

Multi-Model Person Re-identification: Combining Facial and Body Features

by

**Hemani Bharadwaj
202111036**

A Thesis Submitted in Partial Fulfilment of the Requirements for the Degree of

MASTER OF TECHNOLOGY
in
INFORMATION AND COMMUNICATION TECHNOLOGY
to

DHIRUBHAI AMBANI INSTITUTE OF INFORMATION AND COMMUNICATION TECHNOLOGY

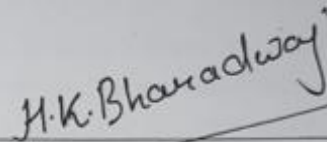


May, 2023

Declaration

I hereby declare that

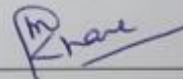
- i) the thesis comprises of my original work towards the degree of Master of Technology in Information and Communication Technology at Dhirubhai Ambani Institute of Information and Communication Technology and has not been submitted elsewhere for a degree,
- ii) due acknowledgment has been made in the text to all the reference material used.



Hemani Bharadwaj

Certificate

This is to certify that the thesis work entitled Multi-model person re-identification: combining facial and body features has been carried out by Hemani Bharadwaj for the degree of Master of Technology in Information and Communication Technology at *Dhirubhai Ambani Institute of Information and Communication Technology* under my/our supervision.



Dr. Manish Khare
Thesis Supervisor



Dr. Bakul Gohel
Co-Thesis Supervisor

Acknowledgments

I sincerely appreciate and would like to express my heartfelt gratitude to everyone who helped complete this thesis. It is a great pleasure to thank them for their important assistance, advice, and inspiration.

I want to start by sincerely thanking my mentor, Dr Manish Khare, for his important guidance and assistance throughout this research. His extensive knowledge, skill, and direction have formed my research. His insightful thoughts and perceptive ideas have improved my critical thinking ability and comprehend my research topic. I feel privileged to have had the opportunity to collaborate with him and learn from his guidance. I am grateful to my guide for providing the laboratory, air conditioning, and GPU, which were essential in enabling me to complete my thesis.

My sincere gratitude to my co-guide, Dr Bakul Gohel, for his invaluable advice and assistance throughout this project. His technical know-how and perceptive remarks were crucial in guiding the research and assuring its calibre. I am thankful for the opportunity to work with him, as his inspiration and enthusiasm for research have been highly motivating.

In addition to the people already mentioned, I'd like to express my gratitude to my juniors and colleagues Ruchita, Krutika, Dhairya, Divya, Harsh, and Naila, who were instrumental in the production of the video dataset used in this study. This thesis was made possible by their efforts. Their commitment, diligence, and assistance have preserved the dataset's quality.

I also appreciate my younger brother, who has professional machine-learning experience. I have benefited from his ongoing assistance in resolving my technological issues and developing new study concepts. In addition, I want to thank my companions on this voyage, Prakhar, Nisarg, and Divya, who have always been by my side throughout the thesis.

I also want to thank my parents for their support, love, and encouragement throughout my academic career. My success has been fueled by their sacrifices and devotion to my education, and I will always appreciate their support.

Contents

Abstract	v
List of Principal Symbols and Acronyms	vi
List of Tables	vii
List of Figures	viii
1 Introduction	1
1.1 Data processing	2
1.2 Motivation	3
1.3 Objective and Problem Statement	3
1.4 Challenges involved in Person-reidentification	4
1.5 Organization of thesis	5
2 Literature Survey	6
2.1 Metric-learning based person Re-ID	6
2.2 Deep learning based person Re-ID	7
2.3 Multi-scale feature learning related works	9
3 A Comprehensive Overview	11
3.1 Multi-feature learning	11
3.2 ResNeXt model	12
3.3 Loss Function	13
3.3.1 Cross-entropy loss	13
3.3.2 Triplet loss	14
4 Proposed Methodology	16
4.1 A multi-scale residual block architecture for Person Re-identification	16
4.2 Depthwise Separable Convolutions	17
4.3 Aggregation gate	18

4.4 Proposed Architecture	20
4.5 Implementation details:	21
4.6 Proposed Framework with Face and Body Fusion for Person Re-identification	22
5 Experimental Results and Analysis	24
5.1 Datasets	24
5.2 Custom video datasets	26
5.3 Results	28
5.3.1 Results on Market-1501 and DukeMTMC-reID datasets . . .	28
5.3.2 MARKET-1501 Rankings(Predicted matching images) for sample query images:	29
5.3.3 DukeMTMC-reID Rankings(Predicted matching images) for sample query images:	30
5.4 Comparison of our model with other methods	31
5.5 Results on custom video Datasets:	31
5.5.1 Results on Custom dataset-1 (Indoor video)	31
5.5.2 Results on Custom dataset-2 (Outdoor video-1)	33
5.5.3 Results on Custom dataset-3 (Outdoor video-2)	35
5.5.4 Accuracy and mAP on Custom datasets	38
6 Conclusions and Future scope	39
References	40

Abstract

In the surveillance industry, Person Re-identification (Re-ID) is significant as it matches a person's appearance across multiple non-overlapping cameras. However, this task is challenging due to changes in camera viewpoints, occlusion, and varying appearances such as clothes, shoes, and pose. To overcome these challenges, discriminative feature learning is necessary. Recently, For this aim, deep convolutional neural networks (CNNs) have been widely employed. This study proposes a lightweight and robust network for person Re-ID, which employs the YOLOv4 object detection model for pedestrian detection and the DeepSORT algorithm for tracking. The proposed model is designed to learn discriminative features at multiple semantic levels, utilizing the ResNeXt architecture as a backbone. Specifically, the network comprises multiple blocks, where channels are concatenated between blocks, and an aggregation gate is used to aggregate the output of multiple channels. The aggregation gate produces channel-wise weights that dynamically fuse the resulting multi-scale feature maps. This layout effectively allows the model to extract discriminative features even under challenging conditions. To determine whether our recommended strategy is effective, we conducted experiments on the widely-used Market1501 dataset and our own custom-made datasets, including indoor, outdoor, and same-dress outdoor datasets, which cover various challenges such as occlusion, lighting variations, and similar body features. The experiment results show that our approach to Person Re-identification is effective in Person Re-identification. These results indicate the potential of our proposed method to be applied in various real-world scenarios, such as surveillance cameras in markets, shopping malls, parking areas, and other public places.

List of Principal Symbols and Acronyms

◦ Convolution between two tensors

*Element wise convolution operation

σ Activation function

AG Aggregation gate

CNN Convolution neural network

DPM Deformable Part Model

DSC Depthwise separable convolution

e Exponential base 10

F Feature maps

G Aggregation gates

mAP Mean average precision

Re-ID Re-identification

ReLU Rectified linear unit is an activation function

YOLO You only look once

List of Tables

5.1 Datasets	25
5.2 Custom Video Datasets	26
5.3 Results of our model on public datasets	28
5.4 Comparison of our method with other methods on Market-1501 and DukeMTMC-reID datasets	31
5.5 Accuracy and mAP on Custom datasets.	38

List of Figures

1.1	An illustration of several camera perspectives	1
1.2	Procedures for the Person Re-ID task	2
1.3	To obtain the local representations of the various features, various patch-wise data processing algorithms are used.	2
2.1	General Approach for the metric-learning based Person Re-identification	7
2.2	General architecture of deep learning model.	8
2.3	Classification of various Deep Learning-based Person Re-identification techniques [13]	9
3.1	ResNeXt architecture [7]	13
3.2	Triplet loss	14
4.1	(a) Standard 3×3 convolution. (b) Lite 3×3 convolution. DW: Depth-Wise.[33]	16
4.2	Explanation of depthwise separable convolution used in our model.	17
4.3	Flow diagram of Aggregation Channel Gate	19
4.4	Our proposed model	20
4.5	Flow diagram of our proposed framework.	23
5.1	Samples of Market-1501 dataset	25
5.2	Samples of Market-1501 dataset	25
5.3	Samples of different viewpoints in our Custom dataset-1 (Indoor video)	26
5.4	Samples of different viewpoints in our Custom dataset-2 (Outdoor video-1)	27
5.5	Samples of different viewpoints in our Custom dataset-3 (Outdoor video-2)	27
5.6	MARKET-1501 Rankings of our model	29
5.7	DukeMTMC-reID Rankings of our model	30
5.8	Results on Custom dataset-1 (Indoor video)	32

5.9	Results on CAMERA-1 Custom dataset-2 (Outdoor video-1)	33
5.10	Results on CAMERA-2 Custom dataset-2 (Outdoor video-1)	34
5.11	Results on CAMERA-3 Custom dataset-2 (Outdoor video-1)	34
5.12	Results on Custom dataset-3 (Outdoor video-2)	37

CHAPTER 1

Introduction

Person Re-identification (Re-ID) plays a significant role in surveillance videos since they frequently occur in multi-camera settings. The primary objective of the person re-id systems is to assign a consistent identification number to individuals visible in each non-overlapping camera view. The background, atmospheric changes, human motions or movements, various camera views, and many other factors can generate considerable intra-class variances in the cameras from which these photographs were shot. Figure 1.1 provides examples of the various camera views. Person Re-ID involves three significant steps. The first important step is to detect



Figure 1.1: An illustration of several camera perspectives

the person from the frame captured by the surveillance camera. After detecting the person, it becomes essential to track the same person through all the frames. There is a need to establish a stable id for tracking each person. If a new the person appears in the frame. The system must retrieve its id from the available database. All this is summarised in Figure 1.2.

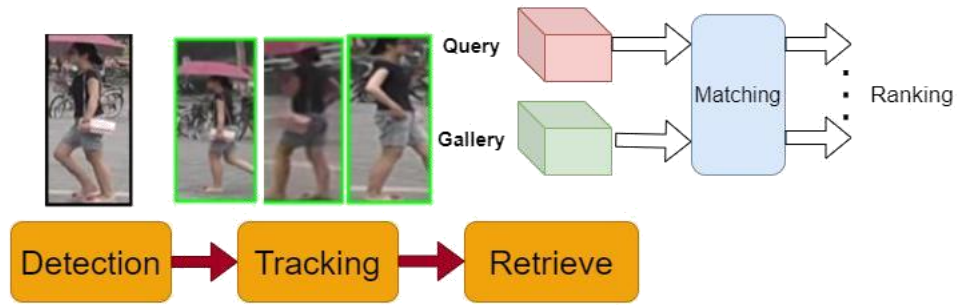


Figure 1.2: Procedures for the Person Re-ID task

If anyone wants to design a Person Re-ID system from scratch, the required steps are highlighted in Figure 1.3. Data collection becomes an important part when the system is designed from scratch. After completing annotating the data, any Person Re-ID methods can be applied.

1.1 Data processing

The goal is to extract the relevant features from the dataset after it has been gathered and processed. The traits of a certain individual can be extracted from their image in many ways. There are two approaches for extracting features. (For example, local or patch-based and global processing). The global operating system focuses on the topology of the camera. In other words, the camera's physical location can make a significant difference. For example, if there are two cameras at the entrance and exit, it is clear that the person first appears on the entrance camera and then on the exit camera. In contrast, the patched process focuses on the minor details of the image. Local or patched-based can help discriminate the intra-class samples. Some of the Patch-wise processing techniques are shown in Figure 1.3.

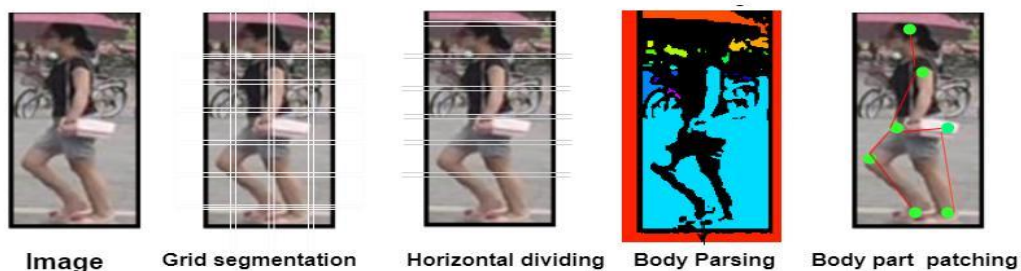


Figure 1.3: To obtain the local representations of the various features, various patch-wise data processing algorithms are used.

1.2 Motivation

Person Re-identification becomes essential when security is a concern. From the security perspective, only person detection and tracking are not sufficient. If the same person appears with different characteristics, these algorithms may fail to identify the person correctly. To deal with this issue, Person Re-identification is essential. For Person Re-identification, it becomes essential that the proposed approach is well generalized to handle different diversities caused by various characteristics of the person. There are many approaches available for doing the same. These days, Deep Learning based approaches are becoming popular.

The traditional methods often struggle with the scale variations in a person's image, where the individual appears at different distances. Our model aims to capture global and local information from person images, allowing accurate matching regardless of scale variation. It targets to achieve efficiency and real-time performance and robustness to occlusion and pose variations. In many scenarios where people can change their accessories, exchange their clothes, or similar appearance, they can fool the traditional methods. But our multiple model approach to using face and body features makes our person re-id system more accurate and efficient.

1.3 Objective and Problem Statement

Problem statement: Person Re-identification (Re-ID) presents significant challenges due to changes in camera perspectives, occlusion, and variations in appearance, such as clothing, footwear, and pose. Overcoming these challenges requires the learning of discriminative features. While deep convolutional neural networks (CNNs) have been widely utilized for this purpose, current Re-ID models often lack the capability to learn features across multiple scales. Consequently, this study addresses the need for an improved Re-ID model to learn discriminative features at multiple semantic levels effectively. By doing this, we want to improve the accuracy of person Re-ID systems in diverse real-world scenarios, including deploying surveillance cameras in markets, shopping malls, parking areas, and other public spaces.

Objective: Our research aims to advance an efficient and robust network for Person Re-identification on deep learning capabilities to learn distinctive features across multiple semantic levels. In our proposed approach, we introduce a novel methodology combining body and face features.

Specifically, body detection is accomplished through YOLOv4 object detection and DeepSORT tracking, while body features are obtained using our model. Simultaneously, face features are extracted using the FaceNet model.

To achieve this, we employ the ResNeXt architecture as the foundation of our model, ensuring effective extraction of discriminative features even in challenging scenarios involving occlusion, lighting variations, and similar body characteristics. Our primary objective is to enhance the accuracy of Person Re-identification in diverse real-world environments, including surveillance cameras in markets, shopping malls, parking areas, and other public spaces.

1.4 Challenges involved in Person-reidentification

Since Person Re-identification is an open-world problem, it is very challenging as CCTV can form indoor and outdoor surveillance videos. The outdoor videos are even more difficult. The several challenges involved in this task are:

Occlusion: Occlusion occurs when a person is partially or entirely hidden from view, making it challenging to extract meaningful features for identification. Occlusion can be caused by objects, other individuals, or even self-occlusion due to body pose.

Illumination Changes: Lighting conditions can vary significantly in different environments, leading to variations in the appearance of individuals.

Similarities Among Individuals: Individuals with a similar style of clothes, body shapes, or appearances can generally fool the network.

Camera Calibration and Viewpoint Variation: Variations in camera views, angles and calibrations directly affect the captured images. And challenging to detect and identify persons in those multiple viewpoints, angles or calibrations.

Scalability: As the number of individuals in a dataset or the complexity of surveillance systems increases, the scalability of Person Re-identification algorithms becomes crucial. Efficient algorithms are required to handle large-scale datasets and real-time processing in practical scenarios.

1.5 Organization of thesis

The thesis flow is going through like this, Chapter2 contains a literature survey related to Person Re-identification. Chapter 3 will explain the comprehensive overview of the thesis with an exploration of multi-feature learning basics, the backbone of our model and the loss functions used in our research. Chapter 4 contains the proposed methodology, which in detail explains the architecture of our proposed model, the components of our proposed models like Lite block, aggregation gate, implementation details, and framework we have used to increase the accuracy of Person Re-identification. In Chapter 5, we discussed the datasets used in experiments and the results on standard benchmark datasets and custom-made datasets by ourselves and compared our methods with other methods. In the Final Chapter 6, the Conclusion and Future Scopes are written.

CHAPTER 2

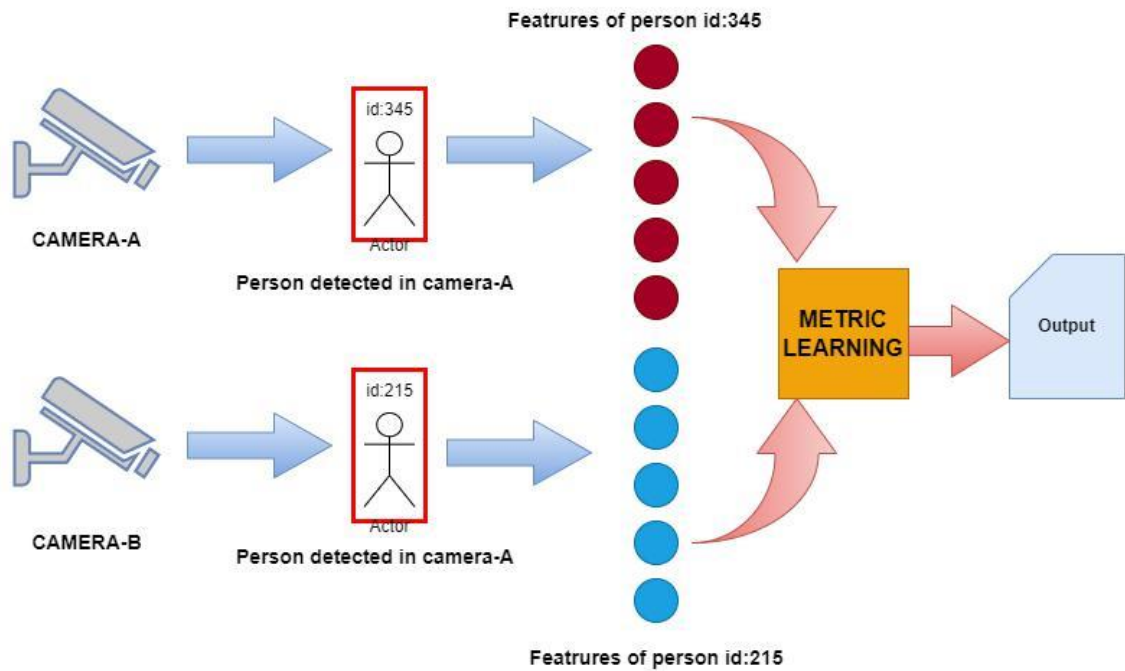
Literature Survey

This chapter delves into a comprehensive overview of the research on Person Re-identification (Re-ID). Over the last two decades, researchers had a significant interest in developing Re-ID systems. Person Re-id systems can generally be categorized into feature-based and metric-based approaches. Feature-based approaches aim to find an efficient representation of a person as a feature. In contrast, metric-based approaches primarily focus on developing effective metrics for computing the similarity between images of two individuals [2]. To provide a thorough understanding of the existing literature in this field, we organize the literature survey based on the methods employed to develop person Re-ID systems. This paper references some prominent works in the field. Notably, recent advancements primarily revolve around deep learning techniques. The utilization of multi-scale feature learning in deep learning has gained popularity in developing Re-ID systems.

2.1 Metric-learning based person Re-ID

This particular approach focuses primarily on the similarity metric. Figure 2.1 illustrates the Metric-learning based approach. In the early 2000s, the Mahalanobis distance[4] was introduced in the domain of Person Re-identification (Re-ID) as a means to measure similarity. However, it was found to be prone to overfitting. To address this issue, Meibin et al.[11] introduced the regularized independent metric. As the demand for surveillance videos grew, multiple cameras were employed to enhance security. Traditional metrics were unable to cope with such scenarios. Chen et al.[2] proposed an asymmetric distance metric to handle these situations. Xiaojing et al.[1] introduced a hypergraph-based metric as an alternative to Cartesian systems. Zhao et al.[26] later presented an improved hypergraph method version incorporating joint learning. Metric learning algorithms can be broadly classified into two categories: classical metric learning algorithms and

deep-learning-based metric learning algorithms. Classical metric learning learns a distance metric (e.g., Mahalanobis Distance), while deep-learning-based metric learning uses neural networks to learn discriminative embeddings (e.g., Siamese Networks, Triplet Networks).



2.2 Deep learning based person Re-ID

In this section, we highlight some of the most recent approaches that address various aspects of Re-ID systems shown in Figure 2.2.

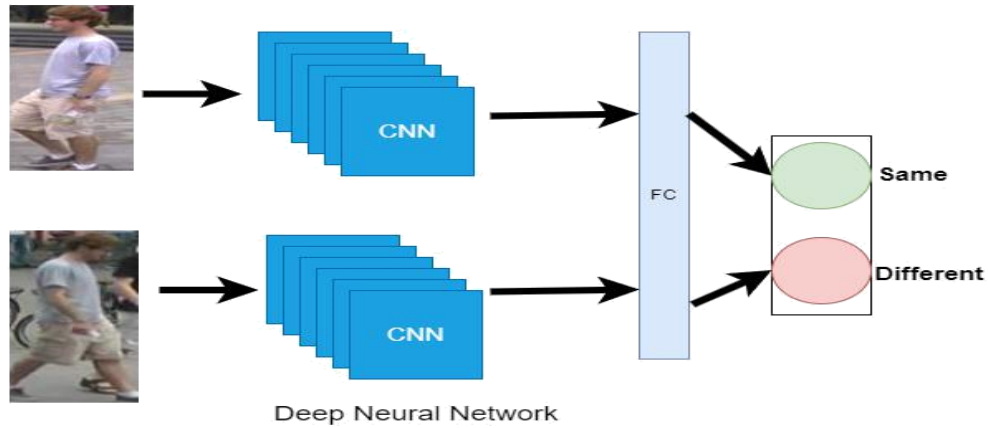


Figure 2.2: General architecture of deep learning model.

These approaches are categorized and well illustrated in Figure 2.3. Furthermore, Pu et al.[15] work addresses the limitations of stationary domain person Re-ID by introducing a novel framework called Adaptive Knowledge Accumulation (AKA) for knowledge representation. To handle the challenges in unsupervised person Re-ID systems, Xuan et al.[23] propose the introduction of intra-inter camera similarity computations to account for variations caused by multiple cameras. The fusion of inter-camera and intra-camera similarities has significantly improved the performance of person Re-ID systems. Additionally, Zheng et al.[27] present a grouping-based approach for enhancing unsupervised person Re-ID, leveraging the concept of unsupervised domain adaptiveness where a system trained on labeled domains can be applied to unlabeled domains without requiring annotations. Moreover, numerous deep learning-based approaches for Person Re-identification frameworks have emerged, which are extensively covered in a recent survey paper by Ye et al. [24], and Z.Ming et al.[12].

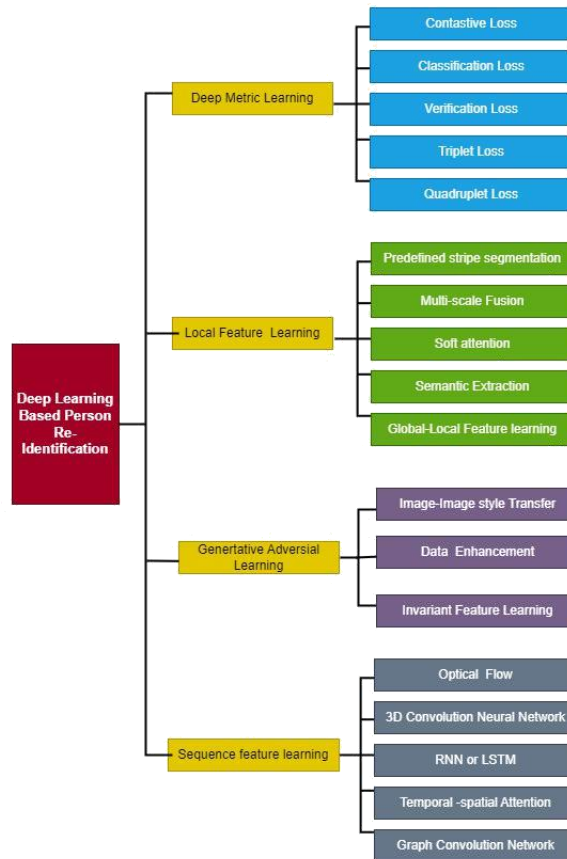


Figure 2.3: Classification of various Deep Learning-based Person Re-identification techniques [13]

2.3 Multi-scale feature learning related works

In recent years, Person Re-identification (Re-ID) has been widely studied and has become a popular topic in computer vision. To address the challenges of this task, researchers have proposed various approaches, including deep learning-based methods and multi-feature learning methods.

Deep learning-based methods have shown impressive performance in Re-ID due to their ability to learn discriminative features from large amounts of data. For instance, the work of Zheng et al.[28] proposed a deep neural network architecture that learns a discriminative embedding space for Re-ID. Similarly, Hermans et al.

[6] proposed the use of triplet loss for learning a deep feature embedding space for Re-ID.

Multi-feature learning methods have also been widely studied in Re-ID. These methods aim to extract complementary features from different modalities, such as

color, texture, and shape, and combine them to improve the Re-ID performance. One such work is OSNet [33], which proposes a lightweight network architecture for efficient person Re-ID. OSNet combines the spatial and channel attention mechanisms and the ResNet backbone for multi-scale feature extraction. However, OSNet focuses only on body features and does not consider face features.

CHAPTER 3

A Comprehensive Overview

In this chapter, we give an overview of our thesis, covering foundational concepts that have played a crucial role in its development. One such concept is multi-feature learning, which we discuss in detail, highlighting its importance in our research. We also delve into the backbone of our model, explaining its architecture and design choices that contribute to its effectiveness. Additionally, we explore the inspirations behind our thesis, shedding light on the research and developments that have influenced our work. Furthermore, we examine various loss functions employed in our model, including triplet loss and softmax loss, elucidating their roles in optimizing feature learning and enhancing the performance of our system. Through this chapter, we aim to establish a strong foundation and provide a comprehensive understanding of the key elements driving our thesis.

3.1 Multi-feature learning

Multi-feature learning is an approach that has gained significant attraction in object detection, aiming to enhance the accuracy and robustness of detection systems. In object detection, multiple features, encompassing aspects like colour, texture, and shape, are extracted from an image. These features are combined to form a comprehensive feature vector, which serves as the basis for object detection algorithms. In the context of Person Re-identification, multi-feature learning shares similarities with object detection in extracting multiple types of features from an image. In Person Re-identification, the focus is on extracting discriminative features that can effectively represent individuals across different camera views or scenarios. These features are carefully selected and combined to construct a feature representation that captures the unique characteristics of an individual. By leveraging multiple features, Person Re-identification systems can mitigate challenges such as variations in lighting conditions, occlusions, and changes in clothing

In Person Re-identification, the goal is to match people across multiple cameras, which is challenging due to the large variations in appearance and pose that can occur. One approach to address this challenge is to use multiple channels that take one image as input and output different feature representations, which are then aggregated to form a final feature vector.

For example, one channel might focus on capturing colour information, another on capturing texture information, and a third on capturing shape information. These channels can be designed using different feature extraction techniques, such as handcrafted or deep learning-based features.

The output of each channel can then be combined using different fusion strategies, such as concatenation, element-wise multiplication, or weighted averaging. The resulting feature vector can then be used for Person Re-identification.

Multi-scale feature learning is also important in Person Re-identification because it helps to capture features at different scales, which is important for dealing with the variations in appearance and pose that can occur. For example, a person's face may provide fine-grained details that are important for identification, while their body shape may provide coarse-grained features that are more robust to changes in pose and clothing.

By extracting features at multiple scales, we can capture both the fine-grained and coarse-grained information that is important for Person Re-identification and combine them to form a more robust feature representation. This approach can help to improve the accuracy and robustness of Person Re-identification systems and reduce their sensitivity to changes in appearance and pose.

3.2 ResNeXt model

ResNeXt [14] is a deep learning model that was introduced by researchers at Face-book AI Research (FAIR) in 2016. It is a variation of the ResNet (Residual Network) model that was originally proposed by Microsoft Research in 2015.

The main idea behind ResNeXt[14] is to increase the representational power of deep neural networks by aggregating the output of multiple parallel pathways, or "cardinality", which allows for more diverse and expressive feature representations.

This is achieved by replacing the single convolutions in the original ResNet model with groups of parallel convolutions, with each group having a different set of filters.

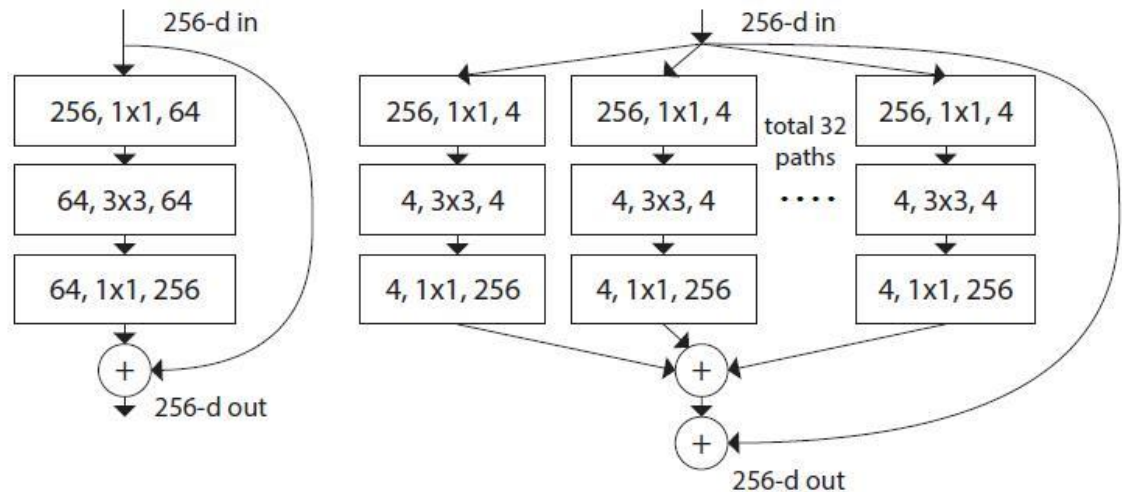


Figure 3.1: ResNeXt architecture [7]

By aggregating the output of multiple groups of convolutions, ResNeXt is able to capture more diverse and complex features, which can lead to improved performance on a wide range of computer vision tasks. In addition to its strong performance, ResNeXt is also highly scalable and can be easily adapted to different network architectures and datasets.

Since its introduction, ResNeXt has become one of the most popular deep learning models in computer vision and has been used to achieve state-of-the-art performance on a wide range of image recognition tasks, such as image classification, object detection, and semantic segmentation.

3.3 Loss Function

3.3.1 Cross-entropy loss

It is used to measure the dissimilarity between the predicted probability distribution and the true distribution of the classes. The Softmax function is employed to convert the logits, which are the unnormalized scores produced by the model, into a normalized probability distribution across classes equation 3.1 is the formula for

cross-entropy loss.

$$\text{Loss} = - \sum_{i=1}^N \log \frac{\exp(W_{y_i}^T \cdot f_i)}{\sum_{j=1}^C \exp(W_j \cdot f_i)} \quad (3.1)$$

where:

N is the number of samples (person images) in the batch. C is the number of classes (persons) in the dataset. f_i represents the extracted feature vector for the i^{th} person image. W_j represents the weight vector for the j^{th} class.

3.3.2 Triplet loss

For deep learning applications like face recognition and picture retrieval, triplet loss is a common loss function. It is intended to learn a feature embedding space where the gaps between samples belonging to the same class are minimized and the gaps between samples belonging to different classes are maximized.

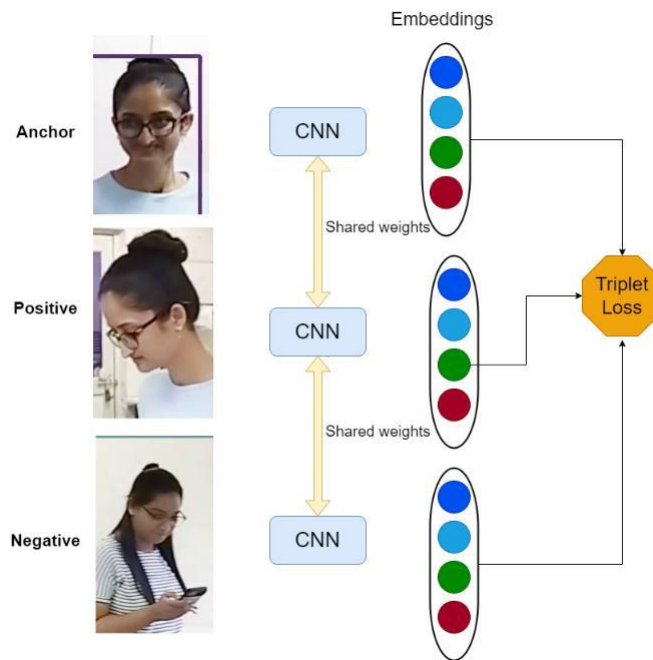


Figure 3.2: Triplet loss

The fundamental concept behind triplet loss is to group three samples together: an anchor, a positive sample, and a negative sample. The positive sample is a different sample from the same class as the anchor sample, and the negative sample is a sample from a different class.

The goal is to make sure that the distance between the anchor and the positive sample is, by a certain margin, smaller than the distance between the anchor and the negative sample refer equation 3.2

$$L = \max(0, d(a, p) - d(a, n) + \text{margin}) \quad (3.2)$$

where:

L is the triplet loss, $d(a, p)$ represents the distance between the anchor sample (a) and the positive sample (p) in the embedding space, $d(a, n)$ represents the distance between the anchor sample (a) and the negative sample (n) in the embedding space, margin is a hyperparameter that represents the desired margin between the positive and negative distances.

triplet loss is used in our model: Our model architecture processes each image through its network to extract features. These features are typically high-dimensional embeddings that represent the visual characteristics of the person in the image. The embeddings are then passed through the triplet loss function to compute the loss given in equation 3.3.

$$L_{\text{triplet}} = \frac{1}{N} \sum_{i=1}^N \max(0, d(a_i, p_i) - d(a_i, n_i) + \alpha) \quad (3.3)$$

In this equation, N represents the total number of triplets in the training set. a_i denotes the anchor sample, p_i represents the positive sample (same class as the anchor), and n_i represents the negative sample (different class from the anchor). $d(a_i, p_i)$ and $d(a_i, n_i)$ denote the distances between the anchor sample and the positive and negative samples, respectively. The term α is the margin, which is a hyperparameter that controls the separation between the positive and negative samples in the embedding space. The equation computes the average triplet loss over all the triplets in the training set.

The distance between anchor and positive pairings needs to be kept minimum while the distance between anchor and negative pairs should be increased, according to the loss function. Learning discriminative embeddings that can effectively differentiate between various people is made easier by this technique.

CHAPTER 4

Proposed Methodology

In this chapter provides a detailed discussion of the architecture used in our Person Re-identification research, implementation details, and the framework employed. It covers the design, components, implementation specifics, and the chosen framework, offering a comprehensive understanding of our approach.

4.1 A multi-scale residual block architecture for Person Re-identification

We present a model specializing in multi-scale feature learning for the re-ID task. The architecture consists of residual blocks equipped with Lite 3x3 layers. The 1x1 layer is used to manipulate feature dimensions, which does not contribute to information aggregation. Residual bottlenecks[33] are the fundamental component



Figure 4.1: (a) Standard 3 × 3 convolution. (b) Lite 3 × 3 convolution. DW: Depth-Wise.[33]

of our architecture, containing the Lite 3 × 3 layer (see Figure. 4.1(a)). Given an input x , The goal of this bottleneck is to discover a residual \tilde{x} with a mapping function F , such that $y = x + \tilde{x}$, where $\tilde{x} = F(x)$. Here, F represents a Lite 3 × 3 layer that learns single-scale features (3 layer that learns characteristics at a single scale (scale = 3). Note that the as 1 × 1 layer are used to change feature dimensions and do not contribute to the aggregation of spatial information, they are omitted

in the notation [5]. In this context, y represents the output of the residual block.

4.2 Depthwise Separable Convolutions

We use depthwise separable convolutions to reduce the number of parameters. [3]. The main idea is to split the convolutional layer. $\text{ReLU}(w * x)$ see Figure 4.2 with kernel $w \in \mathbb{R}^{k \times k \times c \times c_0}$ into two separate layers $\text{ReLU}((v \circ u) * x)$ with depthwise kernel $u \in \mathbb{R}^{k \times k \times 1 \times c_0}$ and pointwise kernel $v \in \mathbb{R}^{1 \times 1 \times c \times c_0}$.

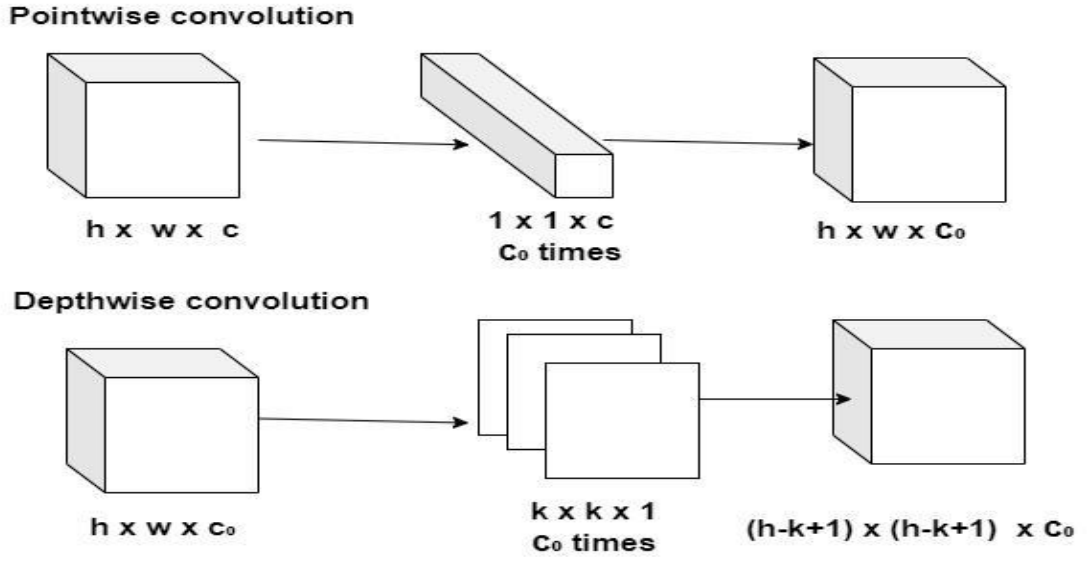


Figure 4.2: Explanation of depthwise separable convolution used in our model.

Where \circ denotes convolution, k is the kernel size, c is the input channel width, and c_0 is the output channel width. Given an input tensor $x \in \mathbb{R}^{h \times w \times c}$ of height h and width w , y computational cost (number of multiplications) without DSC. y' computational cost after using DSC. A number of parameters before DSC is denoted by p , and after DSC, no. of parameters is represented by p' .

$$y = h \cdot w \cdot k^2 \cdot c \cdot c_0 \quad (4.1)$$

$$y' = h \cdot w \cdot (k^2 + c) \cdot c_0 \quad (4.2)$$

$$P = k^2 \cdot c \cdot c_0 \quad (4.3)$$

$$P' = (k^2 + c) \cdot c_0. \quad (4.4)$$

The computational cost is reduced from equation 4.1 to equation 4.2, and the number of parameters from equation 4.3 to equation 4.4.

In our study, $\text{ReLU}((u \circ v) * x)$ pointwise \rightarrow depthwise is used instead of depthwise \rightarrow pointwise, which performed better for omni-scale feature learning [33]. We have called such layer Lite 3×3 hereafter. The implementation is shown in Figure. 4.1 (b).

4.3 Aggregation gate

We propose a unified aggregation gate (AG) for dynamic scale fusion, allowing the combination of multi-scale features. Each feature stream provides information on a single scale, resulting in scale homogeneity. However, by integrating the out-puts of multiple streams with different weights assigned to each scale, omni-scale features can be learned dynamically dependent on the input image. The AG is a trainable neural network shared across all feature streams in a multi-scale residual block. The AG generates channel-wise weights that dynamically fuse the resulting multi-scale feature maps, leading to a more effective representation. The AG has several advantages, including the ability to adjust to input-dependent channelwise weights and efficient model training due to shared parameters. The aggregation gate is a crucial component of our proposed method for learning discriminative features at multiple semantic levels.

In our model, the aggregation gate is a mechanism used to combine and aggregate features from multiple branches or paths in the network. Let's denote the input feature maps from different branches as F_1, F_2, \dots, F_n , where n is the number of branches. The aggregation gate G_i for the i -th branch is computed as follows in equation 4.5:

$$G_i = \sigma(W \cdot g(F_i) + b), \quad (4.5)$$

Where σ represents the activation function, such as sigmoid or softmax, W is the learnable weight matrix, $g(\cdot)$ denotes a transformation function applied to F_i (e.g., global average pooling or 1×1 convolution), and b is the bias term.

The gate values G_i determine the importance or contribution of each branch's features to the final aggregated features. The gate values can be used to scale the feature maps before combining them, typically using element-wise multiplication. The final aggregated feature maps can be computed in equation 4.6:

$$F_{\text{aggregated}} = G_1 \cdot F_1 + G_2 \cdot F_2 + \dots + G_n \cdot F_n. \quad (4.6)$$

The aggregation gate is best explained in figure 4.3. The input tensor (feature maps) enters the aggregation gate. The input tensor undergoes global average pooling, reducing the spatial dimensions to 1x1. The pooled tensor is passed through a 1x1 convolution, which reduces the number of channels while maintaining the spatial dimensions. Optionally, layer normalization is applied to the output of the first convolution. The ReLU activation function is used element-wise to the result of the normalization step. Another 1x1 convolution is performed, generating channel-wise gate values. The specified gate(Relu, sigmoid or linear) by default activation function is sigmoid applied to the gate values. The input tensor is multiplied element-wise with the gate values, selectively amplifying or suppressing features based on the gate values. The final output of the Aggregation module is the result of the element-wise multiplication see Figure 4.3.

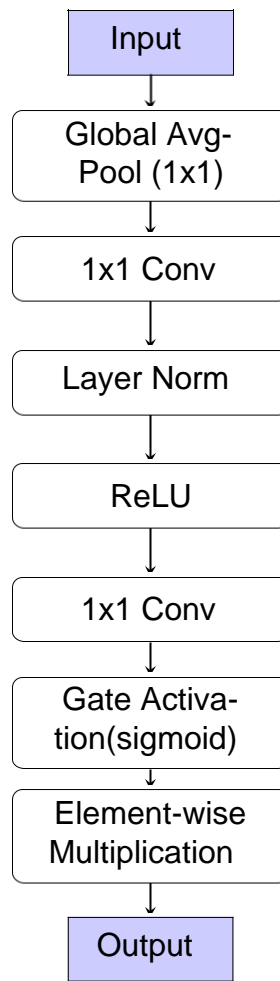


Figure 4.3: Flow diagram of Aggregation Channel Gate

4.4 Proposed Architecture

Our proposed model draws inspiration from both Inception and ResNeXt in terms of their multi-scale design. The architecture of our proposed model is illustrated in Figure 4.4.

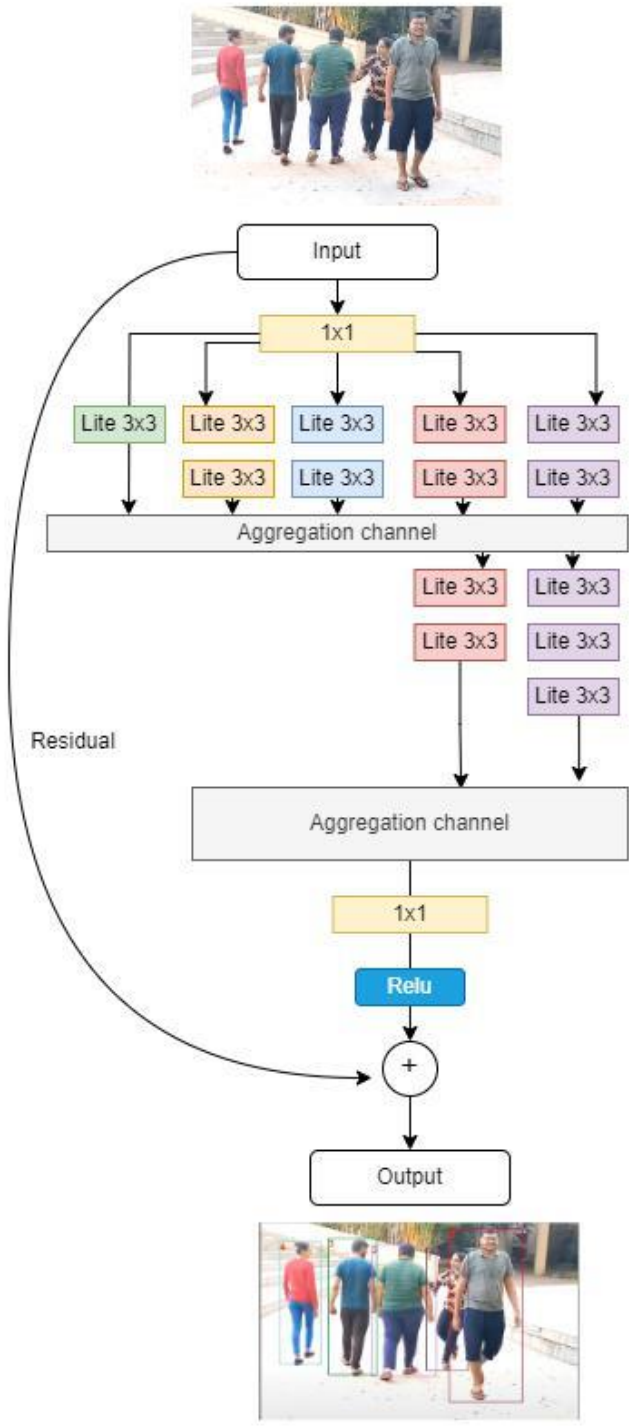


Figure 4.4: Our proposed model

Significance of two aggregation channels:

Introducing double aggregation channels after 2 Lite blocks in multi feature model can have several potential benefits:

Improved Feature Discrimination: Double aggregation channels can help capture more diverse features and improve the model's ability to discriminate between different classes.

Enhanced Model Capacity: The additional channels can increase the model's capacity to learn more complex representations, potentially leading to better performance.

Better Robustness to Variations: By aggregating features from multiple scales and different channels, the model can be more robust to variations in input images, such as changes in lighting conditions or viewing angles.

Reduced Overfitting: With more diverse features and increased model capacity, the risk of overfitting can be reduced, leading to better generalization performance on unseen data.

Overall, introducing double aggregation channels can be an effective strategy to improve the performance of multi-feature model, particularly for challenging computer vision tasks such as Person Re-identification.

4.5 Implementation details:

We set batch size to 64 and weight decay are set to $5e-4$. The person matching is performed based on the L2 distance of 512-D feature vectors extracted from the last fully connected layer. We had fine-tuned our model using imagenet pretrained weights. We train the network with AMSGrad(Adam based optimizer) and an early learning rate of 0.0015 for 150 epochs to fine-tune it. Every 60 epochs, the learning rate decays by 0.1. The ImageNet pre-trained base network is frozen for the initial 10 epochs, leaving only the randomly initialized classifier available for training. Resized images are 256 x 128. We also performed data augmentation methods like random flip and crop. For training, our model has used softmax loss. For testing, we have created three custom video datasets by ourselves captured

from different viewpoints and includes challenges such as occlusion, a shift in the light, a different perspective, and a change in a person's pose and appearance (frontal or back appearance) with the same dresses to fool the model.

4.6 Proposed Framework with Face and Body Fusion for Person Re-identification

In this study, we propose a multi-scale feature learning framework for Person Re-identification that combines body and face features to improve the accuracy of the task. The proposed model utilizes two aggregation channels to extract body features and FaceNet to extract face features.

The combination of these features has been shown to increase accuracy for outdoor scenarios and identical appearances or same dresses, making our model more robust for the task.

For the extraction of face features, we use the MTCNN (Multi-task Cascaded Convolutional Networks)[22] model for face detection. Once a face is detected, it is aligned to a canonical pose, and a 128-dimensional feature vector is extracted using FaceNet[19]. FaceNet[19] is a deep convolutional neural network trained to directly optimize the embedding of face images into a feature space, where the distances between faces correspond to a measure of face similarity see Figure 4.5

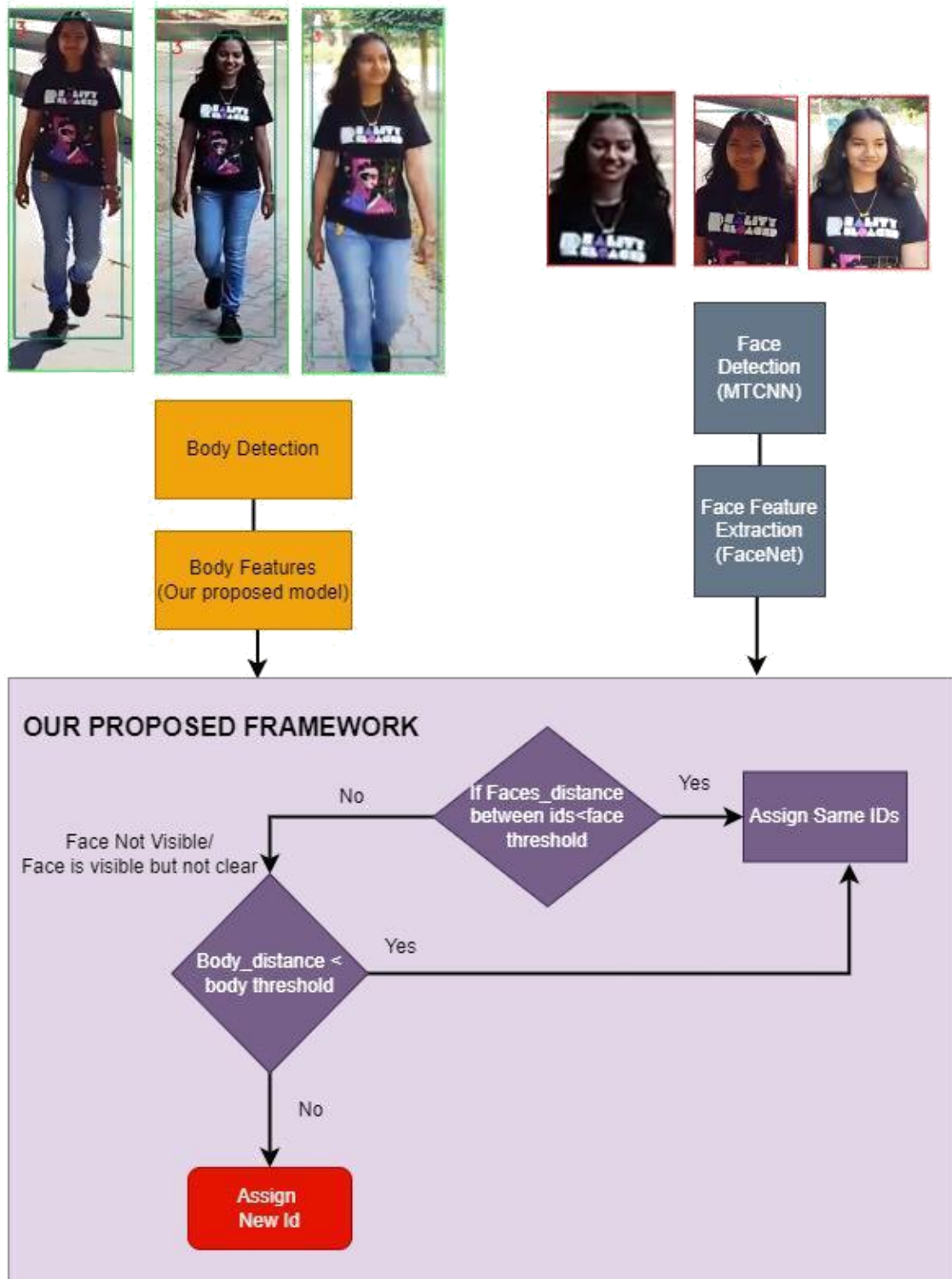


Figure 4.5: Flow diagram of our proposed framework.

Our model is evaluated on standard datasets, including Market1501 and our own same-dress dataset, and the results demonstrate the effectiveness of our proposed method. The combination of face and body features improves the accuracy of the person's re-identification task, particularly in challenging scenarios.

CHAPTER 5

Experimental Results and Analysis

The Experimental Results and Analysis chapter provides an overview of our research findings and analysis. We discuss the datasets used in our experiments, including standard datasets like Market1501 and DukeMTMC-reID and any Custom datasets created for specific evaluations. We present the results obtained by applying our model to these datasets, showcasing the performance of metrics such as mAP, rank-1, rank-5, rank-10, and rank-20. Additionally, we compare the performance of our model with other existing methods in the field, highlighting its strengths and areas of improvement. To enhance the understanding of our results, we provide visual representations of the video results, allowing for a more intuitive evaluation of the performance of our Person Re-identification system. This chapter presents and analyses the experimental outcomes of our research, providing insights into the effectiveness and capabilities of our model.

5.1 Datasets

Market-1501[29] and DukeMTMC-reID[17] are popular benchmark datasets used for Person Re-identification tasks. Market-1501 was introduced in 2015 and contains 32,217 images of 1,501 people captured from six cameras. Each person has an average of 27 images approx, with variations in pose, illumination, and background. The dataset includes manually labelled bounding boxes and a training/testing split.



Figure 5.1: Samples of Market-1501 dataset

DukeMTMC-reID was introduced in 2017 and is larger than Market-1501, containing 36,441 images of 1,812 people captured from eight cameras. Each person has an average of 20 images approx, with variations in clothing, background, and viewpoint.



Figure 5.2: Samples of Market-1501 dataset

The dataset also includes manually labelled bounding boxes and a training/testing split. See Table 5.1 for the datasets description.

Table 5.1: Datasets

Dataset	Year	ID	Boxes	Cameras	Labeled
Market-1501	2015	1,501	32,217	6	DPM+Handcrafted
DukeMTMC-reID	2017	1,812	36,441	8	Handcrafted

Market-1501 and DukeMTMC-reID have become popular benchmark datasets in person re-ID due to their realistic and challenging nature. Many algorithms use these datasets to benchmark their performance and compare it against other state-of-the-art methods. Moreover, both datasets have become widely recognized in the research community and have contributed significantly to advancing the field of person re-ID.

5.2 Custom video datasets

Three custom video datasets were created to test and evaluate Person Re-identification algorithms. The datasets are described below:

Table 5.2: Custom Video Datasets

Dataset	Year	Identities	Camera Views	Difficulty	Challenges
Indoor	2023	3	3	Easy	Controlled environment
Outdoor 1	2023	5	3	Moderate	High light exposure , outdoor, crowd.
Outdoor 2	2023	5	3	Difficult	Identical clothing , outdoor, light exposure.

Custom dataset-1 (Indoor video): The first dataset was an indoor dataset with three distinct identities and three camera views, designed to be relatively easy. The dataset was captured in a controlled environment with three camera views positioned to capture the entire space. The three identities were chosen to be easily distinguishable and easily identifiable, with distinct clothing and physical characteristics. This dataset was designed to be relatively easy as a baseline for testing algorithms see Figure 5.3.



Figure 5.3: Samples of different viewpoints in our Custom dataset-1 (Indoor video)

Custom dataset-2 (Outdoor video-1): The second dataset was an outdoor dataset with five distinct identities and three camera views, designed to be more challenging due to high light exposure, crowded scenes, and varying outdoor lighting conditions. The five identities were chosen to have similar clothing and physical characteristics, making it more difficult for algorithms to match individuals across different camera views accurately samples are shown in Figure 5.4.



Figure 5.4: Samples of different viewpoints in our Custom dataset-2 (Outdoor video-1)

Custom dataset-3 (Outdoor video-2): The third dataset was also an outdoor dataset, but with the added challenge of all individuals wearing identical clothing in an attempt to fool the network. This dataset consisted of five distinct identities and three camera views, with an outdoor setting similar to the second Custom dataset. The added challenge of identical clothing made it more difficult for algorithms to accurately match individuals across camera views as shown in Figure 5.5.



Figure 5.5: Samples of different viewpoints in our Custom dataset-3 (Outdoor video-2)

Overall, these custom video datasets provided a valuable tool for evaluating the performance of Person Re-identification algorithms in various challenging scenarios, and allowed for a more comprehensive assessment of algorithm capabilities in real-world situations. The datasets were designed with different levels of difficulty, including both indoor and outdoor settings, and incorporated challenges such as high light exposure, and identical clothing to test algorithm robustness. The datasets can be used as a benchmark for evaluating the effectiveness of Person Re-identification algorithms in various real-world situations.

5.3 Results

5.3.1 Results on Market-1501 and DukeMTMC-reID datasets

The results that are obtained for our model’s performance on the Market-1501 and DukeMTMC-reID datasets are presented in Table 5.3. Our model achieved an mAP (mean average precision) of 72.1% and 59.5% on Market-1501 and DukeMTMC-reID datasets, respectively. In terms of rank-1 accuracy, our model achieved 89.1% and 79.8% on Market-1501 and DukeMTMC-ReID, respectively.

Table 5.3: Results of our model on public datasets

	Market-1501	DukeMTMC-reID
mAP	72.1%	59.5%
Rank-1	89.1%	79.8%
Rank-5	95.2%	87.9%
Rank-10	96.6%	91.1%
Rank-20	97.7%	93.0%

Our model’s performance was also evaluated at rank-5, rank-10, and rank-20 accuracy levels. At rank-5, our model achieved 95.2% and 87.9% on Market-1501 see Figure 5.6 and DukeMTMC-reID see Figure 5.7, respectively. At rank-10, our model achieved 96.6% and 91.1% on Market-1501 and DukeMTMC-reID, respectively. Finally, at rank-20, our model achieved 97.7% and 93.0% on Market-1501 and DukeMTMC-reID, respectively. Overall, our model performed well on both datasets, particularly regarding rank-1 accuracy and mAP .

5.3.2 MARKET-1501 Rankings(Predicted matching images) for sample query images:

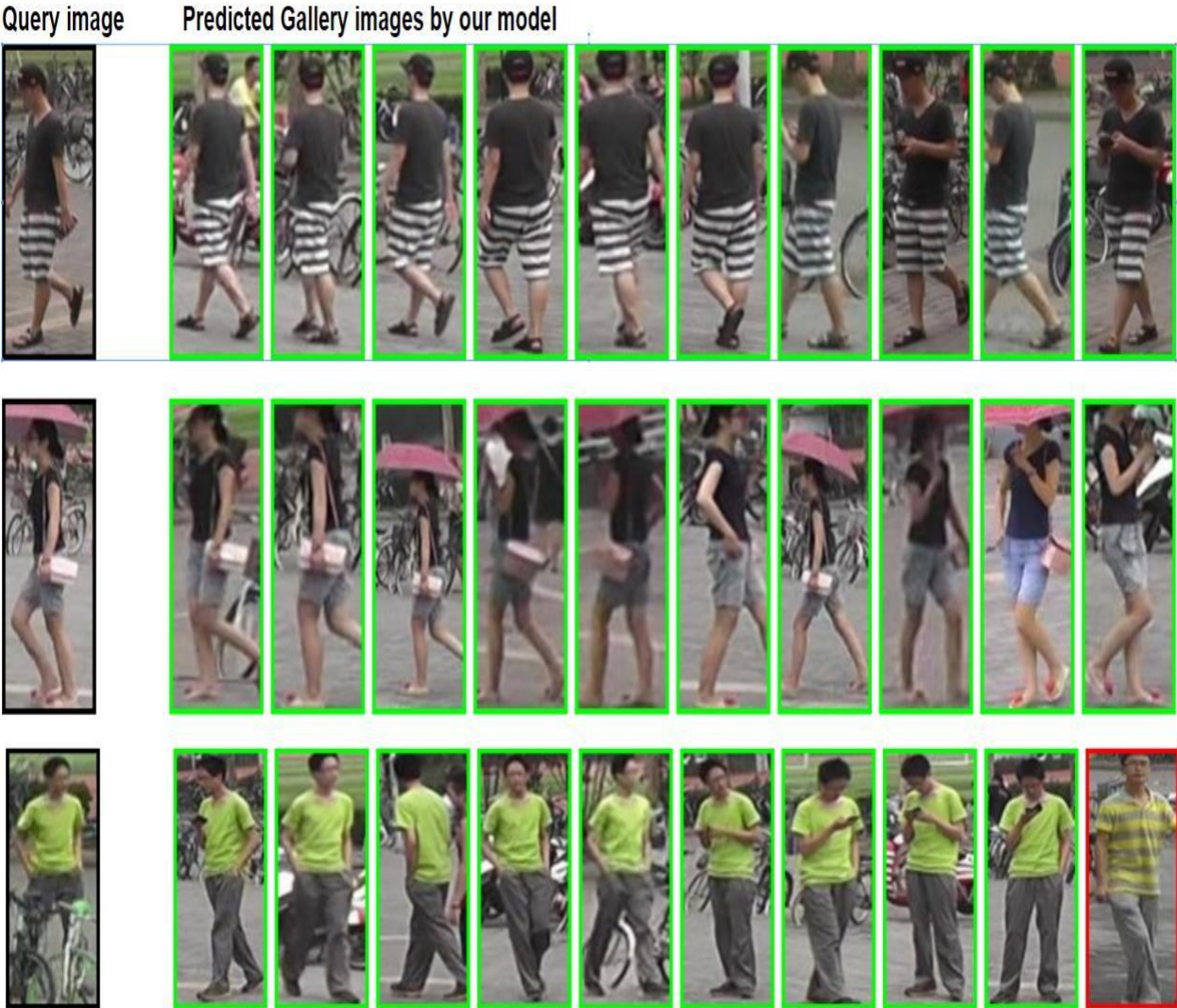


Figure 5.6: MARKET-1501 Rankings of our model

**5.3.3 DukeMTMC-reID Rankings(Predicted matching images)
for sample query images:**



Figure 5.7: DukeMTMC-reID Rankings of our model

5.4 Comparison of our model with other methods

Table 5.4: Comparison of our method with other methods on Market-1501 and DukeMTMC-reID datasets

Method	Rank-1 (Market-1501)	mAP (Market-1501)	Rank-1 (DukeMTMC-ReID)	mAP (DukeMTMC-ReID)
Verif-Identif [30]	79.5	59.9	68.9	49.3
DCF [9]	80.3	57.5	-	-
SVDNet [21]	82.3	62.1	76.7	56.8
PAN [32]	82.8	51.5	71.6	51.5
DeformGAN [20]	80.6	61.3	-	-
LSRO [31]	84.0	66.1	67.7	47.1
PT [10]	87.7	68.9	78.5	56.9
Multi-pseudo [8]	85.8	67.5	76.8	58.6
ShuffleNet [25]	84.8	65.0	71.6	49.9
MobileNetV2 [18]	87.0	69.5	76.2	55.8
Our model	89.1	72.1	79.8	59.5
PN-GAN [16]	89.4	72.6	73.6	53.2

Our model outperforms the other methods, including PN-GAN, in terms of the DukeMTMC-reID dataset, known for its complexity (8 camera views) and many unique identities. Our model has a higher Rank-1: 79.8% and mAP: 59.5% on DukeMTMC-ReID, which is higher than PN-GAN accuracy. However, PN-GAN falls short in comparison, with lower performance scores in DukeMTMC-reID datasets, as the PN-GAN model needs more generalization, i.e. the ability to generalize across different datasets and real scenarios. In comparison, our model results outperform the other methods in terms of metric rank-1 and mAP.

5.5 Results on custom video Datasets:

5.5.1 Results on Custom dataset-1 (Indoor video)

In the indoor dataset, we observed three individuals in three camera views: two girls and one boy. The first camera serves as a reference, and the IDs assigned to the individuals in that camera are considered the ground truth. The boy consistently maintains an ID of 7 across all three cameras. The girl wearing the blue t-shirts is assigned ID 1, while the second girl is consistently assigned ID 13 in all three cameras. This means we are getting pretty good results in the video under a controlled environment like a lab see Figure 5.8.



Figure 5.8: Results on Custom dataset-1 (Indoor video)

Accuracy on on Custom dataset-1 (Indoor video):

$$\text{Accuracy} = \frac{\text{Number of correctly matched individuals}}{\text{Total number of individuals}} \quad (5.1)$$

In the above equation, the "Number of correctly matched individuals" refers to the count of correctly identified and matched individuals across the three videos. The "Total number of individuals" represents the overall number of individuals in the video sequence. Since we have an equal number of identities in all three cameras and in a controlled situation, all three identities are correctly re-identified for this particular dataset; we achieved 100% accuracy.

mAP on Custom Video Datasets:

For calculating mAP on videos, we took a frame every 3secs from the output video(Combined video of all cameras with re-identification task performed). If id is present the first time, that particular frame will be considered the reference frame for that id, which would be regarded as the reference(ground truth id) for it. For the rest of the frames, we calculated the precision of every id in videos, and

then we computed mean average precision as we were doing with image datasets like Market-1501 DukeMTMC-ReID.

For Custom dataset-1(indoor) Total frames count is 11 and mAP is 100%.

5.5.2 Results on Custom dataset-2 (Outdoor video-1)

In-camera 1, see Figure 5.9 the frame which consists of all five people is the first frame which will be served as a reference for other frames as it gives ids to all five people who appear in camera-1 for the first time. Since few people are only visible from the back while few are visible from the front, our algorithm can successfully re-identify them when they are again entered in the view of camera 1; you observe the re-identification in all three figures shown in Figure 5.9. The red sweater girl's first appearance was from backward, but the algorithm correctly re-identified her when she re-appear in the camera from the front. Similarly, one girl(id:1) and one boy(id:5) who appear from the front in the first frame are re-identified correctly in other frames where they reappear from backward.



Figure 5.9: Results on CAMERA-1 Custom dataset-2 (Outdoor video-1)

In camera-2, see Figure 5.10 the people who appear in camera-2 are re-identified correctly and got ids given to them by camera-1. Below are a few frames of the

camera-2 view. In camera-2, lighting and background are also quite challenging as in this camera view we have shadows of trees.



Figure 5.10: Results on CAMERA-2 Custom dataset-2 (Outdoor video-1)

In Camera-3, see Figure 5.11 in this camera view, the light exposure is quite high, still, ids 5, 4, and 1 are re-identified correctly. This shows our algorithm mainly works on global features such as clothes, body shape etc. hence, the feature captured by this body is called body features.



Figure 5.11: Results on CAMERA-3 Custom dataset-2 (Outdoor video-1)

Accuracy on Custom dataset-2 (Outdoor video-1): Using equation 5.1, In this dataset, the number of people is unequal in a particular video, unlike Custom dataset-1, So accuracy is calculated for each video, and then the average accuracy

of Custom dataset-2 (Outdoor video-1) is calculated.

In our study, we consider the initial appearance of each person in camera-1 as the reference point. Subsequently, if a person reappears in camera-1, their identification is included when calculating accuracy. Notably, in the specific camera-1 video, all five persons observed initially have reappeared.

Camera-1: $4/5 \times 100 = 80\%$

Camera-2: $3/3 \times 100 = 100\%$

Camera-3: $3/3 \times 100 = 100\%$

Average accuracy: 93.33%

The average accuracy for the Custom dataset-2 (Outdoor video-1) is 93.33%, the Total frame count is 18 and the mAP is 93.33%.

5.5.3 Results on Custom dataset-3 (Outdoor video-2)

To assess the limitations of our model and explore its drawbacks in some cases, we conducted an experiment where we intentionally created a person identical to appear in the camera image. Surprisingly, it was found that our model did not accurately identify people wearing the same clothes as humans.

To solve this problem and improve the accuracy of our model, we realized the importance of identifying facial and body features for the human experience. Armed with this understanding, we create a framework that takes these two factors into account.

The frameworks sequentially, starting with measuring the faces. Set the face threshold, and if the distance between two people's faces is lower(similar face) than the threshold, they are given the same tag or ID. But if the faces differ, the frame starts comparing people based on their physical characteristics.

For comparison of physical characteristics, a distinction is defined as body features. If the distance between two people is below this threshold, they are considered similar and assigned the same ID. Instead, they will be assigned separate identities if their body features differ, indicating a different identity. By using facial and body features in our decision-making process, we aim to improve accuracy and solve the uniformity challenge. This framework enables further analysis of individuals,

ensuring that many dimensions of their appearance are considered for reliable and accurate identification.

For example, the girl (id:3) with chains, blue jeans, and curly hair appears in all three camera views with the same identity, our framework has correctly captured her (id:3) face and body features, so she (id:3) re-identified correctly. Similarly, the girl with id:8 does not appear by face in camera-3 but is assigned id:8 because of body feature capturing. Similarly, other ids are re-identified correctly, as shown in all five frames in Figure 5.12.

Accuracy on Custom dataset-3 (Outdoor video-2): Using equation 5.1, the number of people in this dataset is unequal in the videos. So accuracy is calculated for each video, and then the average accuracy of Custom dataset-3 (Outdoor video-2) is calculated.

In our study, we consider the initial appearance of each person in camera-1 as the reference point. Subsequently, if a person reappears in camera-1, their identification is included when calculating accuracy. Notably, in the specific camera-1 video, no person re-appeared.

$$\text{Camera-2: } 3/4 \times 100 = 75\%$$

$$\text{Camera-3: } 5/5 \times 100 = 100\%$$

$$\text{Average accuracy: } 87.5\%$$

The average accuracy for the Custom dataset-2 (Outdoor video-1) is 87.5%, The Total frame count is 22 and the mAP is 75.83%.

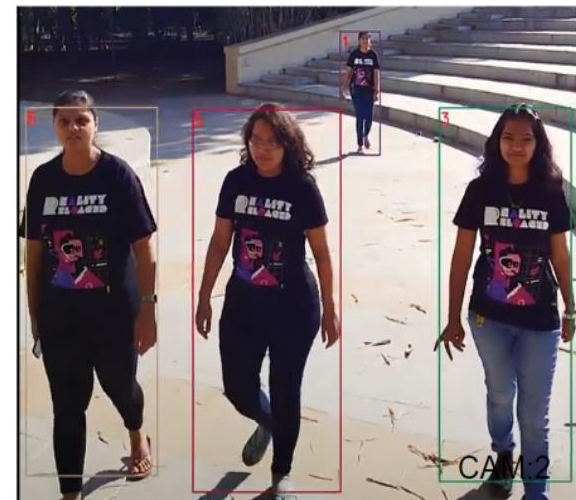


Figure 5.12: Results on Custom dataset-3 (Outdoor video-2)

5.5.4 Accuracy and mAP on Custom datasets

Custom datasets	Accuracy	mAP
Custom dataset-1 (Indoor video)	100%	100%
Custom dataset-2 (Outdoor video-1)	93.33%	93.33%
Custom dataset-3 (Outdoor video-1)	87.5%	75.83%

Table 5.5: Accuracy and mAP on Custom datasets.

We evaluated the performance of our Person Re-identification model on three custom videos. The datasets included Custom dataset-1, which had an equal dis-tribution of individuals and a controlled indoor environment, and two challenging datasets: Custom dataset-2 (Outdoor video-1) and Custom dataset-3 (Outdoor video-2).

Our model achieved impressive accuracy on Custom dataset-1. This indicates that our model performed exceptionally well in accurately matching individuals in an indoor environment with a balanced distribution.

On Custom dataset-2 (Outdoor video-1), our model achieved an accuracy of 93.33%. This dataset presented more challenges due to the outdoor setting, such as varying lighting conditions and potential occlusions. Despite these difficulties, our model demonstrated a strong performance, accurately identifying individuals in most cases.

For Custom dataset-3 (Outdoor video-2), our model achieved an accuracy of 87.5%. This dataset is more challenging, likely due to additional complexities in the outdoor environment, such as similar clothing. Nonetheless, our model still achieved a significant level of accuracy, providing promising results. The mAP 75.83% is low compared to accuracy because the model predicted more ids that were actually present in the video.

Table 5.5 summarizes the accuracy achieved by our model on the Custom datasets, highlighting the varying performance across different environments. These re-sults showcase the effectiveness of our model in Person Re-identification tasks, particularly in controlled and challenging scenarios.

CHAPTER 6

Conclusions and Future scope

In conclusion, we proposed a lightweight and robust model for Person Re-id that uses multi-scale feature learning and fusion of body and face features. The model employs the YOLOv4 object detection model for pedestrian detection and the DeepSORT algorithm for tracking. The ResNeXt[14] architecture is used as a backbone to learn discriminative features at multiple semantic levels.

Our experimental results on the widely-used datasets: Market1501, DukeMTMC-reID, and our own datasets demonstrate the effectiveness of our proposed method in Person Re-id, even in challenging scenarios such as occlusion, lighting variations, and identical appearances. The incorporation of face features using FaceNet[19] further improves the accuracy of our model.

Future work may explore the use of other face recognition models or alternative methods for fusing body and face features. Additionally, efforts may be made to improve the testing time of the model to make it more suitable for real-time applications. Overall, our proposed method has the potential to be applied in various real-world scenarios, such as surveillance cameras in markets, shopping malls, parking areas, and other public places.

Also, the model works efficiently for pre-recorded videos but can be faster and more accurate for real-time person re-identification. There is also scope for improvement in loss function and trying to make the testing of the model faster.

References

- [1] L. An, X. Chen, and S. Yang. Person re-identification via hypergraph-based matching. *Neurocomputing*, 182:247–254, 2016.
- [2] Y.-C. Chen, W.-S. Zheng, J.-H. Lai, and P. C. Yuen. An asymmetric distance model for cross-view feature mapping in person reidentification. *IEEE transactions on circuits and systems for video technology*, 27(8):1661–1675, 2016.
- [3] F. Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258, 2017.
- [4] R. De Maesschalck, D. Jouan-Rimbaud, and D. L. Massart. The mahalanobis distance. *Chemometrics and intelligent laboratory systems*, 50(1):1–18, 2000.
- [5] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *cvpr*. 2016. arXiv preprint arXiv:1512.03385, 2016.
- [6] A. Hermans, L. Beyer, and B. Leibe. In defense of the triplet loss for person re-identification. arXiv preprint arXiv:1703.07737, 2017.
- [7] S. Hitawala. Evaluating resnext model architecture for image classification. arXiv preprint arXiv:1805.08700, 2018.
- [8] Y. Huang, J. Xu, Q. Wu, Z. Zheng, Z. Zhang, and J. Zhang. Multi-pseudo regularized label for generated data in person re-identification. *IEEE Transactions on Image Processing*, 28(3):1391–1403, 2018.
- [9] D. Li, X. Chen, Z. Zhang, and K. Huang. Learning deep context-aware features over body and latent parts for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 384–393, 2017.
- [10] J. Liu, B. Ni, Y. Yan, P. Zhou, S. Cheng, and J. Hu. Pose transferrable person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4099–4108, 2018.

- [11] Q. Meibin, W. Yunxia, T. Shengshun, et al. Person re-identification based on regularization of independent measure matrix. *Pattern Recognition and Artificial Intelligence*, 29(6):511–518, 2016.
- [12] Z. Ming, M. Zhu, X. Wang, J. Zhu, J. Cheng, C. Gao, Y. Yang, and X. Wei. Deep learning-based person re-identification methods: A survey and outlook of recent works. *Image and Vision Computing*, 119:104394, 2022.
- [13] H. S. Oliveira, J. J. Machado, and J. M. R. Tavares. Re-identification in urban scenarios: A review of tools and methods. *Applied Sciences*, 11(22):10809, 2021.
- [14] G. Pant, D. Yadav, and A. Gaur. Resnext convolution neural network topology-based deep learning model for identification and classification of pediastrum. *Algal research*, 48:101932, 2020.
- [15] N. Pu, W. Chen, Y. Liu, E. M. Bakker, and M. S. Lew. Lifelong person re-identification via adaptive knowledge accumulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7901–7910, 2021.
- [16] X. Qian, Y. Fu, T. Xiang, W. Wang, J. Qiu, Y. Wu, Y.-G. Jiang, and X. Xue. Pose-normalized image generation for person re-identification. In *Proceedings of the European conference on computer vision (ECCV)*, pages 650–667, 2018.
- [17] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part II*, pages 17–35. Springer, 2016.
- [18] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.
- [19] 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*, pages 815–823, 2015.
- [20] A. Siarohin, E. Sangineto, S. Lathuiliere, and N. Sebe. Deformable gans for pose-based human image generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3408–3416, 2018.

- [21] Y. Sun, L. Zheng, W. Deng, and S. Wang. Svdnet for pedestrian retrieval. In Proceedings of the IEEE international conference on computer vision, pages 3800–3808, 2017.
- [22] J. Xiang and G. Zhu. Joint face detection and facial expression recognition with mtcnn. In 2017 4th international conference on information science and control engineering (ICISCE), pages 424–427. IEEE, 2017.
- [23] S. Xuan and S. Zhang. Intra-inter camera similarity for unsupervised person re-identification. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 11926–11935, 2021.
- [24] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. C. Hoi. Deep learning for person re-identification: A survey and outlook. IEEE transactions on pattern analysis and machine intelligence, 44(6):2872–2893, 2021.
- [25] X. Zhang, X. Zhou, M. Lin, and J. Sun. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 6848–6856, 2018.
- [26] X. Zhao, N. Wang, Y. Zhang, S. Du, Y. Gao, and J. Sun. Beyond pairwise matching: Person reidentification via high-order relevance learning. IEEE transactions on neural networks and learning systems, 29(8):3701–3714, 2017.
- [27] K. Zheng, W. Liu, L. He, T. Mei, J. Luo, and Z.-J. Zha. Group-aware label transfer for domain adaptive person re-identification. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 5310–5319, 2021.
- [28] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian. Scalable person re-identification: A benchmark. In Proceedings of the IEEE international conference on computer vision, pages 1116–1124, 2015.
- [29] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian. Scalable person re-identification: A benchmark. In Proceedings of the IEEE international conference on computer vision, pages 1116–1124, 2015.
- [30] Z. Zheng, L. Zheng, and Y. Yang. A discriminatively learned cnn embedding for person reidentification. ACM transactions on multimedia computing, communications, and applications (TOMM), 14(1):1–20, 2017.

- [31] Z. Zheng, L. Zheng, and Y. Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In Proceedings of the IEEE international conference on computer vision, pages 3754–3762, 2017.
- [32] Z. Zheng, L. Zheng, and Y. Yang. Pedestrian alignment network for large-scale person re-identification. IEEE Transactions on Circuits and Systems for Video Technology, 29(10):3037–3045, 2018.
- [33] K. Zhou, Y. Yang, A. Cavallaro, and T. Xiang. Omni-scale feature learning for person re-identification. In Proceedings of the IEEE/CVF international conference on computer vision, pages 3702–3712, 2019.