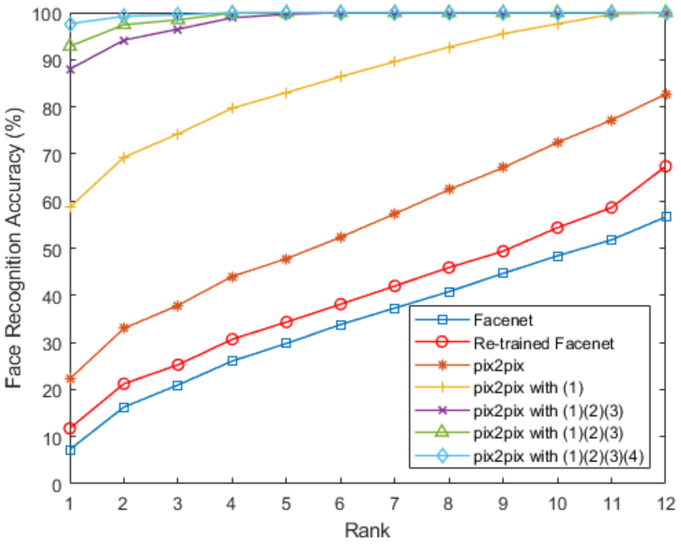


Figure 6.16: Face Recognition Accuracy - (1) is pre-trained feature extractor, (2) is pre-trained discriminator, (3) is identity network

with all the added components is trained using 70% of the data for training using cross-validation and is tested on the remaining 30% of the data and de-noising is applied to the generated images. For this experiment, the Rank-1 accuracy is found out to be 97.58%.

Chapter 7

Conclusions and Future Work

7.1 Conclusions

In this thesis, the problem of availability of data is solved by using transfer learning which uses the pre-trained models. Transfer learning reduces the number of images required to develop a new deep learning model by a large amount. It also reduces the computation cost and training time required for the model to converge. By developing models using transfer learning, knowledge from existing models can be used to train new models instead of training an original model. Different pre-trained state-of-the-art models are combined with other models for feature extraction and are trained using thermal and visible data. Thereby proposing a dual band based deep face detection model that works almost equally well in visible and thermal bands. Faster R-CNN is determined to be the model that works better than any other model trained with our data at a learning rate of 3e-5. The other models trained and tested are R-FCN and SSD. SSD is the fastest among the three models and R-FCN and Faster R-CNN detected at the same speed. The Faster R-CNN and R-FCN models are further tested by reducing the number of proposals of the RPN from 300 to 50 by reducing 50 at a time to arrive at an optimal trade-off between speed and accuracy. The optimal number of proposals that achieved a good speed/accuracy trade-off is 150 where the accuracy is 95.28% for Faster R-CNN and the time for single detection is 12 ms. Beyond that, the accuracy of the detector dropped by a high amount. To follow up, face recognition experiments are conducted using traditional methods for thermal data and using Facenet model for visible images using the cropped faces from the face detectors and using faces that are manually cropped. The results yielded using the automatically cropped images are comparable to that of the manually cropped images.

Addition of data can still be useful to improve the performance of deep learning models. The second part of the thesis focuses on augmenting the data to further improve the Faster R-CNN face detection accuracy, specifically when less number of proposals are used. Traditional data augmentation methods like changing the brightness of the images, contrast enhancement, applying noise and denoising the images are used to increase the size of the dataset significantly. This helped in training a robust model that could detect faces from images that are not perfect. The detection accuracy increased by a small amount by this. Then, a GAN is proposed to generate new face identities that can be used to train the face detector. The generator and discriminator architectures are developed using fractional-strided and strided convolutions respectively. Mode collapse occurs while training a GAN, an issue that arises when the generator finds one or two images that fools the discriminator successfully and keeps generating these despite the randomness of the input noise. This problem is addressed by adding a similarity calculation layer right before the sigmoid layer of the discriminator network which penalizes the generator, if the mode starts to collapse. This addition increased the accuracy of the detector by a great amount, specifically when less number of proposals are used.

There are many state-of-the-art models available to perform face recognition in visible spectrum images, however these models cannot be directly used to develop cross spectral face recognition models due to the modality gap. A feasible solution is to convert images in one spectrum to another. The third contribution is the proposal of a thermal-to-visible image conversion GAN that can be used to convert thermal images to their visible counterparts facilitating the use of the existing state-of-the-art visible-to-visible face recognition models to be used for recognition. The pix2pix GAN is taken as a base for our model which is a conditional GAN that takes an image or any other information as condition for generator and discriminator. Facenet model is used to extract features from original visible images that can be used to condition generator and discriminator networks. A pre-trained Inception ResNet classifier is used as discriminator and is trained to classify images as real or fake. In addition to using a pre-trained discriminator model, Facenet model is used as an identity network that penalizes generator further and is trained using triplet loss. Thermal images are input to the generator and it outputs their visible counterparts. The visible counterparts generated using this model are used as probe set against the original visible images as gallery images. The Rank-1 face recognition accuracy yielded using the facenet face matcher is 97.63%.

7.2 Future Work

In this work, a pre-trained model is used as a discriminator to generate visible images to thermal images, however the GAN used to augment thermal images in 4 does not use a pre-trained model. The training time required for this model to converge is 100 hours as transfer learning is not used here. In future work, this discriminator will be replaced with a pre-trained model, thereby reducing the training time. Also, the boundaries of the GAN in 4 can be pushed further to generate more images. The generated images can be combined with the original images as real images to the discriminator, generating more new face identities. Noisy images etc. can also be included in the dataset to generate visibl eimages from thermal to further improve the performance of GAN.

Bibliography

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

[2] Radford, A., Metz, L., and Chintala, S., "Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks," *CoRR*, Vol. abs/1511.06434, 2015.

[3] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., and Berg, A. C., "SSD: Single Shot MultiBox Detector," *European Conference on Computer Vision*, 2016, pp. 21–37.

[4] Dai, J., Li, Y., He, K., and Sun, J., "R-FCN: Object Detection via Region-Based Fully Convolutional Networks," *Advances in Neural Information Processing Systems*, 2016, pp. 379–387.

[5] Ren, S., He, K., Girshick, R., and Sun, J., "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," *Advances in Neural Information Processing Systems*, 2015, pp. 91–99.

[6] Schroff, F., Kalenichenko, D., and Philbin, J., "Facenet: A Unified Embedding for Face Recognition and Clustering," *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 815–823.

[7] Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A., "Image-to-Image Translation with Conditional Adversarial Networks," *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125–1134.

[8] Ma, C., Trung, N. T., Uchiyama, H., Nagahara, H., Shimada, A., and ichiro Taniguchi, R., "Mixed Features for Face Detection in Thermal Images," *Proceedings of SPIE*, Vol. 10338, 2017, pp. 1–8.

[9] Zheng, Y., "Face Detection and Eyeglasses Detection for Thermal Face Recognition," *Image Processing: Machine Vision Applications V*, Vol. 8300, 02 2012, pp. 9–18.

[10] Murata, T., Matsuno, S., Mito, K., Itakura, N., and Mizuno, T., "Investigation of Facial Region Extraction Algorithm Focusing on Temperature Distribution Characteristics of Facial Thermal Images," *International Conference on Human-Computer Interaction*, 2017, pp. 347–352.

[11] Kopaczka, M., Acar, K., and Merhof, D., "Robust Facial Landmark Detection and Face Tracking in Thermal Infrared Images using Active Appearance Models," *VISIGRAPP*, 2016, pp. 150–158.

[12] Kwaśniewska, A., Rumiński, J., Czuszyński, K., and Szankin, M., "Real-Time Facial Features Detection from Low Resolution Thermal Images with Deep Classification Models," *Journal of Medical Imaging and Health Informatics*, Vol. 8, 2018, pp. 979–987.

[13] Kwaśniewska, A., Rumiński, J., and Rad, P., "Deep Features Class Activation Map for Thermal Face Detection and Tracking," *International Conference on Human System Interactions (HSI)*, IEEE, 2017, pp. 41–47.

[14] Zhang, K., Zhang, Z., Li, Z., and Qiao, Y., "Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks," *IEEE Signal Processing Letters*, Vol. 23, 2016, pp. 1499–1503.

[15] Farfade, S. S., Saberian, M. J., and Li, L.-J., "Multi-View Face Detection Using Deep Convolutional Neural Networks," *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*, 2015, pp. 643–650.

[16] Koestinger, M., Wohlhart, P., Roth, P. M., and Bischof, H., "Annotated Facial Landmarks in the Wild: A Large-Scale, Real-World Database for Facial Landmark Localization," *IEEE international conference on computer vision workshops (ICCV workshops)*, 2011, pp. 2144–2151.

[17] Krizhevsky, A., Sutskever, I., and Hinton, G., "Imagenet Classification with Deep Convolutional Neural Networks," *Neural Information Processing Systems*, 2014, pp. 1–9.

[18] Ranjan, R., Patel, V. M., and Chellappa, R., "HyperFace: A Deep Multi-Task Learning Framework for Face Detection, Landmark Localization, Pose Estimation and Gender Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 41, 2019, pp. 121–135.

[19] Yang, S., Luo, P., Loy, C. C., and Tang, X., "Faceness-Net: Face Detection Through Deep Facial Part Responses," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 40, 2018, pp. 1845–1859.

[20] Wu, Y., Hassner, T., Kim, K., Medioni, G., and Natarajan, P., "Facial Landmark Detection with Tweaked Convolutional Neural Networks," *IEEE transactions on pattern analysis and machine intelligence*, Vol. 40, 2018, pp. 3067–3074.

[21] Sun, X., Wu, P., and Hoi, S. C., "Face Detection using Deep Learning: An Improved Faster RCNN Approach," *Neurocomputing*, Vol. 299, 2018, pp. 42–50.

[22] Cheng, E.-J., Chou, K.-P., Rajora, S., Jin, B.-H., Tanveer, M., Lin, C.-T., Young, K.-Y., Lin, W.-C., and Prasad, M., "Deep Sparse Representation Classifier for Facial Recognition and Detection System," *Pattern Recognition Letters*, 2019.

[23] Sun, Y., Wang, X., and Tang, X., "Deep Convolutional Network Cascade for Facial Point Detection," *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 3476–3483.

[24] Zhang, C. and Zhang, Z., "Improving Multiview Face Detection with Multi-Task Deep Convolutional Neural Networks," *IEEE Winter Conference on Applications of Computer Vision*, 2014, pp. 1036–1041.

[25] Antoniou, A., Storkey, A., and Edwards, H., "Data Augmentation Generative Adversarial Networks," *arXiv preprint arXiv:1711.04340*, 2017.

[26] Ronneberger, O., Fischer, P., and Brox, T., "U-net: Convolutional Networks for Biomedical Image Segmentation," *International Conference on Medical image computing and computer-assisted intervention*, Springer, 2015, pp. 234–241.

[27] He, K., Zhang, X., Ren, S., and Sun, J., "Deep Residual Learning for Image Recognition," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[28] Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q., "Densely Connected Convolutional Networks," *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[29] Zhu, X., Liu, Y., Li, J., Wan, T., and Qin, Z., "Emotion Classification with Data Augmentation using Generative Adversarial Networks," *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 2018, pp. 349–360.

[30] Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A., "Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks," *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.

[31] Zhao, J., Mathieu, M., and LeCun, Y., "EnergyBbased Generative Adversarial Network," *arXiv preprint arXiv:1609.03126*, 2016.

[32] Berthelot, D., Schumm, T., and Metz, L., "BEGAN: Boundary Equilibrium Generative Adversarial Networks," *arXiv preprint arXiv:1703.10717*, 2017.

[33] Frogner, C., Zhang, C., Mobahi, H., Araya, M., and Poggio, T. A., "Learning with a Wasserstein Loss," *Advances in Neural Information Processing Systems*, 2015, pp. 2053–2061.

[34] Shin, H.-C., Tenenholtz, N. A., Rogers, J. K., Schwarz, C. G., Senjem, M. L., Gunter, J. L., Andriole, K. P., and Michalski, M., "Medical Image Synthesis for Data Augmentation and Anonymization using Generative Adversarial Networks," *International Workshop on Simulation and Synthesis in Medical Imaging*, Springer, 2018, pp. 1–11.

[35] Tanaka, F. H. K. d. S. and Aranha, C., "Data Augmentation Using GANs," *arXiv preprint arXiv:1904.09135*, 2019.

[36] Deng, J., Guo, J., Xue, N., and Zafeiriou, S., "Arcface: Additive Angular Margin Loss for Deep Face Recognition," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4690–4699.

[37] Huang, C., Li, Y., Chen, C. L., and Tang, X., "Deep Imbalanced Learning for Face Recognition and Attribute Prediction," *IEEE transactions on pattern analysis and machine intelligence*, 2019.

[38] Wang, H., Wang, Y., Zhou, Z., Ji, X., Gong, D., Zhou, J., Li, Z., and Liu, W., "Cosface: Large Margin Cosine Loss for Deep Face Recognition," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5265–5274.

[39] Masi, I., Chang, F.-J., Choi, J., Harel, S., Kim, J., Kim, K., Leksut, J., Rawls, S., Wu, Y., Hassner, T., et al., "Learning Pose-Aware Models for Pose-Invariant Face Recognition in the

Wild," *IEEE transactions on pattern analysis and machine intelligence*, Vol. 41, No. 2, 2018, pp. 379–393.

[40] Lezama, J., Qiu, Q., and Sapiro, G., "Not Afraid of the Dark: Nir-Vis Face Recognition via Cross-Spectral Hallucination and Low-Rank Embedding ," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6628–6637.

[41] Li, S., Yi, D., Lei, Z., and Liao, S., "The Casia Nir-Vis 2.0 Face Database," *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2013, pp. 348–353.

[42] Narang, N. and Bourlai, T., "Face Recognition in the SWIR Band When using Single Sensor Multi-Wavelength Imaging Systems," *Image and Vision Computing*, Vol. 33, 2015, pp. 26–43.

[43] Osia, N. and Bourlai, T., "A Spectral Independent Approach for Physiological and Geometric based Face Recognition in the Visible, Middle-Wave and Long-Wave Infrared Bands," *Image and Vision Computing*, Vol. 32, No. 11, 2014, pp. 847–859.

[44] Mirza, M. and Osindero, S., "Conditional Generative Adversarial Nets," *arXiv preprint arXiv:1411.1784*, 2014.

[45] Zhang, T., Wiliem, A., Yang, S., and Lovell, B., "TV-GAN: Generative Adversarial Network based Thermal to Visible Face Recognition," *2018 international conference on biometrics (ICB)*, IEEE, 2018, pp. 174–181.

[46] Di, X., Zhang, H., and Patel, V. M., "Polarimetric Thermal to Visible Face Verification via Attribute Preserved Synthesis," *IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, IEEE, 2018, pp. 1–10.

[47] Chen, C. and Ross, A., "Matching Thermal to Visible Face Images Using a Semantic-Guided Generative Adversarial Network," *arXiv preprint arXiv:1903.00963*, 2019.

[48] Wang, Z., Chen, Z., and Wu, F., "Thermal to Visible Facial Image Translation using Generative Adversarial Networks," *IEEE Signal Processing Letters*, Vol. 25, No. 8, 2018, pp. 1161–1165.

[49] Deng, Z., Li, K., Zhao, Q., Zhang, Y., and Chen, H., "Effective Face Landmark Localization via Single Deep Network," *arXiv preprint arXiv:1702.02719*, 2017.

[50] Simonyan, K. and Zisserman, A., "Very Deep Convolutional Networks for Large-Scale Image Recognition," *International Conference on Learning Representations*, 2014.

[51] Ioffe, S. and Szegedy, C., “Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift,” *International Conference on Machine Learning*, 2015.

[52] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z., “Rethinking the Inception Architecture for Computer Vision,” *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2016, pp. 2818–2826.

[53] Szegedy, C., Ioffe, S., Vanhoucke, V., and Alemi, A. A., “Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning.” *AAAI*, Vol. 4, 2017, pp. 1–12.

[54] Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., and Adam, H., “Mobilenets: Efficient Convolutional Neural Networks for Mobile Vision Applications,” *CoRR*, Vol. abs/1704.04861, 2017.

[55] Huang, J., Rathod, V., Sun, C., Zhu, M., Balan, A. K., Fathi, A., Fischer, I., Wojna, Z., Song, Y., Guadarrama, S., and Murphy, K., “Speed/Accuracy Trade-Offs for Modern Convolutional Object Detectors,” *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3296–3297.

[56] Girshick, R., “Fast R-CNN,” *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1440–1448.

[57] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A., “Going Deeper with Convolutions,” *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1–9.

[58] Beck, A. and Teboulle, M., “Mirror Descent and Nonlinear Projected Subgradient Methods for Convex Optimization,” *Operations Research Letters*, Vol. 31, No. 3, 2003, pp. 167–175.

[59] Kingma, D. P. and Ba, J., “Adam: A Method for Stochastic Optimization,” *CoRR*, Vol. abs/1412.6980, 2014.

[60] Verma, R. and Ali, J., “Comparative Study of Various Types of Image Noise and Efficient Noise Removal Techniques,” *International Journal of advanced research in computer science and software engineering*, Vol. 3, No. 10, 2013.

[61] Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., and Chen, X., “Improved Techniques for Training GANs,” *Advances in neural information processing systems*, 2016, pp. 2234–2242.

[62] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al., "Imagenet: Large Scale Visual Recognition Challenge," *International journal of computer vision*, Vol. 115, No. 3, 2015, pp. 211–252.