2023

# Generative Adversarial Network and Its Application in Aerial Vehicle Detection and Biometric Identification System

Moktari Mostofa
*West Virginia University*, mm0251@mix.wvu.edu

# Generative Adversarial Network and Its Application in Aerial Vehicle Detection and Biometric Identification System

Moktari Mostofa

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

in partial fulfillment of the requirements
for the degree of

Doctor of Philosophy
in
Electrical Engineering

Nasser M. Nasrabadi, Ph.D., Chair
Matthew C. Valenti, Ph.D.
Jeremy Dawson, Ph.D.
Omid Dehzangi, Ph.D.
Benjamin S. Riggan, Ph.D.

Lane Department of Computer Science and Electrical Engineering

Morgantown, West Virginia
2023

Keywords: Generative Adversarial Network, Aerial Vehicle Detection, Super-Resolution, Cross Spectral Iris Matching, Pose Attention, Profile to Frontal Face Recognition.

## Abstract

Generative Adversarial Network and Its Application in Aerial Vehicle Detection and Biometric Identification System

Moktari Mostofa

In recent years, generative adversarial networks (GANs) have shown great potential in advancing the state-of-the-art in many areas of computer vision, most notably in image synthesis and manipulation tasks. GAN is a generative model which simultaneously trains a generator and a discriminator in an adversarial manner to produce real-looking synthetic data by capturing the underlying data distribution. Due to its powerful ability to generate high-quality and visually pleasing results, we apply it to super-resolution and image-to-image translation techniques to address vehicle detection in low-resolution aerial images and cross-spectral cross-resolution iris recognition. First, we develop a Multi-scale GAN (MsGAN) with multiple intermediate outputs, which progressively learns the details and features of the high-resolution aerial images at different scales. Then the upscaled super-resolved aerial images are fed to a You Only Look Once-version 3 (YOLO-v3) object detector and the detection loss is jointly optimized along with a super-resolution loss to emphasize target vehicles sensitive to the super-resolution process. There is another problem that remains unsolved when detection takes place at night or in a dark environment, which requires an IR detector. Training such a detector needs a lot of infrared (IR) images. To address these challenges, we develop a GAN-based joint cross-modal super-resolution framework where low-resolution (LR) IR images are translated and super-resolved to high-resolution (HR) visible (VIS) images before applying detection. This approach significantly improves the accuracy of aerial vehicle detection by leveraging the benefits of super-resolution techniques in a cross-modal domain. Second, to increase the performance and reliability of deep learning-based biometric identification systems, we focus on developing conditional GAN (cGAN) based cross-spectral cross-resolution iris recognition and offer two different frameworks. The first approach trains a cGAN to jointly translate and super-resolve LR near-infrared (NIR) iris images to HR VIS iris images to perform cross-spectral cross-resolution iris matching to the same resolution and within the same spectrum. In the second approach, we design a coupled GAN (cpGAN) architecture to project both VIS and NIR iris images into a low-dimensional embedding domain. The goal of this architecture is to ensure maximum pairwise similarity between the feature vectors from the two iris modalities of the same subject. We have also proposed a pose attention-guided coupled profile-to-frontal face recognition network to learn discriminative and pose-invariant features in an embedding subspace. To show that the feature vectors learned by this deep subspace can be used for other tasks beyond recognition, we implement a GAN architecture which is able to reconstruct a frontal face from its corresponding profile face. This capability can be used in various face analysis tasks, such as emotion detection and expression tracking, where having a frontal face image can improve accuracy and reliability. Overall, our research works have shown its efficacy by achieving new state-of-the-art results through extensive experiments on publicly available datasets reported in the literature.

*To my dear parents, adorable siblings, and lovely husband who have been my continuous support, and strength throughout this Ph.D. journey.*

# Acknowledgement

Undertaking a Ph.D. has been a significant decision in my life. At the beginning, it seemed to be very challenging and I could not be able to accomplish it unless there are unwavering support and help from so many people surrounding me. I would like to express my sincere appreciation and thank those beautiful souls.

First and foremost, I would like to express my heartfelt gratitude to my supervisor, Dr. Nasser M. Nasrabadi, for his keen guidance and continuous support throughout my Ph.D. journey. I really feel lucky to conduct my research under such a brilliant scholar like Dr. Nasrabadi. His expertise and insightful feedback on my research have been invaluable and helped me to learn and grow both academically and personally. I am also grateful to the members of my thesis committee, Dr. Matthew Valenti, Dr. Jeremy Dawson, Dr. Omid Dehzangi, and Dr. Benjamin S. Riggan for their constructive feedback and productive ideas to improve the quality of my work.

I would like to thank the funding agency Center for Identification Technology Research (CITeR) for supporting me financially to conduct experiments, attend conferences, and present my research works to the wider academic community.

Next, I would like to thank my colleagues, Sobhan, Hadi, Fariborz, Amol, Poorya, Salman, Saeed, Sahar, Niloufar, Mahedi, Sami, Ali Zafri, Hussain, Poorya, Mohammad, and Atefeh. My warm gratitude goes to Ali Dabouei and Uche for their encouragement, diverse perspectives, expertise, and enthusiasm which enriched my research.

I am deeply grateful to my late father Md. Golam Mostofa, mother Monoara Begum, my siblings Popy, Babu and Shimul for instilling in me a love of learning which brought me here today. I would like to thank my sweet li'l sister Meghla and best friend Shondha for being there when I need them the most. I cannot thank enough my partner Dr. Mahmood Khan for his unwavering support when times were really challenging and felt never-ending. I would also like to thank my in-laws, Maqsuda Khan and Lutfar Khan. Their encouragement and love have been a constant source of inspiration and motivation, which helped me finish such an arduous chapter of my life.

# Table of Contents

# List of Tables

# List of Figures

# Chapter 1

# Introduction

## 1.1 Problem and Motivation

In recent years, deep learning (DL) has achieved a great deal of attention in a variety of application domains of computer-vision. With its inception DL has revolutionary advanced different computer-vision tasks such as object detection (e.g., [3, 4]), motion tracking (e.g., [5, 6]), action recognition (e.g., [7, 8]), and human pose estimation (e.g., [9, 10]). It has also reported state-of-the-art results ever since it adopted several Deep Neural Network (DNN) architectures. There are two major resources that have contributed to the significant progress of DNN. First, advancements in hardware technology have enabled the development of more powerful and efficient computing systems that can process and analyze large amounts of data. More specifically, graphics processing units (GPUs) with a large amount of memory and processing cores have allowed for the efficient optimization of numerous training parameters via parallel processing. In addition to GPUs, recently google's tensor processing units (TPUs) have seen being adapted to train many DNN architectures. These specialized chips are designed to perform matrix multiplication, a key operation in deep learning, much more efficiently than general-purpose CPUs or GPUs. Second, the availability of large datasets has been critical for training DNNs. The abundance of data has enabled the creation of large training datasets, which can be used to enhance the generalization of the learning process. Overall, the progress in DNNs has been driven by a combination of these resources and the development of new techniques and approaches that leverage them. Despite significant contributions of DNN in various directions of computer vision, there still remain challenges in domain

adaptation and image synthesis techniques.

GAN [11] introduced by Ian GoodFellow can be an effective solution, which has dramatically changed the direction of deep learning research. According to Yann LeCun, *"GAN is the best concept proposed in the last ten years in the field of deep learning (neural networks)"*. GAN trains two networks simultaneously called generator and discriminator in an adversarial fashion and forces both networks to achieve different objectives. The generator tries to fool the discriminator by generating real-looking fake images from random latent vectors, while the discriminator learns to evaluate the authenticity of the generated data by comparing it to the real training data. Eventually, the generator learns which features make images real and starts generating more realistic images that confuse the discriminator. Driven by this concept many computer vision applications have been growing and experiencing tremendous success in different domains. In the following sections, we provide our contribution in developing GAN-based cross-domain super-resolution techniques to assist aerial vehicle detection and the sequence of this dissertation. We also briefly outline our proposed conditional GAN-based iris-verifier and face frontalization framework to facilitate biometric identification systems.

### 1.1.1 GAN-based Super-Resolution for Aerial Vehicle Detection

GAN has been extensively used in the field of super-resolution [12] to enhance the quality of low-resolution images. It has created scope to boost the performance of deep learning-based small vehicle detection on aerial imagery. In GANs, a generative model learns to recover high-resolution super-resolved images with high perceptual quality from the low-resolution counterpart via a competitive process between a generator and a discriminator. Conditional GAN (cGAN) [13], an extension of generative model has been enormously applied to image translation, which again fueled aerial vehicle detection at IR imaging environment. Here, the generator learns a mapping in a conditional setting to generate IR images from the corresponding visible samples or vice versa so that we can handle lack of training data for the detector. Given the contribution of GANs to various demanding areas of computer vision, it is necessary to investigate these networks accurately.

In chapter 2, first we have analysed the performance of GAN architecture in preserving high-frequency details while enhancing the resolution of the input aerial image where I contributed

dedicatedly as a co-author. Afterward, we investigate the detection performance of one of the state-of-the-art detectors on these super-resolved images at multiple resolutions.

Based on the results reported in chapter 2, we focus on improving the architecture of GAN to ensure the potential contribution of super-resolution (SR) to vehicle detection in low-resolution aerial and satellite images. Hence, in chapter 3, we propose and develop an end-to-end trainable joint super-resolution and vehicle detection network, which offers multi-tasking paradigm by leveraging complementary information from both tasks.

In chapter 4, we have attempted to solve a major shortcoming of RGB-based vehicle detector. Inspired by the idea of cGAN based domain transformation technique, we design a deep CNN which jointly learns domain specific features along with the super-resolution process to facilitate the detection task in cross-domain.

### 1.1.2 Application of GAN in Biometric Identification System

Recent years have seen significant improvements in the accuracy of different biometric recognition systems based on deep learning models including face recognition [14–21], fingerprint [22–27], iris [28–35], and ear [36–38] recognition. Especially, face and iris have attracted the interest of the biometric research community compared to other physiological traits due to their wide range of applications in security and law enforcement, border control, and immigration, government offices to identify and accurately authenticate one's identity.

More specifically, the texture and patterns in the iris of the eye are highly unique and do not change over time, making it a reliable biometric identifier. Most commercial iris recognition systems are designed to perform iris matching in the same domain. Therefore they suffer from cross-domain challenges when the probe and gallery are in the same spectrum within the same resolution. Hence, in chapter 5, we formulate two different novel approaches to alleviate this cross-domain problem. One of them is designed to be a preprocessing module for cross-spectral iris matching based on GAN, which integrates cross-resolution matching scenario into the cross-spectrum setting and offer a join super-resolution and cross-spectral iris matching framework. In our second approach, we analyze iris recognition performance in latent embedded domain. Therefore, we devise a novel coupled GAN framework to effectively learn cross-domain iris features in the latent

embedded feature subspace.

In addition, we have also focused on the profile-to-frontal face recognition task and generated frontal faces from corresponding profile faces by developing a face frontal GAN. Face recognition (FR) has become a challenging task due to the joint variations of pitch and yaw view angles of the facial images captured in unconstrained environments. Moreover, these pose variations cause appearance distortion, and semantic consistency missing in profile faces, which indicates a loss of useful information for FR. Existing methods either perform 'pose normalization (i.e., frontalization [39–42])' or disentangle poses for pose invariant face recognition [41, 43–46]. In chapter 6, first, we design a novel coupled profile-to-frontal FR network utilizing pose as auxiliary information via an attention mechanism and learn a compact embedding subspace. After that, we develop a *face frontal GAN* to reconstruct the frontal face from the deep features learned in the embedding domain for its corresponding profile face.

# Chapter 2

# Super Resolution-Assisted Deep Aerial Vehicle Detection

## 2.1 Introduction

Real-time vehicle detection in aerial imagery is extremely difficult. Predicting small vehicles from these large images are even difficult for human eyes. The reason lies in the nature of data along with computational constraints, low resolution and complex background of these imagery. Recently research community has paid substantial attention in this area as the outcome of this research can help better plan in transportation system, surveillance and reconnaissance. As these images are taken from top view with varying altitude, the target objects can not contain much information. Also, target objects in these images are very small relative to the entire image that makes it hard to design a detector which distinguishes target from background. Using a deep learning technique; super-resolution, we can accelerate the detection performance.

An extensive study has been conducted for object detection in aerial imagery; however it is still an open problem demanding a high performance solution. Recent approaches have shown promising performances and most of them use convolutional neural network (CNN) based solution to detect objects of interest from aerial imagery. In this work, we propose a new architecture utilizing deep neural network (DNN) that helps to locate vehicles in satellite images ensuring effective performance. Our designed architecture can be divided into two stages. The first stage takes an input image and creates super resolved version of the original image using Generative Adversarial

Network (GAN) [11]. This network helps to augment image features improving image quality that leads to a better object recognition. Also, it ensures to preserve high level features while transferring from low to high resolution domain.The second stage handles target prediction from these super-resolved images using a trained SSD-based [47] vehicle detector. We further investigate the detection performance on multi-resolution images with their super resolved counterpart.

Our proposed method demonstrates promising performance while tracking vehicles in aerial imagery. This paper is organized as follows. Section 2 outlines the review of the previous works related to ours. In Section 3, we discuss our proposed framework. Experimental details and results are demonstrated in Section 4. Finally, in Section 5, we reach the conclusion.

## 2.2 Related Work

### 2.2.1 Single Image Super Resolution

Here, we will discuss the algorithms related to single image super-resolution (SISR). The algorithms can be categorized into four groups–prediction based models, edge based methods, image statistical models and example-based methods. Among them example-based methods [48–59] are known as more powerful methods which aim at mapping between low and high resolution images. They rely on external datasets. Early method was proposed by Freeman et al. [51, 60]. In the work of Glasner et al, they use patch redundancy across the scales within the image to model the image super-resolution problem. In Huang et al. [61], self-similarity based super-resolution (SR) algorithms are expanded by incorporating small transformations and geometric variations. A convolutional sparse coding approach was proposed by Gu et al. [62]. In their work, they ignore consistency of the pixels in overlapped patches by working on whole image that helps generate more robust reconstruction of image local structures.

Tai et al. [63] combine the benefits of edge-directed SR with learning-based SR to reconstruct more realistic texture details in super-resolved images. To recover high quality SR, Zhang et al. [64] introduce a multi-scale dictionary method that simultaneously incorporates local and non-local priors. The local prior removes artifacts from target image and the non-local prior helps produce more perceptually satisfying image.

The use of sparse learned dictionaries in combination with neighbor embedding methods obtain improved quality and speed in the work of Timofte et al. [53, 65]. The authors propose anchored neighborhood regression. Kernel ridge regression (KRR) is adopted to learn a map from input LR images to target SR images in Kim and Kwon [66] works. This idea is based on example pairs of input and output images. In Dai et al. [50] a joint learning of patch-specific regressors is proposed during training. At testing phase, it selects the best regressor which yields the smallest super-resolving error.

Convolutional neural network (CNN) based SR algorithms have attained superior performance. Wang et al. [67] propose a sparse representation in combination with feed-forward network architecture . The underlying idea is based on the learned iterative shrinkage and thresholding algorithm (LISTA) [68, 69]. Dong et al. [70, 71] trained a three layer deep convolutional neural network that achieved state of the art SR performance.

To recover perceptually more convincing HR images, Johnson et al. [72] and Bruna et al. [73] use a loss function closer to perceptual similarity. For our paper, we particularly follow the works of Christian Ledig et al. [12], the authors use a perceptual loss function with MSE loss to generate more realistic SR images.

### 2.2.2 Vehicle Detection in Aerial Imagery

Vehicle detection in aerial imagery has been studied a lot in the object detection literature. Apart from the convolutional neural network, prior works employed other approaches to address this problem. A model proposed in [74] utilized Bayesian Network with handcrafted parameters to identify vehicles in aerial imagery. To find cars in satellite images, the method presented in [75] applied Mean-shift algorithm utilizing shape information of the targets. The framework discussed in [76, 77, 77] trains a Dynamic Bayesian Network (DBN) with features preserving region level information.

With the progressive success of deep learning in object detection, many recent works use CNN for vehicle detection in aerial imagery. A fast detector proposed by Carlet and Abayowa [78] modifies YOLOv2 for locating vehicles in aerial imagery. Terrail et al. [79] apply a modified faster R-CNN algorithm that achieves a breakthrough performance on aerial vehicle detection. In [80],

Soleimani et al. propose a text-guided detection scheme that utilizes both the visual and textual features for detection. Yang et al. [81]present a framework that uses skip connection to merge lower and higher level features utilizing focal loss function. Li et al. [82–84] propose a framework for multi-oriented vehicle detection. In this framework, a rotatable region proposal network is utilized that learns the orientation of vehicles while performing classification on aerial images and videos.

Though vehicle detection in aerial imagery has been a research focus in recent years, comprehensive study of detection performance on super resolved images has not yet been investigated in most of them. The work presented in [85] gives an overview of detection performance on super resolved images considering multiple resolutions. In this paper, a performance gain is reported for different resolutions by applying super resolution technique on the original images. Liujuan Cao et al. [86] proposed a framework that employs super resolution with coupled dictionary learning on the satellite imagery and then a detection algorithm is applied on the generated images.

## 2.3 Proposed Model

In this section, we introduce our method in a detailed way. First, we super resolve an image from its low resolution input image and then feed it to a detector. In this work, we have proposed a two-stage framework that employs SRGAN [12] followed by an SSD (Single Shot Detector) [47] for vehicle detection in aerial imagery.

### 2.3.1 Super-Resolution using GAN

Super-resolution aims at recovering a high resolution super resolved image from a low resolution input image. In [51], the authors propose that there can be two ways to super resolve a low resolution input image: single image based SR and multiple images based SR. In multiple-image super-resolution algorithms [87], a couple of low-resolution images of the same scene is used as input and then a registration algorithm is employed to find the transformation between them. These algorithms can recover higher resolution details, however, their performances are limited by improvement factors close to two [88]. Single-image super-resolution algorithms, like [70], usually have a single input. They train a set of low-resolution images along with their high-resolution

counterparts to learn a relationship between them. The underlying idea is to use this learned relationship to predict the missing high-resolution details of the input low-resolution images. This idea helps generate high-resolution images far better than their low-resolution inputs [88].

Recently, Super-resolution using the concept of GAN [11] has achieved the state-of-the art results. For our work, we use the super-resolution method of [12], which is based on the powerful framework GAN. The algorithm encourages to generate photo-realistic images with high perceptual quality. Using this concept, our goal is to generate high resolution, super resolved image $X^{SR}$ from its low resolution $X^{LR}$ input; however $X^{LR}$ is the low resolution counterpart of its original high resolution $X^{HR}$. The network use high resolution images during training. Here, we train a feed-forward CNN which generates a function G, parametrized by $\theta_G$. $\theta_G$ consists of weights and biases of a deep neural network that is optimized using generative adversarial loss. $\theta_G$ is obtained by the following equation:

$$\theta_G = \arg\min_{\theta_G} \frac{1}{N} \sum_{n=1}^{N} l^{SR}(G_{\theta_G}(X_n^{LR}), X_n^{HR}), \tag{2.1}$$

where $X_n^{LR}$ with corresponding $X_n^{HR}$ are the training images for n = 1.....N. $l^{SR}$ is the perceptual loss [12] designed as a weighted combination of several loss functions which is to be minimized to recover the desired characteristics from the reconstructed super-resolved image.

**Generative Adversarial Network Architecture(GAN)**

We train and optimize both discriminator network $D_{\theta_D}$, along with $G_{\theta_G}$ to solve the adversarial min-max problem [11]:

$$\min_{\theta_G} \max_{\theta_D} \mathbb{E}_{X^{HR} \sim p_{train}(X^{HR})} \left[ \log D_{\theta_D}\left(X^{HR}\right) \right] + \mathbb{E}_{X^{LR} \sim p_G(X^{LR})} \left[ \log \left(1 - D_{\theta_D}\left(G_{\theta_G}\left(X^{LR}\right)\right)\right) \right]. \tag{2.2}$$

The main idea is to train a generative model G so that it can learn to create solutions that are highly similar to real images. Along with this, the discriminator is trained to distinguish the super-resolved image from the real one. So both network is optimized in an alternating manner to find the super-resolved image that looks like the original high resolution image.

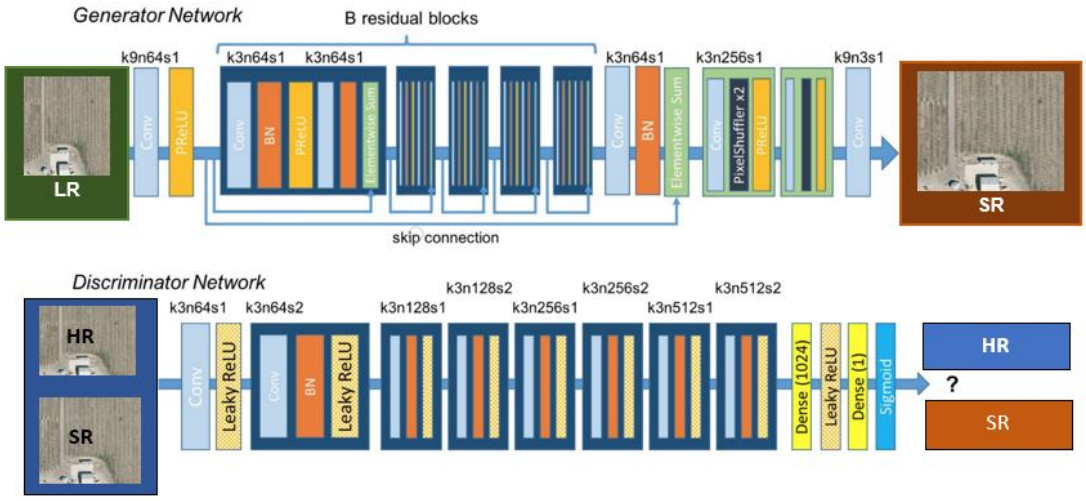



Figure 2.1: Architecture of Generator and Discriminator Network with corresponding kernel size (k), number of feature maps(n) and stride (s) indicated for each convolutional layer.

Following Christian Ledig et al. [12], we use B residual blocks with identical layout in our generator network shown in Figure 2.1. Each block contains two convolutional layers with 3x3 kernels and 64 feature maps followed by batch-normalization layers [89] and Parametric ReLU [90] as the activation function. Here, the idea of sub-pixel fractional convolution layers is used to increase the resolution of the input image as proposed by Shi et al. [91].

To train discriminator, we use eight convolutional layers with an increasing number of 3×3 filter kernels increased by a factor of 2 from 64 to 512 kernels as in the VGG network. Here, we have followed the architecture summmarized by Radford et al. [92]. We also add two dense layers and a sigmoid activation function at the end of the resulting 512 feature maps to obtain a probability for classification.

### Loss functions

In [12], the authors design a loss function that is assessed perceptually. The loss is calculated as the weighted sum of a content loss ($l^{SR}_{content}$) and an adversarial loss ($l^{SR}_{Gen}$) component as:

$$l^{SR} = l^{SR}_{cont} + 10^{-3} l^{SR}_{Gen}. \tag{2.3}$$

They design the content loss using a pre-trained 19 layer VGG network [93] and define VGG

loss as the euclidean distance between the feature representations of a reconstructed image and its corresponding original reference high resolution image that is closer to the perceptual similarity.

$$l^{SR}_{VGG/i.j} = \frac{1}{W_{i,j}H_{i,j}} \sum_{x=1}^{W_{i,j}} \sum_{y=1}^{H_{i,j}} \left( \phi_{i,j}\left(X^{HR}\right)_{x,y} - \phi_{i,j}\left(G_{\theta_G}\left(X^{HR}\right)\right)_{x,y} \right)^2, \tag{2.4}$$

here, $\phi_{i,j}$ indicates feature map at j-th convolutional layer followed by i-th maxpooling layer. $W_{i,j}$ and $H_{i,j}$ represent the dimensions of the respective feature maps within the VGG architecture.

In addition, they also add the GAN loss which encourages the network to obtain natural looking images. It is based on the probability of discriminator, $D_{\theta_D}\left(G_{\theta_G}\left(X^{LR}\right)\right.$ that the reconstructed super-resolved image $G_{\theta_G}\left(X^{LR}\right)$ is an original high resolution image. Instead of $\log\left(1 - D_{\theta_D}\left(G_{\theta_G}\left(X^{LR}\right)\right.\right.$ in Equation2,we minimize the following equation:

$$l^{SR}_{Gen} = \sum_{n=1}^{N} -\log D_{\theta_D}\left(G_{\theta_G}\left(X^{LR}\right)\right). \tag{2.5}$$

**Object Detection using SSD**

Object detection is a computer vision problem which aims to determine the presence of objects in an image or video mimicking human-brain. It also identifies the location as well as the type of object. An extensive research has been conducted through decades to solve this problem. Before the emergence of deep neural network, object detection was performed using feature extractors like SIFT [94], HOG [95] and classifiers such as SVM [96], AdaBoost [97], DPM [98]. A notable increase in performance was observed using deep learning based methods because of the robust learning capability of these nets. The state of the art deep learning based methods can be broadly classified into two categories. The first one is a region-proposal based and the second one is an end-to-end detection based model. RCNN [99], Fast-RCNN [100], Faster-RCNN [101] are the pioneer models with high performance of the first group while YOLO [102], SSD [47] can be considered as significant ones for the second one. The problem with region-proposal based approaches is that the detection is performed by combining multiple stages making the system incompatible for real-time applications. End-to-end detection based methods alleviate this phenomena by eliminating region proposal stage with a unified architecture for detection and classification.

In our approach, we use an SSD (Single Shot Detector) which locates vehicles in an image

in real time. Like other object detection algorithms, SSD extracts features from an image for generating the bounding boxes with class labels. In contrast to classical region based detectors, it does not need region proposals to perform detection making the process extremely fast.

SSD needs input image and the corresponding annotations of objects for detection. The architecture uses the features extracted by VGG-16 [103] with six additional layers for object detection. It predicts objects from multiple feature maps of different resolutions organized in a decreasing fashion that ensures detection of variable size objects. For each position in feature map, SSD generates default boxes of varying sizes using multiple scales and aspect ratios with computed confidence score that designates the presence of objects in those boxes. Later, these boxes are matched with the ground truth annotated boxes. Boxes that have high overlap with the ground truth boxes provided by the annotations are considered as good matches. A threshold factor named Intersection Over Union (IoU) is applied on the generated boxes. Afterwards, a non-maximum suppression algorithm removes the duplicate bounding boxes for the same object.

By default, SSD considers generation of default boxes starting from the convolutional layer 4_3. In our method, we have also considered default boxes from convolutional layer 3_3 as the size of our boxes are really tiny. While computing the loss, SSD uses both localization and confidence loss. The loss is calculated using Equation (2):

$$L_{Total} = \frac{1}{N}(L_{conf}(x, c) + L_{loc}(x, l, g)), \tag{2.6}$$

here, Localization loss ($L_{loc}$) measures the distance between ground-truth and predicted bounding boxes while confidence loss ($L_{conf}$) indicates the presence of an object in the generated bounding box.

Our proposed architecture is demonstrated in Figure 2.2. The generator network super resolves the original low resolution image. Discriminator is designed in such a way that it can distinguish between the real and fake images produced by the generator. These two networks benefit from each other generating more realistic images similar to the original image. Our goal is to train the detector with super resolved images and investigate the performance. Testing of our model is performed using the SSD object detector.

Performance of the detector depends on the design of the default boxes from different feature

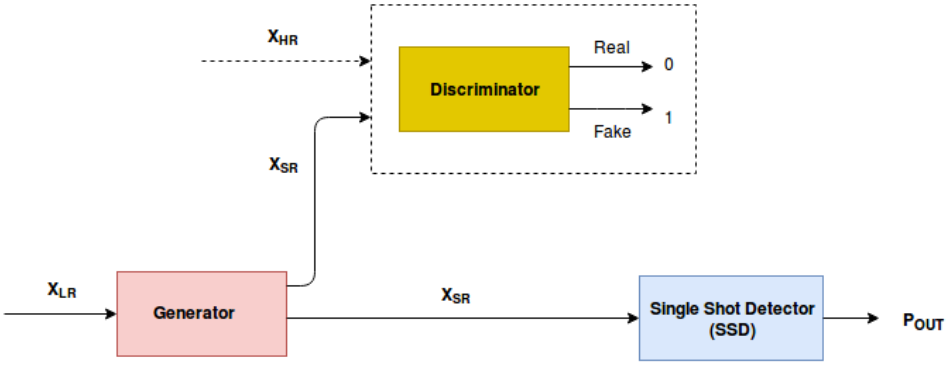



Figure 2.2: Architecture of super resolution with aerial imagery detection

maps, training data and also on the parameters such as IoU, confidence threshold associated with the detection algorithm.

## 2.4 Experiments

### 2.4.1 Dataset and Implementation Details

We perform our experiments on VEDAI (Vehicle Detection in Aerial Imagery) [104] aerial dataset. VEDAI images are taken from Utah database. These images contain multi-oriented vehicles with complex background which makes it to be considered as ideal dataset for any aerial image analysis task. The VEDAI dataset contains around 1200 images. For training and testing, we split the dataset into 1100 and 271 images, respectively. The number of samples in our dataset is small enough to analyze the proposed network. To make the model more robust to different features,we have used different augmentation techniques such as sharpening and flipping.

In our experiments, we use both the scale factor of 2x and 4x between low- and high-resolution images. For training, network requires a low-resolution (LR) image as input. To obtain LR images, the HR images are downsampled using bicubic kernel with downsampling factor of 2 and 4. During implementation, we use the input image of size 128x128 and 256x256 to super-resolve to 256x256 and 512x512, respectively for the scale factor of 2. We further use the input resolution 128x128

Table 2.1: Effects of Super-Resolution for upscale factor 2 on SSD performance

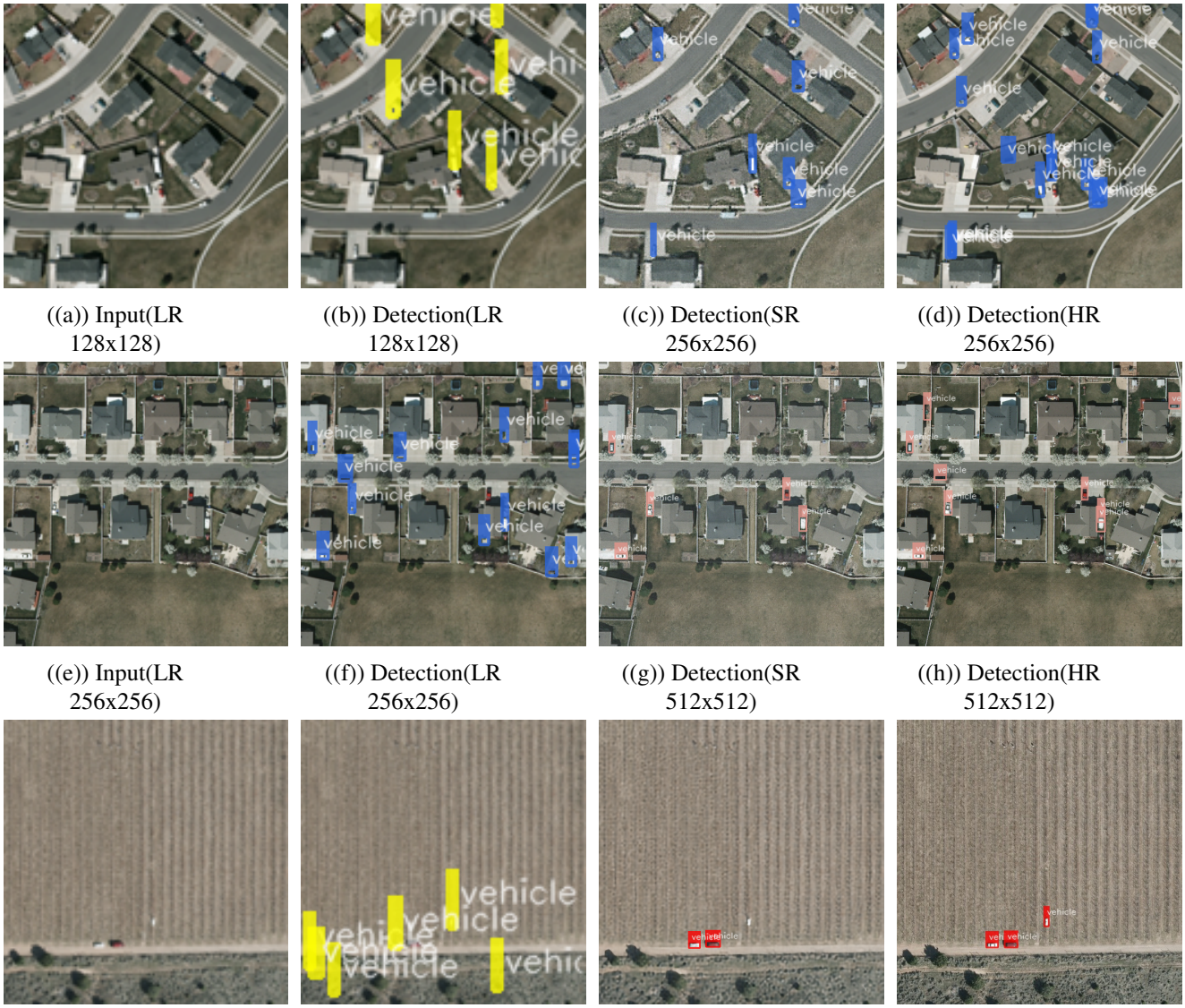
| Input Resolution | mAP(Input) | SR(2x) | mAP(SR 2x) | Original HR | mAP(HR) |
|---|---|---|---|---|---|
| 128x128 | 40.25 | 256x256 | 44.34 | 256x256 | 55.47 |
| 256x256 | 55.47 | 512x512 | 74.56 | 512x512 | 77.81 |

to super-resolve to 512x512 for the upscaling factor of 4. To perceive our detection performance in more details, we have shown the detection result of the low, high and super resolved images in Figure 2.3. More positive objects are detected in Figure 2.3(d), Figure 2.3(h) and Figure 2.3(l) than Figure 2.3(c), Figure 2.3(g) and 2.3(k) indicating high resolution detection is more accurate than the corresponding super resolved detection. Figure 2.3(b), Figure 2.3(f) and 2.3(j) demonstrate poor detection performance with many false detections in low resolution.

As mentioned in section 2.3.1, the generator of our network uses 16 residual blocks. Each of the blocks consists of two convolutional layer with the kernel size of 3x3, stride of 1 and 64 feature maps followed by a batch-normalization and a parametric-relu activation function. There is no max-pool layer throughout the network. To increase the resolution of the input image by 2 and 4 factor, the network adds one and two subpixel convolutional layer, respectively.

We set the network hyperparameters as follows: For super-resolution, we adopt Adam optimizer with a momentum of 0.9 and a learning rate of $10^{-4}$. The model is trained for 2000 iterations with a batch size of 4. For the detection model, we train the SSD architecture with initial learning rate of 0.001 with batch size 16. We optimize the network by Stochastic Gradient Descent (SGD) with a momentum of 0.02 and a weight decay of 0.9. We fine-tune our detection model by setting IoU 0.48, confidence threshold 0.05 and nms with jaccard coefficient of 0.45 to achieve the best performance. The entire network is implemented in tensorflow framework on two NVIDIA TITAN XP GPUs. We follow object detection and GAN based solution on Github* for our network implementation.

*https://github.com/tensorlayer/srgan, https://github.com/ljanyst/ssd-tensorflow



((a)) Input(LR 128x128) ((b)) Detection(LR 128x128) ((c)) Detection(SR 256x256) ((d)) Detection(HR 256x256)

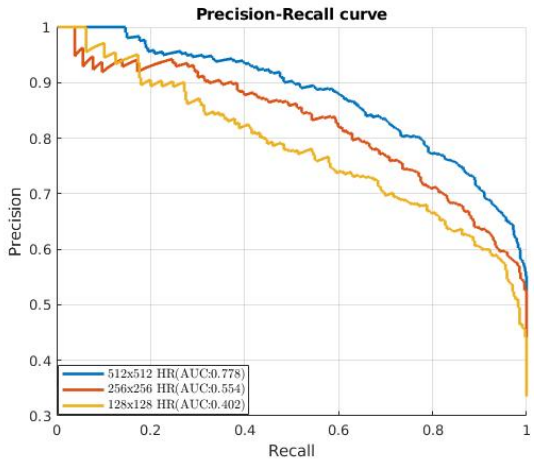
((e)) Input(LR 256x256) ((f)) Detection(LR 256x256) ((g)) Detection(SR 512x512) ((h)) Detection(HR 512x512)

((i)) Input(LR 256x256) ((j)) Detection(LR 256x256) ((k)) Detection(SR 512x512) ((l)) Detection (HR 512x512)

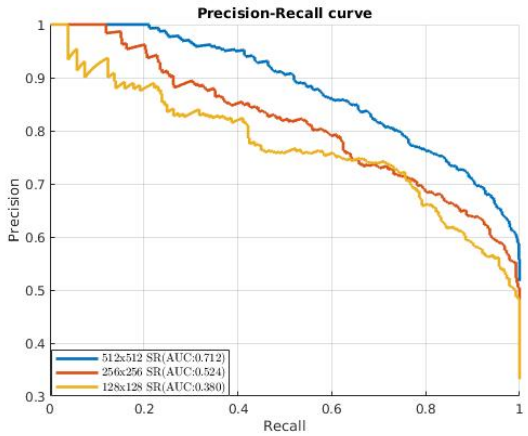
Figure 2.3: Images illustrating the detection performance on super-resolved images of the VEDAI dataset.

Table 2.2: Effects of Super-Resolution for upscale factor 4 on SSD performance

| Input Resolution | mAP(Input) | SR(4x) | mAP(SR 4x) | Original HR | mAP(HR) |
|---|---|---|---|---|---|
| 128x128 | 40.25 | 512x512 | 67.41 | 512x512 | 77.81 |




((a)) Detection performance on original HR aerial images



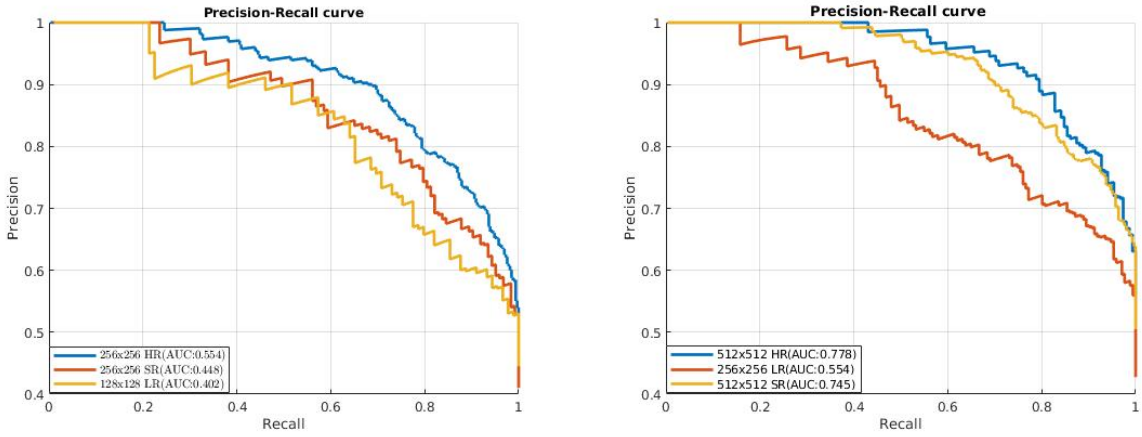

((b)) Detection performance of the model fine-tuned with the SR aerial images

Figure 2.4: Precision-recall curves for high resolution and super-resolved aerial images.

## 2.5 Performance Evaluation of the network

We have performed our experiments for both upscale factor of 2 and 4 and reported our model's performance calculating the total Area Under the Curve (AUC), a metric which falls between 0 and 1, with a higher number indicating better classification performance. In addition, Mean Average Precision (mAP) is used as the metric of evaluation. we have tested our model on 128x128, 256x256 and 512x512 resolutions of the ground truth and their corresponding super-resolved images. The performance of our network for super resolution at upscale factor 2 and 4 is given in Table 2.1 and Table 2.2, respectively.

We also include our experimental results on calculating AUC for detailed explanation. As shown in Figure 2.4(a), the system performance trained with high resolution image degrades with the decrease in resolution. In this Figure, we observe that performance of our proposed scheme is 20% higher in 512x512 resolution than resolution of 256x256 and 30% higher than 128x128 resolution. We investigate the network fine tuned with super-resolved images which is already trained on original high resolution images and report the results in Figure 2.4(b) which shows a little increase in performance. We can compare the detection results in terms of resolution in Figure 5. We observe that when we perform detection on 512x512 super-resolved images from 256x256 input resolution, the performance gain is close to original 512x512 high-resolution image. But when we experiment on 256x256 super-resolved images from 128x128 input resolution, we




((a)) Detection performance on 256x256 SR aerial images compared to its HR and LR input images of size 128x128 ((b)) Detection performance on 512x512 SR aerial images compared to its HR and LR input images of size 256x256

Figure 2.5: Detection performance on super-resolved aerial images for the upscale factor 2

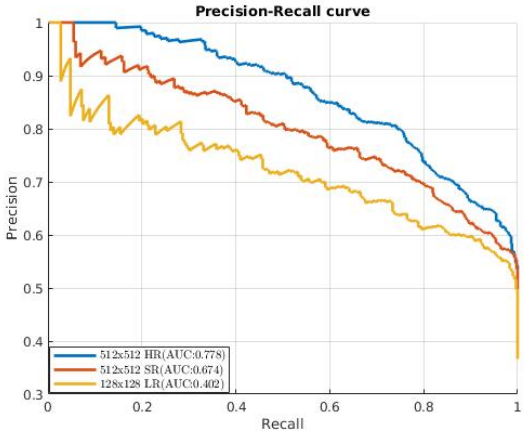



Figure 2.6: Detection performance on super-resolved aerial images for the upscale factor 4

don't achieve similar performance gain though in both case the resolution is increased by two factor. In order to see the effect of upscale factor 4 on the detection performance, we conduct our experiment on 512x512 super resolved images where the input resolution is 128x128. During this experiment, We notice a great improvement in detection performance illustrated in Figure 2.6. We can summarize that super-resolution helps detection mostly in high-resolution.

## 2.6 Conclusion

This paper presents a model for real-time vehicle detection in aerial image combining two deep networks. The first model augments the features of the low resolution images by super-resolution. During experiments of this stage, we have found that ResNet has a great impact in recovering finer texture details; however it still has limitations to recover missing information from low resolution aerial image. Then, to identify objects from aerial images, the second model utilizes these features and performs detection in real time. Our proposed scheme gives faster object recognition with competitive performance. We have also demonstrated our detection performance in a comparative analysis of multiple resolutions.

In future, we intend to modify the existing architecture that helps to generate better prediction results. Other aerial datasets like COWC [105], DOTA [106], X-VIEW [107] can also be explored to get the comparable performances. Furthermore, we can extend our work for vehicle tracking from aerial video. Our proposed system can be rebuilt in such a way that the optimization of the entire network loss takes place in combination for increasing the performance of the detector.

# Chapter 3

# Joint Super-Resolution and Vehicle Detection Network

## 3.1 Introduction

Real-time vehicle detection in aerial imagery has been an active research area in recent years [108–111]. Due to high altitudes in which aerial images are acquired, targets of interest (e.g., vehicles) contain fewer pixels than targets imaged at considerably lower elevations (e.g., building surveillance cameras, or traffic cameras), which significantly degrades detection performance. Moreover, complex background and computational constraints further hinder detection performance. Single image super-resolution (SISR) techniques are commonly used to alleviate poor detection performance by generating a high-resolution counterpart to the original low-resolution image. Recently, generative adversarial networks (GANs) [11] have demonstrated the ability to synthesize high-quality images [12, 112] for many applications, including super-resolution. However, GANs have also been known to be somewhat unstable, frequently lacking discriminability in synthesized imagery. Therefore, we aim to produce and simultaneously train both discriminative and super-resolved images by using multi-task learning to combine correlated tasks such as super-resolution and object detection networks.

The inter-relationship between super-resolution techniques and object detection algorithms has been previously studied to improve detection performance [109, 113, 114]. However, none of them have tried to explore performance of super-resolution if the entire network is trained jointly. One

might presume that the reason there are still misdetections and detection failures is because the super-resolution algorithm is not optimized for target detection task.

In this paper, we propose a deep neural network (DNN) framework to simultaneously generate super-resolved aerial images and locate vehicles in the super-resolved images. Our proposed framework is composed of (i) a Multi-scale Generative Adversarial Network (MsGAN) framework to create super resolved versions of the original images. This network preserves high-level features when mapping between low resolution to high resolution domains, and (ii) locate vehicles using one of the variants of YOLO [102] introduced in [115] as YOLOv3 object detector. We jointly train the entire network at each iteration such that target regions in the super-resolved images become contextually more distinctive from the background. We refer to our proposed algorithm in this paper as the Joint Super-Resolved Vehicle Detection Network (Joint-SRVDNet). Our proposed framework has been evaluated on several extensively used aerial datasets. We train the model on VEDAI, xView and DOTA datasets to evaluate both qualitative and quantitative performances. Moreover, our network shows promising performances compared to a set of state-of-the-art methods. In summary, the key contributions of this paper are as follows:

- In this paper, we propose an end-to-end jointly trainable deep neural network what we named Joint-SRVDNet, which offers a multi-tasking paradigm by handling both super-resolution and vehicle detection for aerial and satellite imagery. To the best of our knowledge, our proposed Joint-SRVDNet is the first multi-task model that leverages complementary information of the two tasks to jointly learn Super-Resolution (SR) and vehicle detection in aerial images. Such a novel framework allows for improved super-resolution reconstructions and more accurate vehicle detection in aerial imagery.

- An MsGAN architecture is proposed for the first time for aerial and satellite image super-resolution, which ensures progressive learning of the statistical distributions of images at multi-scale and significantly improves the performance of SR reconstruction by producing discriminative and high-quality super-resolved images.

- The proposed MsGAN architecture for super-resolution has potential contributions to vehicle detection in low-resolution aerial and satellite images.

- We show remarkable improvements for both super-resolution and vehicle detection for low-resolution aerial imagery with comparable performance to the existing state-of-the-art methods when evaluated on the corresponding high-resolution aerial images.

The rest of this paper is organized in the following manner. Section 3.2 reviews related super-resolution and detection algorithms. It also describes challenges when applied to aerial imagery. We give details of our proposed method in section 3.3. Besides, we also discuss the training loss functions of our network in section 3.4. Section 3.5 presents the datasets and experimental details of our work. Section 3.6 shows comparative results and explains the performance. Finally, we provide a conclusion and state some limitations of our algorithm in section 3.7.

## 3.2 Related Work

### 3.2.1 Deep Learning Based Single Image Super-Resolution

Single Image Super-Resolution (SISR) techniques have been studied extensively in the field of computer vision. Recently, Convolutional Neural Network (CNN) architectures have been widely used in image SR algorithms since they can extract representative features that are useful in recovering high-frequency details in super-resolved images. A three-layer CNN was first proposed by Dong et al. [116] and referred as SRCNN to learn a mapping between Low-Resolution (LR) and High-Resolution (HR) image pairs, which was later modified in VDSR [117] and DRCN [118]. In VDSR [117], Kim et al. implemented an efficient SSIR method, where they showed that increasing the network depth trained by adjustable gradient clipping resulted in a significant improvement in visual quality of super-resolved images. In DRCN [118], they increased recursion depth by adding more weight layers with skip connection to improve the performance of SRCNN. However, all these methods apply interpolation to the LR inputs, which significantly loses some useful information and thereby yields poor results with increased computational cost. Since then these super-resolution architectures have been frequently modified by developing CNN-based architectures like Residual Networks (ResNet) [119], Recurrent Neural Networks (RNNs) [120–122] to extract features from the original LR inputs.

Recently, GANs [11] have replaced these SR algorithms. Ledig et al. [12] introduce ResNet

as the base architecture for image super-resolution and utilize the idea of GAN to reconstruct fine texture details in the super-resolved images. GAN architectures have successfully attained superior performances in many applications of computer vision, such as style transfer, image reconstruction and image SR. SRGAN [12] is the first attempt which utilizes GAN to produce photo-realistic natural looking images close to the original high resolution images. They formulate a loss function which is a combination of a perceptual similarity loss [72, 73, 123] in addition to an adversarial loss [11] so that the network learns to preserve content of images during SR training.

Although SRGAN has shown remarkable performances, still it finds difficulty in generating high-resolution (e.g., $256 \times 256$) images due to training instability and mode collapse. During up-scaling the LR images to the desired HR counterparts, GAN suffers from the training instability due to low chance of sharing hyper-parameters between image distribution and model distribution in a high-dimensional space. To stabilize the training process, Zhang et al. proposed StackGAN [124]. The motivation came from the observation that image distributions are related at multiple scales. StackGAN outperforms significantly other state-of-the-art methods in reconstructing real looking super-resolved images. In StackGAN, they used multiple-generators along with discriminators at each scale to share most of their parameters across the whole network. This structure pushes the resulting solutions towards the original image distributions. For our work, we incorporate the idea of using multiple discriminators at each different scale in addition to the work of Ledig et al. [12] where the authors use a perceptual loss function with Mean Squared Error (MSE) loss to generate more realistic SR images. Our network can be viewed as multi-scale GAN architecture since we are using only one generator instead of multiple generators like StackGAN and stack discriminators at each intermediate outputs to improve the learning of image distributions at multiple scales. As shown in Fig. 3.2, discriminators at intermediate outputs sequentially help generator produce real-looking super-resolved images to the desired size. The prime goal is to approximate highly related image distributions at different scales. So, stacking multiple discriminators helps the network accomplish this goal by continuously giving feedback from image distributions at one scale to another.

### 3.2.2 Deep Learning Based Vehicle Detection Architectures

Vehicle detection recently has become a prominent research area with applications in civilian and military surveillance, traffic monitoring and planning transportation systems. In [74], the authors proposed a method which utilized Bayesian network to integrate the important features for car detection. Choi et al. [75] applied the Mean-shift algorithm to extract car like shape for detecting cars in satellite images. In the work of [76], they trained a Dynamic Bayesian Network (DBN) to preserve region level features.

Carlet and Abayowa [78] proposed a modified YOLOv2 [125] for locating vehicles in aerial imagery. A modified faster R-CNN was applied in the work of Terrail et al. [79] that showed promising performances in aerial vehicle detection. In [80], Soleimani et al. proposed a text-guided detection scheme using both visual and textual features for detection. Yang et al. [81] applied skip connection in their framework to merge lower and higher level features and utilized a focal loss function for vehicle detection. For multi-oriented vehicle detection, Li et al. [82] designed a rotatable region proposal network which learned the orientation of vehicles while performing classification on aerial images and videos.

Vehicle detection in overhead imagery remains a challenging issue due to the low resolution of vehicles. To alleviate this shortcoming, researchers have focused on super-resolution techniques. An overview of detection performance on super-resolved images is reported in [113] considering multiple-resolutions. In this paper, we propose a joint training approach which learns to extract discriminative features from low-resolution images such that it can produce super-resolved images that are as visually similar to the corresponding high-resolution images as possible.

### 3.2.3 Joint Training of Super-Resolution and Detection

Improving object detection performance guided by learning based super-resolution has been a recent research focus. In [113], the impact of super-resolution on object detection has been extensively studied. Haris et al. [126] adopt a task-driven super resolution approach employing a novel compound loss based end-to-end training that enhances the image quality leading to a better recognition. Cansizoglu et al. [127] design an identity preserving face super-resolution framework and achieve outstanding performance for face verification in real time. In this work, the authors

propose to use a two-stage loss minimization technique rather than end-to-end training. They hypothesize that end-to-end training involves higher computational complexity respect to limited data samples. On the other hand, another study in [128] propose a deep model that jointly optimizes face hallucination and verification loss for low resolution face identification. In this study, face hallucination loss is measured in terms of pixel difference between the ground truth HR images and network-generated images and verification loss is estimated by the classification error and intra-class distance. Most of the recent works focus on verification, which is easier from the detection task. For example, verification confirms identity whereas detection involves recognition of desired object (e.g., human face, vehicle, etc.). Again, during verification, the probe face has already been detected, but detection has to minimize different constraints before detecting the target object.

Pang et al. [129] introduce JCS-Net that combines classification and super-resolution task as one for small-scale pedestrian detection. However, these algorithms do not deal with vehicle detection and super-resolution for aerial imagery that deals with more fundamental challenges. For instance, the average height of pedestrians in the benchmark datasets (e.g., Caltech [130], KITTI [131]) ranges from 60 pixels tall to 430 pixels tall, whereas the average resolution for aerial vehicles is $10 \times 15$ pixels in the publicly available benchmark datasets (e.g., VEDAI [132], xVIEW [106], DOTA [107]), which yields poor detection results.

These reviews strongly suggest to use super-resolution technique for developing a robust detection system, which helps to recover detailed information in the low-resolution space. In this paper, we try to investigate the relationship between super-resolution and vehicle detection by proposing a joint training approach so that they can be benefited from each other. We propose to integrate both super-resolution and detection network together. Usually, the super-resolution technique recovers useful detailed information in the low-resolution image, but here it focuses especially on the target regions as detector loss is integrated to SR training. The network gradually learns the input image distributions in the high-resolution space and produce super-resolved version of low-resolution image with distinctive properties of target objects, which also helps detector to achieve better results.

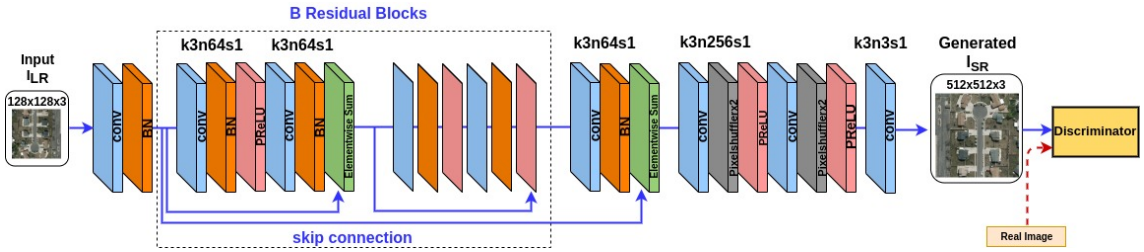



Figure 3.1: Architecture of SRGAN with corresponding kernel size (k), number of feature maps (n) and stride (s) indicated for each convolutional layer.

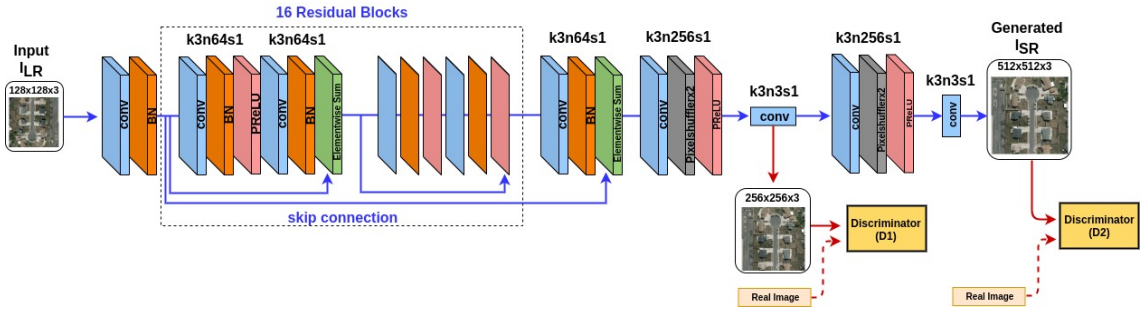



Figure 3.2: Architecture of Multi-scale SRGAN (MsSRGAN) with corresponding kernel size (k), number of feature maps (n) and stride (s) indicated for each convolutional layer.

## 3.3 Proposed Framework

In this section, we describe our proposed framework in detail. The proposed framework is an end-to-end network that generates super-resolved aerial images using an MsGAN architecture and jointly optimized YOLOv3 detector to perform vehicle detection in aerial super-resolved imagery.

### 3.3.1 Generative Adversarial Networks (GANs)

GANs are a special type of generative models which have shown remarkable performances in representation learning and synthesized image generation. They have been widely used in image super-resolution (first applied by Ledig et al. in [12]), image synthesis and image translation using conditional GANs (cGANs) [13] and cyclic GANs (cycleGANs) [112]. Their goal is to learn statistical distribution of the training data to train a mapping $G : x \rightarrow y$ such that image distribution from G(x) is indistinguishable from image distribution of target y. Typically, the generator G is a differentiable function which is trained to learn the distribution $p_{data}$ over data y. To do so, it takes input from the distribution $p_x(x)$ and maps it to the target data space as $G(x; \theta_g)$ where

$\theta_g$ defines the parameters of the generator model. In addition, the discriminator D acts like a classifier which is trained to return probability distributions D(y) and D(G(x)) for both training examples from the distribution $p_{data}(y)$ and samples from G(x), respectively. Basically, D is trained to maximize the probability of assigning the correct label to both training examples and samples from G. Simultaneously G is trained to minimize log(1 – D(G(x))). In other words, D and G play the following two-player minimax game with the adversarial loss $l_{GAN}(G, D)$:

$$\min_G \max_D l_{GAN}(G, D) = \min_G \max_D [E_{y\sim p_{data}}[\log D(y)] + E_{x\sim p_x}[\log(1 - D(G(x)))]]. \tag{3.1}$$

However, it is very difficult to achieve the desired output by training the network only with adversarial loss. Adding a $l_{L1}$ reconstruction loss in addition to adversarial loss may result in high quality super-resolved images. Thus, the final objective function consists of two loss function as follows:

$$G^* = \arg\min_G \max_D l_{GAN}(G, D) + \lambda l_{L1}(G), \tag{3.2}$$

where $l_{L1}(G) = \frac{1}{N}\sum_{i=1}^{N} ||y_i - G(x_i)||_1$, N defines the number of samples in the training set and $\lambda$ is a weighting factor.

### 3.3.2 Multi-scale GAN Architecture for Image Super-Resolution

One of the objectives of our work is to estimate a high resolution version with distinctive features of its low resolution input aerial image. The network is trained to learn a generating function G that aims to output photo-realistic images (according to a large distribution of images). Our basic deep generator network is illustrated in Fig. 3.1 which consists of B(=16) serially connected residual blocks with identical layout. Each residual block uses two convolution layers of 3x3 kernel and 64 feature maps followed by batch-normalization layers [133] and ParametricReLU [90] as the activation function. To increase the resolution of the input image, we employ two sub-pixel convolutional layers [91] in our generator network.

Although this architecture achieved promising results in recovering high-frequency information from low-resolution images; it cannot handle varying condition (sharpness, atmospheric tur-

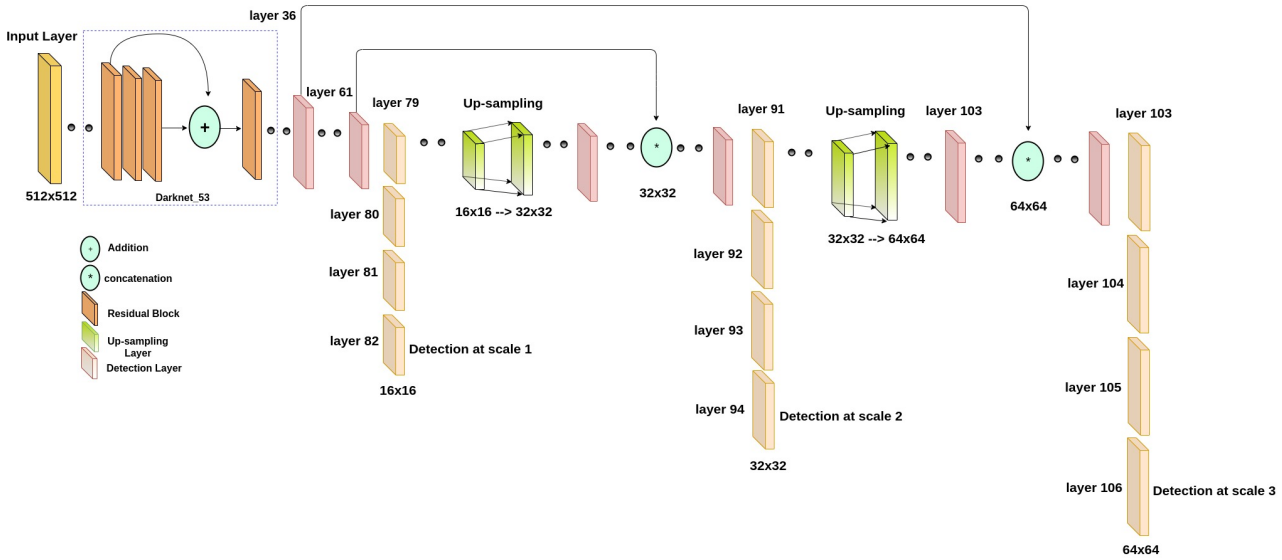



Figure 3.3: YOLOv3 architecture for vehicle detection at three scales showing residual block, upsampling layer as feature extractor.

bulance, motion blur, etc.). Usually, the estimated super-resolved images suffer from image blurriness and shape distortions. Moreover, some details which are vital for producing natural looking images are missing in the super-resolved images.

One application of aerial image super-resolution is vehicle detection, which requires enough visual detail to distinguish vehicles from background (e.g., roads, buildings, trees, etc.) in super-resolved images. Our previous detection results [109] showed that this network is not able to produce a high-detection performance while performing on super-resolved images generated by the classical SRGAN. We follow the framework of Kazemi et al. [134] and Wang et al. [135] to build a progressive generator that learns to reconstruct a multi-stage network through a series of multi-scale image reconstructions. We train our generator model to produce multiple outputs at different resolutions as shown in Fig. 3.2. The main idea is to encourage the network to learn the image distribution at different scales. We enforce constraints on our network at two different image resolutions 256x256 and 512x512. When the network generates images of size 256x256, the first discriminator, D1 is pushing the generator to learn the probability distribution at that scale. Simultaneously, the second discriminator, D2 is contributing to help the generator to learn the distribution of the training images of size 512x512.

Gradually, the network learns to remove blurriness and recover missing object parts as it is

trained at multi scales. Following this approach, it assures information transfer between images of different scales and generate more high-quality images.

We follow similar network structure for both discriminators D1 and D2. We adapt the architectural guidelines from Radford et al. [92] to design our discriminator. We utilize a LeakyReLU activation ($\alpha$ = 0.2) and avoid max-pooling to ignore feature size reduction. Our discriminator has eleven convolutional layers, which use 4x4 filter kernels. Network employs strided convolutions to decrease image resolution while increasing the feature map size. At the end of the network, one dense layer and a final sigmoid activation function is added to obtain a probability for sample classification.

### 3.3.3 Aerial Vehicle Detection

Our goal is to perform vehicle detection on several aerial datasets. The datasets contain vehicles of different sizes which require strong detection algorithm to extract contextual and semantic information of those target objects. In our research work, we use YOLOv3 of the state-of-the-art object detection algorithms to perform vehicle detection in real-time.

#### Architecture Details

The architecture of YOLOv3 shown in Fig. 3.3 is based on the idea of residual network which employs Darknet-53 convolutional network for feature extraction. To retrieve fine-grained information, it concatenates deeper layers with the earlier layers through up-sampling. YOLOv3 takes an image and divides it into M $\times$ M (16 $\times$ 16, 32 $\times$ 32 and 64 $\times$ 64 as in Fig. 3.3) grids. Then it applies classification and localization at each grid size. The grid cell is responsible for detecting object, if the center of the ground truth object falls within a grid cell. For each grid cell, a number of bounding boxes with their confidence scores and their associated class probabilities are generated using a fully convolutional network architecture. YOLOv3 performs multi-scale prediction applying the feature pyramid network (FPN) [136] concept. It predicts objects at three different scales of 16, 32 and 64 for large, medium and small object detection. YOLOv3 uses 9 anchor boxes while predicting objects. Design of the anchor boxes greatly impacts the performance of the detector. We have used k-means clustering to generate these anchors for each database. The final

number of detection results by YOLOv3 is $M \times M \times (B * (4 + 1 + C))$. Here, $M \times M$ is the number of grid cells, B is predicted number of bounding boxes in a cell, 4 denotes the four coordinates of the bounding boxes and 1 is for the objectness score, C is the number of classes ( C=1:'vehicle' in our experiments). It uses multi-label classification. Softmax is replaced by a logistic regression to compute objectness score. Instead of using mean squared error in calculating the classification loss, it uses the binary cross-entropy loss for each label.

### 3.3.4 Our Proposed Joint Super-Resolution and Detection Network

In this paper, we propose an end-to-end multi-task model that jointly does super-resolution and vehicle detection in aerial imagery. Super-Resolution and vehicle detection for low-resolution aerial images have been considered as highly interrelated tasks. Usually, multi-task learning is adapted to address such highly correlated tasks as they can leverage significant information from each other. The vehicles in aerial scenes suffer from appearance ambiguity due to the low resolution characteristics of the images. In addition, it becomes challenging to deal with different sizes of vehicles with varying conditions such as blurry edges and lack of sharpness, etc. Moreover, the similarities between target vehicles and complex background make it even more difficult during detection.

In our previous work [109], super-resolution and vehicle detection networks were developed independently to help each other. We notice that the information extracted from the low-resolution space is not maximized when only one task is performed without utilizing the advantages of the other task (e.g., detection is performed on super-resolved images generated from already trained SR module). In other words, if we apply super-resolution and vehicle detection successively, it does not benefit from multi-tasking. Therefore, our goal is to create a bridge between these highly interrelated tasks so that they can get the maximum benefit from the multi-task learning. Hence, we propose the Joint-SRVDNet to generate distinctive super-resolved images with high perceptual quality and simultaneously locate vehicles on these super-resolved images. We have developed a MsGAN super-resolution module that explicitly incorporates the structural information (edges, sharpness, perceptual features defined by visual deterministic properties of objects) about targets into the super-resolution reconstruction process as well as jointly learns both the super-resolution and object detection modules together as presented in Fig. 3.4. As shown in Fig. 3.4 super-

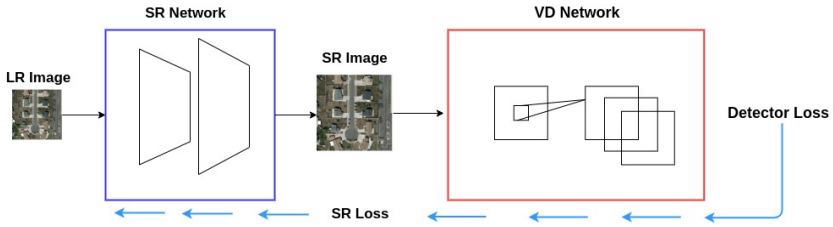



Figure 3.4: Architecture of our proposed model Joint-SRVDNet during the training process where the detector and super-resolution losses are back propagated to the generator.

resolution and detection modules are cascaded to execute the joint training in an end-to-end fashion.

The joint loss optimization of our model is difficult to converge from scratch compared to the training of each module independently. Therefore, we first train super-resolution module given the paired high-resolution and corresponding low-resolution aerial training images. Then we train detection module with high resolution images to obtain network parameters for further training. Finally, we fine tune both modules together and integrate into one unified framework by optimizing (3.7) where super-resolution and detection losses are jointly trained together. Such a training scheme leads to a better convergence. Our proposed network optimizes a combination of four different losses : adversarial loss, pixel-wise mean square error (MSE), perceptual loss, and detection loss. The adversarial loss aims to help generator to create solutions that are close to real images by differentiating between real and generated aerial images. The widely used pixel-wise MSE estimates an overly smoothed solution as it only measures pixel differences between super-resolved images and ground truth high resolution images. A perceptual loss using the pretrained VGG-19 network recovers photo-realistic textures, and a detection loss that aims for locating the target of interests with varying attributes such as lost edge details and structural features.

## 3.4 Loss Function

We combine multiple loss terms to train our proposed joint network. The ultimate final loss function includes pixel-wise MSE loss, perceptual loss, adversarial loss and detection loss.

### 3.4.1 Pixel-Wise MSE Loss

State-of-the-art image SR methods [91, 116] mostly rely on pixel-wise MSE loss to optimize the network. For the training images $I_n^{HR}$ with their corresponding low-resolution $I_n^{LR}$, *n = 1,....,N*, we can calculate the MSE loss also referred to as the content loss $l_{cont}$ using the following equation:

$$L_{cont} = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{WH}\sum_{x=1}^{W}\sum_{y=1}^{H}((I_n^{HR})_{x,y} - G(I_n^{LR})_{x,y})^2, \tag{3.3}$$

where W and H represent width and height of the image and $G(I_n^{LR})$ are the super-resolved images for *N* training samples.

Although MSE loss is the widely used optimization method for super-resolution which achieves high peak signal-to-noise ratios, the resulting estimates often lack fine texture details and are perceptually not convincing because of overly blurry results. In addition, MSE doesn't have ability to capture spatially varying high frequency information, as it is based on pixel-wise image differences.

### 3.4.2 Perceptual Loss

Since optimizing the MSE loss is prone to overfitting when defined over the pixel-wise differences between estimated super-resolved images and ground truth high resolution images, Ledig et al. [12] propose the perceptual loss, which is defined as the MSE loss over high-level features extracted from the corresponding images. These features, which are extracted using a pretrained 19 layer VGG Network [137], map raw images to a lower dimensional and representative subspace. Thus, optimizing the perceptual loss better preserves discriminative information and alleviate overfitting. The perceptual loss can be considered as the L2 distance between the feature representations of the generated super-resolved image and ground truth HR image. For *N* training samples we solve:

$$L_{per} = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{C_j W_j H_j}\sum_{c=1}^{C_j}\sum_{x=1}^{W_j}\sum_{y=1}^{H_j}(\phi_j(I_n^{HR})_{c,x,y} - \phi_j(G(I_n^{LR}))_{c,x,y})^2, \tag{3.4}$$

where $\phi_j$ stands for feature map of *j*-th convolutional layer and $C_j$, $W_j$ and $H_j$ define the dimensions of the respective feature maps within the VGG19 network.

### 3.4.3 Adversarial Loss

Since the network cannot learn to recover all high-frequency information by optimizing only the MSE or the perceptual losses, we also add the adversarial loss to the perceptual and the pixel-wise MSE losses to train our proposed network. The adversarial loss described by (3.1) pushes the solutions move towards the natural image manifold by training the generator to fool the discriminator by generating photo-realistic images, and training the discriminator to accurately classify ”real” images from the generated ones (i.e., fake images). Thus, the estimated solutions reside on the real samples manifold. The adversarial loss $l_{adv}$ defines the probability of the discriminator $D(G(I^{LR}))$ that the reconstructed image $G(I^{LR})$ is a real HR image. Both discriminators as shown in Fig. 3.2, use the following adversarial loss functions to optimize the network.

$$L_{adv} = \min_{G} \max_{D} [E_{I^{HR} \sim P_{train}(I^{HR})}[\log D(I^{HR})] + E_{I^{LR} \sim P_G(I^{LR})}[\log(1 - D(G(I^{LR})))]], \tag{3.5}$$

where $P_{train}(I^{HR})$ and $P_G(I^{LR})$ define the probability distribution of real high-resolution images and corresponding low-resolution images, respectively.

### 3.4.4 Detection loss

YOLOv3 is the combination of three losses: localization, confidence and classification loss. Equation (6) defines this loss. $1_{ij}^{obj}$ means the object is detected by $j^{th}$ boundary box of grid cell i. $x_i$,$y_i$, $w_i$,$h_i$ are the real ground truth bounding box coordinates whereas $\hat{x}_i$,$\hat{y}_i$, $\hat{w}_i$,$\hat{h}_i$ are the predicted bounding box coordinates. $C_i$ is the box confidence score in cell i, $\hat{C}_i$ is the box confidence score for the predicted object:

$$\begin{aligned} L_{detection} = \lambda_{coord} \sum_{i=0}^{S^2} \sum_{j=0}^{B} 1_{ij}^{obj} (x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2 + \lambda_{coord} \sum_{i=0}^{S^2} \sum_{j=0}^{B} 1_{ij}^{obj} (\sqrt{w_i} - \sqrt{\hat{w}_i})^2 + (\sqrt{h_i} - \sqrt{\hat{h}_i})^2 \\ + \sum_{i=0}^{S^2} \sum_{j=0}^{B} 1_{ij}^{obj} l(C_i, \hat{C}_i) + \lambda_{noobj} \sum_{i=0}^{S^2} \sum_{j=0}^{B} 1_{ij}^{noobj} l(C_i, \hat{C}_i) + \sum_{i=0}^{S^2} 1_{i}^{obj} \sum_{c \in classes} l(p_i(c) - \hat{p}_i(c)). \end{aligned} \tag{3.6}$$

### 3.4.5 Joint Loss Optimization

Our proposed model can be viewed as a joint learning approach. The network is learning semantic information about targets from the training distribution so that the appearance of the target looks more clear and obvious in super-resolved images to help the detection module. In this section, we show how we combine the detection loss along with the pixel-wise MSE loss, perceptual loss and adversarial loss through an optimization to produce our desired output with full target details. Therefore, to show the dependency of different loss functions, lets assume $W_{SR}$, $W_{VGG}$, $W_{dis}$ and $W_d$ denote the parameter set for super-resolution model, pre-trained VGG 19 architecture, discriminator model and detection model, respectively. The parameterized version of the final loss function is as follows:

$$L = L_{cont}(I_n^{LR}; W_{SR}) + \alpha L_{per}(I_n^{LR}; W_{SR}, W_{VGG}) + \beta L_{adv}(I_n^{LR}; W_{SR}, W_{dis}) + \gamma L_{detection}(I_n^{LR}; W_{SR}, W_d). \tag{3.7}$$

We apply gradient descent algorithm to find the local minimum, and update the network's parameter by calculating the gradient $\nabla W = [\nabla W_{SR} \nabla W_d]$ with a learning rate $\eta$.

**Gradient with respect to $W_d$**

We calculate $\frac{\partial L}{\partial W_d}$ and use the standard back propagation algorithm as the following chain rule holds:

$$\frac{\partial L}{\partial W_d} = \sum_{n=1}^{N} \frac{\partial L}{\partial o_n} \frac{\partial o_n}{\partial W_d}, \tag{3.8}$$

where $o_n$ defines a vector representation of the bounding box coordinates and confidence score. Again, $\frac{\partial L}{\partial o_n}$ involves three terms according to the definition as below:

$$\frac{\partial L}{\partial o_n} = \frac{\partial L_c}{\partial o_n} + \frac{\partial L_b}{\partial o_n} + \frac{\partial L_{conf}}{\partial o_n}, \tag{3.9}$$

where $L_c$, $L_b$ simply calculate the loss for bounding box coordinates (e.g., center, width and height) and $L_{conf}$ defines bounding box confidence score loss.

**Gradient with respect to $W_{SR}$**

To update the parameter set for SR model, we consider loss terms associated with the SR reconstruction process and apply gradient descent algorithm to find $\frac{\partial L}{\partial W_{SR}}$. The chain rule holds as follows:

$$\frac{\partial L}{\partial W_{SR}} = \sum_{n=1}^{N} \frac{\partial L}{\partial G(I_n^{LR})} \frac{\partial G(I_n^{LR})}{\partial W_{SR}}. \tag{3.10}$$

If we set the partial derivative of the loss function with respect to $G(I_n^{LR})$ and expand L, we get

$$\frac{\partial L}{\partial G(I_n^{LR})} = \frac{\partial L_{cont}}{\partial G(I_n^{LR})} + \alpha \frac{\partial L_{per}}{\partial G(I_n^{LR})} + \beta \frac{\partial L_{adv}}{\partial G(I_n^{LR})} + \gamma \left( \frac{\partial L_c}{\partial G(I_n^{LR})} + \frac{\partial L_b}{\partial G(I_n^{LR})} + \frac{\partial L_{conf}}{\partial G(I_n^{LR})} \right). \tag{3.11}$$

or, we can also express the above equation as follows:

$$\frac{\partial L}{\partial G(I_n^{LR})} = \frac{\partial L_{cont}}{\partial G(I_n^{LR})} + \alpha \frac{\partial L_{per}}{\partial G(I_n^{LR})} + \beta \frac{\partial L_{adv}}{\partial G(I_n^{LR})} + \gamma \frac{\partial L_{detection}}{\partial G(I_n^{LR})}. \tag{3.12}$$

We can summarize the optimization steps in Algorithm 1.

**Algorithm 1** Our proposed Joint-SRVDNet model training

Training samples, I = ¡$I_n^{LR}$, $I_n^{HR}$¿ Model parameters set W = [$W_{SR}$, $W_d$] **while** not converged **do**
t=t+1;
calculate the partial derivative $\frac{\partial L}{\partial W_d}$;
calculate the partial derivative $\frac{\partial L}{\partial o_n}$;
execute back propagation from top layer to the bottom layer of detection to obtain $\frac{\partial L}{\partial W_d}$;
calculate the partial derivative $\frac{\partial L}{\partial G(I_n^{LR})}$;
add the $\frac{\partial L_{cont}}{\partial G(I_n^{LR})}$, $\frac{\partial L_{per}}{\partial G(I_n^{LR})}$ and $\frac{\partial L_{adv}}{\partial G(I_n^{LR})}$ to the derivative $\frac{\partial L}{\partial G(I_n^{LR})}$ obtained in step 6;
execute back propagation from the last layer to the first layer of SR to obtain $\frac{\partial L}{\partial W_{SR}}$;
update the parameter W by $W^{t+1} = W^t + \eta \nabla W$;

## 3.5 Training Details

### 3.5.1 Experimental Data

We evaluate the performance of our proposed method on three publicly available benchmark datasets: Vehicle Detection in Aerial Imagery (VEDAI) dataset [132], xView dataset [107] and

DOTA dataset [106]. In this section, detailed description of the training datasets are provided. Then, we describe the implementation and experimental strategies.

**Vehicle Detection in Aerial Imagery (VEDAI) Dataset**

The VEDAI dataset is a publicly available benchmark for small target recognition especially vehicle detection in aerial images. This dataset has around 1,210 images of two different resolutions such as $1,024 \times 1,024$ pixels and $512 \times 512$ pixels. The images mostly contain small vehicles having diverse backgrounds, multiple orientations, lighting/shadowing changes, specularities or occlusions. In addition, it includes nine different classes of vehicles, namely the plane, boat, camping car, car, pick-up, tractor, truck, van, and the other category. We consider all classes as a single class namely 'vehicle' for our task. For training and testing, we split the dataset into 1,100 and 271 images, respectively. The number of samples in our dataset is small for analyzing the proposed network. Therefore, to make the model more robust to different features, we have used different augmentation techniques such as image sharpening and flipping.

**Dataset for Object detection in Aerial images (DOTA)**

DOTA is a large-scale multi-sensor and multi-resolution aerial dataset. This dataset is challenging because of its immense number of object instances from various categories exhibiting a wide variety of scales, orientations and shapes. The dataset contains 2,806 images of varying size ranging from $800 \times 800$ to $4,000 \times 4,000$ pixels. We have created patches of size $512 \times 512$ from the original images. The complex aerial scenes present in this dataset are collected from Google Earth, satellite JL-1 and satellite GF-2. The dataset has fifteen categories of objects namely plane, ship, storage tank, swimming pool, ground track field, harbor, bridge, large vehicle, small vehicle, helicopter, roundabout, soccer ball field, basketball court, baseball diamond and tennis court. We have omitted class swimming pool, ground track field, harbor, bridge, roundabout, soccer ball field, basketball court, baseball diamond and tennis court and unified the remaining six classes as one class 'vehicle'.

**x-View Dataset**

xView is currently the largest publicly available dataset collected from WorldView-3 satellites. The dataset contains 60 highly imbalanced classes. To overcome the problem of poor detection performance, we have generalized all the classes into one class 'vehicle'. It contains around 1 million objects covering 1,400 $km^2$ of the earth surface. The dataset is cropped into smaller patches of $512 \times 512$. Each pixel corresponds to $0.3 \times 0.3$ $m^2$ area in the ground. The annotation provided is in geoJSON format and contains information about the bounding boxes for objects present in an image.

### 3.5.2 Implementation Strategies and Training Parameters

At the beginning, we separately train both sub-networks: super-resolution and detection modules to obtain their network weights which have been used to initialized the joint network of our proposed model. We perform all experiments using 4x upsampling factor between low- and high-resolution images. To obtain LR images, bicubic kernel is used to downscale the HR images with a scale factor of 4. During implementation, we use input images of size $128 \times 128$ to super-resolve to $512 \times 512$.

To train a deep neural network using a small dataset is troublesome due to the over-fitting problem. One approach to overcome this difficulty is to use data augmentation, specifically sharpening and [horizontal, veritical] flipping.

For the super-resolution network, we adapt the Adam optimizer with a momentum of 0.9 and a batch size of 4. We initially set the learning rate at $10^{-4}$ which decays by a factor of 0.1 after every 5 epochs. For YOLOv3 model, we optimize the network by Adam with a learning rate of $10^{-4}$ and $10^{-6}$ with batch size 16. For non-maximum suppression, the threshold is set to 0.5. Following (3.6), the network calculates bounding box loss, coordinate loss, class confidence scores and objectness score for each detection layer. These losses are offset to predict the object probability, class probability and bounding box coordinates for each grid which together represents an object at that grid. Usually the network generates several bounding boxes and selects the bounding box with the highest Intersection over Union (IoU). For each aerial dataset, we train both networks for 10 epochs and achieve satisfactory results.

Table 3.1: Comparison of super-resolution architectures for upscale factor 4x on aerial datasets.

| Dataset | VEDAI-VISIBLE | | | | VEDAI-IR | | | | XVIEW | | | | DOTA | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Algorithm | PSNR | MSSIM | UQI | VIF | PSNR | MSSIM | UQI | VIF | PSNR | MSSIM | UQI | VIF | PSNR | MSSIM | UQI | VIF |
| Bicubic | 22.060 | 0.912 | 0.945 | 0.560 | 22.513 | 0.920 | 0.980 | 0.597 | 15.856 | 0.419 | 0.663 | 0.416 | 24.617 | 0.936 | 0.963 | 0.349 |
| SRGAN | 25.856 | 0.918 | 0.981 | 0.607 | 25.876 | 0.928 | 0.988 | 0.627 | 17.799 | 0.517 | 0.783 | 0.515 | 24.893 | 0.941 | 0.959 | 0.514 |
| MsSRGAN | 26.899 | 0.927 | 0.991 | 0.653 | 27.890 | 0.939 | 0.995 | 0.683 | 18.838 | 0.541 | 0.794 | 0.550 | 28.474 | 0.975 | 0.971 | 0.623 |
| DenseNet GAN | 29.9 | | | | | - | | | | - | | | | - | | |
| **Joint-SRVDNet (Ours)** | **30.338** | **0.969** | **0.995** | **0.693** | **29.227** | **0.958** | **0.999** | **0.713** | **20.550** | **0.617** | **0.795** | **0.562** | **31.360** | **0.987** | **0.975** | **0.712** |

For joint-training, we consider the sub-networks together and train it as a unified network. To initialize the overall network, we employ the weights from the independently pre-trained models. We choose Adam as the optimizer by setting initial learning rate as $10^{-4}$. The learning rate decays exponentially with moving average decay of 0.9991. After training for 4 epochs with a mini-batch size 1, we observe significant improvement in results which verifies that our proposed method has been successfully implemented. We implement the proposed network using tensorflow framework and train it over two NVIDIA Titan XpGPU. Moreover, we explored the effect of varying the hyperparameters ($\alpha$, $\beta$ and $\gamma$) adapted in (3.7) to further validate the results of our model. The analysis of the hyperparameters has been made on the test dataset and their impact will be discussed in the ablation study.

## 3.6 Experimental Results Analysis

In this section, we present comparative results for both image super-resolution and vehicle detection on several aerial datasets to evaluate the performance of our proposed model. We compare the reconstruction quality of the super-resolved images generated by our proposed network to other methods including bicubic interpolation, SRGAN [12], MsSRGAN [134] and DenseNet GAN [138] on overhead datasets which were described in the previous section. Then we investigate vehicle detection performance of our network in terms of mean Average Precision (mAP) and F1 score. For more comprehensive performance analysis, we provide precision-recall curve and plot true positive rate (TPR) against false positive rate (FPR).

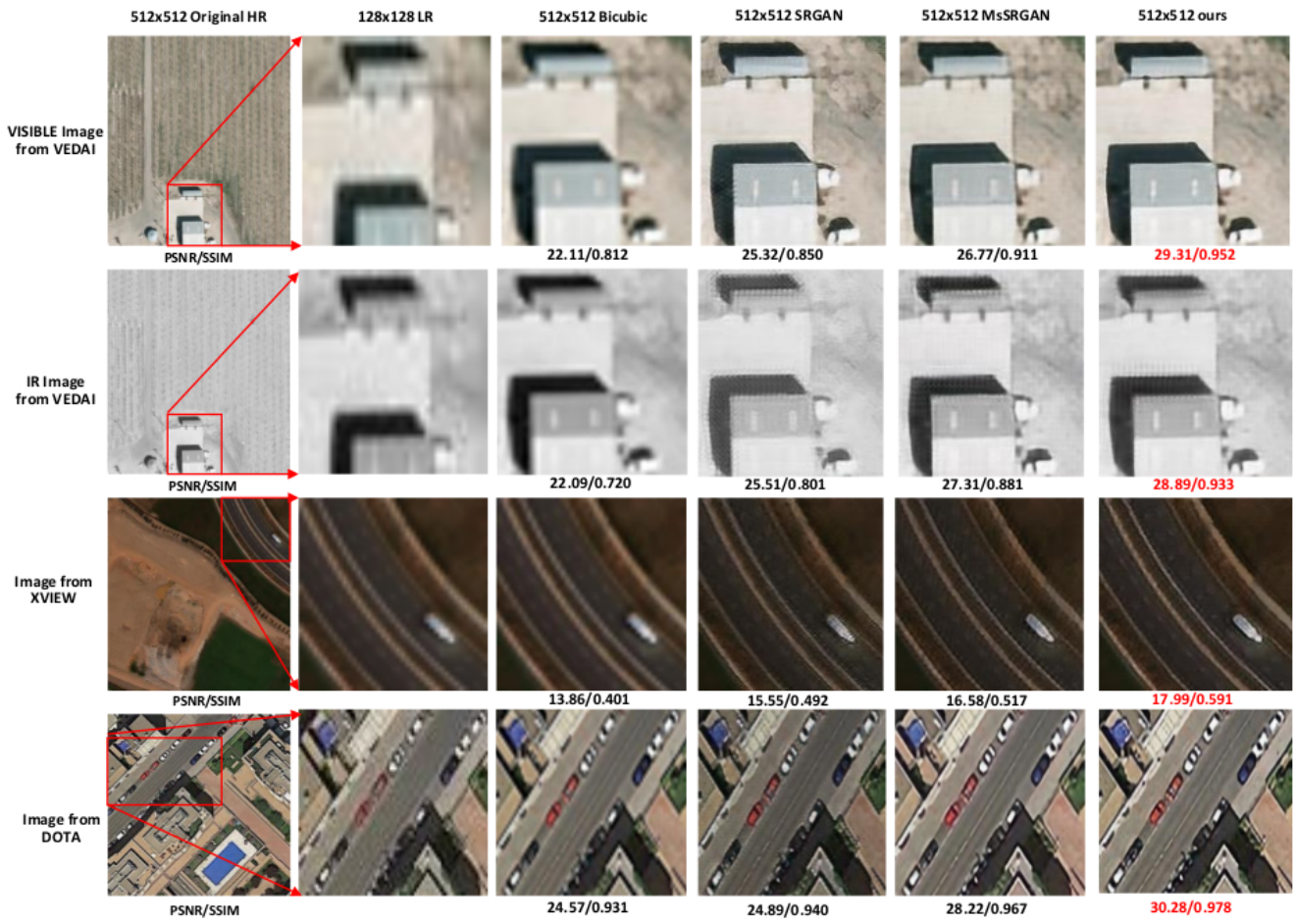



Figure 3.5: Visual results using Bicubic, SRGAN, MsSRGAN and our proposed model Joint-SRVDNet with scaling factor 4 over VEDAI, xView and DOTA datasets.

### 3.6.1 Super-Resolution Results

We have reported the super-resolution results of our experiments using several objective image quality metrics such as Peak Signal-to-Noise ratio (PSNR), Multi-scale Structural Similarity (MSSIM) [143], Universal image Quality Index (UQI) [144] and Visual Information Fidelity (VIF) [145] on a validation subset of images for each dataset. Table 3.1 shows a comparative analysis of our approach with other GAN based state-of-the-art techniques. For comparison, first we include results from bicubic interpolation method. Then we follow SRGAN architecture; one of the pioneering works on super-resolution using GAN introduced by Ledig et al [12]. As expected, the performance of this network is much better than the previous approaches due to addition of the perceptual loss which enables the network to produce images with sharper edges and features. After

Table 3.2: Comparative detection performance in terms of mean average precision (mAP) and F1-score of the proposed network and existing state-of-the-art approaches. Red bold indicates the optimal performance using actual HR imagery and blue bold indicates the second optimal performance using SR images generated by our proposed network.

| Dataset | VEDAI-VISIBLE | | VEDAI-IR | | XVIEW | | DOTA | |
|---|---|---|---|---|---|---|---|---|
| Architectures | mAP@0.5 | F1 score | mAP@0.5 | F1 score | mAP@0.5 | F1 score | mAP@0.5 | F1 score |
| Ren, et al. (Z&F) [101] | 32.00 | 0.212 | - | | - | | - | |
| Girshik, et al. (VGG-16) [100] | 37.30 | 0.224 | - | | - | | - | |
| Ren, et al. (VGG-16) [101] | 40.90 | 0.225 | - | | - | | - | |
| Zhong, et al. [139] | 50.20 | 0.305 | - | | - | | - | |
| Chen, et al. [140] | 59.50 | 0.451 | - | | - | | - | |
| YOLOv3_SRGAN_512 | 62.45 | 0.591 | 70.10 | 0.687 | 53.47 | 0.479 | 86.18 | 0.837 |
| YOLOv3_MsSRGAN_512 | 66.74 | 0.643 | 74.61 | 0.723 | 57.96 | 0.494 | 87.02 | 0.859 |
| YOLOv3_SSSDet_512 [141] | 45.97 | | - | | - | | 79.52 | |
| Ju, et al. [142] | - | | - | | - | | 88.63 | |
| **YOLOv3_Joint-SRVDNet_512 (Ours)** | **72.46** | **0.702** | **80.40** | **0.792** | **61.50** | **0.671** | **90.01** | **0.893** |
| **YOLOv3_HR_512** | **85.33** | **0.826** | **85.66** | **0.876** | **66.02** | **0.687** | **94.56** | **0.933** |

that, we notice, adding multiple intermediatory discriminators to the same generator architecture as the SRGAN helps to generate even higher quality images with more perceptual similarity which often lacks in the generated images from the SRGAN. We refer to this network as MsSRGAN which is actually introduced in [134] to handle super-resolution for facial images. We utilize this concept and conduct experimients on aerial datasets. We observe slight improvements in the reconstructed SR results and report it for comparison. Moreover, we have also compared our results with DenseNet GAN [138] for VEDAI dataset. All these GAN based methods use perceptual loss, MSE loss along with adversarial loss even if they modify their architecture which shows gradual improvement in their solutions. However, they cannot meet the demand of current situation. They are often unable to extract fine texture details of the targets (vehicle) of interest. So, our aim is to produce solutions which contain clear view of our targets with fine-grained details. We design a loss function which incorporates detecton loss along with other losses (perceptual loss, adversarial loss and MSE loss) which helps to reach our goal. Table 3.1 shows that our proposed algorithm obtains the highest PSNR, MSSIM, UQI and VIF scores which proves the quantitative effectiveness of our proposed network. To show the quality of the super-resolved images specifically for the target regions produced by our network, we select a small area around the targets and show the gradual progression of different SR results which are visible in Fig. 3.5. We have conducted our experiments for 4x enhancement ($128 \times 128$ to $512 \times 512$). We can see that in the super-resolved image the selected area around the target and the target itself is getting more close to

the original one as bicubic interpolation, SRGAN, MsSRGAN and our network have been applied successively. Visual results are showing that recovering high frequency details in low-resolution domain is extremely difficult but it is captured by using our proposed network. The ultimate goal of our work is to recover target details which has a great effect on the detection performance.

### 3.6.2 Detection Performance Analysis

Table 3.2 summarizes a comparative performance measures of our proposed model and other leading state-of-the-art algorithms in terms of mAP and F1 score for aerial vehicle detection. The mAP values and F1 scores are reported on VEDAI, xView and DOTA datsets for most of the algorithms based on the availability. We calculate the mAP as the average of the maximum precisions at different recall values in the range ($0.0 \sim 1.0$). For each dataset, we show the precision-recall graphs at different IoU thresholds ($0.3 \sim 0.7$) for YOLOv3 performed on super-resolved images generated from SRGAN, MsSRGAN and our proposed network as shown in Fig. 3.6. We have evaluated all the methods over the same set of test data. we can conclude that our proposed technique is much more stable and robust for aerial vehicle detection in comparison to the current state-of-the-art detection techniques.

Besides, we include detection results of recent CNN-based detectors: Faster R-CNN [101] with Z&F model, Faster R-CNN [101] with VGG-16 model and Fast R-CNN [100] with VGG-16 model for VEDAI dataset. Also, we have compared our detection performance with [139] and most recently proposed detection algorithm [140]. It is easily noticeable from the results presented in Table 3.2 that our proposed model demonstrates the best performance compared to these detection methods and yields the 2nd best mAP (72.46%) and F1-Score (0.702) for VEDAI. For comparison with the current DCNN based approaches, we include the results of SSSDet [141] reported in their publications for VEDAI and DOTA as they claim to achieve the most competitive results on such datasets. We observe that detection performance of our method on VEDAI and DOTA datasets is extremely good compared to [141] in terms of mAP. As shown in Table 3.2, the performance of our proposed scheme is 26.49% and 10.49% higher than [141] for VEDAI-VISIBLE and DOTA datasets respectively.

Again, compared with the detection performance of super-resolved images generated from the

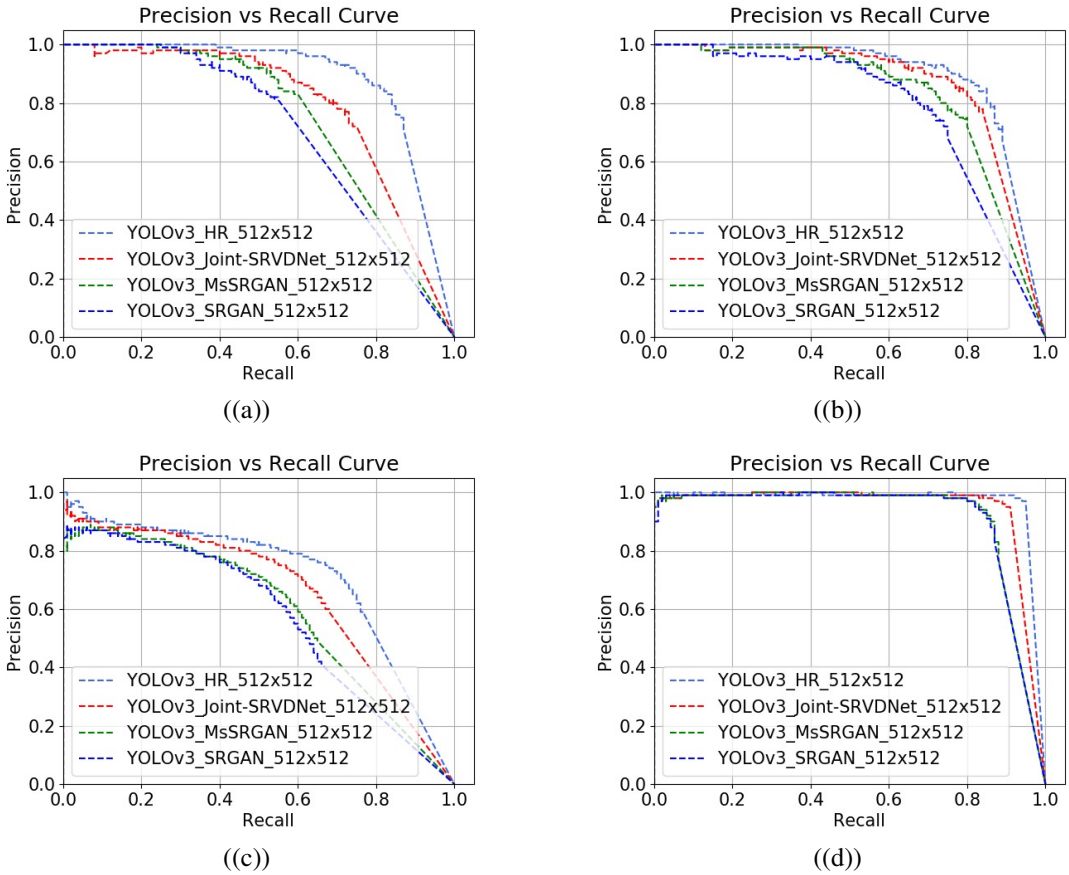



Figure 3.6: Precision-recall graph of the state-of-the-art object detector YOLOv3 performed on the original 512x512 high-resolution test images and the corresponding super-resolved images generated from SRGAN, MsSRGAN and our proposed Joint-SRVDNet over (a) VEDAI-VISIBLE, (b) VEDAI-IR, (c)xView and (d) DOTA.

existing most resent MsSRGAN based SR architecture, our method has achieved almost 5.75% higher mAP and 7% better F1 score for both VEDAI-VISIBLE and VEDAI-IR images. Moreover, for both dataset, we observe that the detection performance of our network (indicated by blue bold in Table 3.2) is also close to the optimal performance of the detector using original HR imagery, which is shown at the bottom row of Table 3.2. We also report mAP and F1 score for the xView satellite images which is very challenging as it contains extremely small targets in the image. Due to the low-resolution, targets do not contain detailed information which might help the detection task. As a result we cannot achieve satisfactory performance like other two datasets. However, still we have achieved 3.54% higher mAP and 2% better F1 score than the performance of super-resolved images from MsSRGAN and it is also close to the detection performance of the original

512x512 high resolution images. We also investigate our model's performance on DOTA dataset. During experiments, we notice a great improvement in detection performance for this dataset as shown in Fig. 3.6 (d) and fourth column of Table 3.2. The targets in this dataset seem to have the best appearance quality among two other datasets which has contributed to secure high detection performance. Therefore, we obtain promising results compared to [146] as well as for all the other algorithms. In addition, Fig. 6 helps to analysis the relationship between precision and recall rate for all datasets. It is obvious from the precision-recall plots that, our proposed method (YOLOv3_Joint-SRVDNet_512x512 in red curve) is significantly better than the other GAN based methods (YOLOv3_MsSRGAN_512x512 in green curve and YOLOv3_SRGAN_512x512 in blue curve) and more specifically, the performance gain is comparable to the detection performance of the original 512x512 high resolution images.

However, some important information might be missing if we only depend on precision-recall metric and F1 scores to determine the performance of our proposed method. For more robust analysis, we focus on plotting receiver operating characteristic curve (ROC) to study the characteristics of detection results. ROC curve can be drawn by plotting TPR against FPR at different thresholds. ROC curve reflects the relationship between TPR and FPR which may help to compare our method to other detection approaches.

## 3.7 Ablation Study

To achieve the best version of our proposed model, we made several experiments through changing the value of hyperparameters to see the impact of the hyperparameter changes on the original version of our work. We have summarized the analysis in Table 3.3 and 3.4.

Table 3.3: Super-resolution results of our proposed model using different hyperparameter settings for upscale factor 4x on the aerial test datasets. Magenta bold indicates the optimal SR results generated by our proposed network.

| Dataset | VEDAI-VISIBLE | | | | VEDAI-IR | | | | XVIEW | | | | DOTA | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Hyperparameter Settings | PSNR | MSSIM | UQI | VIF | PSNR | MSSIM | UQI | VIF | PSNR | MSSIM | UQI | VIF | PSNR | MSSIM | UQI | VIF |
| $\alpha = 2 \times 10^{-6}$, $\beta = 10^{-2}$, $\gamma = 10^{-2}$ | 27.060 | 0.812 | 0.745 | 0.690 | 26.513 | 0.720 | 0.780 | 0.697 | 17.856 | 0.529 | 0.523 | 0.436 | 24.327 | 0.845 | 0.813 | 0.457 |
| **$\alpha = 2 \times 10^{-6}$ , $\beta = 10^{-3}$, $\gamma = 10^{-3}$** | **30.338** | **0.969** | **0.995** | **0.693** | **29.227** | **0.958** | **0.999** | **0.713** | **20.550** | **0.617** | **0.795** | **0.562** | **31.360** | **0.987** | **0.975** | **0.712** |
| $\alpha = 2 \times 10^{-6}$ , $\beta = 10^{-2}$, $\gamma = 10^{-4}$ | 26.746 | 0.723 | 0.716 | 0.705 | 25.976 | 0.723 | 0.789 | 0.778 | 16.799 | 0.427 | 0.654 | 0.515 | 24.212 | 0.841 | 0.849 | 0.524 |

Table 3.4: Vehicle detection results in terms of mean average precision (mAP) and F1-score of our proposed model using different hyperparameter settings on the aerial test datasets. Cyan bold indicates the second optimal performance using SR images generated by our proposed network.

| Dataset | VEDAI-VISIBLE | | VEDAI-IR | | XVIEW | | DOTA | |
|---|---|---|---|---|---|---|---|---|
| Hyperparameters Settings | mAP@0.5 | F1 score | mAP@0.5 | F1 score | mAP@0.5 | F1 score | mAP@0.5 | F1 score |
| $\alpha = 2 \times 10^{-6}$, $\beta = 10^{-2}$, $\gamma = 10^{-2}$ | 68.89 | 0.678 | 77.78 | 0.756 | 59.61 | 0.556 | 88.59 | 0.778 |
| **$\alpha = 2 \times 10^{-6}$, $\beta = 10^{-3}$, $\gamma = 10^{-3}$** | **72.46** | **0.702** | **80.40** | **0.792** | **61.50** | **0.671** | **90.01** | **0.893** |
| $\alpha = 2 \times 10^{-6}$, $\beta = 10^{-2}$, $\gamma = 10^{-4}$ | 69.90 | 0.685 | 78.79 | 0.771 | 58.88 | 0.521 | 89.12 | 0.789 |

### 3.7.1 Hyperparameter analysis

We analyze the values of $\alpha$, $\beta$ and $\gamma$ adapted in (3.7) in order to obtain better quantitative results in aerial datasets. In (3.7), we have used $\alpha$, $\beta$ and $\gamma$ as weight factors to numerically balance the magnitude of different losses which accelerates the total loss convergence. The network can benefit from the relative influence of different loss functions, which is somehow guided by the weight factors. Since there is no rule of choosing the optimum parameters for the model, we conduct a series of experiments to find out the optimal parameters of the proposed model. We observe that the optimal values lead the training to generate real-looking images with full target details (edges, sharpness, perceptual features, etc.), that has been already reported in the experimental result analysis section. In Table 3.3 and 3.4, we show the average accuracy of our model varying these hyperparameters on several aerial test datasets.

Among the above settings, we report the results for the second setting (indicated by bold Magenta, Cyan) in Table 3.1, Table 3.2, and Fig 3.5, Fig 3.6, and Fig 3.7 as it provides the best results that is almost comparable to the original HR.

## 3.8 Conclusion

To address the challenge of detecting small targets (vehicles) in aerial images, we propose an approach that jointly optimizes super-resolution and detection modules. The purpose of our algorithm is to generate high quality super-resolved images from lower-resolution images, so that larger areas can be surveilled with minimal degradation in detection performance. With extensive experiments we demonstrated that our proposed joint network is able to learn and extract features from low-resolution domain which reflects in the generated super-resolved images produced by the network

and helps to improve detection performance. Most importantly, the proposed network has two vital contributions: for super-resolution task, using multi-scale GAN approach instead of classical SRGAN approach makes the detection task easier by adding more details in the super-resolved images which is essential to locate objects in the aerial images. Second, network's total loss integrates detection loss during super-resolution training which helps the SR module to specially learn the target area so that those specific area gets more obvious in the final super-resolution results. To evaluate our model's performance we conduct experiments on several publicly available datasets and the results indicate that compared with the leading state-of-the-art super-resolution and detection approaches, our proposed network achieves impressive results and it may have great impact on remote sensing community.

# Chapter 4

# Joint Cross-Modal Super-Resolution for Aerial Vehicle Detection

## 4.1 Introduction

Recently, vehicle detection in aerial imagery has become a well-studied problem in order to facilitate military and aerial surveillance based applications [141, 147]. Due to the proliferation of deep learning techniques, a significant development has been noticed in the aerial vehicle detection task. However, still the existing state-of-the art algorithms cannot achieve the expected high-level of accuracy. There are several major factors which should be taken into consideration during modeling the detection algorithm. Most of the aerial scenes comprises of a varieties of vehicle types with extremely small in size. Moreover, they have variable shapes, multiple scales, orientations, and with complex background which lead to interclass similarities between target and nontarget objects. We have already addressed this problem in our previous works [109, 148] by adapting Generative Adversarial Network (GAN) [11] based super-resolution technique, which aims to estimate a high-quality super-resolved versions of the low-resolution counterparts. Yet, the struggle remains when the detection takes place in a dark environment since most of the algorithms focus on the RGB-based detection due to the availability of the high resolution visible surveillance cameras. However, for low lighting or dark environment surveillance applications, these RGB images are not well suited. Therefore, infrared (IR) imaging becomes necessary which again necessitates modality transformation to alleviate the detection problem in such environment.

To address these challenges, we propose a joint cross-domain translation with super-resolution framework based on the GAN for vehicle detection in aerial images. The contribution of our work is three-fold: (1) We design a cross-modal & super-resolution network which jointly learns image-to-image modality transformation and super-resolution across two different domains. (2) We perform detection on these translated super-resolved images using one of the state-of-the-art detectors, You Only Look Once-version 3 (YOLOv3) which provides significant performance gain in vehicle detection for cross-domain framework. (3) In addition, we quantitatively prove that the domain adaptation technique along with the super-resolution can certainly meet the demand of training images for a detector.

The rest of the paper is organized as follows: Section 4.2 outlines the review of the previous works related to our proposed method. In Section 4.3, we discuss our proposed framework. We explain the loss-functions in Section 4.4, which have been optimized during training the network. Experimental details and results are discussed in Section 4.5 and 4.6. Finally, we provide the conclusion in Section 4.7.

## 4.2 Related Work

### 4.2.1 GAN Based Image Modality Transformation

In the last few years, GAN has been vigorously studied for modeling image-to-image translation problems. In [149], Isola et al. proposed the pix2pix algorithm; a GAN based synthesized image generation framework which investigated conditional adversarial network to learn a mapping from the input image (”source domain”) to the output image (”target domain”). Their approach effectively synthesized images from multiple domains. A Domain Transfer Network (DTN) is proposed by Taigman et al. [150] which employed a compound loss function including multiclass GAN loss, f-constancy and a regularizing component to learn a mapping from the source domain to the target domain in an unsupervised manner. For multi-task learning, Choi et al. [151] proposed the StarGAN, which utilized a mask-vector to translate images in multiple domains. The DiscoGAN [152] is another example of cross-domain learning approach, that successfully transfers style from one domain to another while preserving attributes. Liu et al. presented high-quality image translation

results in [153]. The authors hypothesized a shared-latent space between different domains and proposed an unsupervised image-to-image translation framework based on a coupled GAN.

### 4.2.2 Aerial Vehicle Detection in Super-Resolved Imagery

With the advent of deep learning techniques, vehicle detection algorithms have been significantly improved, specifically in aerial surveillance based applications. However, owing to the limitations in the resolution of surveillance cameras, it has become challenging to achieve satisfactory detection performance in aerial images. To overcome this challenge, Molina et al. [154] employed a super resolution algorithm to improve the vehicle type recognition performance. In their work, they used super-resolved images to train the neural network, which obtained an improved recognition performance. In [155], Ji et al. applied the GAN framework to perform simultaneous super-resolution and vehicle detection in an end-to-end manner, where the detection loss is back propagated into the super-resolution convolutional neural network during training to assist the detection task. The effects of super-resolution techniques on vehicle detection algorithm performance have been explored in the work of Shermeyer et al. [113]. In their work, the authors enhanced satellite images by an upscale factor of 2x, 4x, and 8x using the Very Deep Super-Resolution (VDSR) framework and a custom Random Forest Super-Resolution (RFSR) framework over five distinct resolutions to identify various types of vehicles. Inspired by the positive effects of super-resolution on vehicle detection, we proposed vehicle detection approach utilizing the GAN-based super-resolution architecture proposed in [109]. In that architecture, we used a Single Shot Multi Box Detector (SSD) for vehicle detection on super-resolved images, which achieved a significantly higher performance compared to that of the low-resolution images. Recently, we have proposed a Joint Super Resolution and Vehicle Detection Network (Joint-SRVDNet) [148] that tries to generate discriminative, high-resolution images of vehicles from low-resolution aerial images. The validity of the proposed method is ensured by showing extensive experiments on several aerial datasets, which has provided superior results over the other state-of-the-art methods for aerial super-resolution and vehicle detection.

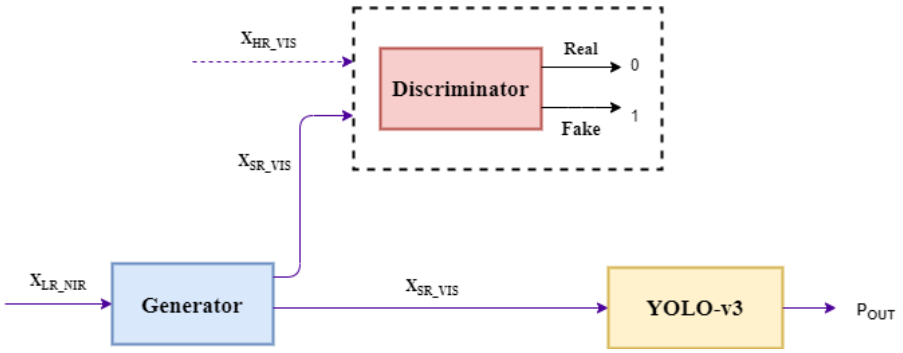



Figure 4.1: Our proposed network

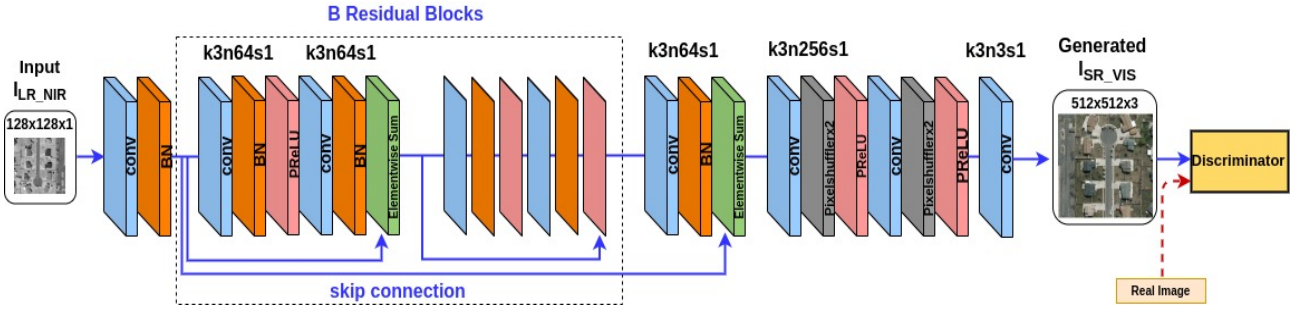



Figure 4.2: Architecture of the generator network with corresponding kernel size (k), number of feature maps (n) and stride (s) indicated for each convolutional layer.

## 4.3 Proposed Model

In this paper, we proposed a joint cross-modal and super-resolution approach for vehicle detection in aerial imagery shown in Fig. 4.1 which is based on the idea of conditional Generative Adversarial Network (cGAN). Hence, to facilitate the detection task in cross-domain, we exploit the benefit of domain transformation technique along with the super-resolution using a GAN-based architecture. More specifically, we enforce a Deep Convolutional Neural Network (DCNN) to jointly learn domain-specific features and a mapping between the LR images and the corresponding HR counterparts in a cross-domain framework, such that it can alleviate the domain mismatch and improve the performance of vehicle detection in aerial images.

### 4.3.1 Cross-Modal Super-Resolution Using GAN

GANs have shown remarkable performances in representation learning and synthesized image generation. They have been widely applied in image super-resolution (first introduced by Ledig et al. in [12]), image synthesis and image translation using the conditional GANs [13] and cyclic GANs (cycleGANs) [112]. As one of our objectives is to estimate a HR version with distinctive features of its LR input aerial image in a cross-domain framework, we utilize the cGAN architecture and train it to learn a mapping function $G : x \rightarrow y$ which is conditioned on a source domain x. The goal of G is to learn statistical distribution of the training data x which is in the LR IR domain for our case and generate the corresponding SR VIS counterparts such that their image distribution is indistinguishable from the image distribution of the target y that is in the HR VIS domain. Typically, the generator G is a differentiable function which is trained to learn the distribution $p_{data}$ over HR VIS data y. To do so, it takes LR IR input from the distribution $p_x(x)$ and maps it to the target data space as $G(x; \theta_g)$ where $\theta_g$ defines the parameters of the generator model. In addition, the discriminator D acts like a classifier which is trained to return the probability distributions D(y) and D(G(x)) for both training HR VIS examples from the distribution $p_{data}(y)$ and SR VIS samples from G(x), respectively. Basically, D is trained to maximize the probability of assigning the correct label to both the training examples and samples from G. Simultaneously G is trained to minimize log(1 – D(G(x))). In other words, D and G play the following two-player minimax game with the adversarial loss $l_{GAN}(G, D)$:

$$\min_G \max_D l_{GAN}(G, D) = \min_G \max_D [E_{y \sim p_{data}}[\log D(y)] + E_{x \sim p_x}[\log(1 - D(G(x)))]]. \quad (4.1)$$

Following the architecture by Ledig et al. [12], we design our basic deep generator network illustrated in Fig. 2 which consists of B(=16) serially connected residual blocks with identical layout. Each residual block uses two convolution layers of $3 \times 3$ kernel and 64 feature maps followed by batch-normalization layers [89] and Parametric ReLU [90] as the activation function. To increase the resolution of the input image, we employ two sub-pixel convolutional layers [91] in our generator network.

The discriminator network is a standard convolutional network as shown in Fig. 3 that can categorize the images fed to it as real or fake. It has been implemented following the architecture

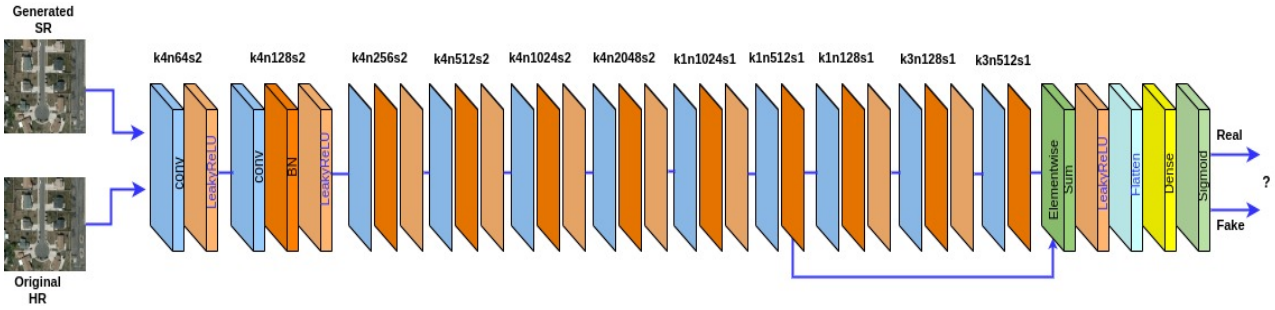



Figure 4.3: Architecture of the discriminator network with corresponding kernel size (k), number of feature maps (n) and stride (s) indicated for each convolutional layer.

summarized by Radford et al. [92]. To train discriminator, we use eight convolutional layers with an increasing number of 4x4 filter kernels increased by a factor of 2 from 64 to 512 kernels. We also add a dense layer as well as a sigmoid activation function at the end of the resulting 512 feature maps to obtain a probability for classification.

### 4.3.2 Vehicle Detection in Aerial Imagery Using YOLOv3

Our goal is to perform vehicle detection in the SR VIS aerial images which are generated from our proposed network using the LR IR input images. Generally, the aerial datasets contain vehicles of different sizes which require a strong detection algorithm to extract contextual and semantic information of those target objects. Hence, in this work, we use YOLOv3 the state-of-the-art object detection algorithm for aerial vehicle detection.

#### Architecture Details

The architecture of YOLOv3 shown in Fig. 4.4 utilizes the idea of residual network when employing the Darknet-53 convolutional network to extract features from the input image. It concatenates deeper layers with the earlier layers through up-sampling along with the additional convolutional layers to retrieve fine-grained feature information. As shown in Fig. 4.4, YOLOv3 takes an image and divides it into multiple grids, such as $16 \times 16$, $32 \times 32$ and $64 \times 64$ when input image is $512 \times 512$. Then classification and localization are applied at each grid size. If the center of any ground truth object falls within a grid cell, that grid cell is then considered to be responsible for

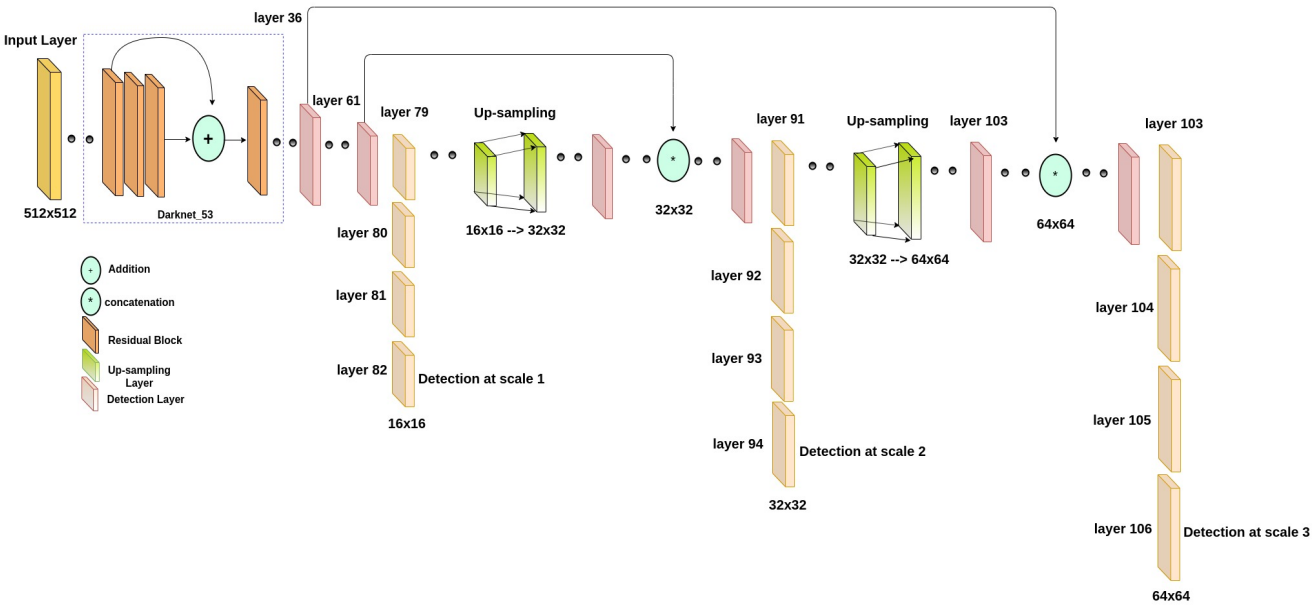



Figure 4.4: The YOLOv3 architecture for vehicle detection at three scales showing residual block, upsampling layer as feature extractor.

detecting object. For such grid cell, a number of bounding boxes with their confidence scores and their associated class probabilities are generated using a fully convolutional network architecture. The most salient feature of YOLOv3 is that it makes detections at three different scales and three different places in the network applying the feature pyramid network (FPN) [136] concept. It predicts objects at three different scales of 16, 32 and 64 for large, medium and small object detection. YOLOv3 uses 9 anchor boxes while predicting objects. Design of the anchor boxes greatly impacts the performance of the detector. We have used k-means clustering to generate these anchors for this database. The final number of detections by YOLOv3 is $M \times M \times (B * (4 + 1 + C))$. Here, $M \times M$ is the number of grid cells, B is predicted number of bounding boxes in a cell, 4 denotes the four coordinates of the bounding boxes and 1 is for the objectness score, C is the number of classes ( C=1 : 'vehicle' in our experiments). It uses multi-label classification. To compute objectness score, it replaces softmax with logistic regression. In addition, for each label binary cross-entropy loss is used instead of mean squarred error for the calculation of the classification loss.

## 4.4 Loss Functions

### 4.4.1 Pixel-Wise MSE Loss

State-of-the-art image SR methods [91, 116] mostly rely on pixel-wise MSE loss to optimize the network. For the training images $I_n^{HR_{VIS}}$ in the VIS domain with their corresponding low-resolution $I_n^{LR_{NIR}}$ in the NIR domain, $n = 1,....,N$, we can calculate the MSE loss $l_{MSE}$ using the following equation:

$$L_{MSE} = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{WH}\sum_{x=1}^{W}\sum_{y=1}^{H}((I_n^{HR_{VIS}})_{x,y} - G(I_n^{LR_{NIR}})_{x,y})^2, \tag{4.2}$$

where W and H represent width and height of the image and $G(I_n^{LR_{NIR}})$ are the super-resolved and translated images into the VIS domain for $n = 1,....,N$ training samples.

### 4.4.2 Perceptual Loss

Ledig et al. [12] proposed the perceptual loss, which is defined as the MSE loss over the high-level features extracted from the corresponding images. These features are extracted using a pretrained 19 layer VGG Network [137]. The perceptual loss can be considered as the L2 distance between the high-level feature representations of the synthesized super-resolved VIS image generated from the LR NIR input image and the ground truth HR VIS image. For $N$ training samples we solve:

$$L_{per} = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{C_j W_j H_j}\sum_{c=1}^{C_j}\sum_{x=1}^{W_j}\sum_{y=1}^{H_j}(\phi_j(I_n^{HR_{VIS}})_{c,x,y} - \phi_j(G(I_n^{LR_{NIR}}))_{c,x,y})^2, \tag{4.3}$$

where $\phi_j$ stands for feature maps of $j$-th convolutional layer and $C_j$, $W_j$ and $H_j$ define the dimensions of the respective feature maps within the VGG19 network.

### 4.4.3 Adversarial Loss

The adversarial loss described by (1) forces the solutions to move towards the natural image manifold by training the generator to fool the discriminator and thus generating photo-realistic images, and training the discriminator to accurately classify "real" images from the generated ones (i.e., fake images). The adversarial loss $L_{adv}$ defines the probability of the discriminator $D(G(I^{LR_{NIR}}))$ that the reconstructed image $G(I^{LR_{NIR}})$ is a real HR VIS image. Both the generator and discrimina-

tor as shown in Fig. 2 and 3, respectively, use the following adversarial loss functions to optimize the network.

$$L_{adv} = \min_{G} \max_{D} [E_{I^{HR_{VIS}} \sim P_{train}(I^{HR_{VIS}})}[\log D(I^{HR_{VIS}})] + E_{I^{LR_{NIR}} \sim P_G(I^{LR_{NIR}})}[\log(1 - D(G(I^{LR_{NIR}})))]], \quad (4.4)$$

where $P_{train}(I^{HR_{VIS}})$ and $P_G(I^{LR_{NIR}})$ define the probability distribution of the real high-resolution VIS images and the corresponding low-resolution NIR images, respectively.

### 4.4.4 Total Loss

We optimize equations (2), (3), and (4) to train our joint cross-modal super-resolution network. Therefore, the final loss function can be defined as follows:

$$L_{Total} = L_{adv} + L_{MSE} + L_{per}. \quad (4.5)$$

### 4.4.5 Detection loss

YOLOv3 optimizes three different losses: the localization, confidence and classification loss. Equation (2) defines this loss. $1_{ij}^{obj}$ means the object is detected by $j^{th}$ boundary box of grid cell i. $x_i$, $y_i$, $w_i$, $h_i$ are the real ground truth bounding box coordinates whereas $\hat{x}_i$, $\hat{y}_i$, $\hat{w}_i$, $\hat{h}_i$ are the predicted bounding box coordinates. $C_i$ is the box confidence score in cell i, $\hat{C}_i$ is the box confidence score for the predicted object:

$$\begin{aligned} L_{detection} = \lambda_{coord} \sum_{i=0}^{S^2} \sum_{j=0}^{B} 1_{ij}^{obj} (x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2 + \lambda_{coord} \sum_{i=0}^{S^2} \sum_{j=0}^{B} 1_{ij}^{obj} (\sqrt{w_i} - \sqrt{\hat{w}_i})^2 + (\sqrt{h_i} - \sqrt{\hat{h}_i})^2 \\ + \sum_{i=0}^{S^2} \sum_{j=0}^{B} 1_{ij}^{obj} l(C_i, \hat{C}_i) + \lambda_{noobj} \sum_{i=0}^{S^2} \sum_{j=0}^{B} 1_{ij}^{noobj} l(C_i, \hat{C}_i) + \sum_{i=0}^{S^2} 1_{i}^{obj} \sum_{c \in classes} l(p_i(c) - \hat{p}_i(c)). \end{aligned} \quad (4.6)$$

## 4.5 Experiments

We evaluate our proposed method on a publicly available Vehicle Detection in Aerial Imagery (VEDAI) dataset. In this section, we discuss the dataset and the network implementation strategies

in details.

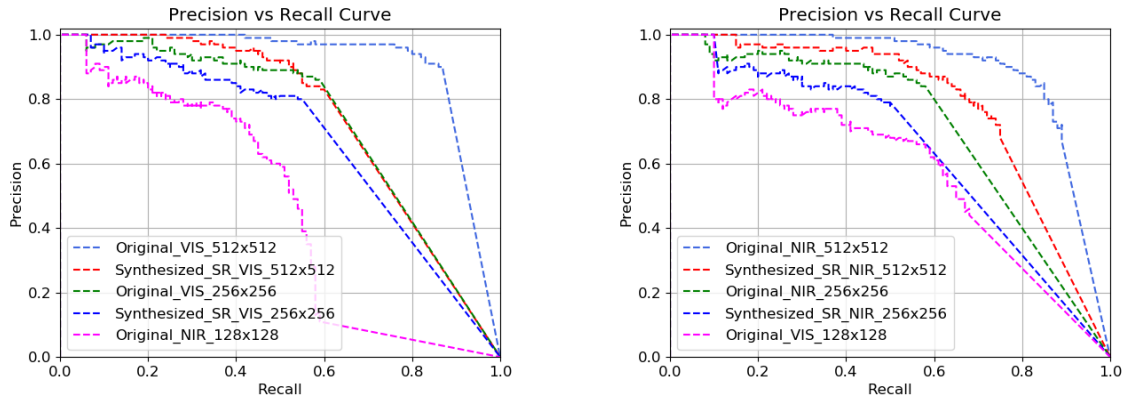
### 4.5.1 Vehicle Detection in Aerial Imagery (VEDAI) Dataset

The VEDAI dataset is a benchmark for small target recognition especially vehicle detection in aerial images. This dataset has around 1,210 images of two different resolutions such as $1,024 \times 1,024$ pixels and $512 \times 512$ pixels. The images contained in this dataset are captured under both near-infrared and visual light spectrum. The vehicles in these aerial images have extremely small size, diverse backgrounds, multiples shapes, variations in scale, lighting/shadowing changes, and specularities or occlusions. In addition, it includes nine different classes of vehicles, namely the plane, boat, camping car, car, pick-up, tractor, truck, van, and the other category. We consider all classes as a single class namely 'vehicle' for our task. For training and testing, we split the dataset into 1,100 and 166 images, respectively. To increase the size of the training data, we apply different augmentation techniques such as horizontal flipping and sharpening that make the model more robust to different features.

### 4.5.2 Training Details

In our experiments, we use both an upscale factor of 2x and 4x between low- and high-resolution images. To obtain LR images, the HR images are down sampled using bicubic kernel with a down sampling factor of 2 or 4. During implementation, we use an input image of size $128 \times 128$ either in the NIR or VIS domain and super-resolve it to $256 \times 256$ and $512 \times 512$, respectively in a cross-domain framework.

We set the network hyper-parameters as follows: For cross-modal super-resolution, we adopt the Adam optimizer with a momentum of 0.9 and a learning rate of $10^{-4}$. The model is trained for 10 iterations with a batch size of 4. For the detection model, we train the YOLO-v3 network for 10 epochs with an initial learning rate of $10^{-4}$ for the first 2 epochs and $10^{-6}$ for the last 8 epochs with batch size 6. We optimize the network by the Adam optimizer with a momentum of 0.9 and a weight decay of 0.9995. We fine-tune our detection model by setting IoU to 0.60, confidence threshold to 0.45 and NMS with the Jaccard coefficient of 0.45 to achieve the best performance. We implement the entire network using the tensorflow framework and train it over two NVIDIA




((a)) ((b))

Figure 4.5: Precision-recall graph of the state-of-the-art object detector YOLOv3 performed on the (a) HR VIS images of size 256 × 256, 512 × 512, and the corresponding SR VIS images generated from our proposed network for the case of scenario 1 over the VEDAI-VISIBLE dataset and (b) HR NIR images of size 256 × 256, 512 × 512, and the corresponding SR NIR images generated from our proposed network for the case of scenario 2 over the VEDAI-NIR dataset.

Titan XpGPU.

## 4.6 Performance Evaluation

In this section, we present comparative results for vehicle detection on the VEDAI dataset to evaluate the performance of our proposed joint cross-modal super-resolution model. We investigate vehicle detection performance of our network in terms of mean Average Precision (mAP). For more comprehensive performance analysis, we also provide precision recall curve. To prove the efficacy of our proposed method, we conduct several experiments considering three different scenarios. The scenarios are as follows:

### 4.6.1 Scenario 1

In this case, we assume that we have a LR NIR detector and two well trained HR VIS detectors to perform detection on the synthesized super-resolved images generated from our proposed network. Detection results can be summarized as follows.

- We apply the LR NIR detector on the LR NIR (128 × 128) images and plot precision-recall curve (see Fig. 4.5 (a)) to show the performance of this detector.

Table 4.1: Detection performance for scenario 1 in terms of mean average precision (mAP) of the proposed network.

| **Input (LR NIR)** | **mAP (LR NIR)** | **SR (2x) VIS** | **mAP (SR (2x) VIS)** | **mAP (Original VIS)** | **SR (4x) VIS** | **mAP (SR (4x) VIS)** | **mAP (Original VIS)** |
|---|---|---|---|---|---|---|---|
| 128 × 128 | 43.84% | 256 × 256 | 51.81% | 55.33% | 512 × 512 | 61.31% | 85.33% |

Table 4.2: Detection performance for scenario 2 in terms of mean average precision (mAP) of the proposed network.

| **Input (LR VIS)** | **mAP (LR VIS)** | **SR (2x) NIR** | **mAP (SR (2x) NIR)** | **mAP (Original NIR)** | **SR (4x) NIR** | **mAP (SR (4x) NIR)** | **mAP (Original NIR)** |
|---|---|---|---|---|---|---|---|
| 128 × 128 | 44.18% | 256 × 256 | 49.24% | 54.39% | 512 × 512 | 69.45% | 85.66% |

- To evaluate the efficacy of our proposed method, we perform detection using the HR VIS detectors on synthesized SR VIS images of size 256 × 256 and 512 × 512, respectively generated from our network which has been trained with the NIR input images of size 128 × 128. We plot the corresponding precision recall curves in Fig. 4.5. From this figure, we observe that at a given recall, the precision values of the HR VIS detectors tested on the SR VIS images of size 256 × 256 (blue curve in Fig. 4.5(a)) and 512 × 512 (red curve in Fig. 4.5(a)) are much higher than the LR NIR detector that is trained with the images of comparatively lower resolution (magenta curve in Fig. 5(a)) and also comparable to the original ones.

To evaluate the performance of our detection model we use mAP as the evaluation metric which measures mean of the average precision. Table 4.1 summarizes the mAP results of our experiments from where we can see that our proposed network has achieved 7.97% better precision for the SR VIS images of size 256 × 256 than the LR NIR 128 × 128 images. Moreover, the precision is increased by a significant margin of 17.47% for the SR VIS images of size 512 × 512 compared to the LR NIR input images.

### 4.6.2 Scenario 2

We have conducted experiments for the reverse scenario, where we train our proposed network with the VIS input images of size 128 × 128 and super-resolve them into the NIR domain by an upscale factor of 2x and 4x. We generate the corresponding SR NIR images of size 256 × 256 and

512 × 512, respectively. In this case, we have a LR VIS detector and two well-trained HR NIR detectors. Therefore, we perform detection on these synthesized super-resolved images using the HR NIR detectors to varify the adequacy of our proposed algorithm. The results are discussed as follows:

- We perform detection on the LR VIS images of size 128 × 128 using the trained LR VIS detector and plot the corresponding precision recall curve in Fig. 4.5(b).
- Similar to the previous case, we also invesitigate the detection performance of the trained HR NIR detectors on the synthesized SR NIR images of two different sizes 256 × 256 and 512 × 512. The results in Fig. 4.5(b) shows the robustness of the proposed method by providing better results even for the synthesized NIR images.

Table 4.2 briefly summarizes the mAP results for scenario 2 where we can see that when we use the HR NIR detector on the SR NIR images of size 256 × 256, it achieved 5.06% better precision than the LR VIS counterparts. Moreover, for the SR NIR images of size 512×512, the network achieved significant improvement in precision which is 25.27% higher than the LR NIR counterparts.
From the above-mentioned quantitative results for the detection tasks, we observe that the NIR detector is performing better than the VIS detector specifically in the case of high resolution. The mAP result for the detection on the SR NIR images of size 512 × 512 is 8.14% higher than the SR VIS versions. These findings focus on the encouragement of using the NIR detector in dark environment instead of using the RGB detector which might be a great help to the military and surveillance based applications.

### 4.6.3 Scenario 3

In Case 3, we conduct experiment using a LR VIS detector as well as a well trained HR VIS detector to perform detection on the synthesized images generated from our proposed network. The difference between this case and the previous ones is that, we don't perform super-resolution. We only perform detection on the synthesized images.

- To see the effect of the domain transformation technique on detection results, we apply the LR VIS detector on the synthesized LR VIS (128×128) images from the LR NIR (128×128)

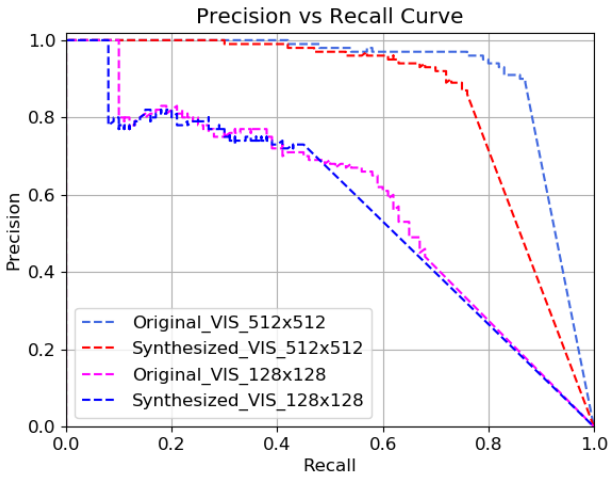



Figure 4.6: Precision-recall graph of the state-of-the-art object detector YOLOv3 performed on the HR VIS images of size 128 × 128, 512 × 512, and the corresponding synthesized VIS images generated from our proposed network for the case of scenario 3 over the VEDAI dataset.

Table 4.3: Detection performance when using the LR NIR input for case 3 in terms of mean average precision (mAp) of the proposed network.

| **Input (LR NIR)** | **Synthesized VIS** | **mAP (Synthesized VIS)** | **mAP (Original VIS)** |
|---|---|---|---|
| 128 × 128 | 128 × 128 | 39.24% | 44.18% |

Table 4.4: Detection performance when using the HR NIR input for case 3 in terms of mean average precision (mAp) of the proposed network.

| **Input (HR NIR)** | **Synthesized VIS** | **mAP (Synthesized VIS)** | **mAP (Original VIS)** |
|---|---|---|---|
| 512 × 512 | 512 × 512 | 74.19% | 85.33% |

images. The results are summarized in Table 4.3 and Fig. 4.6. In this case, the network achieved 39.24% average precision which is close to the performance of the original VIS images.

- In addition, we also utilize the HR VIS detector to perform detection on the synthesized HR VIS (512 × 512) images generated from the HR NIR (512 × 512) input images. From the results shown in Fig. 4.6 and mAP scores in Table 4.4, it is clearly seen that the proposed network obtained 74.19% average precision for the synthesized images that can be considered as good as to the original one. From the experimental results, we can say that for both scales our proposed network ensures the significant effect of modality transformation on the detection task.

## 4.7 Conclusion

In this work, to address the challenge of RGB-IR vehicle detection problem in aerial images, we propose an approach that jointly optimizes a unified network for both domain transformation and super-resolution technique. The purpose of our algorithm is to generate high quality super-resolved images in cross-domain from the lower-resolution counterparts, so that the larger areas can be surveilled with minimal degradation in detection performance. To extend our research work, currently, we are trying to test our proposed framework on other aerial datasets. We are also trying to optimize our network for different spectral bands such as long-wave to mid-wave to investigate the detection performance.

# Chapter 5

# Deep GAN-Based Cross-Spectral Cross-Resolution Iris Recognition

## 5.1 Introduction

Identity verification based on the analysis of a person's physiological properties is believed to be more reliable than other means of identification such as a PIN or password, username, or access card. Fingerprint, palmprint, face, and iris biometric modalities have seen widespread use for human identification [156–159]. Among all biological traits, iris is well suited for the most accurate and secure personal authentication because of the distinctive patterns present in the iris textures for individuals [159, 160]. The human iris pattern is observed to have unique and different textures due to the process of chaotic morphogenesis that causes its formation in early childhood, exhibiting variation even among identical twins. Therefore, in recent decades, iris recognition has received significant attention as an identity verification method in the biometric community [161, 162] for civilian and surveillance applications.

Conventional iris recognition biometric systems are based on iris images obtained under near-infrared (NIR) illumination due to the optical properties of the human iris in the NIR band of the electromagnetic spectrum. Broadly speaking, the NIR light absorption characteristics of the pigment melanin within the iris tissue determines the visibility of iris texture details in NIR imaging. As a result, the iris textures appear much better under illumination in the $700 \sim 900$ nm wavelength range compared to illumination within the visible wavelengths in the $400 \sim 700$ nm range.

For this reason, in most commercial iris recognition systems, single-band near-infrared (NIR) iris matching techniques have been extensively used for identity authentication tasks [161, 163]. These systems use well-established algorithms and protocols to perform identification when the probe and gallery are in the same domain, which has resulted in highly-accurate performance. However, the majority of these methods require close-distance iris imaging to ensure that the acquired images are in good quality with minimum acceptable iris diameter [159]. To eliminate these constraints in the NIR-based iris recognition, several visible wavelength based iris recognition systems have been developed [164, 165] in the last few years, which has expanded the scope of investigating the capabilities of the iris matching techniques under visible light illumination. In addition, several competitions such as the Noisy Iris Challenge Evaluation (NICE) [166], and the Mobile Iris Challenge Evaluation [167] focus on the realistic acquisition process of visible iris images. The major factors behind this attention to visible wavelength-based iris recognition are (1) visible range cameras are capable of acquiring images from long distance, and (2) they are low-cost compared to NIR cameras. Emerging dual imaging technology in recent smartphones offer image capture in the visible and NIR illuminations. As a result, now police and every law enforcement officer, customs and border protection officer, and special operator has an agency-issued cellphone to perform multi-modal biometric captures (face, fingerprint, and iris), which are used later for authentication. In this context, effective usage of this opportunistic visible iris images requires accurate iris matching with the corresponding NIR images enrolled in the national ID databases.

Moreover, recent advances in video surveillance technology have enabled the capture of very high-resolution iris images in the visible spectrum using low-cost camera technologies, which can be used for identification purposes within the same domain or across different spectra. However, most large-scale galleries of iris images have been acquired in the lower resolution near-infrared (NIR) domain. Therefore, cross-spectral and cross-resolution iris matching has emerged as a major challenge [168, 169]. It is essential to address both cross-spectral as well as cross-resolution methods for matching these opportunistic visible iris query probes against the enrolled NIR iris images in a gallery.

In the last few years, deep neural network architectures, such as a convolutional neural network (CNN), have dramatically improved the capabilities in automatically learning the deep representation of specific image features for object detection and classification of visual patterns. These

Table 5.1: A summary of the recent related works on cross-spectral iris recognition, accuracy is reported at a given False Acceptance Rate (FAR)

| Reference | Method | Database | Iris comparison | Iris matching | Features | Matching accuracy |
|---|---|---|---|---|---|---|
| [170] | A predictive NIR iris image is used from the color image | WVU Multi-spectral iris database | Genuine = 280<br>Impostor = 20,745 | Cross-spectral | Hand-crafted | 95.2%<br>(FAR = 0.001) |
| [171] | IrisCode using 1D Log-Gabor filter | (1) PolyU bi-spectral iris database<br>(2) Cross-eyed-cross-spectral iris recognition database | Genuine = 2800<br>Impostor = 1,953,000<br>Genuine = 2160<br>Impostors = 516,240 | Cross-spectral | Hand-crafted | (1) 52.6%<br>(FAR = 0.1)<br>(2) 70.3%<br>(FAR = 0.1) |
| [169] | NIR to VIS texture synthesis using MRF model | (1) PolyU bi-spectral iris database<br>(2) Cross-eyed-cross-spectral iris recognition database | Genuine = 2800<br>Impostor = 1,953,000<br>Genuine = 2160<br>Impostors = 516,240 | Cross-spectral | Hand-crafted | (1) 64.91%<br>(FAR = 0.1)<br>(2) 78.13%<br>(FAR = 0.1) |
| [168] | CNN with softmax cross-entropy loss for feature extraction and SDH for compression and classification | (1) PolyU bi-spectral iris database<br>(2) Cross-eyed-cross-spectral iris recognition database | Genuine = 2800<br>Impostor = 1,953,000<br>Genuine = 2160<br>Impostors = 516,240 | Cross-spectral | self-learned | (1) 90.71%<br>(FAR = 0.01)<br>(2) 87.18%<br>(FAR = 0.01) |
| cpGAN [1] | Conditional coupled generative adversarial network (cpGAN) | (1) PolyU bi-spectral iris database | Genuine = 2800<br>Impostor = 1,953,000 | Cross-spectral and cross-resolution | self-learned in the embedded domain | (1) **92.38%**<br>(FAR=0.01) |

algorithms have also shown superior results when compared to classical techniques based on hand-crafted features. Recently, successful deployment of deep learning architectures for the task of the same or cross-domain iris recognition has gained attention in the literature. Generative adversarial networks (GANs) [11], among other deep neural network architectures, have shown outstanding capabilities in different areas of computer vision and biometric applications [18, 35, 92, 172–179]. A range of applications of GANs for iris recognition has been presented, including data augmentation, synthesis of NIR periocular images, synthesizing iris images and iris super-resolution [180, 181]. In this paper, our main contribution is the extensive application of our novel algorithms on three publicly available iris datasets comparing two different GAN-based frameworks for cross-spectral (VIS vs NIR) and cross-resolution (low-resolution (LR) NIR to high-resolution (HR) VIS) iris matching, which resulted in a new state-of-the-art approach in the area of ocular biometrics.

We have developed two approaches by which we apply a family of deep learning frameworks for different cross-spectral iris matching scenarios. In our first approach, we employ a conditional GAN (cGAN) [149] architecture to map the cross-spectral data to the same spectral domain. We apply it at the same resolution and extend it to the cross-resolution iris matching problem. We have designed our first method based on a scenario when one already has access to an Open-source or an off-the-shelf iris matcher (e.g., Open-Source OSIRIS [182] matcher) to conduct the iris veri-

fication process. The key idea in our first method is to synthesize the VIS iris images from their corresponding NIR iris images in a gallery at the same resolution or higher resolution through a joint cross-modal super-resolution process. Our first method is assumed to be a preprocessing module that translates a NIR image into its corresponding VIS iris image before using a commercial iris matcher. In our work, we have used OSIRIS software to conduct the matching between the synthesized VIS iris images from a gallery of NIR iris images and a probe VIS iris image. In a summary, our first approach offers four contributions to the field of iris recognition:

• A new domain adaptation framework, which acts as a preprocessing module for cross-spectral iris matching based on generative adversarial networks to transform the cross-domain problem to the same domain and achieves comparable performance when compared to several state-of-the-art methods.

• Integrating the cross-resolution matching scenario into the cross-spectrum setting and redefining the matching framework as a joint super-resolution and cross-spectral matching architecture.

• Introducing a new WVU face and iris dataset, which will contribute to the biometric field for cross-spectral face and iris recognition.

• Performing substantial experiments on the PolyU Bi-Spectral dataset [168, 169], WVU face and iris dataset and cross-eyed-cross-spectral iris recognition database [183].

We observe a significant improvement in the cross-spectral iris matching accuracy from the experimental results of our first approach, which validates that our domain adaptation technique requiring self-learned features extracted from the raw data can achieve remarkable performance gains for iris verification tasks similar to the previous research presented in the literature. However, it is still essential to explore a more compressed and distinctive representation of the raw data. In earlier works, researchers have widely used subspace learning for data representation [184–186]. Basically, it has been proven that the most relevant and useful inner characteristics of an image can be mapped to a reduced low-dimensional latent subspace.

Motivated by this, in our second method, we focus on the idea of designing a dedicated cross-spectral iris matcher completely avoiding the use of any off-the-shelf iris matcher. We hypothesize that iris images in the VIS domain possess a latent connection with iris images in the NIR domain in a latent feature subspace. Therefore, we exploited this latent connection by projecting the VIS

and NIR iris images into a common latent embedding subspace, even if they are at different resolutions. Furthermore, we posit that, if we perform matching in this latent domain, verification performance would be more accurate than our first method due to the inherent direct iris matching in a shared common feature domain. Moreover, our second method is designed to circumvent several shortcomings of the commercial iris matchers that our first method depends on. The idea can be elaborated on with a use of a case scenario. For example, the matching engine of a commercial iris matcher cannot be adjusted to the resolution of the iris images captured at a distance. It cannot operate directly on the lower resolution images while enrolled images in the gallery, are comparatively, at a higher resolution. In addition, effective usage of opportunistic high-resolution VIS iris images captured by smartphones, surveillance cameras, etc. requires an accurate, fast, stable, and secure iris matcher. This can be achieved with the real-valued feature representation in the common embedded latent subspace instead of a binarized feature representation like the one used in other classical iris recognition approaches in the literature. These underlying reasons motivated us to develop such a dedicated cross-spectral iris matcher, which is highly desirable. Hence, we have proposed a deep coupled learning framework for cross-spectral iris recognition, which utilizes a conditional coupled generative adversarial network (cpGAN) [1] to learn a common embedded feature vector via exploring the correlation between the NIR and VIS iris images in a reduced dimensional latent embedding feature subspace. The key benefits from our second iris recognition approach can be summarized as the following:

• We analyze an effective method to learn the subspace embedded features and develop a novel framework for cross-spectral iris matching using our cpGAN architecture.

• Comprehensive experiments on three different benchmark datasets (1) PolyU Bi-Spectral dataset (2) WVU face and Iris dataset and (3) Cross-eyed-cross-spectral database with superior results over the baseline approaches ascertain the validity of our cpGAN framework.

• To the best of our knowledge, this is the first study that has investigated two different techniques utilizing the potential capabilities of a GAN to improve the performance of existing cross-spectral iris recognition methods reported in the literature.

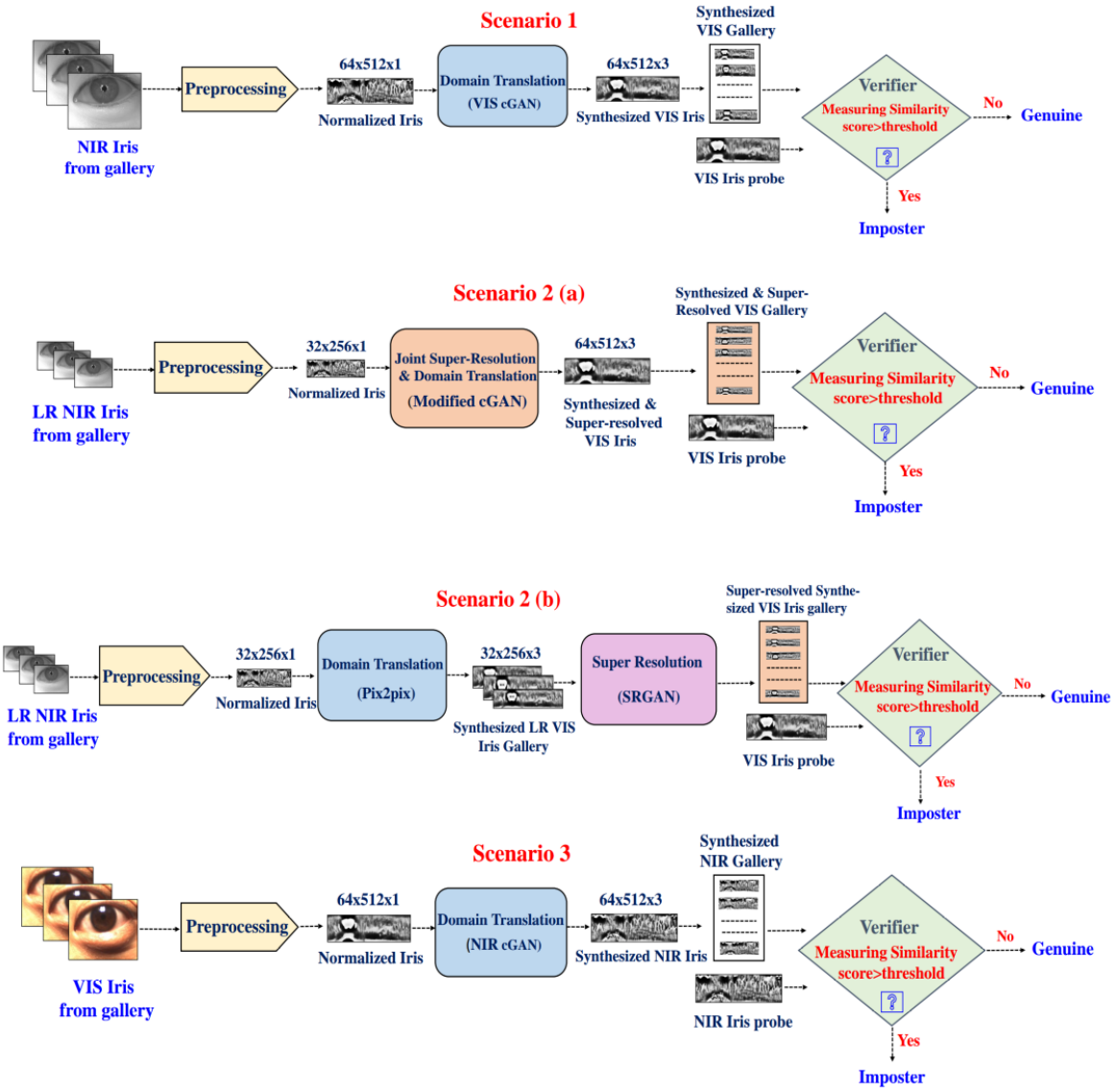



Figure 5.1: Cross-domain and cross-resolution iris recognition framework; Scenario 1: NIR to visible translation; Scenario 2: NIR to visible joint/separate translation and Super-resolution , Scenario 3: Visible to NIR translation.

## 5.2 Related Work

Cross-spectral iris recognition requires a VIS iris probe to be matched against a gallery of NIR iris images. While conventional iris recognition methods have achieved high matching accuracy, cross-spectral iris matching algorithms have not yet reached a high level of performance and pose a greater challenge for real-world applications.

Table 1 summarizes recent cross-spectral iris algorithms that are based on two strategies 1) extracting information from both spectral domains and then combining such information for the final decision, or 2) synthesizing a NIR image from its corresponding visible image and then matching against a NIR gallery. Using the first strategy, Vyas and Kanumuri [187] proposed a new feature descriptor using template partitioning based on variations in the iris texture. In their work, they have applied a 2D Gabor filter bank to obtain the iris pattern at various scales and orientations. They utilize the difference of variance (DoV) features to divide the filtered iris template into sub-blocks, as the DoV features are invariant to noise caused by illumination occlusion and position shifting. However, this method could not achieve the high accuracy required for practical applications (high equal error rate (EER) of 31.08%) because it is unable to relate the information comprised in the NIR and VIS images. Tan et al. [188] describe a framework for segmenting iris images in both domains which is helpful for further multi-spectral fusion of information. According to Oktiana et al. [189] local binary pattern (LBP) and binary statistical image feature (BSIF) are the best feature descriptors based on the VIS and NIR imaging systems, which are able to accurately extract the texture patterns of the iris for cross-spectral matching.

Another recent work [190] also used BSIF along with the $\chi^2$ distance metric to obtain match scores between the VIS probe and NIR reference ocular images. They then fuse all the scores to make the final decision. To encourage advances in cross-spectral iris and periocular recognition, there has been a competition [191] held among five teams, which is considered as an extension of 1st competition that was arranged for a similar task (*more recently, Sequeira et al. [183] released a new cross-eyed and cross-spectral iris dataset to advance research on the challenging cross-spectral iris matching problem*). They submitted twelve methods for the periocular task and five for the iris task. In the work of Alonso-Fernandez et al. [192] fusion of periocular and iris information achieved considerable match performance improvement, where iris information is obtained

by using different iris features extraction techniques. Wild et al. [193] used information from iris images captured at multiple bands of the electromagnetic spectrum and presents an efficient feature-level fusion to improve cross-spectral iris recognition performance. Sharma et al. [194] proposed an algorithm, that consists of two neural network architectures, and trained it on a cross spectral periocular dataset. It resulted in an improved matching accuracy compared to the existing feature descriptors previously mentioned above.

On the other hand, using the second strategy, several efforts toward estimating NIR images from visible images have been proposed recently. For instance, researchers in [170] have explored an adaptive learning method to predict NIR images to address the performance shortcomings, which was considered below the benchmarks caused by cross-spectral matching. Similarly, in [169], authors develop a domain adaption framework using Markov random fields (MRF) to estimate a NIR iris image from its corresponding VIS iris image and perform matching against a NIR image gallery. In the same direction, Burge and Monaco [195, 196] implemented a model which utilized features derived from the color and structure of the VIS iris images to predict the corresponding synthesized NIR iris images. We have also noticed similar works in the ocular biometric field for the task of cross-spectral periocular image recognition. Recently, Reja et al. [197] proposed a novel image transformation technique using cascaded refinement networks to synthesize a NIR periocular image from the corresponding VIS periocular image. Another study [198] reported that feature-based approaches are prone to changes during the feature extraction process. Therefore, they have adopted phase-only correlation and band-limited phase-only correlation techniques to develop a phase-based iris recognition system.

Although the approaches mentioned above have advanced cross-spectral iris matching one step ahead by achieving good results, but to keep pace with the increasing demand for more robust biometric systems, researchers have recently concentrated their efforts towards CNN-based iris verification system [168]. In this study, the authors observed that CNN-based features offer a significantly compact representation for the iris template along with sparse information, which potentially helps to improve the accuracy of the iris recognition system. Moreover, this approach incorporates a supervised discrete hashing (SDH) on the learned features, which achieved an EER of 5.39%.

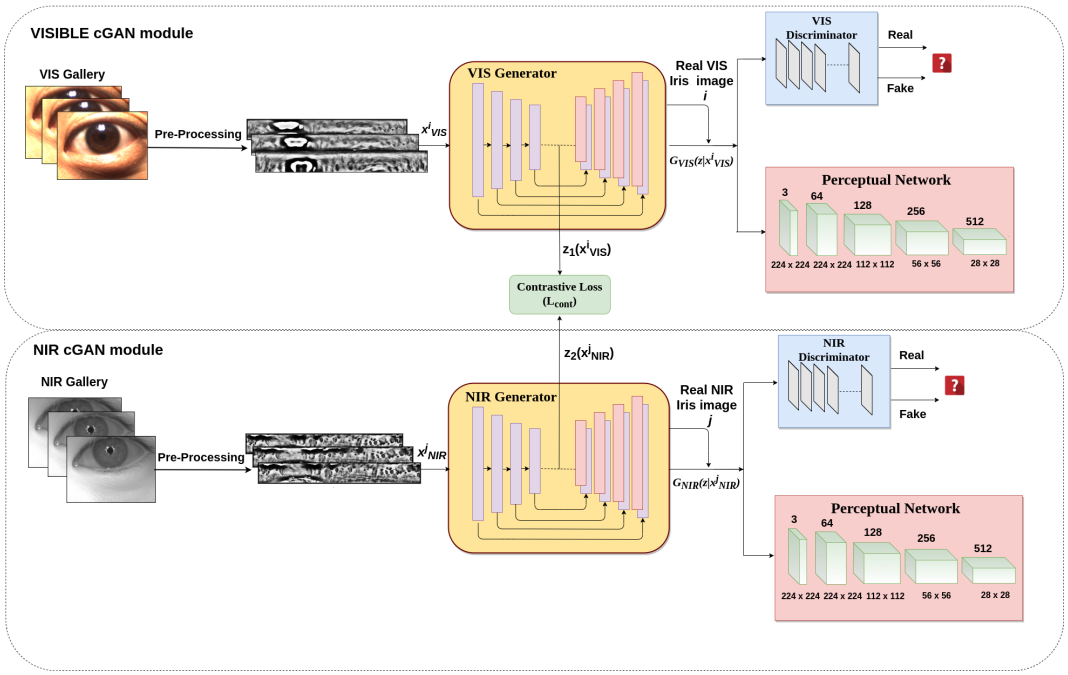



Figure 5.2: Architecture of our proposed conditional cpGAN framework. During training, the contrastive loss function is used in the latent embedding subspace to optimize the network parameters so that latent features of iris images from different spectral domain of the same identity are close to each other while the features of different identities are pushed further apart [1].

Another interesting approach, iris image super-resolution, has also gained attention due to its impact on iris verification methods. The authors in [199] explored deep learning architectures such as stacked auto-encoders and CNN for single-image iris super-resolution. Wang et al. [200] proposed a framework based on an adversarial training with triplet networks in order to improve iris image resolution for further recognition.

## 5.3 Methodology

To address the performance degradation reported in cross-spectral iris matching, our primary goal is to develop an algorithm that minimizes the distance between the VIS iris image and NIR iris image distributions belonging to the same person. Therefore, we have developed two different domain adaptation techniques. These two techniques are based on finding a mapping, or a low-dimensional shared latent subspace, between the VIS and NIR iris modalities to significantly reduce the cross-spectral iris matching discrepancy and provide a new state-of-the-art result. The

techniques developed in this paper are shown in Fig. 5.1 and Fig. 5.2, respectively. Since most of the available iris galleries are acquired under NIR illumination and the opportunistic iris images are obtained under the VIS domain at higher resolution, in our first technique we find a mapping between the NIR and VIS iris images and design a new framework (see Fig. 5.1) based on joint cross-domain and cross-resolution matching to enable cross-spectrum iris recognition for pairs of images with the same and different resolutions. In greater detail, we address two challenges: 1) performing cross-domain mapping for the purpose of the intra-domain iris matching, and 2) doing the same when the images from each spectral domain have different resolutions. As a result, there should be a joint transformation of spectrum and resolution, which will be discussed in detail in subsections 5.3.1.

We consider three main scenarios, as shown in Fig. 5.1, and develop our iris matching framework around them based on different cross-spectral scenarios. The first scenario is matching a visible probe against an NIR gallery translated to an equivalent visible gallery. The second scenario is matching a high-resolution visible probe against a translated and super-resolved NIR gallery to a high-resolution visible gallery. The third scenario is matching an NIR probe against a visible gallery translated to an equivalent NIR gallery. The reason behind including the cross-resolution setting in our framework, is that, as an emerging problem domain, current opportunistic visible iris images extracted from high-resolution face images are typically at a higher resolution than the NIR images.

Apart from being focused solely on the generation of a synthesized VIS image from its NIR counterpart, in our second approach, we emphasize the idea of learning a latent subspace to extract meaningful representative features from the VIS and NIR iris images. Thus, we develop our second approach as shown in Fig. 5.2, which projects both the NIR and VIS iris images to a common latent low-dimensional embedding subspace using two generative networks. The key reason behind developing this architecture is to learn the semantic similarity between two samples of the same subject but in different spectral domains. Therefore, inspired by our previous cpGAN architecture [1], we trained this network using a similarity measure based on a contrastive loss [201] to ensure that the distance between the images corresponding to the genuine pairs (VIS iris image and NIR iris image of the same person) is minimized, and that of the imposter pairs (VIS iris image and NIR iris image of the different persons) is maximized.

To summarize our two approaches, we have studied and developed two different deep convolutional GAN-based architectures to ascertain the adaptive learning potential for cross-spectral iris matching, i.e., cGAN and cpGAN. Like other approaches, before training both networks, preprocessing steps require iris images from both spectra to be subjected to segmentation, normalization and image enhancement processes. Then, normalized image samples are fed to each network. The network in our first approach mainly utilizes the adversarial loss to synthesize VIS iris image from its NIR counterpart before performing different verification scenarios, while in our second approach, the network integrates the contrastive loss along with the adversarial learning [11] to generate matching scores. The following sections provide the details of our approaches and introduce the cGAN and cpGAN architectures along with the associated loss functions that are implemented in our framework to investigate the cross-spectral iris matching problem.

### 5.3.1 Deep Conditional Adversarial Framework

Recently, GANs have received considerable attention from the deep learning research community due to their significant contributions in the field of image generation. The basic GAN framework consists of two modules– a generator module, G, and a discriminator module, D. The objective of the generator, G, is to learn a mapping, $G : z \rightarrow y$, so that it can produce synthesized samples from a noise variable, z, with a prior noise distribution, $p_z(z)$, which is difficult for the discriminator, D, to distinguish from the real data distribution, $p_{data}$, over y. The generator, $G(z; \theta_g)$ is a differentiable function which is trained with parameters $\theta_g$ when mapping the noise variable, z, to the actual data space, y. Simultaneously, the discriminator, D, is trained as a binary classifier with parameters $\theta_d$ such that it can distinguish the real samples, y, from the fake ones, G(z). Both the generator and discriminator networks compete with each other in a two-player minimax game. We calculate the following loss function, L(D, G), for the GAN:

$$L(D, G) = E_{y \sim P_{data}(y)}[\log D(y)] + E_{z \sim P_z(z)}[\log(1 - D(G(z)))]. \tag{5.1}$$

The objective function of GAN defines the term "two-player minimax game" by optimizing

the loss function, L(D, G), as follows:

$$\min_{G}\max_{D} L(D,G) = \min_{G}\max_{D}[E_{y\sim P_{data}(y)}[\log D(y)] + E_{z\sim P_z(z)}[\log(1 - D(G(z)))]]. \tag{5.2}$$

One of the variants of GAN, the cGAN is introduced in [149], which expands the scope of synthesized image generation by setting a condition for both the generative and discriminative networks. The cGAN applies an auxiliary variable, x, as a condition which could be any kind of useful information such as texts [202], images [13] or discrete labels [149]. The loss function for the cGAN, $L_c(D, G)$, can be represented as follows:

$$L_c(D,G) = E_{y\sim P_{data}(y)}[\log D(y|x)] + E_{z\sim P_z(z)}[\log(1 - D(G(z|x)))]. \tag{5.3}$$

Similar to (2), the objective function of the cGAN is minimized in a two-player minimax manner, which is denoted as $L_{cGAN}(D, G, y, x)$ and defined by:

$$L_{cGAN}(D,G,y,x) = \min_{G}\max_{D}[E_{y\sim P_{data}(y)}[\log D(y|x)] + E_{z\sim P_z(z)}[\log(1 - D(G(z|x)))]]. \tag{5.4}$$

**Domain Translation Using cGAN**

A more recent algorithm in the field of ocular biometrics [169,203] has shown success in estimating NIR iris images from VIS iris images and then matching them against the NIR instances in the gallery. However, they did not use CNN-based algorithms, even though many of the recent iris recognition systems have investigated the capabilities of CNN in learning anatomical properties. Therefore, we have developed a deep CNN-based domain translation network in our first method. We proposed to translate the iris images from the NIR domain to visible, or vice versa. Therefore, image translation plays an important role as one of two integral parts of our frameworks.

Recent advances in deep learning reported in the literature have provided very powerful tools for the task of image-to-image translation [13]. Such translations can be interpreted as image domain transformations, where the task is to learn a mapping from one modality to another modality. In our first method, we use the conditional GAN (cGAN) architecture [13] for the task of NIR to VIS iris image translation or vice versa. The cGAN architecture has been successful in a va-

riety of image-to-image translation tasks in the computer vision research community. It includes Sketch → Portrait, Sketch→ Pokemon, Depth→ Streetview, pose transfer, etc. Such deployment of cGAN in image translation tasks has inspired us to explore its performance in synthesizing corresponding VIS iris images from the NIR iris gallery, to be used as a preprocessing module for the cross-spectral iris image translation.

During training the cGAN, we condition on an NIR iris image and generate a corresponding synthesized VIS iris output image or vice versa as shown in our proposed framework (see Scenario 1 and Scenario 3 in Fig. 5.1). Here, we have demonstrated that a simplified cGAN framework is sufficient to achieve adequate synthesized results through adversarial learning. In addition, our analysis shows that this method is effective at conducting cross-spectral iris matching under the same spectrum setting (a VIS iris probe is matched against a synthesized VIS gallery generated from its corresponding NIR gallery or vice versa.) with impressive results.

**Joint Translation and Super-Resolution Using Modified cGAN**

Leveraging the benefits of the cGAN architecture, we have investigated the possibility of iris domain translation by using a structured loss [13] to penalize any probable structural mismatch between the synthesized output and target. Successful deployment of this network helps us to overcome the challenge faced in cross-spectral iris matching. However, in Scenarios 2(a) and 2(b), representing the additional cross-resolution case (see Fig. 1), the size of the output image should be larger than the size of the input image, i.e., the network should learn domain translation to a higher resolution. In this context, we modify the architecture of our cGAN generator by integrating the concept of super-resolution during the cross-domain translation. Super-Resolution (SR) estimates a HR super-resolved image from its LR counterpart, which has been vigorously applied to various computer vision applications. Although reconstructing an accurate HR image from its LR version is a very difficult task, multiple SR algorithms have been developed in recent years [204] to address this challenge.

Recently, the GAN-based SRGAN [12] approach has shown excellent results with high perceptual image quality by retrieving the fine textural details from a LR input image. Following their approach of up-sampling the LR input image, we improve our cGAN-based translation architec-

ture and incorporate a super-resolution layer as part of our cross-spectral framework to deal with the cross-resolution task considered in our basic Scenario 2. To synthesize high-quality VIS iris images, we train our network with a perceptual loss [72], which helps to generate a more accurate VIS iris images along with the widely used $L_2$ reconstruction loss [91, 116] and the adversarial loss [12] functions. A similar iris super-resolution method has been proposed in [200], which integrates adversarial training into triplet networks in order to develop a super-resolution architecture for low-quality iris images.However, the ability of their SR network is limited to super-resolving iris images in the same spectral domain. On the other hand, we jointly perform super-resolution and domain transformation in one shot to overcome the limitations of acquiring high-resolution NIR iris images. More specifically, our network produces a gallery of super-resolved HR VIS iris images from a gallery of LR NIR iris images, which is then used to match a HR VIS iris probe against it.

### 5.3.2 Verification

In this article, we have proposed to perform cross-spectral iris matching under the same spectra and the same resolution setting by adopting joint translation and super-resolution technique followed by the verification process. To accomplish this, we train our network on unrolled iris images of one spectral domain as input and generate unrolled iris images of the other spectral domain at the same resolution or higher resolution based on the scenarios described in the earlier sections. To perform verification, we employ a commercially available software, Open Source for IRIS (OSIRIS), which was developed within the BioSecure project [182] and offered by its authors as a free, open-source iris matcher. OSIRIS follows the iris matching concept proposed in the works of Daugman [159]. It applies Daugman's rubber sheet model for unwrapping the iris image from polar coordinates onto a Cartesian rectangle to process image segmentation and normalization tasks. Hence, during verification we match a normalized VIS iris probe against a gallery of synthesized normalized VIS iris images generated from our network using this OSIRIS software. It first generates iris codes by applying phase quantization of multiple Gabor wavelet filtering outcomes, while matching is performed using XOR operation, with normalized Hamming distance as an output dissimilarity metric. For genuine comparisons, we expect values close to zero, while we expect scores around

0.5 for imposter comparisons.

### 5.3.3 Deep Coupled Adversarial Framework

Our second proposed technique is a cpGAN architecture that consists of two coupled cGAN modules with the same architecture, as shown in Fig. 5.2. One of them is dedicated to synthesizing the VIS iris images, and hence, we refer to as the VIS cGAN module. Similarly, the other module is dedicated to synthesizing the NIR iris images, which is referred to as the NIR cGAN module. Our cpGAN network is inspired by the Siamese network [205], which ensures pairwise learning, where all the parameters are simultaneously updated throughout the network. We have followed a more recent U-Net-based, densely-connected encoder-decoder structure proposed in [206] to design our generator, which helps to achieve the low-dimensional embedded subspace for cross-spectral iris matching via a contrastive loss along with the standard adversarial loss. In addition to the adversarial loss and contrastive loss [201], the perceptual loss [72], and $L_2$ reconstruction loss are also used to guide the generators towards the optimal solutions. Perceptual loss is measured via a pre-trained VGG 16 network [103], which helps in sharp and realistic reconstruction of the images. In realistic opportunistic iris recognition scenarios, a VIS iris probe is usually matched against a gallery of NIR iris images. To create such application scenario, we focus on matching a VIS iris probe against a gallery of NIR iris images, that have not been seen before by the network. To perform this matching in a cross-spectral domain setting, a discriminative model is required to produce a domain-invariant representation. Therefore, we force the network to learn iris feature representations in a common embedding subspace by utilizing a U-Net auto-encoder architecture that uses class-specific contrastive loss to match the iris patterns in the latent domain. As previously mentioned, we use a U-Net auto-encoder architecture for our generator due to its structural ability of extracting features in the latent embedding subspace. More specifically, the contracting path of the "U shaped" structure of the U-Net captures contextual information, which is passed directly across all the layers, including the bottleneck. In neural networks, the bottleneck forces the network to learn the compressed version of the input data that only contains useful information to preserve the structural integrity of the image required to reconstruct the input. Along with the bottleneck, the high-dimensional features of the contracting path of the U-Net, combined with the corresponding

upsampled features of the symmetric expanding path, provides a means to share the useful information throughout the network. Moreover, during domain transformation, a significant amount of low-level information needs to be shared between input and output, which can be accomplished by leveraging a U-Net-like architecture.

We have followed the architecture of patch-based discriminators [13] to design the discriminators of our proposed model. The discriminators are trained simultaneously along with the respective generators. It is worthwhile to mention that the $L_1$ loss performs very well when applied to preserve the low-frequency details but fails to preserve the high-frequency information, whereas patch-based discriminator penalizes the structure at the patch scale to ensure the preservation of high-frequency details.

The main idea behind using the U-Net shaped generator is to gradually build a connection between the VIS and NIR iris images in the common embedding feature subspace. Since the features are domain invariant in the embedded subspace, it provides credibility to discriminate images based on identity. Therefore, our final objective is to find a set of domain invariant features in a common latent embedding subspace by coupling the two generators via a contrastive loss function, $L_{cont}$ [201].

The contrastive loss function, $L_{cont}$, is defined as a distance-based loss metric, which is computed over a set of pairs in the common embedding subspace such that images belonging to the same identity (genuine pairs i.e., a VIS iris image of a subject with its corresponding NIR iris image) are embedded as close as possible, and images of different identities (imposter pairs i.e., a VIS iris image of a subject with a NIR iris image of a different subject) are pushed further apart from each other. The contrastive loss function is formulated as:

$$L_{cont}(z_1(x^i_{VIS}), z_2(x^j_{NIR}), Y) = (1 - Y)\frac{1}{2}(D_z)^2 + (Y)\frac{1}{2}(\max(0, m - D_z))^2, \tag{5.5}$$

where $x^i_{VIS}$ and $x^j_{NIR}$ denote the input VIS and NIR iris images, respectively. The variable, Y, is a binary label, which is set to 0 if $x^i_{VIS}$ and $x^j_{NIR}$ belong to the same class (i.e., genuine pair), and equal to 1 if $x^i_{VIS}$ and $x^j_{NIR}$ belong to different classes (i.e., impostor pair). $z_1(.)$ and $z_2(.)$ are denoted as the encoding functions of the U-Net auto-encoder, which transform both $x^i_{VIS}$ and $x^j_{NIR}$, respectively into a common latent embedding subspace. Here, m, is used as the contrastive margin to "tighten"

the constraint. The Euclidean distance, $D_z$, between the outputs of the functions, $z_1(x^i_{VIS})$, and $z_2(x^j_{NIR})$, is given by:

$$D_z = \left\| z_1(x^i_{VIS}) - z_2(x^j_{NIR}) \right\|_2 . \tag{5.6}$$

Therefore, if Y = 0 (i.e., genuine pair), then the contrastive loss function, ($L_{cont}$), is given as:

$$L_{cont}(z_1(x^i_{VIS}), z_2(x^j_{NIR}), Y) = \frac{1}{2} \left\| z_1(x^i_{VIS}) - z_2(x^j_{NIR}) \right\|^2_2, \tag{5.7}$$

and if Y = 1 (i.e., impostor pair), then the contrastive loss function, ($L_{cont}$), is :

$$L_{cont}(z_1(x^i_{VIS}),z_2(x^j_{NIR}), Y) = \frac{1}{2}\max\left(0, m - \left\| z_1(x^i_{VIS}) - z_2(x^j_{NIR}) \right\|^2_2\right). \tag{5.8}$$

Thus, the total loss for coupling the VIS generator and NIR generator is denoted by $L_{cpl}$ and is given as:

$$L_{cpl} = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} L_{cont}(z_1(x^i_{VIS}), z_2(x^j_{NIR}), Y), \tag{5.9}$$

where N is the number of training samples. The contrastive loss in the above equation can also be replaced by some other distance-based metric, such as the Euclidean distance. However, the main aim of using the contrastive loss is to be able to use the class labels implicitly and find a discriminative embedding subspace, which may not be the case with some other metric such as the Euclidean distance. This discriminative embedding subspace would be useful for matching the VIS iris images against the gallery of NIR iris images.

## 5.4 Loss Functions

### 5.4.1 Generative Adversarial Loss

The VIS and NIR generators are denoted as $G_{VIS}$ and $G_{NIR}$, as they will synthesize the corresponding VIS and NIR iris images from the input VIS and NIR iris images, respectively. The patch-based discriminators used for the VIS and NIR iris GANs are denoted as $D_{VIS}$ and $D_{NIR}$,

respectively. To implement our proposed methods, we have used the conditional GAN, where the generator networks $G_{VIS}$ and $G_{NIR}$ are conditioned on the input VIS and NIR iris images, respectively. In addition, we have trained the generators and the corresponding discriminators with the cGAN loss function [149] to ensure a real-looking natural image reconstruction such that the discriminators cannot distinguish the generated images from the real ones. Let $L_{VIS}$ and $L_{NIR}$ denote the cGAN loss functions for the VIS and NIR GANs, respectively. Therefore, the loss function for the cGAN which is considered as the backbone architecture in our first approach, can be defined as following:

$$L_{VIS} = L_{cGAN}(D_{VIS}, G_{VIS}, y^i_{VIS}, x^i_{VIS}), \tag{5.10}$$

$$L_{NIR} = L_{cGAN}(D_{NIR}, G_{NIR}, y^j_{NIR}, x^j_{NIR}), \tag{5.11}$$

where $L_{cGAN}$ is defined as the cGAN objective function in (5.4). The term, $x^i_{VIS}$, is used to denote the VIS iris image, which is defined as a condition for the VIS cGAN, and $y^i_{VIS}$, is denoted as the real VIS iris image. It is worth mentioning that the real VIS iris image, $y^i_{VIS}$, is same as the network condition given by $x^i_{VIS}$. Similarly, $x^j_{NIR}$, denotes the NIR iris image that is used as a condition for the NIR cGAN. Again, like $y^i_{VIS}$, the real NIR iris image, $y^j_{NIR}$, is same as the network condition given by $x^j_{NIR}$. The total adversarial loss for our proposed cpGAN is given by:

$$L_{GAN} = L_{VIS} + L_{NIR}. \tag{5.12}$$

### 5.4.2 $L_2$ Reconstruction Loss

We consider the $L_2$ reconstruction loss as a classical constraint for both the VIS cGAN and NIR cGAN to ensure better results. The $L_2$ reconstruction loss measures the reconstruction error in terms of the Euclidean distance between the reconstructed iris image and the corresponding real iris image. We denote the reconstruction loss for the VIS cGAN as $L_{2_{VIS}}$ and define it as:

$$L_{2_{VIS}} = \left|\left|G_{VIS}(z|x^i_{VIS}) - y^i_{VIS}\right|\right|^2_2, \tag{5.13}$$

where $y^i_{VIS}$ is the ground truth VIS iris image, and $G_{VIS}(z|x^i_{VIS})$, is the output of the VIS generator.

Similarly, we denote the reconstruction loss for the NIR cGAN as $L_{2_{NIR}}$:

$$L_{2_{NIR}} = \left\| G_{NIR}(z|x^j_{NIR}) - y^j_{NIR} \right\|_2^2, \tag{5.14}$$

where $y^j_{NIR}$ is the ground truth NIR iris image, and $G_{NIR}(z|x^j_{NIR})$, is the output of the NIR generator. Depending on the different cross-spectral iris matching scenarios, we use either $L_{2_{VIS}}$ or $L_{2_{NIR}}$ as the reconstruction loss, which is again generally termed as $L_{2_{cGAN}}$ for the method proposed in our first approach.

For the cpGAN architecture proposed in our second approach, the total $L_{2_{cpGAN}}$ reconstruction loss can be defined by the following equation:

$$L_{2_{cpGAN}} = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} (L_{2_{VIS}} + L_{2_{NIR}}). \tag{5.15}$$

### 5.4.3 Perceptual Loss

Although the GAN loss and the reconstruction loss are used to guide the generators, they fail to reconstruct perceptual features in the generated images. Perceptual features are defined by the visual characteristics of objects, which provide a perceptually pleasing look to the image. Hence, we have also used the perceptual loss, introduced in [72], for style transfer and super-resolution. The perceptual loss function basically measures high-level differences, such as content and style dissimilarity, between images. The perceptual loss is based on high-level representations from a pre-trained VGG-16 [103] like CNN. Moreover, it helps the network to generate better and sharper high-quality images [72]. As a result, it can be a significant alternative to solely using the $L_1$ or $L_2$ reconstruction error. Recently, Zhang et al. [207] introduced the LPIPS loss metric, which has been adapted in several deep learning architectures for image reconstruction. Therefore, it can be considered as an alternative loss function for perceptual fidelity instead of the well-known ImageNet pre-trained VGG-based perceptual loss [72].

In both of our approaches, we have added perceptual loss to both the VIS and NIR cGAN modules using a pre-trained VGG-16 network. It involves extracting the high-level features (ReLU3-3 layer) of VGG-16 for both the real input image and the reconstructed output of the generator. The perceptual loss calculates the $L_1$ distance between the features of real and reconstructed images to guide the generators $G_{VIS}$ and $G_{NIR}$. The perceptual loss for the VIS cGAN network is defined as:

$$L_{P_{VIS}} = \frac{1}{C_p W_p H_p} \sum_{c=1}^{C_p} \sum_{w=1}^{W_p} \sum_{h=1}^{H_p} \left\| V(G_{VIS}(z|x^i_{VIS}))^{c,w,h} - V(y^i_{VIS})^{c,w,h} \right\|, \tag{5.16}$$

where V(.) is used to denote a particular layer of the VGG-16 and $C_p$, $W_p$, and $H_p$ denote the layer dimensions.

Likewise, the perceptual loss for the NIR cGAN network is:

$$L_{P_{NIR}} = \frac{1}{C_p W_p H_p} \sum_{c=1}^{C_p} \sum_{w=1}^{W_p} \sum_{h=1}^{H_p} \left\| V(G_{NIR}(z|x^j_{NIR}))^{c,w,h} - V(y^j_{NIR})^{c,w,h} \right\|. \tag{5.17}$$

Here, we simply define $L_{P_{VIS}}$ or $L_{P_{NIR}}$ as $L_{P_{cGAN}}$ to calculate perceptual loss for our first approach. The total perceptual loss function for the cpGAN is given by:

$$L_{P_{cpGAN}} = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} (L_{P_{VIS}} + L_{P_{NIR}}). \tag{5.18}$$

### 5.4.4 Overall Objective Function

We sum all the loss functions defined above to calculate the overall objective function $L_{tot_{cGAN}}$ and $L_{tot_{cpGAN}}$ for our proposed cGAN and cpGAN architectures, respectively:

$$L_{tot_{cGAN}} = L_{2_{cGAN}} + \lambda_1 L_{cGAN} + \lambda_2 L_{P_{cGAN}}, \tag{5.19}$$

where $L_{2_{cGAN}}$ is the total reconstruction error, $L_{cGAN}$ is the total conditional generative adversarial loss function, and $L_{P_{cGAN}}$ is the total perceptual loss for our proposed cGAN model. Variables $\lambda_1$, and $\lambda_2$ are the adjustable hyper-parameters used to weigh the different loss terms. The total loss for cpGAN is given as:

$$L_{tot_{cpGAN}} = L_{cpl} + \lambda_3 L_{GAN} + \lambda_4 L_{P_{cpGAN}} + \lambda_5 L_{2_{cpGAN}}, \tag{5.20}$$

where $L_{cpl}$ is the coupling loss, $L_{GAN}$ is the total generative adversarial loss, $L_{P_{cpGAN}}$ is the total perceptual loss, and $L_{2_{cpGAN}}$ is the total reconstruction error. Variables $\lambda_3$, $\lambda_4$, and $\lambda_5$ are the hyper-parameters used as a weight factor to numerically balance the magnitude of different loss terms.

## 5.5 Experiments

We first briefly introduce the publicly available datasets that we have used in our experiments and discuss the implementation details of our proposed cGAN and cpGAN architectures along with their training setup. To evaluate the performance of our methods, we perform a range of experiments for different cross-spectral iris matching scenarios and compare their performance with other state-of-the-art iris recognition methods in the cross-domain setting. We provide detailed comparative experimental results in the following sections. Finally, in order to ascertain the usefulness of our cross-spectral iris recognition frameworks, we conduct additional experiments for cross-device iris matching scenarios.

### 5.5.1 Database

Three available cross-spectral database, PolyU bi-spectral iris database [169], WVU Face and Iris Dataset*,† and Cross-eyed-cross-spectral iris recognition database [183] are employed to validate our proposed methods.

#### PolyU Bi-Spectral iris database

The PolyU Bi-Spectral iris database contains iris images of 209 subjects acquired simultaneously under both the VIS and NIR illuminations. Each subject consists of 15 different instances of right and left-eye images with a resolution of 640 $\times$ 480 pixels for both VIS and NIR spectrum. Therefore, the total number of images in this dataset is 12,540 (209 $\times$ 2 $\times$ 2 $\times$ 15). We used a publicly-available segmentation algorithm [165] to accurately segment and normalize iris images for the experiments. This segmentation algorithm provides normalized iris images of 512 $\times$ 64

*This data was collected at WVU under IRB # 1805125982 with appropriate human subjects' approval.
†This dataset is available upon request at biic.wvu.edu.

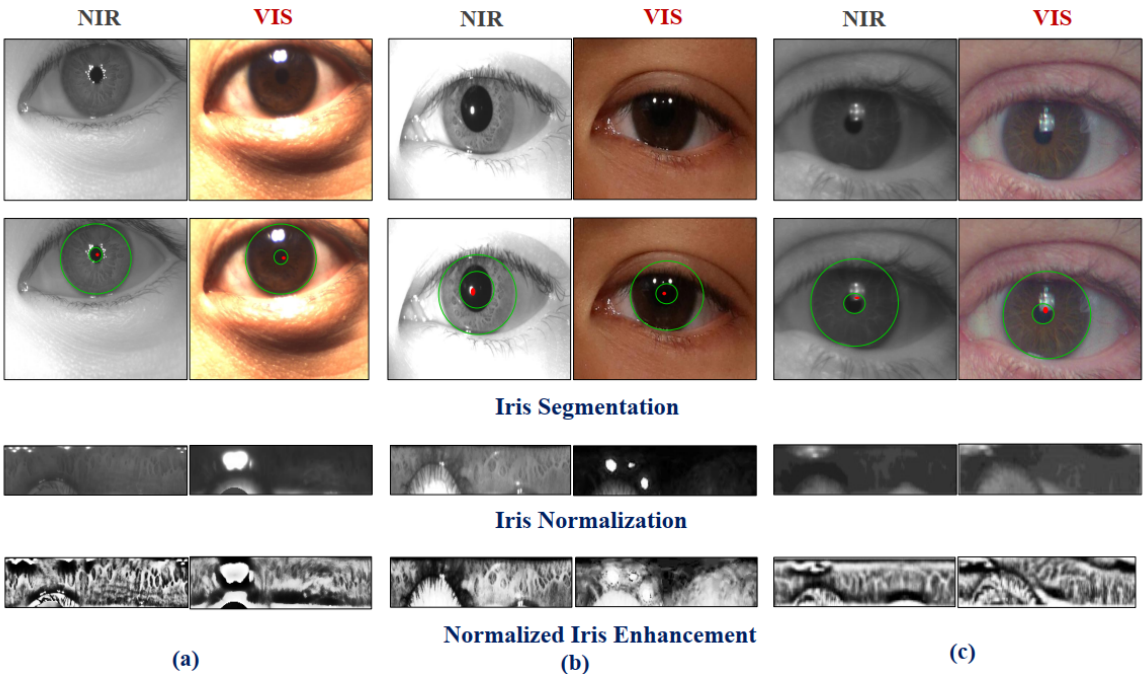
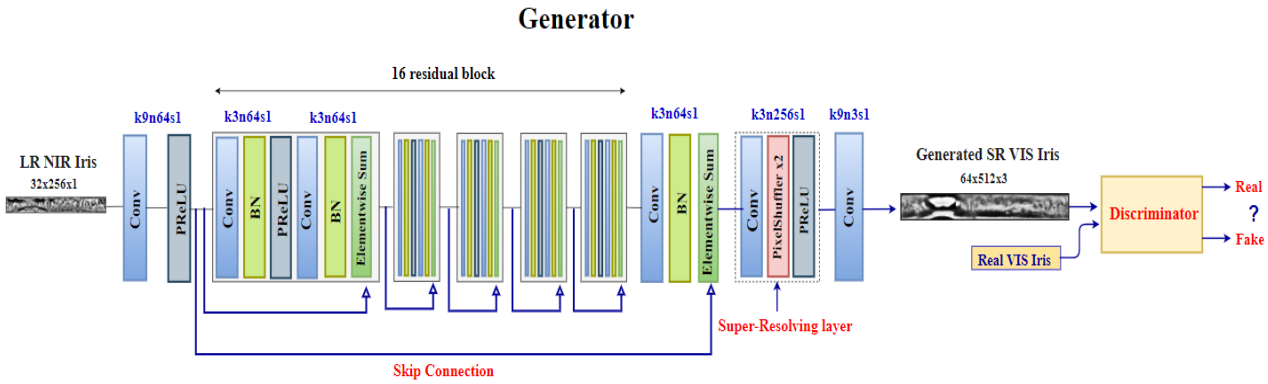



Figure 5.3: Iris image preprocessing steps (Segmentation, Normalization, Enhancement) for (a) PolyU bi-spectral iris database and (b) WVU face and iris dataset (c) Cross-eyed-cross-spectral iris recognition database.

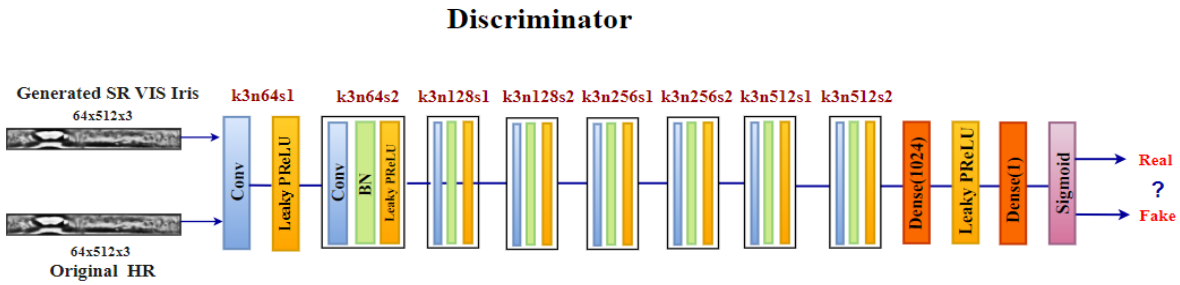
pixels, samples of which are shown in Fig. 5.3(a). Following the approach used in [169], we selected the first ten instances for our network training and the remaining five instances for the testing. The all-to-all matching protocol generated 2,800 genuine scores and 1,953,000 imposter scores.

### WVU face and iris dataset

The West Virginia University (WVU) Face and Iris dataset is particularly developed for cross-spectral opportunistic iris recognition. It contains 1,248 subjects, which provides a total of 2,496 left and right NIR as well as VIS iris images ($1,248 \times 2$). We use the method presented in [165] to extract the normalized iris images ($512 \times 64$) from the original iris images of size $640 \times 480$ pixels. Sample images from this dataset are shown in Fig. 5.3(b). Again, following the same train-test protocol used in reference [169] for this dataset, we attained 750 genuine scores and 561,750 imposter scores for 375 test subjects.




((a))




((b))

Figure 5.4: Architecture of our proposed cGAN (a) generator and (b) discriminator with corresponding kernel size (k), number of feature maps (n) and stride (s) indicated for each convolutional layer.

**Cross-eyed-cross-spectral iris recognition database**

The Cross-eyed-cross-spectral iris recognition database provides 3,840 iris images from 240 classes for both spectra obtained from 120 subjects. Each of the classes from every subject has eight sample of 400 $\times$ 300 pixels for both spectra. We use the same iris segmentation and normalization algorithm used in [165] to normalize all the iris images. The dimension of all the segmented and normalized iris images from this dataset is 512 $\times$ 64 pixels. Sample images from the cross-eyed cross-spectral database are shown in Fig. 5.3(c). In order to ensure fair comparison, we follow the train-test protocol used in [169] and choose five image samples for training and the remaining three samples for testing. Applying an all-to-all matching protocol, the network generated 2,160 genuine scores and 516,240 imposter scores.

### 5.5.2 cGAN Architecture Implementation

We adopted our proposed cGAN network structure from reference [12] as depicted in Fig. 4 for our domain translation technique, and formulated the overall loss function inspired by references [12, 13, 72]. In more detail, for our generator (see Fig. 5.4(a)), we have implemented the ResNet16 architecture [119], with 16 identical residual blocks. A single residual block is composed of two convolutional layers with 3 $\times$ 3 kernels, 64 feature maps, batch-normalization layers and a Parametric Rectified Linear Unit (ReLU) [90, 133] activation function. We use this network for iris domain translation based on two different cross-spectral situations (see Scenario 1 and Scenario 3 in Fig. 1) that we have proposed in our first approach. We also integrate the super-resolution process in the translation network by adding a sub-pixel convolution layer with the layout explained in [91], which has been illustrated in Scenario 2(a) of Fig. 5.1. Like [12], for our discriminator architecture, we follow what is presented in [92], which consists of eight convolutional layers with 3 $\times$ 3 kernel size. The number of kernels increases from 64 to 512, similar to VGGNet [103]. Rather than max-pooling, strided convolution is employed for resolution reduction. As shown in Fig. 5.4(b), after that, we add a dense layer, a Leaky RELU, another dense layer, and finally, a sigmoid activation function. In summary, the generator gets a low-resolution (or high-resolution) image from one of the domains and translates it or jointly translates and super-resolves it to the other domain, and the discriminator is fed with the output of the generator and

also a high-resolution image of the other domain.

### 5.5.3 cpGAN Architecture Implementation

We have implemented our cpGAN architecture using the U-Net architecture as the generator module. We have followed the architecture of ResNet-18 [119] to implement both the encoder and decoder sections of the U-Net model. In encoder, each block is designed by applying two $3 \times 3$ convolutions, each followed by a ReLU. For downsampling, it uses a $2 \times 2$ max pooling operation with stride 2. We double the number of feature channels at each downsampling step. Similarly, each step in the decoder section upscales the feature map by applying a $2 \times 2$ transpose convolution [208], upsampling the dimension of the feature map. Each feature map is concatenated with the corresponding feature map from the encoder, followed by two $3 \times 3$ convolutions with a ReLU activation function.

### 5.5.4 Training details

Both of our frameworks have been implemented in Pytorch. We trained the network with a batch size of 16 and a learning rate of $2 \times 10^{-4}$. We used the Adam optimizer [209] with a first-order momentum of 0.5, and a second-order momentum of 0.999. We have used Leaky ReLU as the activation function with a slope of 0.35 for the discriminator. To find the optimal hyper-parameters for our learning algorithms, we have used a random search strategy [210]. Following their technique, we experiment with different scaling heuristics to find the optimal hyper-parameter multiplier, which results in the best verification accuracy. Accordingly, for the network convergence, we set $\lambda_3$ to 1, and $\lambda_4$, and $\lambda_5$ to 0.3. In addition, $\lambda_1$, and $\lambda_2$, are set to $10^{-6}$ and $2 \times 10^{-3}$, respectively.
For training, genuine/impostor pairs are created from the VIS and NIR iris images of the same/different subjects. During the experiments, we ensure that the training set is balanced by using the same number of genuine and impostor pairs.

### 5.5.5 Evaluation on PolyU Bi-Spectral Database

We perform our first set of experiments on the PolyU Bi-Spectral database considering many different cross-spectral iris matching cases for both previously-mentioned approaches. In all the

experiments, each iris probe image is matched against a gallery of iris images, which generates genuine and imposter scores. Using these matching scenarios, we calculate the key recognition performance parameters, such as genuine acceptance rate (GAR), false acceptance rate (FAR), and equal error rate (EER). In addition, we also plot receiver operating characteristics (ROC) curves to analyze the GAR with respect to FAR. In addition, we compare our results over other considered state-of-the-art cross-spectral iris recognition methods described in [168, 169, 171] and [211] . We use the same train-test protocol provided in their original paper for fair comparison.

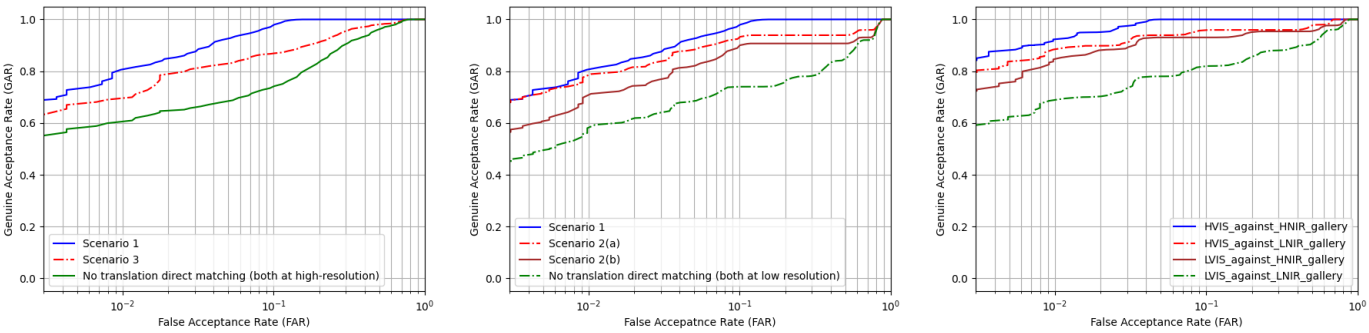
**To evaluate the effectiveness of our proposed cGAN architecture, we conduct the following experiments:**

**(a) Scenario 1 : NIR to VIS domain translation**

In this experiment, we train the network to translate a gallery of NIR iris images to its corresponding gallery of synthesized visible iris images at the same resolution (see Fig. 5.1 Scenario 1). Then, each VIS iris probe of the test set is matched against this synthesized VIS iris gallery. We have shown the ROC result from this experiment in Fig. 5.5(a) and report the EER in Table 5.2. We observe that our proposed algorithm achieves 99.50% and 80.50% GAR at 0.1 and 0.01 FAR, respectively, and obtains an EER of 1.5%, which outperform the results reported for the algorithms evaluated in [169,171], and [211] using the same train-test protocol. The network shows significant improvement in cross-spectral iris matching by obtaining 15.53% and 25.18% less EER compared to the results in [211], and [169], respectively.

**(b) Scenario 2(a) : Joint translation and super-resolution from the LR NIR to HR VIS domain**

Recently, with the emergence of new biometrics applications on smartphones, there is a strong demand for acquiring high-resolution visible iris images at low cost. However, while the availability of higher resolution visible iris images will eventually lead to a cross-resolution mismatch in the problem of cross-spectral iris matching, almost no attention has been turned toward it yet. Although there would be higher noise levels in the visible domain compared to the NIR domain, hopefully the higher resolution can compensate for the effect of this noise. To address the reso-




((a)) ((b)) ((c))

Figure 5.5: ROC plots showing the performance of our approach (a) Scenario 1 and 3 with cGAN architecture, (b) Scenario 2(a) and 2(b) with cGAN architecture and (c) cpGAN architecture obtained on the PolyU Bi-Spectral database for the different cross-spectral matching scenarios [1].

lution differences, we determined how to match LR NIR iris images against the HR visible iris images (i.e., unrolled NIR image size: 32x256, unrolled visible image size: 64x512). We train the network to translate the LR NIR images to HR VIS images in such a way that it jointly transforms the image domain and super-resolves it. Therefore, the network simultaneously learns both image translation and super-resolution tasks. The network super resolves the input image by a factor of two, and then the output can be used as a gallery of visible iris images for visible-to-visible iris verification. Fig. 5.5(b) and Table 5.2 illustrates that our proposed joint translation and super-resolution technique outperforms the baseline approach. It is worth mentioning that we separately train both networks and report the results as a baseline approach to show the comparative performance of the joint learning. We notice that the joint training significantly increases the matching accuracy by 3.94%, 5.60% and 13.8% GAR at FAR of 0.1, 0.01 and 0.001, respectively.

**(c) Scenario 2(b) : Separate translation and super-resolution from the LR NIR to HR VIS domain**

We have also fed the low-resolution NIR images to a cross-domain translation network from reference [13] and then the low-resolution output is fed to a super-resolution GAN (SRGAN) from reference [12]. This is the Scenario 2(b) in Fig. 5.1, and results are shown in Fig. 5.5(b) and Table 5.2. The separate training achieves 88.89%, 70.10%, and 56.10% GAR at FAR of 0.1, 0.01 and 0.001, respectively, which are significantly lower compared to the joint training. These results

Table 5.2: Comparative performances on the PolyU Bi-Spectral database. Symbol '-' indicates that the metric is not available for that protocol.

| Algorithm | Matching | GAR@FAR=0.1 | GAR@FAR=0.01 | GAR@FAR=0.001 | EER (%) |
|---|---|---|---|---|---|
| Wang et al. [211] | HR VIS vs HR NIR | — | 59.10 | 37.00 | 17.03 |
| CNN with SDH [168] | HR VIS vs HR NIR | — | 90.71 | 84.50 | 5.39 |
| Nalla et al. [169] | HR VIS vs HR NIR | 58.8 | — | — | 26.68 |
| NIR to VIS texture synthesis using MRF model [169] | HR VIS vs HR NIR | 61.91 | — | — | 23.87 |
| IrisCode using 1D Log-Gabor filter [171] | HR VIS vs HR NIR | 52.6 | — | — | 17.03 |
| **cpGAN** [1] | **HR VIS vs HR NIR** | **99.99** | **92.38** | **84.98** | **1.02** |
| cpGAN [1] | HR VIS vs LR NIR | 96.5 | 89.89 | 81.21 | 1.21 |
| cpGAN [1] | HR NIR vs LR VIS | 93.30 | 84.75 | 73.45 | 1.26 |
| cpGAN [1] | LR NIR vs LR VIS | 82.60 | 70.10 | 59.97 | 2.51 |
| NIR to VIS domain translation (Ours cGAN) | Scenario1 | 99.50 | 80.50 | 70.1 | 1.5 |
| Joint domain translation & super-resolution (Ours Modified cGAN) | Scenario 2(a) | 92.83 | 75.70 | 69.9 | 1.6 |
| Domain Translation & super-resolution (Separate Training) | Scenario 2(b) | 88.89 | 70.10 | 56.10 | 1.9 |
| VIS to NIR domain translation (Ours cGAN) | Scenario 3 | 87.49 | 69.50 | 64.90 | 1.4 |
| Iriscode (OSIRIS) [182] | HR VIS vs HR NIR | 74.60 | 61.10 | 54.50 | 2.59 |
| Iriscode (OSIRIS) [182] | LR VIS vs LR NIR | 71.05 | 55.60 | 43.10 | 3.0 |

validate our idea of joint transformation and super-resolution.

**(d) Scenario 3 : VIS to NIR Domain Transformation**

In order to examine whether or not the NIR-to-visible image translation is a more effective solution than translating the visible to NIR, both at the same resolution, we have trained a network to map the visible images to the NIR domain and performed verification on the synthesized NIR iris images (i.e., matching the synthesized NIR images against a gallery of NIR images). We feed a given visible iris probe image to the network, which is trained to map visible to NIR images, and then use the output image to compare with an existing gallery of NIR images. We report the ROC result obtained from this experiment in Fig. 5.5(b) along with the comparative results from other approaches. We consider the algorithm used in [182] as comparable benchmark for this scenario. It proves the efficacy of our proposed approach by acquiring 2.19% less EER compared to the baseline result mentioned above.

**Similarly, to ascertain true cross-spectral matching ability of our proposed cpGAN network, we experiment with different types of cross-comparisons as follows:**

**(a) Matching HR VIS probe against a HR NIR gallery**:

To perform this verification, we train our coupled learning network with the unrolled HR $64 \times 512$ VIS and NIR iris images such that VIS and NIR generators are trained to obtain domain invariant features in a common latent embedding subspace using a contrastive loss. We plot ROC curves comparing our approach with other state-of-the-art deep learning methods presented in [168, 211],

which apply different types of feature extraction techniques. From Fig. 5.5(c) and Table 5.2, we notice that our cpGAN framework performs much better than the baseline matching algorithms mentioned above. In this setting, our method achieves 1.67% more identification accuracy with 4.37% decrease in EER compared to the most recent cross-spectral iris recognition method [168]. Additionally, it outperforms the method described in [169, 182] by a significant decrease of 1.57% and 22.85% in EER, respectively. This significant improvement clearly indicates that using a cpGAN framework for projecting both the VIS and NIR iris images into a common latent embedding subspace to retrieve the domain invariant features is better than the other existing deep learning methods.

**(b) Matching HR VIS probe against a LR NIR gallery:**

Here, we consider a realistic iris matching scenario to analyze the cross-spectral matching accuracy of our network. Due to the advances in imaging technology, opportunistic iris images extracted from faces in the visible spectrum are at a higher resolution, while images already stored in the gallery are in the low-resolution NIR domain. It has become a challenging task to build a correlation between iris images in different resolutions as well as in different spectra. Many algorithms fail to retrieve accurate semantic similarity among iris images of different resolutions and spectra, which has resulted in a significant performance degradation in existing iris verification systems. Therefore, we resolve this issue by training our cpGAN with the unrolled HR ($64 \times 512$) VIS and LR ($32 \times 256$) NIR iris images, which ensures the retrieval of contextual and semantic features of the iris images in a common embedding subspace. The results summarized in Fig. 5.5(c) and Table 5.2 indicate that the cpGAN network remains robust enough to provide superior results compared to our matching Scenario 2(a) that was shown in Fig. 5.5(b). It has increased the GAR almost by 14% at 0.01 FAR.

**(c) Matching LR VIS iris images against a gallery of HR NIR iris images:**

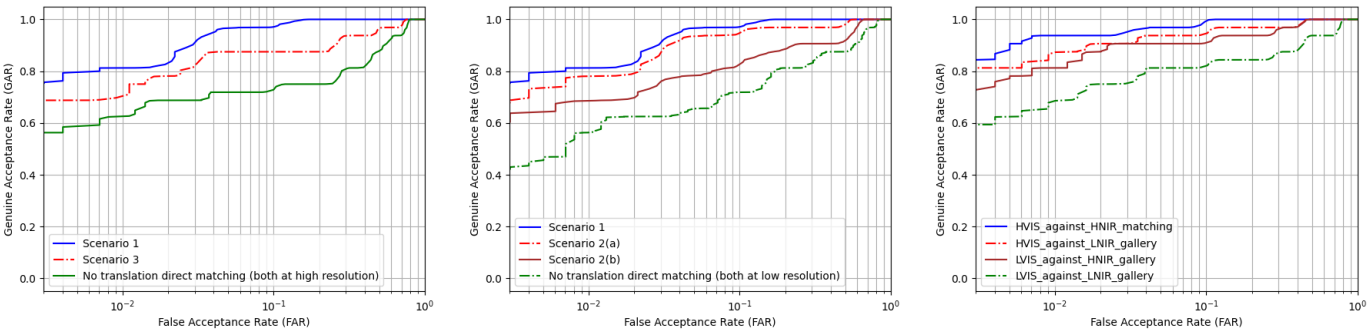
In addition to the study mentioned above, we have also focused on matching LR VIS iris probe against a gallery of HR NIR iris images. We consider a fact when subjects are at a large standoff distance from the camera. Consequently, captured faces are assumed to be suffering from poor quality due to low-resolution. On the other hand, the gallery images have comparatively higher resolution which are usually taken in the NIR spectrum. Therefore, the modality gap between

probe and gallery images makes the cross-spectral matching even more challenging. Hence, we train the VIS and NIR generator of our network with the unrolled LR VIS iris images ($32 \times 256$) and HR NIR iris images ($64 \times 512$), respectively, and perform matching in the latent embedded subspace, that contains basic information about the iris texture patterns irrespective of the resolution. The experimental results reported in Table 5.2 show that our proposed scheme has produced EER with a value of 1.26% which proves the adequacy of our approach even in low-quality videos.

**(d) Matching LR VIS iris images against a gallery of LR NIR iris images :**
We also perform additional experiments where our gallery images are in the low-resolution NIR domain. To investigate the matching performance of our network, we feed both the VIS and NIR generator with the unrolled LR VIS and NIR iris images. The experimental results reported in Table 5.2 and Fig. 5.5(c) indicate the matching accuracy of our network for this cross-spectral setting compared to the approach used in [182]. Even though we achieve an EER of 2.51% that is much lower than several comparable methods, there is a tradeoff with verification performance, which is not as satisfactory as our previous experiments outlined above.

### 5.5.6 Evaluation on WVU Face and Iris Database

To assess the effectiveness of our proposed approaches, we conduct a number of extensive experiments on the WVU face and iris database for different cross-spectral matching scenarios similar to the experiments performed on the PolyU bi-spectral database. To the best of our knowledge, there is no other baseline algorithm in the literature that have performed cross-spectral iris matching on this dataset. Therefore, our evaluation on the WVU face and iris dataset yields a new state-of-the-art cross-spectral iris matching result, which will further encourage the biometric research community to investigate the performance of other existing algorithms on this dataset. In this context, we first report on the evaluation of the method in our first approach for matching cross-spectral iris images under the same spectral domain. Then we discuss experimental results obtained from our second method, which performs matching in the embedded domain.




((a)) ((b)) ((c))

Figure 5.6: ROC plots showing the performance of our approach (a) Scenario 1 and 3 with cGAN architecture, (b) Scenario 2(a) and 2(b) with cGAN architecture and (c) cpGAN architecture obtained on the WVU face and iris database for the different cross-spectral matching scenarios.

### Matching Results Obtained From cGAN Architecture

We consider similar experimental scenarios as stated in Section 5.5.5 when we trained our cGAN model with the WVU face and iris database. We plot ROC results in Fig. 5.6(a) obtained from the cGAN network that has been trained and tested for Scenario 1 and Scenario 3. In addition, we summarize the EER in Table 5.3. Fig. 5.6(a) and Table 5.3 demonstrate that our proposed algorithm achieves 99.50%, 80.50% and 70.1% GAR at 0.1, 0.01 and 0.001 FAR, respectively for Scenario 1, where each VIS iris probe image of the test set is matched against a gallery of synthesized VIS iris images. For comparison, we report recognition accuracy for this database which has been obtained from the algorithm used in [182] for matching the HR VIS iris probe image against a gallery of HR NIR iris images. It is obvious that our proposed cGAN algorithm significantly increases the recognition accuracy by 18.8% for the FAR of 0.01 with 1.14% decrease in EER compared to the cross-spectral iris matching result reported as a baseline approach (Matching HR VIS iris probe image against a gallery of NIR iris images).

We also report identification accuracy for the reverse case as described by Scenario 3. In this case, we train a network to map the VIS iris images to the NIR domain and perform matching between the synthesized NIR iris images and a gallery of NIR iris images. The experimental results summarized in Table 5.3 prove that even for the reverse scenario our approach achieves 8.1% higher recognition accuracy and 0.76% lower EER compared to the baseline result.

Again, to ascertain the comparative performance of our joint network described in Scenario 2(a), which simultaneously translates and super-resolves a LR NIR iris image to a HR VIS image, we separately train both networks and use the result as baseline. Additionally, we apply the popular IrisCode approach [182] to generate comparative matching scores (i.e., matching the LR VIS iris probe against a gallery of LR NIR iris images). The ROC results from this set of experiments are shown in Fig. 5.6(b), which indicate the superiority of our proposed joint learning method over other benchmark results. Table 5.3 summarizes that our joint translation and super-resolution technique significantly outperforms the result obtained from separate training by 8.20% recognition rate at 0.01 FAR.

**Matching Results Obtained From cpGAN Architecture**

To evaluate the verification performance of our coupled learning framework, we follow similar experimental settings that were previously discussed in the earlier section for the PolyU bi-spectral database. We experiment with four different cross-spectral and cross-resolution iris matching scenarios for this dataset and plot ROC results in Fig. 5.6(c) to show the recognition accuracy of our proposed network. We also provide EER results in Table 5.3.

The experimental results illustrated in Table 5.3 indicate that our cpGAN network, which performs verification in the embedding subspace, achieves a lower EER of 0.90% with a higher GAR of 93% at 0.01 FAR, when matching HR VIS iris probe image against a gallery of HR NIR iris images. Moreover, it significantly improves the matching accuracy by 31% GAR at 0.01 FAR compared to the reported baseline result [182] using the same test data for the same cross-spectral matching scenario.

Next, we consider a real-life cross-resolution matching scenario within the cross-spectral domain and train our cpGAN with the unrolled HR ($64\times512$) VIS and LR ($32\times256$) NIR iris images, which gradually learns the inherent hidden correlation between iris images in the cross-resolution and cross-spectral domains. The matching results briefly presented in Fig. 5.6(c) and Table 5.3 show that our cpGAN network ensures an accurate retrieval by outperforming the matching Scenario 2(a) in Fig. 5.6(b) with 10.9% higher recognition accuracy at 0.01 FAR.

Also, we conduct experiments for the scenario with low-quality videos. ROC results and EER

Table 5.3: Comparative performances on the WVU face and iris database.

| Algorithm | Matching | GAR@FAR=0.1 | GAR@FAR=0.01 | GAR@FAR=0.001 | EER (%) |
|---|---|---|---|---|---|
| **cpGAN(Ours)** | **HR VIS vs HR NIR** | **99.54** | **93** | **84** | **0.90** |
| cpGAN(Ours) | HR VIS vs LR NIR | 97.04 | 87.7 | 80.8 | 1.15 |
| cpGAN(Ours) | HR NIR vs LR VIS | 92.89 | 83.50 | 72.70 | 1.20 |
| cpGAN(Ours) | LR NIR vs LR VIS | 82.52 | 69.2 | 59.70 | 1.85 |
| NIR to VIS domain translation (Ours cGAN) | Scenario 1 | 97.79 | 80.8 | 75.1 | 1.0 |
| Joint domain translation & super-resolution (Ours Modified cGAN) | Scenario 2(a) | 94.97 | 77.8 | 69.5 | 1.34 |
| Domain translation & super-resolution (Separate Training) | Scenario 2(b) | 83.50 | 69.60 | 60.0 | 1.97 |
| VIS to NIR domain translation (Ours cGAN) | Scenario 3 | 88.53 | 70.10 | 67.70 | 1.38 |
| Iriscode (OSIRIS) [182] | HR VIS vs HR NIR | 76.02 | 62.0 | 56.1 | 2.14 |
| Iriscode (OSIRIS) [182] | LR VIS vs LR NIR | 71.7 | 55.5 | 42.7 | 3.01 |

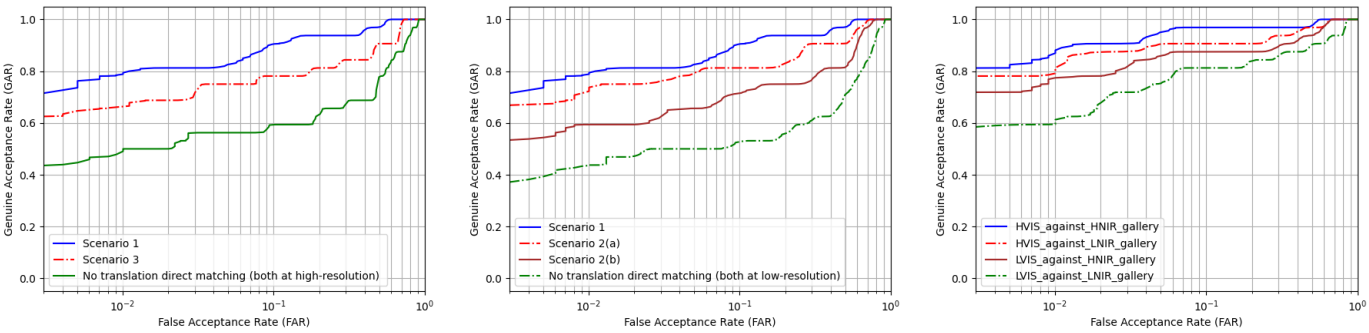
scores detailed in Table 5.3 prove that our proposed scheme maintains robust performance even when matching a LR VIS iris probe against an available HR NIR gallery. It has generated an EER of 1.20%, which is considered as a lower EER value for an ideal biometric system.

Finally, we investigate the verification performance of our proposed cpGAN network when iris images in the gallery are in low-resolution NIR domain. Therefore, we force the cpGAN network to learn invariant features in the common embedding subspace from both the LR ($32 \times 256$) VIS and NIR iris images. The experimental results in Table 5.3 show that our proposed algorithm obtains 3.7% more recognition accuracy at 0.01 FAR than the approach used in [182] on the same test data for this cross-spectral setting.

### 5.5.7 Evaluation on Cross-Eyed-Cross-Spectral Iris Recognition Database

We perform another set of experiments using the cross-eyed database to quantify the cross-spectral iris recognition accuracy for both of the approaches developed for this paper. We follow the same experimental settings conducted for the other two datasets for different cross-spectral matching scenarios that have been described in the previous sections. It is worth noting that while comparing our results obtained for this dataset over existing algorithms [168, 169, 171, 182], we follow the same train-test protocol used in their paper to show fair evaluation.

The comparative matching results from our cGAN and cpGAN architectures are shown in Fig. 5.7, while the corresponding EER results are summarized in Table 5.4. For comparison we use several highly competitive benchmark MRF approach [169], polpular gabor filter based IrisCode [182], SDH method [168] and another 1D log-gabor filter based IrisCode [171] to ascertain the superiority of our proposed approaches.




((a)) ((b)) ((c))

Figure 5.7: ROC plots showing the performance of our approach (a) Scenario 1 and 3 with cGAN architecture, (b) Scenario 2(a) and 2(b) with cGAN architecture and (c) cpGAN architecture obtained on the Cross-eyed-cross-spectral iris recognition database for the different cross-spectral matching scenarios.

Fig. 5.7(a) depicts the experimental results for Scenario 1 and Scenario 3 from our cGAN architecture compared to the baseline result using the most widely deployed IrisCode [182] approach. The results from Scenario 1 indicate that our proposed domain translation technique using the cGAN architecture significantly improves the cross-spectral iris matching accuracy by 28.59% at 0.01 FAR compared to the benchmark result using the IrisCode [182] approach. In addition, it also achieves 15.3% higher GAR at 0.01 FAR and 0.73% lower EER even when we experiment matching for Scenario 3.

In Fig. 5.7(b), we present ROC results for showing the performance of our proposed joint network Scenario 2(a) where the network learns to translate and super-resolve simultaneously from the LR NIR to HR VIS iris image, and compare this result to the approach when both techniques are applied separately (see Fig. 5.1 Scenario 2(b)). Table 5.4 shows that joint training obtains 74.8% GAR at 0.01 FAR, which outperforms the separate training considered as baseline by 14.8% GAR.

We also investigate the performance of our coupled learning framework for four different cross-spectral and cross-resolution scenarios. We plot the resulting ROC curves in Fig. 5.7(c). Table 5.4 summarizes the EER results comparing our proposed approach with other state-of-the-art deep learning iris recognition method proposed in [168, 169, 171, 182] for the same train-test protocol. We notice that when we match the HR VIS iris probe image against a HR NIR iris gallery, our cpGAN achieves superior recognition performance over the other baseline matching algorithms.

Table 5.4: Comparative performances on the Cross-eyed-cross-spectral iris recognition database. Symbol '-' indicates that the metric is not available for that protocol.

| Algorithm | Matching | GAR@FAR=0.1 | GAR@FAR=0.01 | GAR@FAR=0.001 | EER (%) |
|---|---|---|---|---|---|
| CNN with SDH [168] | HR VIS vs HR NIR | — | 87.18 | — | 6.34 |
| NIR to VIS texture synthesis using MRF model [169] | HR VIS vs HR NIR | 78.13 | — | — | 18.40 |
| IrisCode using 1D Log-Gabor filter [171] | HR VIS vs HR NIR | 70.3 | — | — | 19.48 |
| **cpGAN(Ours)** | **HR VIS vs HR NIR** | **96.30** | **89.4** | **81.8** | **1.1** |
| cpGAN(Ours) | HR VIS vs LR NIR | 90.3 | 81.7 | 79.6 | 1.28 |
| cpGAN(Ours) | HR NIR vs LR VIS | 86.40 | 78.4 | 72.3 | 1.31 |
| cpGAN(Ours) | LR NIR vs LR VIS | 81.80 | 62.0 | 59.0 | 2.55 |
| NIR to VIS domain translation (Ours cGAN) | Scenario 1 | 90.30 | 80.09 | 70.1 | 1.54 |
| Joint domain translation & super-resolution (Ours Modified cGAN) | Scenario 2(a) | 80.8 | 74.8 | 67.02 | 1.71 |
| Domain translation & super-resolution (Separate Training) | Scenario 2(b) | 71.30 | 60.0 | 54.90 | 3.04 |
| VIS to NIR domain translation (Ours cGAN) | Scenario 3 | 79.0 | 66.8 | 63.8 | 2.17 |
| Iriscode (OSIRIS) [182] | HR VIS vs HR NIR | 60.0 | 51.5 | 44.8 | 3.9 |
| Iriscode (OSIRIS) [182] | LR VIS vs LR NIR | 53.1 | 44.2 | 38.8 | 5.67 |

It obtains almost 26% and 18.17% more identification accuracy compared to the approach used in [171] and [169], respectively. In addition, it also outperforms the most competitive cross-spectral iris recognition approach [168] in the literature by a remarkable decrease of 5.24% in EER. All the other scenarios achieve EER less than 2%, which reveals the robustness of our coupled network. Again, even if we consider a LR NIR iris probe matched against a LR NIR iris gallery, we observe it performs much better than the benchmark using IrisCode [182] for the same scenario.

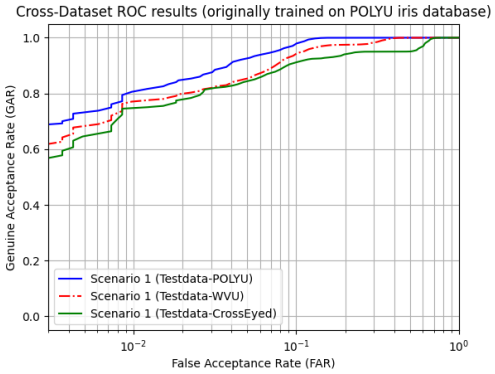
## 5.6 Cross-Database Performance Evaluation

One of the most promising benefits of deep-learning-based iris recognition is its generalization capability, which offers high matching performance even when using the model trained on completely different iris database. Therefore, we also evaluate cross-database matching performance to validate the generalization capability of both of our approaches.

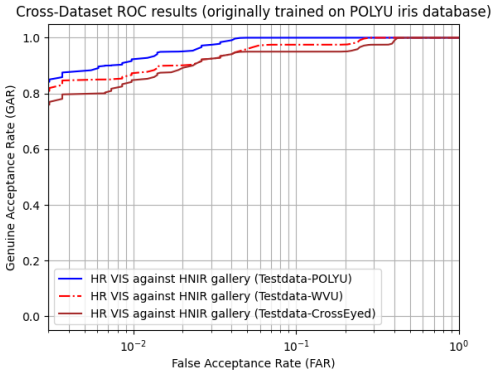
During this cross-database performance evaluation, first, we directly employ one of our models that has been trained on the PolyU bi-spectral database to ascertain the verification performance for the WVU face and iris database and Cross-eyed-cross-spectral iris recognition database without any fine-tuning. More specifically, we have used one dataset for training, and disjoint dataset for testing. Next, we follow the same technique to perform cross-database matching for the other two datasets: we use a model trained on the WVU face and iris image database to evaluate the recognition performance for the PolyU and Cross-eyed database, and similarly, for a model that is trained using the Cross-eyed dataset. We maintain the same test-protocol as described for the

respective databases in previous sections. For matching we consider only Scenario 1 when evaluating the performance of the cGAN architecture. To report evaluation of the cpGAN network, we specifically consider the scenario where the HR VIS iris probe is matched against a HR NIR gallery. We have already introduced both of these scenarios in the earlier sections.

The aim of this evaluation is to validate the generalization capability of our proposed frameworks when the target iris database has limited training samples. We show the comparative performance from the respective databases in Fig. 5.8-5.10 and report respective EER values in Table 5.5-5.7 from this cross-database performance evaluation. These results for the cross-database matching also indicate the performance improvement gained by employing our framework and reveal its generalization capability.




((a))

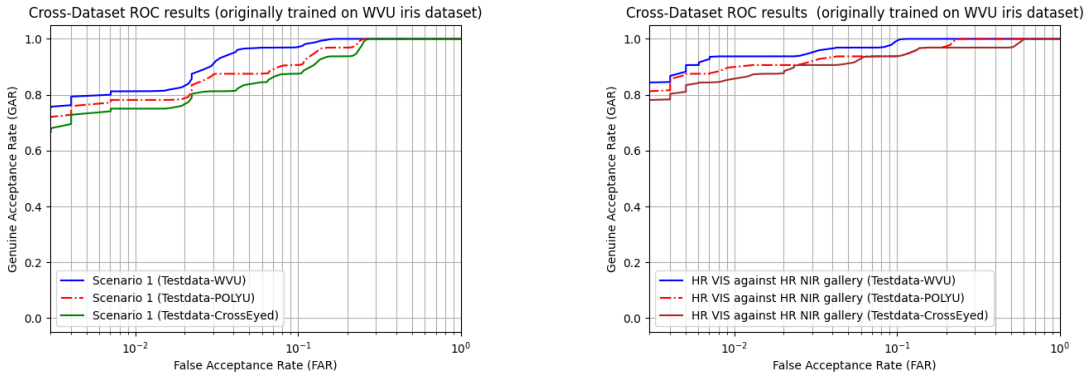



((b))

Figure 5.8: Comparative ROC results showing the cross-database matching of our approach (a) Scenario 1 with cGAN architecture (b) cpGAN architecture (matching the HR VIS iris probe against a HR NIR gallery) where both networks were trained only on the PolyU bi-spectral dataset.

Table 5.5: Cross-dataset matching performance evaluation. We trained both networks only on the PolyU bi-spectral dataset.

| Approach | Iris Comparison | Test Dataset | GAR@FAR=0.01 | GAR@FAR=0.001 | EER (%) |
|---|---|---|---|---|---|
| **cpGAN(ours)** | HR VIS vs HR NIR | **PolyU Bi-Spectral** | **92.38** | **84.98** | **1.02** |
| | | WVU Face and Iris | 88.9 | 81.7 | 1.13 |
| | | Cross-eyed-cross-spectral | 85.2 | 77.5 | 1.20 |
| **cGAN(ours)** | Scenario 1 | PolyU Bi-Spectral | 80.5 | 70.1 | 1.5 |
| | | WVU Face and Iris | 77.8 | 62.9 | 1.64 |
| | | Cross-eyed-cross-spectral | 75.5 | 57.9 | 1.72 |




((a)) ((b))

Figure 5.9: Comparative ROC results showing the cross-database matching of our approach (a) Scenario 1 with cGAN architecture (b) cpGAN architecture (matching the HR VIS iris probe against a HR NIR gallery) where both networks were trained only on the WVU face and iris dataset.

Table 5.6: Cross-dataset matching performance evaluation. We trained both networks only on the WVU face and iris dataset.

| Approach | Iris Comparison | Test Dataset | GAR@FAR=0.01 | GAR@FAR=0.001 | EER (%) |
|---|---|---|---|---|---|
| **cpGAN(ours)** | HR VIS vs HR NIR | **WVU Face and Iris** | **93.0** | **84.0** | **0.90** |
| | | PolyU Bi-Spectral | 89.40 | 81.9 | 1.13 |
| | | Cross-eyed-cross-spectral | 85.8 | 78.5 | 1.18 |
| **cGAN(ours)** | Scenario 1 | WVU Face and Iris | 80.8 | 75.1 | 1.54 |
| | | PolyU Bi-Spectral | 79.6 | 68.9 | 1.60 |
| | | Cross-eyed-cross-spectral | 76.0 | 67.5 | 1.66 |

Table 5.7: Cross-dataset matching performance evaluation. We trained both networks only on the Cross-eyed-cross-spectral iris recognition dataset.

| Approach | Iris Comparison | Test Dataset | GAR@FAR=0.01 | GAR@FAR=0.001 | EER (%) |
|---|---|---|---|---|---|
| **cpGAN(ours)** | HR VIS vs HR NIR | **Cross-eyed-cross-spectral** | **89.4** | **81.8** | **1.1** |
| | | PolyU Bi-Spectral | 82.30 | 74.80 | 1.21 |
| | | WVU Face and Iris | 81.5 | 71.8 | 1.26 |
| **cGAN(ours)** | Scenario 1 | Cross-eyed-cross-spectral | 80.09 | 70.1 | 1.54 |
| | | PolyU Bi-Spectral | 71.5 | 68.9 | 1.75 |
| | | WVU Face and Iris | 69.2 | 64.4 | 1.9 |

## 5.7 Ablation Study

Training a GAN-based architecture is always difficult due to the GAN's natural instability. Additional loss functions in guiding the GAN training can significantly improve the performance. However, these loss terms in the total combined loss are inconsistent on a numerical scale. Therefore, we use hyperparameters as weight factors to numerically balance the magnitude of different

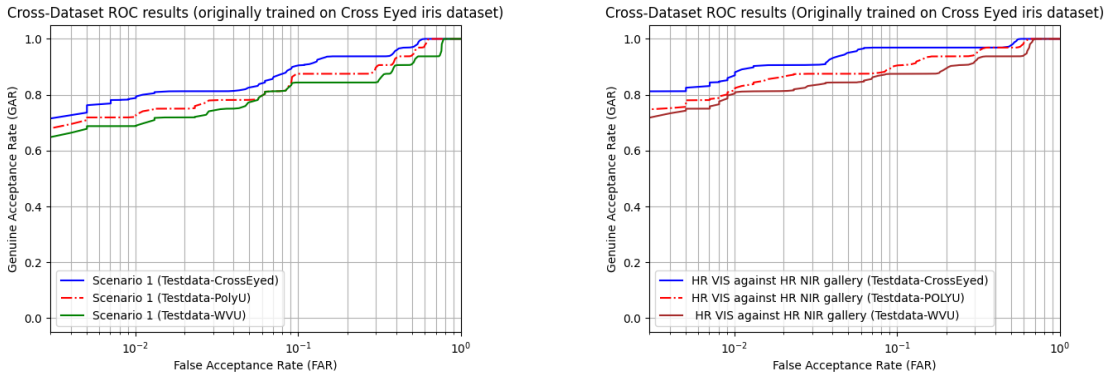



Figure 5.10: Comparative ROC results showing the cross-database matching of our approach (a) Scenario 1 with cGAN architecture (b) cpGAN architecture (matching the HR VIS iris probe against a HR NIR gallery) where both networks were trained only on the Cross-eyed-cross-spectral iris recognition dataset.

| Dataset | PolyU Bi-Spectral | |
|---|---|---|
| Iris Comparison | Scenario 1 (HR VIS vs Synthesized VIS) | |
| Hyperparameter Settings | GAR@FAR=0.01 | GAR@FAR=0.001 |
| $\lambda_1 = 10^0, \lambda_2 = 2 \times 10^{-3}$ | 68.9 | 54.6 |
| $\lambda_1 = 10^{-2}, \lambda_2 = 2 \times 10^{-3}$ | 75.7 | 63.5 |
| $\lambda_1 = 10^{-4}, \lambda_2 = 2 \times 10^{-3}$ | 78.1 | 66.8 |
| $\mathbf{\lambda_1 = 10^{-6}, \lambda_2 = 2 \times 10^{-3}}$ | **80.5** | **70.1** |
| $\lambda_1 = 10^{-6}, \lambda_2 = 2 \times 10^{-2}$ | 72.7 | 61.4 |
| $\lambda_1 = 10^{-6}, \lambda_2 = 2 \times 10^{-1}$ | 70.1 | 58.5 |

((a))

| Dataset | PolyU Bi-Spectral | |
|---|---|---|
| Iris Comparison | HR VIS vs HR NIR | |
| Hyperparameter Settings | GAR@FAR=0.01 | GAR@FAR=0.001 |
| $\lambda_3 = 1, \lambda_4 = 1, \lambda_5 = 0.3$ | 87.3 | 74.2 |
| $\lambda_3 = 1, \lambda_4 = 0.7, \lambda_5 = 0.3$ | 89.7 | 78.9 |
| $\lambda_3 = 1, \lambda_4 = 0.5, \lambda_5 = 0.3$ | 90.1 | 81.8 |
| $\mathbf{\lambda_3 = 1, \lambda_4 = 0.3, \lambda_5 = 0.3}$ | **92.38** | **84.98** |
| $\lambda_3 = 1, \lambda_4 = 0.3, \lambda_5 = 0.5$ | 89.3 | 76.6 |
| $\lambda_3 = 1, \lambda_4 = 0.3, \lambda_5 = 0.7$ | 85.4 | 74.0 |
| $\lambda_3 = 1, \lambda_4 = 0.3, \lambda_5 = 0.1$ | 87.1 | 71.9 |

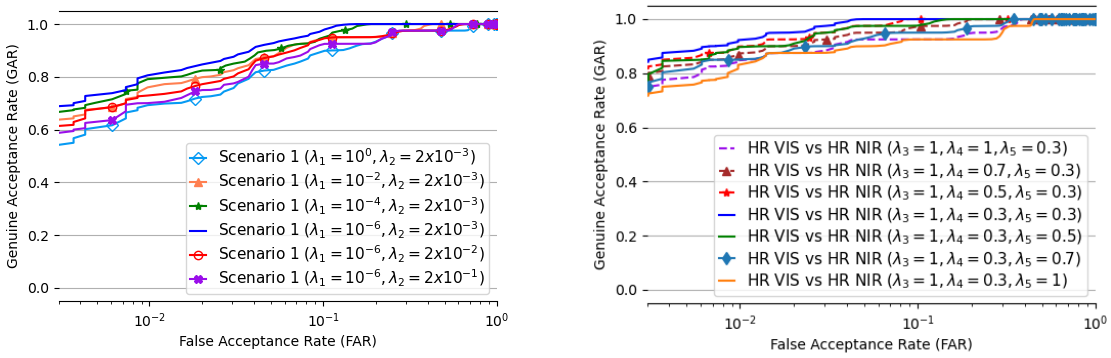
((b))

Table 5.8: Matching performance of our proposed (a) cGAN (b) cpGAN using different hyperparameters settings on the PolyU Bi-Spectral test dataset

losses which accelerates the total loss convergence. To determine the optimal hyperparameters for our both cGAN and cpGAN models, we conduct an ablation study through changing the value of hyperparameters : $\lambda_1, \lambda_2$, and $\lambda_3, \lambda_4, \lambda_5$ adapted in equation (5.19) and (5.20), respectively. We have summarized the analysis in Table 5.8, and show the match performance in Fig. 5.11.

### 5.7.1 Hyperparameter Analysis

We evaluate the sensitivity of match performance when hyperparameters are varied across a range for training our proposed cGAN module. Training the cGAN with an $L_2$ term alone might lead to blurry results, since this loss penalizes the squared distance between ground truth outputs and synthesized outputs at pixel level. Since synthesized image quality is our top priority, we have added the ImageNet trained VGG-based perceptual loss, which is effective at generating realistic




((a)) ((b))

Figure 5.11: Comparative ROC results showing the sensitivity of matching performance on the hyperparameters of our proposed (a) cGAN, when it is trained only on the PolyU Bi-Spectral dataset for Scenario 1 (b) cpGAN, when it is trained only on the PolyU Bi-Spectral dataset for matching the HR VIS iris probe against a HR NIR gallery.

synthesized images by including more recognizable structure. Therefore, we keep the weight factor of the $L_2$ loss term 1 and train cGAN at $\lambda_1 \in \{10^0, 10^{-6}\}$, and $\lambda_2 = 2 \times 10^{-3}$, which are used as weight factors for adversarial loss term and perceptual loss term, respectively. We have also trained the network for a varied range of $\lambda_2 \in \{2 \times 10^{-3}, 2 \times 10^{-1}\}$, when $\lambda_1 = 10^{-6}$. From the analysis of hyperparameters, as shown in Fig. 5.11(a), and Table 5.8(a), we notice that our proposed cGAN achieves the best matching performance for Scenario 1, when it is trained with $\lambda_1 = 10^{-6}$, and $\lambda_2 = 2 \times 10^{-3}$ on the PolyU Bi-Spectral dataset. We have used this setting to perform all the experiments for cGAN and reported the obtained results in this paper.

For training the cpGAN, we have considered additional constraints, such as $L_2$ loss and VGG-based perceptual loss along with adversarial, and contrastive loss functions. Since we have developed our 2nd method to perform cross-spectral iris matching in the common embedded latent feature subspace, we put more emphasis on contrastive loss, which cares about the distance between genuine pairs and also penalizes mismatch between imposter pairs. Therefore, the weight factor for this loss term remains 1, and other hyperparameters have been changed to stabilize the cpGAN training, which allows it to converge faster, and thoroughly improve performance.

As seen in Fig. 5.11(b), and Table 5.8(b), we keep the adversarial weight factor, $\lambda_3 = 1$, when changing the values $\lambda_4$, and $\lambda_5$ from 0.3 to 1.0, which define weight factors for perceptual, and $L_2$ reconstruction loss term, respectively. From this ablation study, we have observed that $\lambda_4 = \lambda_5 = 0.3$

obtains the best matching accuracy, when the HR VIS iris probe is matched against the HR NIR iris gallery for the PolyU Bi-Spectral dataset (see Fig. 5.11(b) and Table 5.8(b)). For fair comparison, we have used these settings to train the cpGAN for other datasets and reported the results in this paper.

## 5.8 Limitation of the iris image acquisition method on the observed results

The quality of iris images affects the matching performance of any iris recognition system, which indicates the significant role of the iris acquisition process. It is the most initial part of any typical iris recognition system. During the acquisition of iris images, one must maintain an ISO standard iris image format (iris diameter has to be 150 pixels [212]), which is not easy to achieve in many data acquisition environments. Most of the commercial iris image acquisition systems are designed to work at a close range and maintain a small operating distance, which is less than 1 meter [213]. Moreover, all of them need users' cooperation. Therefore, it has become troublesome to capture iris images at a distance to generate low-resolution iris images in realistic environments. Therefore, there are no datasets available to study the effect cross-resolution and cross-spectral mismatch on iris recognition systems in the literature. To overcome this limitation to some extent, the researchers developing state-of-the-art iris recognition systems have resized the original high-resolution iris images to their desired low-resolution images. In our work, we first apply a Gaussian filter and then resize the iris image using a bicubic interpolation method. We assume that these artificially-generated low-resolution images have similar characteristics as the original low-resolution images. However, we cannot certainly say that we would have achieved exactly similar performance if we used the original low-resolution images. We have tried to obtain low-resolution iris images as close as possible to a realistic setting. These results can be considered as a baseline for further improvement if the low-resolution iris images can be acquired in a realistic setting.

## 5.9 Conclusion

In this paper, we have described the development of two different deep learning frameworks for cross-spectral and cross- resolution iris recognition. While both frameworks are developed based on domain transformation, one of them functions by translating from one domain to the another (NIR to VIS or vice versa), and the other framework transforms both domains to a latent embedding subspace. Briefly stated, in our first approach, we have introduced a domain translation network which can be considered as preprocessing step for any off-the-shelf iris recognition system. In addition, we have proposed a new joint translation and super-resolution technique to address cross-resolution iris matching under the cross-domain problem. Experimental results on three publicly available cross-spectral datasets indicate the superiority of our proposed method over the earlier methods presented in the literature. This paper also investigates the domain invariant capability of our proposed cpGAN framework, which projects both the VIS and NIR iris texture features into a common latent embedding subspace to perform matching in the embedded domain. The goal of this network is to maximize the pair-wise correlation via contrastive loss during projection for more accurate cross-spectral iris matching. Results reported in Section 5.5 show significant improvement in the matching accuracy compared to other deep learning cross-spectral iris recognition algorithms. For instance, cpGAN achieves improvements of approximately 33%, when compared to the results reported in [211] for the PolyU Bi- Spectral dataset. Finally, we perform cross-database iris matching under the cross-spectral domain to evaluate the generalization capability of our methods.

# Chapter 6

# Pose Attention-Guided Profile-to-Frontal Face Recognition and Reconstruction of Frontal Faces

## 6.1 Introduction

The advent of deep convolutional neural networks (CNNs) has led to promising achievements in unconstrained face recognition and verification techniques [14, 15]. It has even surpassed human performance on several benchmark datasets [17]. However, a challenge that still remains to be solved, is that of extreme pose variations, which degrade frontal-to-profile face verification accuracy by more than 10% compared to frontal-to-frontal matching accuracy. [214]. The most prominent factors contributing to this performance degradation can be classified into three categories:

- Facial appearance distortion: In comparison to controlled environment, real-world profile faces have different imaging conditions besides pose such as expression, occlusion, and illumination variations as shown in Figure 1. These variations cause substantial changes in facial appearance, which indicates a loss of consistent information useful for face recognition.
- Missing Semantic Consistency: When a face view is changed from frontal to profile, the position and shape of facial texture varies nonlinearly, which inevitably introduce loss of semantic correspondence in 2D images along with confusion in interpersonal texture differ-

Figure 6.1: Frontal and Profile faces in the IJB-A Dataset under full pose variation, expression, and different imaging conditions.

ences [215]. In consequence, extracted features from two images at different poses are no longer similar and cannot provide high matching accuracy as expected in conventional face recognition methods.

- Imbalance pose distributions: Deep learning based face recognition algorithms extensively rely on very large datasets, which usually suffer from uneven pose distributions. This data imbalance continues to impose challenges for face recognition, which forces the model to lean towards frontal images than profile face image of a person resulting in poor matching accuracy during frontal-to-profile verification. In contrast, human can easily identify faces with extreme pose variations without significant drop in accuracy.

To address this performance gap between human and automatic models, traditional methods apply several local descriptors such as Gabor [216], Haar [217], and LBP [218] to measure local distortions and then adopt metric learning techniques [219, 220]. On the other hand, another research community emphasizes on frontal view synthesis across poses. They utilize 3D geometrical transformations [14, 39, 221] to reduce pose variations. Moreover, multiple novel architectures have been proposed [46, 222, 223] for face normalization, which aligns faces to a canonical pose. Although, they show impressive performance at normalizing small pose faces, their accuracy drops severely under extreme pose conditions. To handle this problem, some researchers opt on learning pose-robust features [41, 224, 225] for multi-view face images. Among them, Cao et al. [224] propose a lightweight DREAM block which learns specific feature transformation to perform frontalization in feature space, while others explore multi-task learning to perform pose-invariant face recognition (PIFR) [41, 44].

In this paper, we introduce a novel method to learn discriminative pose-invariant representation

in a deep feature embedding subspace without performing profile face normalization (frontalization) or learning disentangled features. Instead, we explicitly deal with the pose variability by incorporating it as an auxiliary information to the feature extraction network. We hypothesize that learning with side information allows for better generalization of the primary task by assisting it to focus on the current context and ignore unnecessary information. To this end, we first develop a deep coupled profile to frontal network using the contrastive loss, which is able to learn mapping internal representation from faces into a compact 512-dimensional embedding subspace. Second, to incorporate pose as an auxiliary signal, we propose an easy-to-implement pose attention block (PAB), which automatically infers significant features from profile faces along channel and spatial axes in deeper layers of the network. In other words, PAB is designed to empirically guide to learn discriminative and pose-invariant features in an embedding subspace. Moreover, we also investigate the capability of these learned embedding features via a generative adversarial network (GAN). In a summary, this paper offers the following contributions:

- A novel coupled profile to frontal PIFR model utilizing pose as an auxiliary information is developed.
- A pose attention block (PAB) module using a pretrained pose-estimation network is proposed to guide a discriminative and pose-invariant feature learning framework in an embedding subspace.
- Extensive experiments on different benchmark datasets and comparison to other state-of-the-art methods have been performed to validate the effectiveness of our proposed PIFR network.
- Capability of the embedding features learned in our proposed network is explored for frontal face synthesis via a GAN model, which indicates its usefulness in different face analysis tasks apart from face recognition.

## 6.2 Related Work

### 6.2.1 Face Frontalization

Face frontalization has become an extremely challenging task due to the self-occlusion that exists in 2D projections of the input face with large pose variations. To address this problem, traditional methods use 3D based models [39, 221, 226], statistical approaches [40], and deep learning based methods [227–230] for face frontalization. Hassner et al. [39] used a 3D face model to generate frontal shape of all input faces. Although it is proved to be efficient in face frontalization task, it cannot achieve expected accuracy for profile and near profile faces, specifically faces with yaw angle greater than 60°. A statistical model is proposed in [40], which solves a constrained low-rank minimization problem to jointly perform frontal view reconstruction and landmark detection. Recently, deep learning based methods have shown outstanding performances in frontal face synthesis. In [228], a recurrent transform unit is proposed to reconstruct discrete 3D views. Yim et al. [229] applied a concatenated network structure to rotate a non-frontal face, where they regularize the output by image level reconstruction loss. With the emergence of GAN, researchers have concentrated more on GAN-based methods, which has advanced the performance of face frontalization methods. However, face frontalization is considered as an image-level pose-invariant representation, which can improve PIFR performance mostly for face images at near frontal or half profile.

### 6.2.2 Pose Invariant Representation Learning

Pose-invariant feature representation has been recently used as a mainstay of many face recognition tasks. Earlier works apply canonical correlation analysis (CCA) [231] to analyze the shared characteristics among pose-invariant samples. Recent deep learning based approaches focus on several aspects while training a network. To name, in [230], a deep neural network is trained to separate face identity from viewpoints. Kan et al. [232] propose feature pooling across different poses to allow a single network structure for inputs at multiple pose views. To disentangle poses in feature representation, several methods [41, 233] carefully factorize out the non-identity part. Authors in [234, 235] mostly consider fusing information at the feature level or distance metric level.

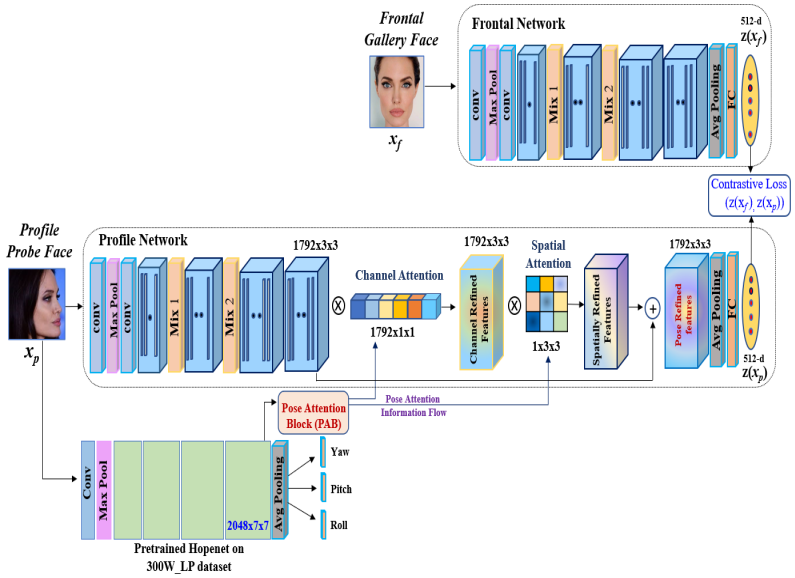



Figure 6.2: Block diagram of our proposed deep coupled profile-to-frontal PIFR network

On the other hand, Cao et al. [224] propose a pose discrepancy corrector module, recently followed by many researchers [45, 236] to empirically perform frontalization in feature space. Contrary to these approaches, we mostly concentrate to utilize pose as side information via attention mechanism and guide the network to learn discriminative, and pose-invariant features in an embedding subspace.

## 6.3 Proposed Method

Here, we describe our proposed method which offers a new perspective of learning pose-invariant feature representation via incorporating pose specific auxiliary information into deep subspace learning profile to frontal face verification network. Inspired by the success of unified face recognition system [17] in learning embedding from faces, we develop a deep coupled framework as shown in Figure 6.2 to learn mapping from both frontal and profile faces to a compact feature embedding subspace. Since profile faces have large pose variations, we exploit this angular knowledge to explore how auxiliary pose information can improve the embedding feature representation for profile faces. To implement this perspective, we propose the PAB module, which helps the network to sequentially refine features in both channel and spatial dimension. In this section, first we discuss the implementation technique of our PAB module, and then we detail how we integrate

the auxiliary pose information to our deep coupled network.

### 6.3.1 Pose Attention Block (PAB)

We adopt a robust pose estimation network, i.e., Hopenet [237], which has been trained on a large synthetically expanded dataset 300W-LP [238]. Hopenet uses ResNet50 as backbone of their architecture and adds three fully connected layers to predict intrinsic Euler angles (yaw, pitch and roll) directly from input off-angle face images as illustrated in Figure 6.2. To implement our proposed PAB module, we take a feature map of size $2048 \times 7 \times 7$ from the last convolutional layer of Hopenet, that already provide us with more complex abstract features such as overall shapes, pose and texture of the input face.

Our proposed PAB module consists of two sequential attention modules: adaptive channel attention module (ACAM), and spatial attention module (SpAM) to emphasize meaningful features and suppress unnecessary ones along channel and spatial axes, respectively. Briefly, for a given input feature map, ACAM and SpaM attend to all channel and spatial locations through inferring a 1D channel attention map and a 2D spatial attention map. Figure 6.3 illustrates the framework of the proposed PAB, that is integrated with our deep coupled learning framework in Figure 6.2. We now discuss each component in detail.

#### Adaptive Channel Attention Module (ACAM)

Given an input feature map, $x \in \mathbb{R}^{C\times H\times W}$ of size $2048 \times 7 \times 7$, ACAM applies average-pooling and max-pooling operation like CBAM [239] to learn inter-channel dependencies and generates two different spatial context descriptors: $x^c_{avg}$, and $x^c_{max}$, respectively. To integrate spatial information, they are forwarded to a shared multilayer perceptron ($MLP_1$), which is typically a two layer fully connected network as shown in Figure 6.3(a). Since our ultimate goal is to distill features for the face recognition task along the channel dimension to highlight *which of the feature maps are relevant to shape*, we set hidden size 128 and output in a way such that it can generate a 1D channel attention map of 1792 consistent with the feature map depth size in our FR network. After that, we use element-wise summation to merge feature vectors obtained from the shared $MLP_1$ network. In a summary, the channel attention is computed as follows:

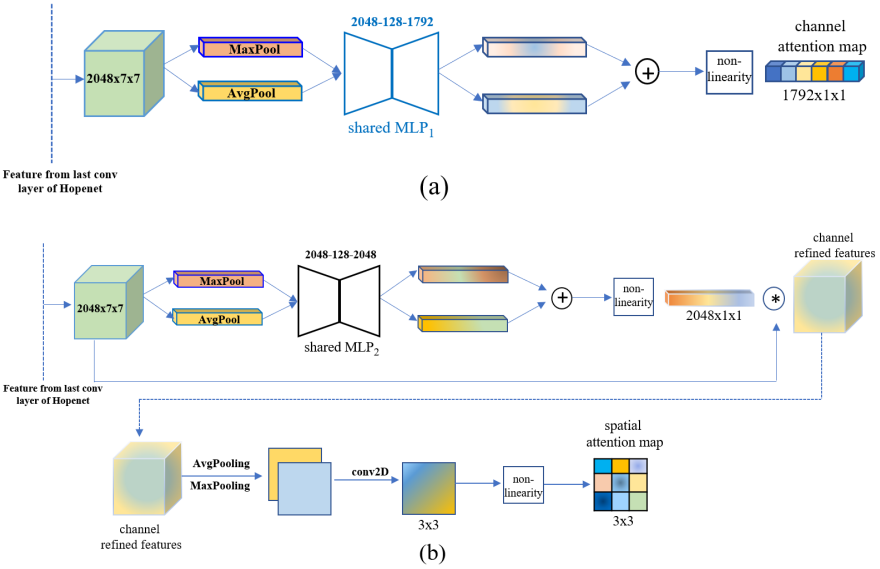



Figure 6.3: Block diagram of our PAB (a) Adaptive Channel attention module (ACAM) (b) Spatial Attention Module (SpAM)

$$M_c(x) = \sigma(MLP_1(x^c_{avg}) + MLP_1(x^c_{max})), \tag{6.1}$$

where $\sigma$ is the non-linear activation function.

**Spatial Attention Module (SpAM)**

Following CBAM [239], we design our spatial attention module (SpAM) to focus on the informative region in the spatial domain. In our SpAM, we compute a 2D spatial attention map of size $3 \times 3$ to be consistent with the feature map size of FR network as follows:

$$F_c(x) = \sigma(MLP_2(x^c_{avg}) + MLP_2(x^c_{max})), \tag{6.2}$$

$$M_s(x) = \sigma(conv2D[F^s_c(x)_{avg}, F^s_c(x)_{max}]), \tag{6.3}$$

where $F_c(x)$ denotes channel refined features in Figure 6.3(b), which is obtained using another shared $MLP_2$, and conv2D refers to the convolution on the concatenation of the average pooled feature map, $F^s_c(x)_{avg}$ and max pooled feature map, $F^s_c(x)_{max}$ along the channel axes, which ensures inter-spatial relationship of features.

### 6.3.2 Profile to Frontal Coupled Subspace Learning Network

Our goal is to learn a discriminative, pose-invariant feature representation from a pair of face images to a compact embedding subspace such that we can perform recognition of profile face images with respect to a gallery of frontal face images in the embedded domain. Therefore, to learn a rich feature representation, we propose a coupled deep convolutional network guided by a pose attention module. Our method adopts a variant of InceptionResnet [240] architecture that is used in FaceNet [17] as the core element of our network, which is pretrained on VGGFace2 [241].

The model structure in Figure 6.2 illustrates that a pair of images goes through the coupled network consisting of two dedicated branches to extract features from both frontal and profile images. Since there exists pose variations in profile faces, we hypothesis that we can leverage pose as an auxiliary information to improve the ability of extracting highly discriminative features from these profile faces. To accomplish this, the profile image is also fed to a pretrained Hopenet pose estimation network, which provides pose attended information via our PAB module to sequentially distill features along both channel and spatial dimension of this network.

Previous section explains the block design of our PAB module. It consists of two sub modules: (1) ACAM, that generates a 1D channel attention map of 1792 to refine the feature map ($1792 \times 3 \times 3$) of our profile coupled network along the channel dimension, and (2) SpAM, which produces a 2D spatial attention map of size 3x3 to spatially attend the informative region in the feature maps of our profile network as in Figure 6.2. For more details note that, both 1D and 2D attention maps are multiplied with the feature maps of the profile network for adaptive feature refinement. In addition, identity mapping after the pose informative feature aggregation results in purely distilled relevant features before average pooling, which we term as *pose refined features*.

Such sharing of pose as auxiliary information during feature extraction from profile faces results in informative and task relevant features, which otherwise would not have been attained from training only with a massive labelled faces. In addition, it also allows for better generalization of the PIFR task by looking at new interpretations of the features. Once the embeddings are established as feature vectors, we optimize the network via class-specific contrastive loss, which tries to minimize squared Euclidean distance between the features of positive pairs (i.e., when profile and frontal image share the same identity) and maximize it for negative pairs (i.e., when profile and

frontal image comes from different identities).

## 6.4 Loss Function

Our goal is to learn a compact 512-D embedding subspace by coupling two mapping networks, one for frontal and another one for profile face image, via a distance-based contrastive loss, $L_{cont}$ [201]. We compute this loss metric, $L_{cont}$ over a set of genuine (i.e., a profile face image of a subject with its corresponding frontal face image) and imposter (i.e., a profile face image of a subject and a frontal face image of a different subject) pairs such that images belonging to the same identity (genuine pair) are embedded as close as possible, and, simultaneously, images of different identities are pushed away from each other in the common embedded subspace. The contrastive loss function is formulated as:

$$L_{cont}(z(x_p^i), z(x_f^j), Y) = (1 - Y)\frac{1}{2}(D_z)^2 + (Y)\frac{1}{2}(\max(0, m - D_z))^2, \tag{6.4}$$

where $x_p^i$ and $x_f^j$ denote the input profile and frontal face images, respectively. The variable Y is a binary label, which is equal to 0 if $x_p^i$ and $x_f^j$ belong to the same class (i.e., genuine pair), and equal to 1 if $x_p^i$ and $x_f^j$ belong to the different class (i.e., impostor pair). z(.) is used to denote the mapping function for $x_p^i$ and $x_f^j$ into a compact embedding subspace. To tighten the constraint, m is used as contrastive margin.

The Euclidean distance, $D_z$, between the embedding features, $z(x_p^i)$ and $z(x_f^j)$, is given by:

$$D_z = \left\| z(x_p^i) - z(x_f^j) \right\|_2 . \tag{6.5}$$

Therefore, if Y = 0 (i.e., genuine pair), then the contrastive loss function ($L_{cont}$) is given as:

$$L_{cont}(z(x_p^i), z(x_f^j), Y) = \frac{1}{2} \left\| z(x_p^i) - z(x_f^j) \right\|_2^2 , \tag{6.6}$$

and if Y = 1 (i.e., impostor pair), then contrastive loss function ($L_{cont}$) is :

$$L_{cont}(z(x_p^i),z(x_f^j), Y) = \frac{1}{2}\max\left(0, m - \left\|z(x_p^i) - z(x_f^j)\right\|_2^2\right). \tag{6.7}$$

Thus, the total loss to optimize the entire network is denoted by $L_{total}$ for coupling both the profile and frontal face in the embedded domain:

$$L_{total} = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} L_{cont}(z(x_p^i), z(x_f^j), Y), \tag{6.8}$$

where N is the number of training samples. The main purpose of using the contrastive loss is to be able to use the class labels, and margin to ensure discriminative embedding subspace, which may not be obtained with some other metric such as the Euclidean distance. Finally, we use this pose attended discriminative embedding subspace for matching of the profile images with the frontal images.

## 6.5 Experiments

In this section, we describe our implementation details and the datasets that we have used to conduct our experiments. To evaluate the performance of our proposed PIFR network, we experiment under two settings: (1) face identification on controlled Multi-PIE [242] face dataset, and (2) face verification/identification on in-the-wild datasets including CFP [214] and IJB-A [243] with their official evaluation protocols. In addition, we also report face recognition accuracy compared to several state-of-the-art results on these datasets.

### 6.5.1 Datasets

**Multi-PIE** : The Multi-PIE dataset is the largest dataset released for multi-view face recognition with respect to controlled variations in illumination and expressions across different poses. It contains 754,204 images of 337 identities, captured at 15 view points ranging from –90° $\sim$ +90°, over 20 illumination conditions. To evaluate our proposed method for identification task, we conduct

experiments under two settings following protocol used in [46] for fair comparison. **Setting 1** includes images only from session 1 in the Multi-PIE dataset, which has 250 subjects. For training, we choose first 150 identities with 11 poses within $\pm 90^{\circ}$ and 20 illuminations. During testing, one frontal view with neutral expression and illumination (i.e., ID 07) is used in the gallery for each of the remaining 100 identities and other images are considered as probe images. **Setting 2** includes images only with neutral expression over all four sessions providing 337 identities. To train our network, we use first 200 identities, while rest of the 137 IDs have been used for testing. We maintain similar setup as setting 1 for our gallery and probe.

**CFP** : The Celebrities in Frontal-Profile (CFP) dataset is introduced to handle large-pose variations. It contains identities of 500 celebrities, which have been collected under constrained (i.e., images at different pose, illumination and expression) and unconstrained (i.e., images collected from the Internet) settings. For each celebrity, it includes 10 frontal and 4 profile images. Following their standard 10-fold evaluation protocol [214], we split the dataset into 10 folds, each with 350 genuine and 350 imposter pairs to perform both frontal-to-frontal (FF), and frontal-to-profile (FP) verification task.

**IJB-C**: The IARPA Janus Benchmark–C (IJB-C) [244] face dataset has been released to advance the unconstrained face recognition by modeling more practical face recognition use cases. It is an extension to the publicly available IJB-B [245] dataset, which contains $3,531$ subjects with extreme variations in expression, illumination, geographic origin, and more. In total, it has $31,334$ still images and $1,17,542$ video frames collected in unconstrained settings with different protocols. To evaluate our algorithm's ability, we perform both face verification (1:1), and identification (1:N) tasks following their protocol.

**M2FPA**: Traditional face recognition databases typically rely on a two-dimensional image of a person's face, which can make it difficult to accurately identify individuals when they are not facing the camera directly. However, by incorporating information about the pitch and yaw angles of the face, the system can more accurately identify an individual even when they are not facing the camera directly. In this direction, Multi-yaw Multi-pitch high-quality database for Facial Pose Analysis (M2FPA) [246] introduces images with joint yaw and pitch variations. M2FPA includes a total of 397,544 images of 229 subjects with 62 poses, 4 attributes, and 7 illuminations. Pose

variations include 13 yaw angles (-90°to + 90°), 5 pitch angles (-30°to +45°), and 44 joint yaw and pitch variations. We have considered 162 subjects for training and 62 subjects for testing.

### 6.5.2 Implementation Details

To implement our proposed coupled learning framework, we have used InceptionResnet-v1 [240] pretrained on VGG-Face2 dataset. Since it is difficult to train the entire network from scratch, we freeze all the trained layers before average pooling for both frontal and profile mapping modules as shown in Figure 6.2. At the same time, our PAB module provides a pose attended 1D channel attention map, and a 2D spatial attention map to assist the network to use only the relevant features while extracting deep features from the profile faces. Therefore, the gradient also flows through this PAB module to update its weights during optimization. Note that, since misleading pose information can misguide the training, we don't train Hopenet, which has been already trained on a very large dataset, and proved to be an efficient pose estimation model. The entire framework has been implemented in Pytorch. We used a batch size of 32 and the Adam optimizer [247] with first-order momentum of 0.5, and learning rate of $10^{-3}$. For training, we generate same number of genuine, and imposter pairs from frontal, and profile images of the same/different subjects to avoid biasness towards positive pairs.

### 6.5.3 Evaluations on the Multi-PIE Benchmark

To show the effectiveness of our proposed method, first we evaluate our model on a controlled database, Multi-PIE for profile to frontal pose-invariant face recognition task in the deep embedding subspace under two different settings. We compare our method with several state-of-the-art PIFR algorithms including HPN [43], c-CNN [248], PIM [46], FNM [46], and competitive GAN-based methods : TP-GAN [249], CAPG-GAN [250], and PF-cpGAN [251].

Table 6.1 shows our rank-1 recognition accuracy compared to other approaches across full yaw variations and illuminations under setting-1. For this experimental setup, we consistently achieve 100% accuracy over yaw angles $< 75°$, while outperforming other baselines. Even under extreme pose (i.e., $\pm 75°$, and $\pm 90°$), when compared to CAPG-GAN, and PF-cpGAN, we significantly outperform them by achieving average 11.85%, and 3% higher accuracy, respectively.

Table 6.1: Rank-1 recognition rates (%) across poses and illuminations under Multi-PIE Setting-1.

| Method | ±90° | ±75° | ±60° | ±45° | ±30° | ±15° |
|---|---|---|---|---|---|---|
| HPN [43] | 29.8 | 47.5 | 61.2 | 72.7 | 78.2 | 84.2 |
| c-CNN [248] | 47.2 | 60.7 | 74.4 | 89.0 | 94.1 | 97.0 |
| TP-GAN [249] | 64.0 | 84.1 | 92.9 | 98.6 | 99.99 | 99.8 |
| PIM [46] | 75.0 | 91.2 | 97.7 | 98.3 | 99.4 | 99.8 |
| CAPG-GAN [250] | 77.1 | 87.4 | 93.7 | 98.3 | 99.4 | 99.9 |
| FNM+VGG-Face [222] | 41.1 | 67.3 | 83.6 | 93.6 | 97.2 | 99.0 |
| FNM+Light CNN [222] | 55.8 | 81.3 | 93.7 | 98.2 | 99.5 | 99.9 |
| PF-cpGAN [251] | 88.1 | 94.2 | 97.6 | 98.9 | 99.9 | 99.9 |
| Backbone(without attention) | 75.7 | 98.2 | **100.0** | **100.0** | **100.0** | **100.0** |
| **Ours** | **89.5** | **98.7** | **100.0** | **100.0** | **100.0** | **100.0** |

Table 6.2: Rank-1 recognition rates (%) across poses and illuminations under Multi-PIE Setting-2.

| Method | ±90° | ±75° | ±60° | ±45° | ±30° | ±15° |
|---|---|---|---|---|---|---|
| FF-GAN [252] | 61.2 | 77.2 | 85.2 | 89.7 | 92.5 | 94.6 |
| TP-GAN [249] | 64.6 | 77.4 | 87.7 | 95.4 | 98.0 | 98.6 |
| CAPG-GAN [250] | 66.0 | 83.05 | 90.6 | 97.3 | 99.5 | 99.8 |
| DA-GAN [42] | 81.5 | 93.2 | 97.2 | 99.1 | 99.8 | 99.9 |
| PIM [46] | 86.5 | 95.0 | 98.1 | 98.3 | 98.5 | 99.0 |
| Backbone (without attention) | 74.8 | 96.8 | **100.0** | **100.0** | **100.0** | **100.0** |
| **Ours** | **88.8** | **97.7** | **100.0** | **100.0** | **100.0** | **100.0** |

We also assess the performance of our proposed network on faces in Multi-PIE under setting-2, which consists of more challenging face identities than setting-1. Evaluation results, shown in Table 6.2 suggests that our proposed PAB purely assists the face recognition network to achieve 2.3%, and 2.7% increase over the best performing method, PIM [46] in the large pose variations; ±90°, and ±75°, respectively. Apart from this, our network achieves superior performance over the other baseline models [46, 249, 252] in all yaw angles. Similar to setting 1, we note 100% recognition accuracy for near-profile faces. These improvements indicate the efficacy of our method for PIFR in constrained environment.

### 6.5.4 Evaluations on the CFP Benchmark

We evaluate our proposed method on the Celebrities in Frontal-Profile (CFP) dataset to analysis face verification in unconstrained environment. To perform evaluation, we follow the standard 10-

Table 6.3: Performance comparison on CFP dataset. Mean Accuracy and equal error rate (EER) with standard deviation over 10 folds.

| | Frontal-Profile (FP) | | Frontal-Frontal (FF) | |
|---|---|---|---|---|
| Algorithm | Accuracy | EER | Accuracy | EER |
| FV+DML [214] | 58.47(3.51) | 38.54(1.59) | 91.18(1.34) | 8.62(1.19) |
| LBP+Sub-SML [214] | 70.02(2.14) | 29.60(2.11) | 77.98(1.86) | 16.00(1.74) |
| HoG+Sub-SML [214] | 77.31(1.61) | 22.20(1.18) | 85.97(1.03) | 11.45(1.35) |
| FV+Sub-SML [214] | 80.63(2.12) | 19.28(1.60) | 88.53(1.58) | 8.85(0.74) |
| Deep Features [214] | 84.91(1.82) | 14.97(1.98) | 93.00(1.55) | 3.48(0.67) |
| Triplet Embedding [253] | 89.17(2.35) | 8.85(0.99) | 98.88(1.56) | 2.51(0.81) |
| Light CNN-29 [254] | 92.47(1.44) | 8.71(1.80) | 99.64(0.32) | 0.57(0.40) |
| PIM (Light CNN-29) [254] | 93.10(1.01) | 7.69(1.29) | 99.44(0.36) | 0.86(0.49) |
| PR-REM [224] | 93.25(2.23) | 7.92(0.98) | 98.10(2.19) | 1.10(0.22) |
| PF-cpGAN [251] | 93.78(2.46) | 7.21(0.65) | 98.88(1.56) | 0.93(0.14) |
| Backbone (without attention) | 92.57(1.10) | 4.24(0.54) | 97.10(0.11) | 1.5(0.25) |
| **Ours** | **95.67(1.64)** | **2.02**(0.62) | **99.70(0.21)** | **0.55(0.35)** |
| Human | 96.57(1.10) | 5.02(1.07) | 96.24(0.67) | 5.34(1.79) |

fold protocol like other approaches in the literature. We report the mean and standard deviation of accuracy, and Equal Error Rate (EER) over the 10 splits for both frontal-frontal and frontal-profile face verification settings.

Table 6.3 shows a comparison of our method with other state-of-the-art face recognition performance on the CFP benchmark datatset. For fair comparison, first, we consider three different hand-crafted feature extraction techniques: Hog [95], LBP [218], and Fisher Vector [255] along with metric learning techniques Sub-SML [256], and diagonal metric learning (DML) [256]. To compare against deep learning based approaches, we include Deep Features [234], Triplet Embedding [253], Light CNN-29 [254], and recently proposed GAN-based latent feature learning framework, PF-cpGAN [251].

From the results summarized in Table 6.3, we observe that our proposed method slightly improves the FF verification performance compared to other deep learning based techniques. However, it outperforms human performance for this setting, and makes substantial improvement over the conventional hand-crafted features by achieving average 18% higher accuracy with 24% decrease in EER for more challenging FP setting. In addition, when compared to best performing PF-cpGAN, our proposed method improves the accuracy by 1.89% and reduce EER significantly

Table 6.4: Performance evaluation on IJB-C benchmark. Symbol '-' indicates that the metric is not available for that protocol.

| Method | 1:1 Verification | | 1:N Identification | |
|---|---|---|---|---|
| | GAR@ FAR= 0.01 | GAR@ FAR= 0.001 | @ Rank-1 | @ Rank-5 |
| GOTS [244] | 61.99 | 33.4 | 38.5 | 53.8 |
| FaceNet [17] | 81.76 | 66.45 | 69.22 | 78.7 |
| VGGFace [18] | 87.13 | 74.79 | 78.60 | 87.2 |
| CFR-GAN [257] | 86.46 | 74.81 | - | - |
| FNM [222] | 91.2 | 80.4 | 78.6 | 88.7 |
| PR-REM [224] | 90.6 | 80.2 | 77.1 | 87.6 |
| Backbone(without attention) | 89.1 | 79.9 | 71.8 | 81.2 |
| **Ours** | **92.8** | **82.5** | **80.33** | **90.42** |

by 5% for FP verification. We also improve on the performance of PR-REM [224] by 2.5% higher accuracy with approximately 6% lower EER. These findings show that the facial respresentations learned by our coupled deep subspace learning framework with the help of PAB are discriminative, and able to obtain pose-invariance at large pose variations.

### 6.5.5 Evaluations on the IJB-C Benchmark

We further evaluate face recognition (i.e., verification and identification) on another challenging benchmark IJB-C, to validate the superiority of our proposed method in unconstrained environment. We compare with the recent state-of-the-art algorithms CFR-GAN [257], FNM [222], and PR-REM [224], along with prior works [17, 18] in [244] for fair evaluation. As shown in Table 6.4, for profile to frontal verification, we improve the genuine accept rate (GAR) by approximately 7.69%, and 2.1% at the false accept rate (FAR) of 0.001 compared to recent works [222, 257]. Moreover, we also obtain outstanding performances on identification. Specifically, we achieve 1.73%, and 3.23% higher recognition accuracy for rank-1 in comparison to the FNM, and PR-REM, respectively. To show the significant contribution of our proposed PAB module, we compare our results with the backbone network (without attention). It shows our idea of incorporating pose information boosts FP verification performance by 2.60% at 0.001 FAR, and identification accuracy by 8.53% for rank-1.

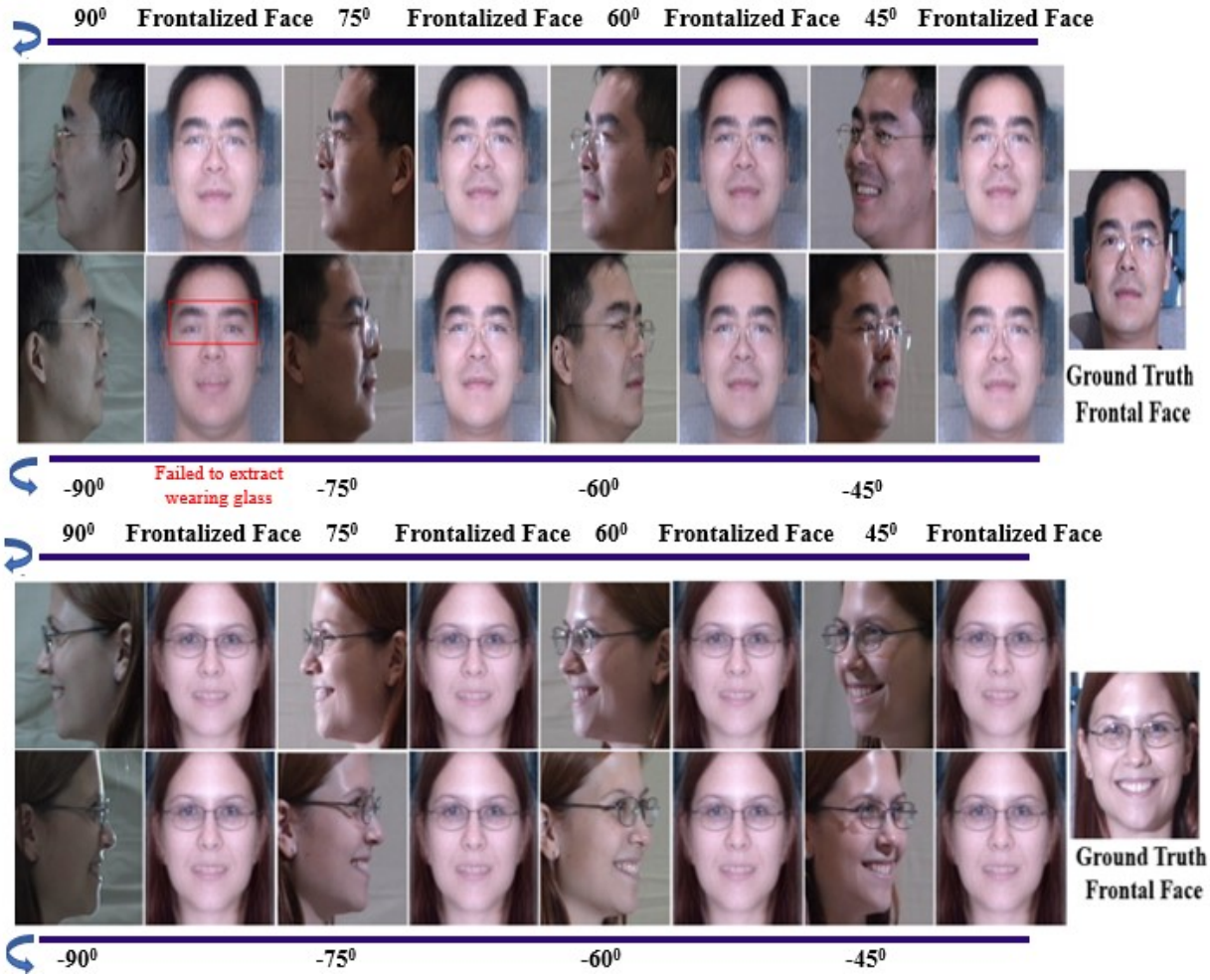



Figure 6.4: Reconstruction results via a GAN [2] model on Multi-PIE across different pose, illumination, and expression using the compact 512-D embedding features learned using our proposed network.

### 6.5.6 Frontal Face Reconstruction from Pose-Invariant Features Learned in Deep Subspace

The purpose of our proposed PAB is to enhance the recognition performance of our coupled deep subspace learning framework via contributing in feature refinement. In addition, class-specific contrastive loss has been used to push the network achieve pose-invariance in the embedding feature domain. To validate our hypothesis, previous sections show comprehensive analysis on verification and identification task for both constrained and in the wild conditions. Apart from recognition task, there are many other scopes to utilize the feature vector learned in the deep subspace. For in-

stance, if we could reconstruct frontal face from the deep features of its corresponding profile face, it can be used in many face analysis tasks including emotion detection, expression tracking etc. Moreover, it has broad applications in vision, graphics, and robotics. Therefore, to demonstrate the usefulness of our proposed method, we adopt a GAN model [2] for reconstruction.

To accomplish this task, we modify residual network used in the two-pathway encode-decoder architecture proposed by Tian et al. [2]. We consider their decoder module with the discriminator network for frontal face synthesis. For training, we select profile and corresponding frontal images from setting 1 of Multi-PIE dataset, which consists of 150 IDs with full yaw variations (-90°to +90°), and remaining 100 IDs for test. First, we extract 512-d embedding feature vector from our proposed PAB guided pose-invariant face recognition network for each of the profile image in the trainset. After that, these profile feature vectors are given as input to the decoder and corresponding frontal faces with no expression and neutral illumination are used as target, which force the network adversarially learn the image distribution of the frontal faces. To generate identity preserving, high visual quality frontal faces from its profile deep features, we incorporate pixel-wise $L_1$ reconstruction error, VGG-16 based Perceptual loss [72], and Light CNN-29 [254] network for identification loss along with adversarial loss. In Figure 6.4, we show some representative results on Multi-PIE test samples. Reconstruction results indicate that our proposed pose attention-guided coupled framework is able to provide robust, and discriminative features in the deep subspace for multiple use of profile to frontal matching in the embedded domain as well as high-fidelity frontal face synthesis.

## 6.6 Ablation Study

### 6.6.1 Embedding Dimensionality

To represent each face into a tightly compact embedding subspace, we explore different embedding dimensionalities: 128, 256, and 512. Experimental results reported in Table 6.5 illustrates that the network is able to extract features enriched with relevant information in 512 dimension. We select 512-d for all experiments reported in this paper.

Table 6.5: Rank-1 recognition rates (%) on Multi-PIE (Setting-1) for different embedding dimensions

| Dims | ±90° | ±75° |
|---|---|---|
| 128 | 60.8 | 67.5 |
| 256 | 70.2 | 78.1 |
| 512 | **89.5** | **98.7** |

Table 6.6: Rank-1 recognition rates (%) on Multi-PIE (Setting-1) for different approaches of spatial attention

| Description | ±90° | ±75° |
|---|---|---|
| Channel Refined Features + 1 × 1 conv + Max Pool | 87.8 | 94.5 |
| Channel Refined Features + 3 × 3 conv + Stride-2 | **89.5** | **98.7** |

### 6.6.2 Attention Maps

In this section, we show the effective design approach of our proposed pose attention mechanism to efficiently guide the face recognition network. We first focus on computing different approaches of SpAM in the pose attention block, PAB. Finally, we observe different combination of channel and spatial attention in deep profile feature extraction. Each experiment has been explained in the following sections.

**Spatial attention**

Given the channel-wise refined features, we explore two different approaches to generate a 2D spatial attention map: (1) first, we use average-and max-pooling across the channel axes, which generates two 2D descriptors, then apply standard 1 × 1 convolution followed by a max pool layer. (2) second, we similarly generate two 2D descriptors, and apply 3 × 3 convolution with stride 2, which proves to be outperforming the first approach. We report the comparison of two methods in Table 6.6.

**Arrangement of Spatial and Channel Attention**

In this experiment, to refine pose informative features from the deep profile branch of our PIFR network, we apply channel and spatial attention in two different ways. From a spatial viewpoint, the channel attention works to infer global information whereas the spatial attention focuses on local neighbourhood. However, the network response can be different upon the sequential order of

Table 6.7: Rank-1 recognition rates (%) on Multi-PIE (Setting-1) for different arrangements in attention mechanism

| Description | ±90° | ±75° |
|---|---|---|
| InceptionResnet + spatial + channel | 87.2 | 95.6 |
| InceptionResnet + channel + spatial | **89.5** | **98.7** |

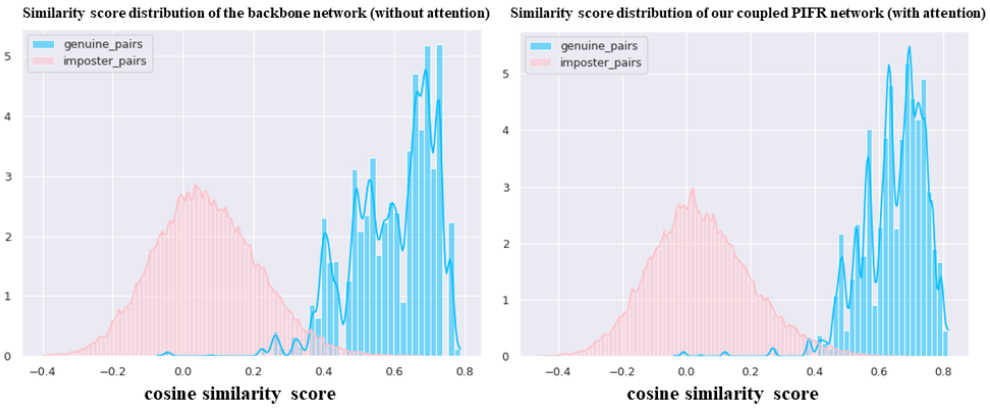



Figure 6.5: Comparing Cosine similarity distributions of the genuine pairs and imposter pairs for full profile faces (±90°) of Multi-Pie Setting 1 between the backbone network (without attention) and our coupled PIFR network (with attention)

each attention.

Table 7 summarizes the recognition performance on Multi-PIE for different attention sequences. The results show that the we achieve better performance when we use channel-spatial order rather than the vise versa.

### 6.6.3 Visualization

As shown in Figure 6.5, when compared to the backbone network (without attention), the similarity distributions of the genuine pairs and the imposter pairs in our proposed coupled PIFR network are more compact and distinct for full profile variations (±90°). Moreover, the area of similarity between genuine pairs spread more and overlap with the area of imposter pairs when we only train backbone network with constrastive loss without imposing attention on it. It further supports our proposed idea of pose refinement via PAB attention module.

Table 6.8: Rank-1 recognition rates (%) across poses at ±15° and ±30° pitch angles on M2FPA dataset with **yaw zero**.

| yaw | 0° | | | |
|---|---|---|---|---|
| pitch | –30° | –15° | +15° | +30° |
| Inception Resnet (w/o pose attention) | 98.0 | 100.0 | 100.0 | 98.2 |
| Inception Resnet (with pose attention) | 98.6 | 100.0 | 100.0 | 98.9 |
| **Resnet50 (with pose attention)** | **100.0** | **100.0** | **100.0** | **100.0** |

Table 6.9: Rank-1 recognition rates (%) across poses at 0° pitch angle on M2FPA dataset with yaw variations (–90° to +90°)

| pitch | 0° | | | | | |
|---|---|---|---|---|---|---|
| yaw | ±15° | ±30° | ±45° | ±60° | ±75° | ±90° |
| Inception Resnet (w/o pose attention) | 100.0 | 100.0 | 100.0 | 97.01 | 85.07 | 62.12 |
| Inception Resnet (with pose attention) | 100.0 | 100.0 | 100.0 | 98.51 | 88.1 | 78.68 |
| **Resnet50 (with pose attention)** | **100.0** | **100.0** | **100.0** | **100.0** | 97.01 | 86.36 |

Table 6.10: Rank-1 recognition rates (%) across poses at +15° pitch angle on M2FPA dataset with yaw variations (–90° to +90°)

| pitch | +15° | | | | | |
|---|---|---|---|---|---|---|
| yaw | ±15° | ±30° | ±45° | ±60° | ±75° | ±90° |
| Inception Resnet (w/o pose attention) | 100.0 | 100.0 | 100.0 | 97.0 | 88.0 | 59.7 |
| Inception Resnet (with pose attention) | 100.0 | 100.0 | 100.0 | 98.5 | 94.01 | 66.7 |
| **Resnet50 (with pose attention)** | **100.0** | **100.0** | **100.0** | **100.0** | 96.75 | 87.5 |

### 6.6.4 Evaluation on M2FPA dataset for joint yaw and pitch variations

To investigate the FR performance of our proposed network for faces at pitch variations along with yaw pose views, we select M2FPA dataset. We have also investigated different stem (i.e., Resnet50) and trained with Supervised Contrastive Loss (SCL) along with Memory Bank to increase the number of positive samples during training. Before analyse the results we briefly discuss SCL and memory buffer technique in the following sections.

#### Supervised Contrastive Loss (SCL)

Supervised contrastive loss (SCL) [258] builds on the standard contrastive loss function, which computes the contrastive loss between a pair of samples, where one sample is the target and the other is a positive or negative example. The positive example is another sample from the same

Table 6.11: Rank-1 recognition rates (%) across poses at –15° pitch angle on M2FPA dataset with yaw variations (–90° to +90°)

| pitch | +15° | | | | | |
|---|---|---|---|---|---|---|
| yaw | ±15° | ±30° | ±45° | ±60° | ±75° | ±90° |
| Inception Resnet (w/o pose attention) | 100.0 | 100.0 | 98.1 | 95.52 | 69.1 | 40.29 |
| Inception Resnet (with pose attention) | 100.0 | 100.0 | 98.51 | 97.01 | 71.64 | 55.68 |
| **Resnet50 (with pose attention)** | **100.0** | **100.0** | **100.0** | **98.72** | 95.52 | 84.85 |

Table 6.12: Rank-1 recognition rates (%) across poses at +30° pitch angles on M2FPA dataset with **yaw zero**.

| pitch | +30° | | | |
|---|---|---|---|---|
| yaw | ±22.5° | ±45° | ±67.5° | ±90° |
| Inception Resnet (w/o pose attention) | 98.3 | 98.01 | 86.0 | 27.0 |
| Inception Resnet (with pose attention) | 98.57 | 98.50 | 86.57 | 47.0 |
| **Resnet50 (with pose attention)** | **100.0** | **100.0** | **97.97** | **75.41** |

Table 6.13: Rank-1 recognition rates (%) across poses at –30° pitch angles on M2FPA dataset with **yaw zero**.

| pitch | –30° | | | |
|---|---|---|---|---|
| yaw | ±22.5° | ±45° | ±67.5° | ±90° |
| Inception Resnet (w/o pose attention) | 98.0 | 94.03 | 67.16 | 26.87 |
| Inception Resnet (with pose attention) | 98.5 | 97.01 | 70.0 | 40.0 |
| **Resnet50 (with pose attention)** | **100.0** | **98.98** | **91.04** | **65.15** |

class as the target, while the negative example is a sample from a different class. The loss function then incorporates the class labels of the target and the positive example to provide additional supervision. By incorporating the class labels, SCL ensures that the learned representations are not only discriminative but also meaningful in terms of the underlying class structure of the data.

### Memory Bank

Memory Bank [259] is a data structure used in various machine learning models, particularly in the context of contrastive learning. In contrastive learning, the aim is to learn representations that can discriminate between different classes or categories in a dataset. The Memory Bank is used to store representations of past examples, and the stored representations are used as positive examples during training. Specifically, when a new example is presented during training, its representation is

compared with the representations in the Memory Bank to find the most similar examples, which are used as positive examples for contrastive loss computation.

The Memory Bank is typically updated periodically during training, with new representations being added and old representations being discarded. This ensures that the Memory Bank contains a diverse set of positive examples that are representative of the dataset. The use of a Memory Bank in contrastive learning can lead to improved performance by providing a richer source of positive examples during training. This can help to overcome the problem of limited positive examples, which is particularly relevant in scenarios where the number of labeled examples is limited.

**Result Analysis**

We evaluate our proposed PIFR network on M2FPA dataset for extreme joint yaw and pitch angles. Table 6.8 reports 100% rank-1 recognition accuracy across poses at $\pm 15°$ and $\pm 30°$ pitch angles with yaw zero, when Resnet50 is used as stem and optimized using SCL loss with memory bank. For similar training procedure, Table 6.9 also shows that when faces are at $0°$ pitch with yaw variations ranging from $\pm 15°$ to $\pm 90°$ our PIFR network achieved around 12.14% higher accuracy compared to the result obtained through using Inception Resnet for extreme yaw angles ($\pm 75°$, and $\pm 90°$). We also summarize the performance of our proposed method for faces posed at joint pitch ($\pm 15°$ to $\pm 30°$) and yaw variations ($\pm 15°$ to $\pm 90°$) in Table 6.10-6.13. We observe promising results in extreme joint pitch and yaw views for Resnet50 optimized for SCL loss, which refers to improving the quality of representation in the embedded domain.

## 6.7 Conclusion

In this paper, we propose a novel perspective of leveraging pose as auxiliary information to guide a coupled profile to frontal deep subspace learning framework for PIFR. A pose attended feature block (PAB) is designed to distill pose-specific useful features from profile faces in deep convolutional layers. To ensure discriminative, pose-invariant feature representation into a compact embedding subspace, we couple both profile and frontal face images via a class-specific contrastive loss, which maximizes the pair-wise similarity in the embedded domain. We perform a comprehensive experiments on several benchmark datasets both in controlled and uncontrolled environmental settings to evaluate the robustness of our model. The results indicate that our model remarkably

outperform other state-of-the-art algorithms for profile to frontal pose-invariant face recognition. In addition, we conduct a quick experiment to explore the generative capability of the embedding features learned in deep subspace of our network. Moreover, we also investigate embedding dimensionality and attention mechanisms from different perspectives to offer an effective design choice of our proposed network.

# Chapter 7

# Conclusion

In this dissertation, we analyzed generative adversarial network and its applications in aerial vehicle detection and biometrics. In the first part of the dissertation, we investigated the problem of vehicle detection in low-resolution aerial images. In addition, we also addressed the difficulty of detection on infrared aerial images which have low-resolution, low-contrast, and few texture features present in the image. We proposed a multi-scale super-resolution technique and a cross-modal super-resolution approach to improve vehicle detection performance on aerial images. In the second part, we developed methods for applications in biometrics. We have provided conclusive analyses as follows:

**Joint Training of Multi-Scale Super-Resolution and Vehicle Detection Network** We empirically demonstrated that joint training of two interrelated tasks such as super-resolution and vehicle detection can mutually benefit each other, which addresses our prime concern of performing vehicle detection on low-resolution aerial images. According to this observation, first, a Multi-scale (MsGAN) is developed to gradually restore the high-frequency components from the low-resolution data at the multi-stage of the network. It shows that the proposed MsGAN achieves superior performance in retrieving high-frequency details compared to conventional up-sampling methods, such as interpolation and CNN-based methods. Then, we jointly train MsGAN and YOLOv3 object detector end-to-end to enhance target regions in the super-resolved images to become contextually more distinctive from the background, which in return improves detection performance. We validated our proposed algorithm through extensive experiments on publicly available aerial (i.e., VEDAI, DOTA) and satellite (i.e., xVIEW) datasets.

**Joint Cross-Modal Super-Resolution for Aerial vehicle detection** A major limitation of the conventional detection algorithm is most of them are RGB image-based detectors, which are not suitable for detection being performed at night or in a dark environment. Again IR detector requires a large amount of annotated training data and computational resources to train and deploy the model. To alleviate this shortcoming, we developed a joint cross-modal super-resolution network using cGAN, which has learned to translate LR IR images to HR super-resolved VIS images. After that, We have applied our existing detector pre-trained on original HR VIS images on these HR super-resolved and translated VIS images. Due to the limited availability of IR training images, we experimented only on the VEDAI dataset which has both VIS and corresponding IR images.

**GAN-Based Cross-Domain Iris Recognition** We have investigated cross-domain (cross-spectral cross-resolution) iris recognition problems and presented two different novel approaches to perform matching iris images acquired under different domains. To summarize, our first method involves synthesizing HR VIS iris images from LR NIR iris images in a gallery or vice versa using a cGAN-based domain adaption framework. By developing this framework, we integrated the cross-resolution matching scenario into a cross-spectrum setting that is intended to be used as a preprocessing step before conducting iris verification using an off-the-shelf iris matcher (i.e.,(OSIRIS) assuming that the user already has access to such a matcher. our second method is designed to address several shortcomings of commercial iris matcher that our first method depends on, specifically in scenarios where the resolution of captured iris images varies. They may not be able to operate directly on lower-resolution images while enrolled images in the gallery are at a higher resolution. This can lead to accuracy and robustness issues in the matching process. Therefore, our second method involves learning subspace embedded features, which are a set of features that can effectively capture the discriminative information of iris images in a low-dimensional subspace, irrespective of the resolution of the input image. We have designed and implemented cpGAN, which captures the correlation between NIR and VIS iris images in a reduced-dimensional latent embedding feature subspace. We validated both approaches for cross-spectral iris recognition using publicly available datasets (i.e., PolyU Bi-Spectral, Cross-Eyed-Cross-Spectral dataset). We introduced a new WVU Face and Iris dataset. Both approaches achieved promising results which suggest new state-of-the-art results in the cross-spectral iris matching domain.

**Pose-Attention Guided Profile-to-Frontal Face Recognition and reconstruction of frontal faces by GAN** In uncontrolled environments, the variation of face pose can reduce the performance of such a system and make recognition very difficult. Face pose variations refer to changes in the orientation of the face, such as tilting, rotating, or looking up or down. These variations can make it challenging for the system to accurately detect and match the face. We investigated the contribution of these pose variations as side information through the implementation of a PAB module. To develop our proposed framework, we implemented a novel coupled learning profile-to-frontal FR network using different architectures and loss functions (i.e., Inception Resnet, Resnet-50, contrastive loss (CL), supervised CL with Memory Bank). Then we infuse pose as complementary information to explicitly help the FR network to focus on important features along both the "channel" and "spatial" dimensions. We validated our proposed network on both controlled (i.e., M2FPA, Multi-PIE) and uncontrolled datasets(i.e., CFP, IJB-C) and achieved outperforming results compared to other pose-invariant face recognition networks. Our proposed PIFR network proved to be robust even for faces with extreme joint yaw and pitch variations. We utilized the pose invariant features that we have learned in the embedding subspace of our PIFR network to develop the Face Frontal GAN, which generates high-quality, identity-preserving frontalized faces even when we have profile faces at extreme angles (+90°/ – 90°). It provides an efficient way for face data augmentation.

## 7.1 List of Publications

- Joint Super-Resolution and Head Pose Estimation for Extreme Low-Resolution Faces, Malakshan, Saadabadi, Solaymani, Moktari, and Nasrabadi, *IEEE Access, 2023.*
- A Quality Aware Sample-to-Sample Comparison for Face Recognition, Saadabadi, Malakshan, Zafari, Moktari, Nasrabadi, *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023.*
- Pose attention-guided profile-to-frontal face recognition, Mostofa, Saadabadi, Malakshan, and Nasrabadi, *IEEE International Joint Conference on Biometrics (IJCB), 2022.*
- Information maximization for extreme pose face recognition., Saadabadi, Malakshan, So-

leymani, Mostofa, Nasrabadi, *IEEE International Joint Conference on Biometrics (IJCB), 2022.*

- Deep gan-based cross-spectral cross-resolution iris recognition, Mostofa, Mohamadi, Dawson, Nasrabadi, *IEEE Transactions on Biometrics, Behavior, and Identity Science (TBIOM), 2021.*

- Cross-spectral iris matching using conditional coupled GAN, Mostofa, Taherkhani, Dawson, and Nasrabadi, *IEEE International Joint Conference on Biometrics (IJCB), 2020.*

- Joint-Srvdnet: Joint super resolution and vehicle detection network, Moktari, Ferdous, Riggan, and Nasrabadi, *IEEE Access, 2020.*

- A joint cross-modal super-resolution approach for vehicle detection in aerial imagery, Mostofa, Ferdous, and Nasrabadi, *In Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications II, SPIE, 2020.*

- Target classification in infrared imagery by cross-spectral synthesis using GAN, Ferdous, Mostofa, Osahor, and Nasrabadi, *In Automatic Target Recognition XXX, SPIE, 2020.*

- Super resolution-assisted deep aerial vehicle detection, Ferdous, Mostofa, and Nasrabadi, *In Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications, SPIE, 2019.*

# References

[1] M. Mostofa, F. Taherkhani, J. Dawson, and N. M. Nasrabadi, "Cross-spectral iris matching using conditional coupled GAN," in *2020 IEEE International Joint Conference on Biometrics (IJCB)*. IEEE, 2020, pp. 1–9. xii, 62, 64, 68, 69, 86, 87

[2] Y. Tian, X. Peng, L. Zhao, S. Zhang, and D. N. Metaxas, "CR-GAN: learning complete representations for multi-view generation," *arXiv preprint arXiv:1806.11191*, 2018. xiii, 116, 117

[3] W. Ouyang, X. Zeng, X. Wang, S. Qiu, P. Luo, Y. Tian, H. Li, S. Yang, Z. Wang, H. Li *et al.*, "DeepID-Net: Object detection with deformable part based convolutional neural networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 7, pp. 1320–1334, 2016. 1

[4] A. Diba, V. Sharma, A. Pazandeh, H. Pirsiavash, and L. Van Gool, "Weakly supervised cascaded convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 914–922. 1

[5] N. Doulamis and A. Voulodimos, "FAST-MDL: Fast adaptive supervised training of multi-layered deep learning models for consistent object tracking and classification," in *2016 IEEE International Conference on Imaging Systems and Techniques (IST)*. IEEE, 2016, pp. 318–323. 1

[6] N. Doulamis, "Adaptable deep learning structures for object labeling/tracking under dynamic visual environments," *Multimedia Tools and Applications*, vol. 77, no. 8, pp. 9651–9689, 2018. 1

[7] L. Lin, K. Wang, W. Zuo, M. Wang, J. Luo, and L. Zhang, "A deep structured model with radius–margin bound for 3d human activity recognition," *International Journal of Computer Vision*, vol. 118, no. 2, pp. 256–273, 2016. 1

[8] S. Cao and R. Nevatia, "Exploring deep learning based solutions in fine grained activity recognition in the wild," in *2016 23rd International Conference on Pattern Recognition (ICPR)*. IEEE, 2016, pp. 384–389. 1

[9] A. Toshev and C. Szegedy, "Deeppose: Human pose estimation via deep neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1653–1660. 1

[10] X. Chen and A. Yuille, "Articulated pose estimation by a graphical model with image dependent pairwise relations," *arXiv preprint arXiv:1407.3399*, 2014. 1

[11] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680. 2, 6, 9, 19, 21, 22, 45, 62, 70

[12] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4681–4690. 2, 7, 8, 9, 10, 19, 21, 22, 25, 31, 37, 38, 49, 52, 72, 73, 83, 86

[13] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125–1134. 2, 25, 49, 71, 72, 75, 83, 86

[14] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1701–1708. 3, 101, 102

[15] Y. Sun, X. Wang, and X. Tang, "Deep learning face representation from predicting 10,000 classes," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1891–1898. 3, 101

[16] Y. Sun, D. Liang, X. Wang, and X. Tang, "Deepid3: Face recognition with very deep neural networks," *arXiv preprint arXiv:1502.00873*, 2015. 3

[17] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 815–823. 3, 101, 105, 108, 115

[18] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," *British Machine Vision Association*, 2015. 3, 62, 115

[19] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "Sphereface: Deep hypersphere embedding for face recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 212–220. 3

[20] F. Wang, J. Cheng, W. Liu, and H. Liu, "Additive margin softmax for face verification," *IEEE Signal Processing Letters*, vol. 25, no. 7, pp. 926–930, 2018. 3

[21] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu, "Cosface: Large margin cosine loss for deep face recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5265–5274. 3

[22] L. N. Darlow and B. Rosman, "Fingerprint minutiae extraction using deep learning," in *2017 IEEE International Joint Conference on Biometrics (IJCB)*. IEEE, 2017, pp. 22–30. 3

[23] Y. Tang, F. Gao, J. Feng, and Y. Liu, "Fingernet: An unified deep network for fingerprint minutiae extraction," in *2017 IEEE International Joint Conference on Biometrics (IJCB)*. IEEE, 2017, pp. 108–116. 3

[24] C. Lin and A. Kumar, "Contactless and partial 3d fingerprint recognition using multi-view deep representation," *Pattern Recognition*, vol. 83, pp. 314–327, 2018. 3

[25] R. R. Omar, T. Han, S. A. Al-Sumaidaee, and T. Chen, "Deep finger texture learning for verifying people," *IET Biometrics*, vol. 8, no. 1, pp. 40–48, 2019. 3

[26] S. Minaee, E. Azimi, and A. Abdolrashidi, "Fingernet: Pushing the limits of fingerprint recognition using convolutional neural network," *arXiv preprint arXiv:1907.12956*, 2019. 3

[27] C. Lin and A. Kumar, "Multi-siamese networks to accurately match contactless to contact-based fingerprint images," in *2017 IEEE International Joint Conference on Biometrics (IJCB)*. IEEE, 2017, pp. 277–285. 3

[28] S. Minaee, A. Abdolrashidiy, and Y. Wang, "An experimental study of deep convolutional features for iris recognition," in *2016 IEEE signal processing in medicine and biology symposium (SPMB)*. IEEE, 2016, pp. 1–6. 3

[29] A. Gangwar and A. Joshi, "Deepirisnet: Deep iris representation with applications in iris recognition and cross-sensor iris recognition," in *2016 IEEE international conference on image processing (ICIP)*. IEEE, 2016, pp. 2301–2305. 3

[30] M. Baqar, A. Ghani, A. Aftab, S. Arbab, and S. Yasin, "Deep belief networks for iris recognition based on contour detection," in *2016 International Conference on Open Source Systems & Technologies (ICOSST)*. IEEE, 2016, pp. 72–77. 3

[31] M. G. Alaslani, "Convolutional neural network based feature extraction for iris recognition," *International Journal of Computer Science & Information Technology (IJCSIT) Vol*, vol. 10, 2018. 3

[32] H. Menon and A. Mukherjee, "Iris biometrics using deep convolutional networks," in *2018 IEEE International Instrumentation and Measurement Technology Conference (I2MTC)*. IEEE, 2018, pp. 1–5. 3

[33] H. Hofbauer, E. Jalilian, and A. Uhl, "Exploiting superior cnn-based iris segmentation for better recognition accuracy," *Pattern Recognition Letters*, vol. 120, pp. 17–23, 2019. 3

[34] S. Ahmad and B. Fuller, "Thirdeye: Triplet based iris recognition without normalization," in *2019 IEEE 10th International Conference on Biometrics Theory, Applications and Systems (BTAS)*. IEEE, 2019, pp. 1–9. 3

[35] S. Minaee and A. Abdolrashidi, "Deepiris: Iris recognition using a deep learning approach," *arXiv preprint arXiv:1907.09380*, 2019. 3, 62

[36] S. Dodge, J. Mounsef, and L. Karam, "Unconstrained ear recognition using deep neural networks," *IET Biometrics*, vol. 7, no. 3, pp. 207–214, 2018. 3

[37] H. Sinha, R. Manekar, Y. Sinha, and P. K. Ajmera, "Convolutional neural network-based human identification using outer ear images," in *Soft Computing for Problem Solving: SocProS 2017, Volume 2*. Springer, 2019, pp. 707–719. 3

[38] I. Omara, X. Wu, H. Zhang, Y. Du, and W. Zuo, "Learning pairwise svm on hierarchical deep features for ear recognition," *IET Biometrics*, vol. 7, no. 6, pp. 557–566, 2018. 3

[39] T. Hassner, S. Harel, E. Paz, and R. Enbar, "Effective face frontalization in unconstrained images," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 4295–4304. 4, 102, 104

[40] C. Sagonas, Y. Panagakis, S. Zafeiriou, and M. Pantic, "Robust statistical face frontalization," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 3871–3879. 4, 104

[41] X. Yin and X. Liu, "Multi-task convolutional neural network for pose-invariant face recognition," *IEEE Transactions on Image Processing*, vol. 27, no. 2, pp. 964–975, 2017. 4, 102, 104

[42] Y. Yin, S. Jiang, J. P. Robinson, and Y. Fu, "Dual-attention GAN for large-pose face frontalization," in *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*. IEEE, 2020, pp. 249–256. 4, 113

[43] C. Ding and D. Tao, "Pose-invariant face recognition with homography-based normalization," *Pattern Recognition*, vol. 66, pp. 144–152, 2017. 4, 112, 113

[44] C. Ding, C. Xu, and D. Tao, "Multi-task pose-invariant face recognition," *IEEE Transactions on Image Processing*, vol. 24, no. 3, pp. 980–993, 2015. 4, 102

[45] E.-J. Tsai and W.-C. Yeh, "PAM: Pose attention module for pose-invariant face recognition," *arXiv preprint arXiv:2111.11940*, 2021. 4, 105

[46] J. Zhao, Y. Cheng, Y. Xu, L. Xiong, J. Li, F. Zhao, K. Jayashree, S. Pranata, S. Shen, J. Xing *et al.*, "Towards pose invariant face recognition in the wild," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2207–2216. 4, 102, 111, 112, 113

[47] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *European conference on computer vision*. Springer, 2016, pp. 21–37. 6, 8, 11

[48] M. Bevilacqua, A. Roumy, C. Guillemot, and M. L. Alberi-Morel, "Low-complexity single-image super-resolution based on nonnegative neighbor embedding," 2012. 6

[49] H. Chang, D.-Y. Yeung, and Y. Xiong, "Super-resolution through neighbor embedding," in *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, vol. 1. IEEE, 2004, pp. I–I. 6

[50] D. Dai, R. Timofte, and L. Van Gool, "Jointly optimized regressors for image super-resolution," in *Computer Graphics Forum*, vol. 34, no. 2. Wiley Online Library, 2015, pp. 95–104. 6, 7

[51] W. T. Freeman, E. C. Pasztor, and O. T. Carmichael, "Learning low-level vision," *International journal of computer vision*, vol. 40, no. 1, pp. 25–47, 2000. 6, 8

[52] S. Schulter, C. Leistner, and H. Bischof, "Fast and accurate image upscaling with super-resolution forests," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3791–3799. 6

[53] R. Timofte, V. De Smet, and L. Van Gool, "Anchored neighborhood regression for fast example-based super-resolution," in *Proceedings of the IEEE international conference on computer vision*, 2013, pp. 1920–1927. 6, 7

[54] J. Yang, Z. Wang, Z. Lin, S. Cohen, and T. Huang, "Coupled dictionary training for image super-resolution," *IEEE transactions on image processing*, vol. 21, no. 8, pp. 3467–3478, 2012. 6

[55] J. Yang, J. Wright, T. Huang, and Y. Ma, "Image super-resolution as sparse representation of raw image patches," in *2008 IEEE Conference on Computer Vision and Pattern Recognition*. Citeseer, 2008, pp. 1–8. 6

[56] J. Yang, J. Wright, T. S. Huang, and Y. Ma, "Image super-resolution via sparse representation," *IEEE transactions on image processing*, vol. 19, no. 11, pp. 2861–2873, 2010. 6

[57] F. Taherkhani, N. M. Nasrabadi, and J. Dawson, "A deep face identification network enhanced by facial attributes prediction," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 553–560. 6

[58] F. Taherkhani and M. Jamzad, "Restoring highly corrupted images by impulse noise using radial basis functions interpolation," *IET Image Processing*, vol. 12, no. 1, pp. 20–30, 2017. 6

[59] R. Zeyde, M. Elad, and M. Protter, "On single image scale-up using sparse-representations," in *International conference on curves and surfaces*. Springer, 2010, pp. 711–730. 6

[60] W. T. Freeman, T. R. Jones, and E. C. Pasztor, "Example-based super-resolution," *IEEE Computer graphics and Applications*, no. 2, pp. 56–65, 2002. 6

[61] J.-B. Huang, A. Singh, and N. Ahuja, "Single image super-resolution from transformed self-exemplars," *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5197–5206, 2015. 6

[62] S. Gu, W. Zuo, Q. Xie, D. Meng, X. Feng, and L. Zhang, "Convolutional sparse coding for image super-resolution," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1823–1831. 6

[63] Y.-W. Tai, S. Liu, M. S. Brown, and S. Lin, "Super resolution using edge prior and single image detail synthesis," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, 2010, pp. 2400–2407. 6

[64] K. Zhang, X. Gao, D. Tao, and X. Li, "Multi-scale dictionary for single image super-resolution," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 1114–1121. 6

[65] R. Timofte, V. De Smet, and L. Van Gool, "A+: Adjusted anchored neighborhood regression for fast super-resolution," in *Asian conference on computer vision*. Springer, 2014, pp. 111–126. 7

[66] K. I. Kim and Y. Kwon, "Single-image super-resolution using sparse regression and natural image prior," *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, no. 6, pp. 1127–1133, 2010. 7

[67] Z. Wang, D. Liu, J. Yang, W. Han, and T. Huang, "Deep networks for image super-resolution with sparse prior," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 370–378. 7

[68] K. Gregor and Y. LeCun, "Learning fast approximations of sparse coding," in *Proceedings of the 27th International Conference on International Conference on Machine Learning*. Omnipress, 2010, pp. 399–406. 7

[69] V. Talreja, M. C. Valenti, and N. M. Nasrabadi, "Multibiometric secure system based on deep learning," in *2017 IEEE Global conference on signal and information processing (globalSIP)*. IEEE, 2017, pp. 298–302. 7

[70] C. Dong, C. C. Loy, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution," in *European conference on computer vision*. Springer, 2014, pp. 184–199. 7, 8

[71] C. Dong', C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 2, pp. 295–307, 2016. 7

[72] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European conference on computer vision*. Springer, 2016, pp. 694–711. 7, 22, 73, 74, 78, 83, 117

[73] J. Bruna, P. Sprechmann, and Y. LeCun, "Super-resolution with deep convolutional sufficient statistics," *arXiv preprint arXiv:1511.05666*, 2015. 7, 22

[74] T. Zhao and R. Nevatia, "Car detection in low resolution aerial images," *Image and Vision Computing*, vol. 21, no. 8, pp. 693–703, 2003. 7, 23

[75] J.-Y. Choi and Y.-K. Yang, "Vehicle detection from aerial images using local shape information," in *Pacific-Rim Symposium on Image and Video Technology*. Springer, 2009, pp. 227–236. 7, 23

[76] H.-Y. Cheng, C.-C. Weng, and Y.-Y. Chen, "Vehicle detection in aerial surveillance using dynamic bayesian networks," *IEEE transactions on image processing*, vol. 21, no. 4, pp. 2152–2159, 2012. 7, 23

[77] H. Kazemi, S. Soleymani, F. Taherkhani, S. Iranmanesh, and N. Nasrabadi, "Unsupervised image-to-image translation using domain-specific variational information bound," in *Advances in Neural Information Processing Systems*, 2018, pp. 10 369–10 379. 7

[78] J. Carlet and B. Abayowa, "Fast vehicle detection in aerial imagery," *CoRR*, vol. abs/1709.08666, 2017. [Online]. Available: http://arxiv.org/abs/1709.08666 7, 23

[79] J. O. d. Terrail and F. Jurie, "Faster RER-CNN: application to the detection of vehicles in aerial images," *arXiv preprint arXiv:1809.07628*, 2018. 7, 23

[80] A. Soleimani, N. M. Nasrabadi, E. Griffith, J. Ralph, and S. Maskell, "Convolutional neural networks for aerial vehicle detection and recognition," in *NAECON 2018-IEEE National Aerospace and Electronics Conference*. IEEE, 2018, pp. 186–191. 7, 23

[81] M. Y. Yang, W. Liao, X. Li, and B. Rosenhahn, "Vehicle detection in aerial images," *arXiv preprint arXiv:1801.07339*, 2018. 8, 23

[82] Z. Deng, X. Hu, L. Zhu, X. Xu, J. Qin, G. Han, and P.-A. Heng, "R3net: Recurrent residual refinement network for saliency detection," in *Proceedings of the 27th International Joint Conference on Artificial Intelligence*. AAAI Press, 2018, pp. 684–690. 8, 23

[83] F. Taherkhani, V. Talreja, H. Kazemi, and N. Nasrabadi, "Facial attribute guided deep cross-modal hashing for face image retrieval," in *2018 International Conference of the Biometrics Special Interest Group (BIOSIG)*. IEEE, 2018, pp. 1–6. 8

[84] V. Talreja, F. Taherkhani, M. C. Valenti, and N. M. Nasrabadi, "Using deep cross modal hashing and error correcting codes for improving the efficiency of attribute guided facial image retrieval," in *2018 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*. IEEE, 2018, pp. 564–568. 8

[85] J. Shermeyer and A. Van Etten, "The effects of super-resolution on object detection performance in satellite imagery," *arXiv preprint arXiv:1812.04098*, 2018. 8

[86] L. Cao, R. Ji, C. Wang, and J. Li, "Towards domain adaptive vehicle detection in satellite image by supervised super-resolution transfer," in *Thirtieth AAAI Conference on Artificial Intelligence*, 2016. 8

[87] G. Polatkan, M. Zhou, L. Carin, D. Blei, and I. Daubechies, "A bayesian nonparametric approach to image super-resolution," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 2, pp. 346–358, 2015. 8

[88] K. Nasrollahi and T. B. Moeslund, "Super-resolution: a comprehensive survey," *Machine vision and applications*, vol. 25, no. 6, pp. 1423–1468, 2014. 8, 9

[89] B. Normalization, "Accelerating deep network training by reducing internal covariate shift," *CoRR.–2015.–Vol. abs/1502.03167.–URL: http://arxiv. org/abs/1502.03167*, 2015. 10, 49

[90] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1026–1034. 10, 26, 49, 83

[91] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1874–1883. 10, 26, 31, 49, 52, 73, 83

[92] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015. 10, 28, 50, 62, 83

[93] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations*, 2015. 10

[94] J. Wu, Z. Cui, V. S. Sheng, P. Zhao, D. Su, and S. Gong, "A comparative study of sift and its variants," *Measurement science review*, vol. 13, no. 3, pp. 122–131, 2013. 11

[95] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *international Conference on computer vision & Pattern Recognition (CVPR'05)*, vol. 1. IEEE Computer Society, 2005, pp. 886–893. 11, 114

[96] L. Wang, *Support vector machines: theory and applications*. Springer Science & Business Media, 2005, vol. 177. 11

[97] R. E. Schapire, "Explaining adaboost," in *Empirical inference*. Springer, 2013, pp. 37–52. 11

[98] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, no. 9, pp. 1627–1645, 2010. 11

[99] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587. 11

[100] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448. 11, 39, 40

[101] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99. 11, 39, 40

[102] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788. 11, 20

[103] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014. 12, 74, 78, 83

[104] S. Razakarivony and F. Jurie, "Vehicle detection in aerial imagery (vedai): a benchmark," *Tech. Rep.*, 2015. 13

[105] T. N. Mundhenk, G. Konjevod, W. A. Sakla, and K. Boakye, "A large contextual dataset for classification, detection and counting of cars with deep learning," in *European Conference on Computer Vision*. Springer, 2016, pp. 785–800. 18

[106] G.-S. Xia, X. Bai, J. Ding, Z. Zhu, S. Belongie, J. Luo, M. Datcu, M. Pelillo, and L. Zhang, "Dota: A large-scale dataset for object detection in aerial images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3974–3983. 18, 24, 35

[107] D. Lam, R. Kuzma, K. McGee, S. Dooley, M. Laielli, M. Klaric, Y. Bulatov, and B. McCord, "xView: Objects in context in overhead imagery," *arXiv preprint arXiv:1802.07856*, 2018. 18, 24, 34

[108] S. Cao, Y. Yu, H. Guan, D. Peng, and W. Yan, "Affine-function transformation-based object matching for vehicle detection from unmanned aerial vehicle imagery," *Remote Sensing*, vol. 11, no. 14, p. 1708, 2019. 19

[109] S. N. Ferdous, M. Mostofa, and N. M. Nasrabadi, "Super resolution-assisted deep aerial vehicle detection," in *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications*, vol. 11006. International Society for Optics and Photonics, 2019, p. 1100617. 19, 27, 29, 45, 47

[110] T. Aspiras, R. Liu, and V. K. Asari, "Convolutional auto-encoder for vehicle detection in aerial imagery (conference presentation)," in *Pattern Recognition and Tracking XXX*, vol. 10995. International Society for Optics and Photonics, 2019, p. 109950D. 19

[111] M. Y. Yang, W. Liao, X. Li, Y. Cao, and B. Rosenhahn, "Vehicle detection in aerial images," *Photogrammetric Engineering & Remote Sensing*, vol. 85, no. 4, pp. 297–304, 2019. 19

[112] J. Hoffman, E. Tzeng, T. Park, J.-Y. Zhu, P. Isola, K. Saenko, A. A. Efros, and T. Darrell, "Cycada: Cycle-consistent adversarial domain adaptation," *arXiv preprint arXiv:1711.03213*, 2017. 19, 25, 49

[113] J. Shermeyer and A. Van Etten, "The effects of super-resolution on object detection performance in satellite imagery," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 0–0. 19, 23, 47

[114] C. Borel-Donohue and S. S. Young, "Image quality and super resolution effects on object recognition using deep neural networks," in *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications*, vol. 11006. International Society for Optics and Photonics, 2019, p. 110061M. 19

[115] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018. 20

[116] C. Dong', C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 2, pp. 295–307, 2015. 21, 31, 52, 73

[117] J. Kim, J. Kwon Lee, and K. Mu Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1646–1654. 21

[118] J. Kim', J. Kwon Lee, and K. Mu Lee, "Deeply-recursive convolutional network for image super-resolution," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1637–1645. 21

[119] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778. 21, 83, 84

[120] D. Eigen, J. Rolfe, R. Fergus, and Y. LeCun, "Understanding deep architectures using a recursive convolutional network," *arXiv preprint arXiv:1312.1847*, 2013. 21

[121] M. Liang and X. Hu, "Recurrent convolutional neural network for object recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3367–3375. 21

[122] R. Socher, B. Huval, B. Bath, C. D. Manning, and A. Y. Ng, "Convolutional-recursive deep learning for 3D object classification," in *Advances in neural information processing systems*, 2012, pp. 656–664. 21

[123] L. Gatys, A. S. Ecker, and M. Bethge, "Texture synthesis using convolutional neural networks," in *Advances in neural information processing systems*, 2015, pp. 262–270. 22

[124] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. N. Metaxas, "Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 5907–5915. 22

[125] J. Sang, Z. Wu, P. Guo, H. Hu, H. Xiang, Q. Zhang, and B. Cai, "An improved yolov2 for vehicle detection," *Sensors*, vol. 18, no. 12, p. 4272, 2018. 23

[126] M. Haris, G. Shakhnarovich, and N. Ukita, "Task-driven super resolution: Object detection in low-resolution images," *arXiv preprint arXiv:1803.11316*, 2018. 23

[127] E. Ataer-Cansizoglu, M. Jones, Z. Zhang, and A. Sullivan, "Verification of very low-resolution faces using an identity-preserving deep face super-resolution network," *arXiv preprint arXiv:1903.10974*, 2019. 23

[128] J. Wu, S. Ding, W. Xu, and H. Chao, "Deep joint face hallucination and recognition," *arXiv preprint arXiv:1611.08091*, 2016. 24

[129] Y. Pang, J. Cao, J. Wang, and J. Han, "JCS-Net: Joint classification and super-resolution network for small-scale pedestrian detection in surveillance images," *IEEE Transactions on Information Forensics and Security*, 2019. 24

[130] P. Dollár, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: A benchmark," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2009, pp. 304–311. 24

[131] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 3354–3361. 24

[132] S. Razakarivony and F. Jurie, "Vehicle detection in aerial imagery: A small target detection benchmark," *Journal of Visual Communication and Image Representation*, vol. 34, pp. 187–203, 2016. 24, 34

[133] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015. 26, 83

[134] H. Kazemi, F. Taherkhani, and N. M. Nasrabadi, "Identity-aware deep face hallucination via adversarial face verification," *arXiv preprint arXiv:1909.08130*, 2019. 27, 37, 39

[135] T. Wang, W. Sun, H. Qi, and P. Ren, "Aerial image super resolution via wavelet multiscale convolutional neural networks," *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 5, pp. 769–773, 2018. 27

[136] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125. 28, 51

[137] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition. arxiv [cs. cv]. 2014," 2018. 31, 52

[138] M. Bosch, C. M. Gifford, and P. A. Rodriguez, "Super-resolution for overhead imagery using densenets and adversarial learning," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2018, pp. 1414–1422. 37, 39

[139] J. Zhong, T. Lei, and G. Yao, "Robust vehicle detection in aerial images based on cascaded convolutional neural networks," *Sensors*, vol. 17, no. 12, p. 2720, 2017. 39, 40

[140] C. Chen, J. Zhong, and Y. Tan, "Multiple-oriented and small object detection with convolutional neural networks for aerial image," *Remote Sensing*, vol. 11, no. 18, p. 2176, 2019. 39, 40

[141] M. Mandal, M. Shah, P. Meena, and S. K. Vipparthi, "SSSDET: Simple short and shallow network for resource efficient vehicle detection in aerial scenes," in *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2019, pp. 3098–3102. 39, 40, 45

[142] M. Ju, J. Luo, P. Zhang, M. He, and H. Luo, "A simple and efficient network for small target detection," *IEEE Access*, vol. 7, pp. 85 771–85 781, 2019. 39

[143] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, vol. 2. IEEE, 2003, pp. 1398–1402. 38

[144] Z. Wang and A. C. Bovik, "A universal image quality index," *IEEE signal processing letters*, vol. 9, no. 3, pp. 81–84, 2002. 38

[145] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Transactions on image processing*, vol. 15, no. 2, pp. 430–444, 2006. 38

[146] E. Koester and C. S. Sahin, "A comparison of super-resolution and nearest neighbors interpolation applied to object detection on satellite data," *arXiv preprint arXiv:1907.05283*, 2019. 42

[147] S. Maiti, P. Gidde, S. Saurav, S. Singh, S. Chaudhury *et al.*, "Real-time vehicle detection in aerial images using skip-connected convolution network with region proposal networks," in *International Conference on Pattern Recognition and Machine Intelligence*. Springer, 2019, pp. 200–208. 45

[148] M. Mostofa, S. N. Ferdous, B. S. Riggan, and N. M. Nasrabadi, "Joint-srvdnet: Joint super resolution and vehicle detection network," *arXiv preprint arXiv:2005.00983*, 2020. 45, 47

[149] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014. 46, 62, 71, 77

[150] Y. Taigman, A. Polyak, and L. Wolf, "Unsupervised cross-domain image generation," *arXiv preprint arXiv:1611.02200*, 2016. 46

[151] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "Stargan: Unified generative adversarial networks for multi-domain image-to-image translation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8789–8797. 46

[152] T. Kim, M. Cha, H. Kim, J. K. Lee, and J. Kim, "Learning to discover cross-domain relations with generative adversarial networks," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017, pp. 1857–1865. 46

[153] M.-Y. Liu, T. Breuel, and J. Kautz, "Unsupervised image-to-image translation networks," in *Advances in neural information processing systems*, 2017, pp. 700–708. 47

[154] M. A. Molina-Cabello, R. M. Luque-Baena, E. López-Rubio, and K. Thurnhofer-Hemsi, "Vehicle type detection by ensembles of convolutional neural networks operating on super resolved images," *Integrated Computer-Aided Engineering*, vol. 25, no. 4, pp. 321–333, 2018. 47

[155] H. Ji, Z. Gao, T. Mei, and B. Ramesh, "Vehicle detection in remote sensing images leveraging on simultaneous super-resolution," *IEEE Geoscience and Remote Sensing Letters*, 2019. 47

[156] D. M. Uliyan, S. Sadeghi, and H. A. Jalab, "Anti-spoofing method for fingerprint recognition using patch based deep learning machine," *Engineering Science and Technology, an International Journal*, vol. 23, no. 2, pp. 264–273, 2020. 60

[157] S. Zhao and B. Zhang, "Deep discriminative representation for generic palmprint recognition," *Pattern Recognition*, vol. 98, p. 107071, 2020. 60

[158] A. Dabouei, F. Taherkhani, S. Soleymani, J. Dawson, and N. Nasrabadi, "Boosting deep face recognition via disentangling appearance and geometry," in *The IEEE Winter Conference on Applications of Computer Vision*, 2020, pp. 320–329. 60

[159] J. Daugman, "How iris recognition works," in *The essential guide to image processing*. Elsevier, 2009, pp. 715–739. 60, 61, 73

[160] R. P. Wildes, "Iris recognition: an emerging biometric technology," *Proceedings of the IEEE*, vol. 85, no. 9, pp. 1348–1363, 1997. 60

[161] K. W. Bowyer, K. Hollingsworth, and P. J. Flynn, "Image understanding for iris biometrics: A survey," *Computer vision and image understanding*, vol. 110, no. 2, pp. 281–307, 2008. 60, 61

[162] A. K. Jain, K. Nandakumar, and A. Ross, "50 years of biometric research: Accomplishments, challenges, and opportunities," *Pattern Recognition Letters*, vol. 79, pp. 80–105, 2016. 60

[163] J. G. Daugman, "High confidence visual recognition of persons by a test of statistical independence," *IEEE transactions on pattern analysis and machine intelligence*, vol. 15, no. 11, pp. 1148–1161, 1993. 61

[164] C.-W. Tan and A. Kumar, "Towards online iris and periocular recognition under relaxed imaging constraints," *IEEE Transactions on Image Processing*, vol. 22, no. 10, pp. 3751–3765, 2013. 61

[165] Z. Zhao and K. Ajay, "An accurate iris segmentation framework under relaxed imaging constraints using total variation model," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 3828–3836. 61, 80, 81, 83

[166] K. W. Bowyer, "The results of the nice. ii iris biometrics competition," *Pattern Recognition Letters*, vol. 33, no. 8, pp. 965–969, 2012. 61

[167] M. De Marsico, M. Nappi, D. Riccio, and H. Wechsler, "Mobile iris challenge evaluation (miche)-i, biometric iris dataset and protocols," *Pattern Recognition Letters*, vol. 57, pp. 17–23, 2015. 61

[168] K. Wang and A. Kumar, "Cross-spectral iris recognition using CNN and supervised discrete hashing," *Pattern Recognition*, vol. 86, pp. 85–98, 2019. 61, 62, 63, 67, 85, 87, 88, 92, 93, 94

[169] P. R. Nalla and A. Kumar, "Toward more accurate iris recognition using cross-spectral matching," *IEEE transactions on Image processing*, vol. 26, no. 1, pp. 208–221, 2016. 61, 62, 63, 67, 71, 80, 81, 83, 85, 87, 88, 92, 93, 94

[170] J. Zuo, F. Nicolo, and N. A. Schmid, "Cross spectral iris matching based on predictive image mapping," in *2010 Fourth IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS)*. IEEE, 2010, pp. 1–5. 62, 67

[171] L. Masek, "Matlab source code for a biometric identification system based on iris patterns," *School of Computer Science and Software Eng., Univ. Western Australia*, 2003. 62, 85, 87, 92, 93, 94

[172] T. Zhang, A. Wiliem, S. Yang, and B. Lovell, "TV-GAN: Generative adversarial network based thermal to visible face recognition," in *2018 international conference on biometrics (ICB)*. IEEE, 2018, pp. 174–181. 62

[173] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "VGGFACE2: A dataset for recognising faces across pose and age. corr abs/1710.08092 (2017)," *arXiv preprint arXiv:1710.08092*, 2017. 62

[174] Z. Zhao and A. Kumar, "Towards more accurate iris recognition using deeply learned spatially corresponding features," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 3809–3818. 62

[175] K. Nguyen, C. Fookes, A. Ross, and S. Sridharan, "Iris recognition with off-the-shelf CNN features: A deep learning perspective," *IEEE Access*, vol. 6, pp. 18 848–18 855, 2017. 62

[176] C. S. Bezerra, R. Laroca, D. R. Lucio, E. Severo, L. F. Oliveira, A. S. Britto, and D. Menotti, "Robust iris segmentation based on fully convolutional networks and generative adversarial networks," in *2018 31st SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*. IEEE, 2018, pp. 281–288. 62

[177] G. Wang, W. Kang, Q. Wu, Z. Wang, and J. Gao, "Generative adversarial network (GAN) based data augmentation for palmprint recognition," in *2018 Digital Image Computing: Techniques and Applications (DICTA)*. IEEE, 2018, pp. 1–7. 62

[178] S. Minaee and A. Abdolrashidi, "Finger-GAN: Generating realistic fingerprint images using connectivity imposed GAN," *arXiv preprint arXiv:1812.10482*, 2018. 62

[179] S. Minaee‘ and A. Abdolrashidi, "Iris-GAN: Learning to generate realistic iris images using convolutional GAN," *arXiv preprint arXiv:1812.04822*, 2018. 62

[180] M. B. Lee, Y. H. Kim, and K. R. Park, "Conditional generative adversarial network-based data augmentation for enhancement of iris recognition accuracy," *IEEE Access*, vol. 7, pp. 122 134–122 152, 2019. 62

[181] J. E. Tapia and C. Arellano, "Soft-biometrics encoding conditional GAN for synthesis of NIR periocular images," *Future Generation Computer Systems*, vol. 97, pp. 503–511, 2019. 62

[182] N. Othman, B. Dorizzi, and S. Garcia-Salicetti, "Osiris: An open source iris recognition software," *Pattern Recognition Letters*, vol. 82, pp. 124–131, 2016. 62, 73, 87, 88, 89, 90, 91, 92, 93, 94

[183] A. Sequeira, L. Chen, P. Wild, J. Ferryman, F. Alonso-Fernandez, K. B. Raja, R. Raghavendra, C. Busch, and J. Bigun, "Cross-eyed-cross-spectral iris/periocular recognition database and competition," in *2016 International Conference of the Biometrics Special Interest Group (BIOSIG)*. IEEE, 2016, pp. 1–5. 63, 66, 80

[184] X.-C. Lian, Z. Li, B.-L. Lu, and L. Zhang, "Max-margin dictionary learning for multiclass image categorization," in *European Conference on Computer Vision*. Springer, 2010, pp. 157–170. 63

[185] Z. Ding, M. Shao, and Y. Fu, "Missing modality transfer learning via latent low-rank constraint," *IEEE transactions on Image Processing*, vol. 24, no. 11, pp. 4322–4334, 2015. 63

[186] Y. Fu, S. Yan, and T. S. Huang, "Correlation metric for generalized feature extraction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 12, pp. 2229–2235, 2008. 63

[187] R. Vyas, T. Kanumuri, and G. Sheoran, "Cross spectral iris recognition for surveillance based applications," *Multimedia Tools and Applications*, vol. 78, no. 5, pp. 5681–5699, 2019. 66

[188] C.-W. Tan and A. Kumar, "Unified framework for automated iris segmentation using distantly acquired face images," *IEEE Transactions on Image Processing*, vol. 21, no. 9, pp. 4068–4079, 2012. 66

[189] M. Oktiana, F. Arnia, Y. Away, and K. Munadi, "Features for cross spectral image matching: a survey," *Bulletin of Electrical Engineering and Informatics*, vol. 7, no. 4, pp. 552–560, 2018. 66

[190] K. B. Raja, R. Raghavendra, and C. Busch, "Cross-spectrum periocular authentication for NIR and visible images using bank of statistical filters," in *2016 IEEE International Conference on Imaging Systems and Techniques (IST)*. IEEE, 2016, pp. 227–231. 66

[191] A. F. Sequeira, L. Chen, J. Ferryman, P. Wild, F. Alonso-Fernandez, J. Bigun, K. B. Raja, R. Raghavendra, C. Busch, T. de Freitas Pereira *et al.*, "Cross-eyed 2017: Cross-spectral iris/periocular recognition competition," in *2017 IEEE International Joint Conference on Biometrics (IJCB)*. IEEE, 2017, pp. 725–732. 66

[192] F. Alonso-Fernandez and J. Bigun, "Near-infrared and visible-light periocular recognition with Gabor features using frequency-adaptive automatic eye detection," *IET Biometrics*, vol. 4, no. 2, pp. 74–89, 2015. 66

[193] P. Wild, P. Radu, and J. Ferryman, "On fusion for multispectral iris recognition," in *2015 international conference on biometrics (ICB)*. IEEE, 2015, pp. 31–37. 67

[194] A. Sharma, S. Verma, M. Vatsa, and R. Singh, "On cross spectral periocular recognition," in *2014 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2014, pp. 5007–5011. 67

[195] M. J. Burge and M. K. Monaco, "Multispectral iris fusion for enhancement, interoperability, and cross wavelength matching," in *Algorithms and Technologies for Multispectral, Hyperspectral, and Ultraspectral Imagery XV*, vol. 7334. International Society for Optics and Photonics, 2009, p. 73341D. 67

[196] M. J. Burge and M. Monaco, "Multispectral iris fusion and cross-spectrum matching," in *Handbook of Iris Recognition*. Springer, 2013, pp. 171–181. 67

[197] K. Raja, N. Damer, R. Ramachandra, F. Boutros, and C. Busch, "Cross-spectral periocular recognition by cascaded spectral image transformation," in *2019 IEEE International Conference on Imaging Systems and Techniques (IST)*. IEEE, 2019, pp. 1–7. 67

[198] M. Oktiana, T. Horiuchi, K. Hirai, K. Saddami, F. Arnia, Y. Away, and K. Munadi, "Cross-spectral iris recognition using phase-based matching and homomorphic filtering," *Heliyon*, vol. 6, no. 2, p. e03407, 2020. 67

[199] E. Ribeiro, A. Uhl, F. Alonso-Fernandez, and R. A. Farrugia, "Exploring deep learning image super-resolution for iris recognition," in *2017 25th European Signal Processing Conference (EUSIPCO)*. IEEE, 2017, pp. 2176–2180. 68

[200] X. Wang, H. Zhang, J. Liu, L. Xiao, Z. He, L. Liu, and P. Duan, "Iris image super resolution based on gans with adversarial triplets," in *Chinese Conference on Biometric Recognition*. Springer, 2019, pp. 346–353. 68, 73

[201] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1. IEEE, 2005, pp. 539–546. 69, 74, 75, 109

[202] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis," *arXiv preprint arXiv:1605.05396*, 2016. 71

[203] N. P. Ramaiah and A. Kumar, "Advancing cross-spectral iris recognition research using bi-spectral imaging," in *Machine Intelligence and Signal Processing*. Springer, 2016, pp. 1–10. 71

[204] Z. Wang, J. Chen, and S. C. Hoi, "Deep learning for image super-resolution: A survey," *IEEE transactions on pattern analysis and machine intelligence*, 2020. 72

[205] J. Bromley, J. W. Bentz, L. Bottou, I. Guyon, Y. LeCun, C. Moore, E. Säckinger, and R. Shah, "Signature verification using a "siamese" time delay neural network," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 7, no. 04, pp. 669–688, 1993. 74

[206] H. Zhang and V. M. Patel, "Densely connected pyramid dehazing network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 3194–3203. 74

[207] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 586–595. 78

[208] W. Shi, J. Caballero, L. Theis, F. Huszar, A. Aitken, C. Ledig, and Z. Wang, "Is the deconvolution layer the same as a convolutional layer?" *arXiv preprint arXiv:1609.07009*, 2016. 84

[209] K. Diederik and J. A. Ba, "A method for stochastic optimization. arxiv 2014," *arXiv preprint arXiv:1412.6980*, 2014. 84

[210] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization." *Journal of machine learning research*, vol. 13, no. 2, 2012. 84

[211] K. Wang and A. Kumar, "Toward more accurate iris recognition using dilated residual features," *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 12, pp. 3233–3245, 2019. 85, 87, 100

[212] http://www.biometrics.gov/Documents/irisrec.pdf/. 99

[213] Y. Liu, Y. He, C. Gan, J. Zhu, and L. Li, "A review of advances in iris image acquisition system," in *Chinese Conference on Biometric Recognition*. Springer, 2012, pp. 210–218. 99

[214] S. Sengupta, J.-C. Chen, C. Castillo, V. M. Patel, R. Chellappa, and D. W. Jacobs, "Frontal to profile face verification in the wild," in *2016 IEEE winter conference on applications of computer vision (WACV)*. IEEE, 2016, pp. 1–9. 101, 110, 111, 114

[215] C. Ding and D. Tao, "A comprehensive survey on pose-invariant face recognition," *ACM Transactions on intelligent systems and technology (TIST)*, vol. 7, no. 3, pp. 1–42, 2016. 102

[216] J. G. Daugman, "Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters," *JOSA A*, vol. 2, no. 7, pp. 1160–1169, 1985. 102

[217] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001*, vol. 1. Ieee, 2001, pp. I–I. 102

[218] T. Ahonen, A. Hadid, and M. Pietikainen, "Face description with local binary patterns: Application to face recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 28, no. 12, pp. 2037–2041, 2006. 102, 114

[219] D. Chen, X. Cao, F. Wen, and J. Sun, "Blessing of dimensionality: High-dimensional feature and its efficient compression for face verification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 3025–3032. 102

[220] K. Q. Weinberger and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification." *Journal of machine learning research*, vol. 10, no. 2, 2009. 102

[221] X. Zhu, Z. Lei, J. Yan, D. Yi, and S. Z. Li, "High-fidelity pose and expression normalization for face recognition in the wild," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 787–796. 102, 104

[222] Y. Qian, W. Deng, and J. Hu, "Unsupervised face normalization with extreme pose and expression in the wild," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9851–9858. 102, 113, 115

[223] L. Tran, X. Yin, and X. Liu, "Disentangled representation learning GAN for pose-invariant face recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1415–1424. 102

[224] K. Cao, Y. Rong, C. Li, X. Tang, and C. C. Loy, "Pose-robust face recognition via deep residual equivariant mapping," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5187–5196. 102, 105, 114, 115

[225] Y. Shi, X. Yu, K. Sohn, M. Chandraker, and A. K. Jain, "Towards universal representation learning for deep face recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 6817–6826. 102

[226] S. Li, X. Liu, X. Chai, H. Zhang, S. Lao, and S. Shan, "Morphable displacement field based image matching for face recognition across pose," in *European conference on computer vision*. Springer, 2012, pp. 102–115. 104

[227] M. Kan, S. Shan, H. Chang, and X. Chen, "Stacked progressive auto-encoders (spae) for face recognition across poses," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1883–1890. 104

[228] J. Yang, S. E. Reed, M.-H. Yang, and H. Lee, "Weakly-supervised disentangling with recurrent transformations for 3d view synthesis," in *Advances in neural information processing systems*, 2015, pp. 1099–1107. 104

[229] J. Yim, H. Jung, B. Yoo, C. Choi, D. Park, and J. Kim, "Rotating your face using multi-task deep neural network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 676–684. 104

[230] Z. Zhu, P. Luo, X. Wang, and X. Tang, "Multi-view perceptron: a deep model for learning face identity and view representations," in *Advances in Neural Information Processing Systems*, 2014, pp. 217–225. 104

[231] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural computation*, vol. 16, no. 12, pp. 2639–2664, 2004. 104

[232] M. Kan, S. Shan, and X. Chen, "Multi-view deep network for cross-view classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4847–4855. 104

[233] Z. Zhu, P. Luo, X. Wang, and X. Tang, "Deep learning identity-preserving face space," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 113–120. 104

[234] J.-C. Chen, V. M. Patel, and R. Chellappa, "Unconstrained face verification using deep CNN features," in *2016 IEEE winter conference on applications of computer vision (WACV)*. IEEE, 2016, pp. 1–9. 104, 114

[235] I. Masi, S. Rawls, G. Medioni, and P. Natarajan, "Pose-aware face recognition in the wild," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4838–4846. 104

[236] J. Huang and C. Ding, "Attention-guided progressive mapping for profile face recognition," in *2021 IEEE International Joint Conference on Biometrics (IJCB)*. IEEE, 2021, pp. 1–8. 105

[237] N. Ruiz, E. Chong, and J. M. Rehg, "Fine-grained head pose estimation without keypoints," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2018, pp. 2074–2083. 106

[238] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li, "Face alignment across large poses: A 3d solution," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 146–155. 106

[239] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19. 106, 107

[240] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Thirty-first AAAI conference on artificial intelligence*, 2017. 108, 112

[241] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "VGGFace2: A dataset for recognising faces across pose and age," in *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*. IEEE, 2018, pp. 67–74. 108

[242] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-pie," *Image and vision computing*, vol. 28, no. 5, pp. 807–813, 2010. 110

[243] B. F. Klare, B. Klein, E. Taborsky, A. Blanton, J. Cheney, K. Allen, P. Grother, A. Mah, and A. K. Jain, "Pushing the frontiers of unconstrained face detection and recognition: Iarpa janus benchmark a," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1931–1939. 110

[244] B. Maze, J. Adams, J. A. Duncan, N. Kalka, T. Miller, C. Otto, A. K. Jain, W. T. Niggel, J. Anderson, J. Cheney *et al.*, "Iarpa Janus benchmark-c: Face dataset and protocol," in *2018 international conference on biometrics (ICB)*. IEEE, 2018, pp. 158–165. 111, 115

[245] C. Whitelam, E. Taborsky, A. Blanton, B. Maze, J. Adams, T. Miller, N. Kalka, A. K. Jain, J. A. Duncan, K. Allen *et al.*, "Iarpa Janus benchmark-b face dataset," in *proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2017, pp. 90–98. 111

[246] P. Li, X. Wu, Y. Hu, R. He, and Z. Sun, "M2fpa: A multi-yaw multi-pitch high-quality dataset and benchmark for facial pose analysis," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 10 043–10 051. 111

[247] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2015. 112

[248] C. Xiong, X. Zhao, D. Tang, K. Jayashree, S. Yan, and T.-K. Kim, "Conditional convolutional neural network for modality-aware face recognition," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 3667–3675. 112, 113

[249] R. Huang, S. Zhang, T. Li, and R. He, "Beyond face rotation: Global and local perception gan for photorealistic and identity preserving frontal view synthesis," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2439–2448. 112, 113

[250] Y. Hu, X. Wu, B. Yu, R. He, and Z. Sun, "Pose-guided photorealistic face rotation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8398–8406. 112, 113

[251] F. Taherkhani, V. Talreja, J. Dawson, M. C. Valenti, and N. M. Nasrabadi, "Pf-cpgan: Profile to frontal coupled gan for face recognition in the wild," in *2020 IEEE International Joint Conference on Biometrics (IJCB)*. IEEE, 2020, pp. 1–10. 112, 113, 114

[252] X. Yin, X. Yu, K. Sohn, X. Liu, and M. Chandraker, "Towards large-pose face frontalization in the wild," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 3990–3999. 113

[253] S. Sankaranarayanan, A. Alavi, C. D. Castillo, and R. Chellappa, "Triplet probabilistic embedding for face verification and clustering," in *2016 IEEE 8th international conference on biometrics theory, applications and systems (BTAS)*. IEEE, 2016, pp. 1–8. 114

[254] X. Wu, R. He, Z. Sun, and T. Tan, "A light CNN for deep face representation with noisy labels," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 11, pp. 2884–2896, 2018. 114, 117

[255] K. Simonyan, O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Fisher vector faces in the wild." in *BMVC*, vol. 2, no. 3, 2013, p. 4. 114

[256] Q. Cao, Y. Ying, and P. Li, "Similarity metric learning for face recognition," in *Proceedings of the IEEE international conference on computer vision*, 2013, pp. 2408–2415. 114

[257] Y.-J. Ju, G.-H. Lee, J.-H. Hong, and S.-W. Lee, "Complete face recovery gan: Unsupervised joint face rotation and de-occlusion from a single-view image," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 3711–3721. 115

[258] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," *Advances in neural information processing systems*, vol. 33, pp. 18 661–18 673, 2020. 120

[259] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, "Unsupervised feature learning via non-parametric instance discrimination," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 3733–3742. 121