Facial and Body features based Multi-Model Person Re-Identification: MPRe-ID

Nikhil Kumar Singh
Research Scholar,
Gujarat Technological University,
Ahmedabad, Gujarat

Manish Khare
Assistant Professor,
DAIICT Gandhinagar,
Gujarat

Hemani Bharadwaj
DAIICT Gandhinagar,
Gujarat

Harikrishna B. Jethva
Associate Professor,
Government Engineering College, Patan,
Gujarat

ABSTRACT

Within the surveillance area, Person Re-identification (Re-ID) holds considerable importance by enabling the matching of a person's appearance across multiple non-overlapping cameras. Nonetheless, this task poses challenges due to factors like changes in camera viewpoints, occlusion, and variations in appearance, including clothing, shoes, and pose. Overcoming these challenges requires discriminative feature learning. Deep convolutional neural networks (CNNs) have recently gained widespread usage to address this objective. This study introduces a lightweight and robust deep learning framework Multi-Model person Re-ID (MPRe-ID) for person re-identification. It incorporates the YOLOv4 object detection model for pedestrian detection and utilizes SORT with deep metric association (DeepSORT) algorithm for tracking. MPRe-ID uses novel body feature extraction model to learn discriminative features at various semantic levels, leveraging the ResNeXt architecture as its backbone. The proposed body feature extraction model contains multiple blocks where channels are concatenated between blocks, and an aggregation gate is employed to aggregate the output of multiple channels. The aggregation gate produces channel-wise weights dynamically, facilitating the fusion of resulting multi-scale feature maps. This layout effectively enables the model to extract discriminative features even in challenging conditions. To evaluate the efficacy of our proposed MPRe-ID framework including body and Face features, we conducted experiments on the widely-used Market1501 and DukeMTMC-reID dataset. The experimental results compared with state-of-the-art approaches demonstrate the effectiveness of our MPRe-ID approach.

Keywords

Person Re-identification, Facial features, YOLO, ResNeXt, Convolution Neural Network, DeepSORT, Body features

1. INTRODUCTION

Person re-identification, a significant application in video surveillance, has gained popularity in the Computer Vision and Image Processing research communities over the past decade due to its potential for enhancing safety and security. It involves identifying a person of interest across distributed, non-overlapping camera views, making it valuable for security purposes, particularly in identifying potential threats in public spaces such as shopping malls, railway stations, airports, and large events. The process faces challenges like variations in lighting conditions, different poses, viewpoints, blurring effects, image resolution, and background changes [1].

Person Re-identification (Re-ID) holds a crucial role in surveillance videos, particularly in multi-camera settings. The main goal of person re-ID systems is to consistently assign identification numbers to individuals captured in each non-overlapping camera view. Intra-class variances can arise due to factors such as background variations, atmospheric changes, human motions, different camera perspectives, and other elements. Figure 1 illustrates examples of diverse camera views.

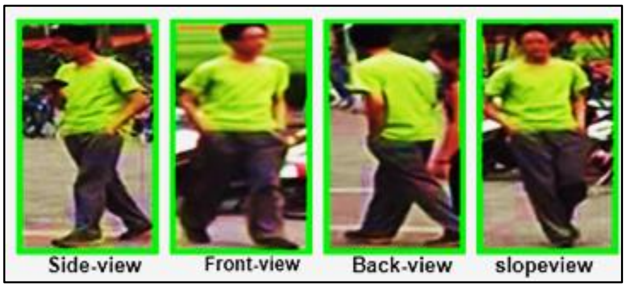

Figure 1: An illustration of several camera perspectives

Person Re-ID encompasses three key stages. The initial crucial step involves detecting the person within the frame captured by the surveillance camera. Following the detection, it becomes imperative to track the same individual across all frames, establishing a stable ID for tracking purposes. Figure 2 provides a summary of these processes.

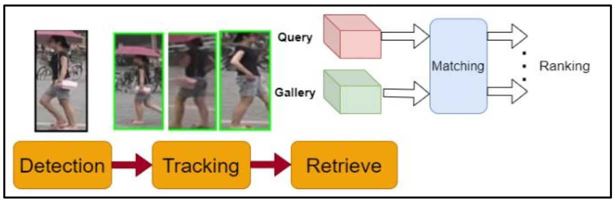

Figure 2: Procedures for the Person Re-ID task

The objective is to derive pertinent features from the dataset post its collection and processing. Various traits of an individual's image can be extracted through different methods. There are two primary approaches for feature extraction, namely local or patch-based and global processing. The global approach centres on the camera's topology, considering factors like the physical location of the camera. For instance, if there are cameras at the entrance and exit, it is evident that a person initially appears on the entrance camera before being captured on the exit camera. Conversely, the patch-based approach focuses on intricate details within the image. Local or patch-based methods are valuable for discriminating intra-class samples. Figure 3 illustrates some of the patch-wise processing techniques.

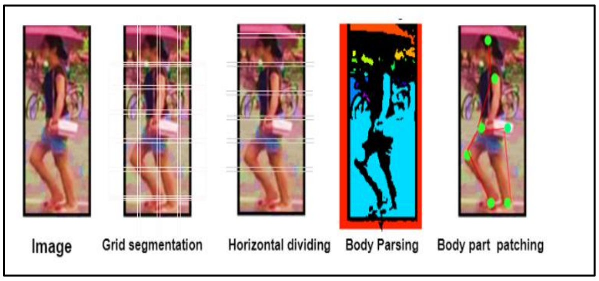

Figure 3: Various patch-wise data processing algorithms to obtain the local features

Person Re-identification (Re-ID) poses significant challenges stemming from alterations in camera perspectives, occlusion, and diverse appearances, including clothing, footwear, and pose. Overcoming these challenges necessitates the acquisition of discriminative features. While deep convolutional neural networks (CNNs) have been extensively employed for this purpose, existing Re-ID models often fall short in their ability to learn features comprehensively across various scales. This study aims to address this limitation by proposing an enhanced Re-ID model capable of effectively learning discriminative features at multiple semantic levels. Numerous obstacles accompany this task, including:

- **Occlusion:** Occurring when individuals are partially or entirely concealed from view, occlusion presents a significant challenge in extracting meaningful identification features. It may result from objects, other individuals, or self-occlusion due to body pose.
- **Illumination Changes:** Drastic variations in lighting conditions across different environments lead to alterations in individuals' appearances, posing a challenge for accurate identification.
- **Similarities among Individuals:** Individuals sharing similarities in clothing style, body shapes, or general appearances can potentially mislead the network during the identification process.
- **Camera Calibration and Viewpoint Variation:** Variations in camera views, angles, and calibrations directly impact captured images, making it challenging to detect and identify individuals across different viewpoints, angles, or calibrations.
- **Scalability:** The scalability of Person Re-identification algorithms becomes crucial as dataset sizes or surveillance system complexities increase. Efficient algorithms are essential to handle large-scale datasets and ensure real-time processing in practical scenarios.

Our research aims to propel the development of an efficient and resilient network for Person Re-identification, leveraging the capabilities of deep learning to acquire distinctive features across multiple semantic levels. In our innovative MPRe-ID approach, we introduce a methodology that combines body and face features. Body detection is achieved through YOLOv4 object detection [2] and SORT with deep metric association (DeepSORT) tracking [3], with novel body feature extraction model for extracting body features. Simultaneously, face features are obtained using the FaceNet model. Employing the ResNeXt architecture as the backbone ensures the effective extraction of discriminative features, even in challenging scenarios involving occlusion, lighting variations, and similar body characteristics.

The rest of the paper is structured as follows: The "Literature Review" section provides an overview of person re-identification methodologies employing diverse deep learning models. "Datasets" section discusses the used datasets in evaluating performance, including Market 1501, DukeMTMC-ReID. The "Proposed Approach" section delineates the architecture and methodologies employed in our work. Following that, the "Implementation" section outlines the experimental methodology. Subsequently, the "Results and Analysis" section presents the experimental findings and their analysis across the utilized datasets. Finally, the "Conclusion" section offers a summary of the paper.

2. LITERATURE SURVEY

In the past two decades, researchers have shown significant interest in advancing Re-ID systems. Person Re-id systems generally fall into two categories: feature-based and metric-based approaches. Feature-based approaches aim to efficiently represent a person using distinctive features, while metric-based approaches concentrate on developing effective metrics for measuring the similarity between images of two individuals [4].

To comprehensively explore the existing literature in this domain, we organize the literature survey based on the methodologies employed in developing person Re-ID systems. This paper references noteworthy works in the field, highlighting recent advancements that predominantly centred on deep learning techniques. The adoption of multi-scale feature learning in deep learning has gained popularity for the development of Re-ID systems.

In 2022, Nikhil et. al. [1] explores popular datasets like ViPER, iLIDS, Market1501, DukeMTMC4ReID, CUHK01, CHUK02, CHUK03, PRID2011, detailing parameters such as the number of persons, images, cameras, frame sizes, and associated challenges. They elaborated various aspects of re-identification approaches, including temporal, spatial, feature, distance metric, machine learning, and automation to uncover solutions for addressing the numerous difficulties and challenges prevalent in the field, based on the latest research works.

The emphasis of the metric learning approach is primarily on the similarity metric. In the early 2000s, the introduction of the Mahalanobis distance [5] was a significant development in the field of Person Re-identification (Re-ID) for measuring similarity. However, it was observed to be susceptible to over fitting. In response to this challenge, Meibin et. al. [6] proposed the regularized independent metric. With the increasing demand for surveillance videos and the use of multiple cameras to bolster security, conventional metrics struggled to adapt to these evolving scenarios.

Chen et. al. [4] proposed an asymmetric distance metric to address the challenges of person re-identification. Xiaojing et. al. [7] introduced a hyper graph based metric as an alternative to Cartesian systems. Zhao et. al. [8] later presented an enhanced version of the hyper graph method, incorporating joint learning. Metric learning algorithms can be broadly categorized into classical metric learning algorithms and deep-learning-based metric learning algorithms. Classical metric learning involves learning a distance metric (e.g., Mahalanobis Distance), while deep-learning-based metric learning uses neural networks to learn discriminative embedding’s (e.g., Siamese Networks, Triplet Networks). Pu et. al. [9] addressed the limitations of stationary domain person Re-ID by introducing a novel framework called Adaptive Knowledge Accumulation (AKA) for knowledge representation.

To tackle challenges in unsupervised person Re-ID systems, Xuan et. al. [10] proposed incorporating intra-inter camera similarity computations to account for variations caused by multiple cameras. The fusion of inter-camera and intra-camera similarities has significantly enhanced the performance of person Re-ID systems. Furthermore, Zheng et. al. [11] introduced a grouping-based approach to improve unsupervised person Re-ID, leveraging the concept of unsupervised domain adaptiveness, where a system trained on labelled domains can be applied to unlabelled domains without requiring annotations. Additionally, various deep learning-based approaches for Person Re-identification frameworks have emerged, extensively covered in a recent survey paper by Ye et al. [12] and Ming et. al. [13].

In recent years, multi-scale research has gained popularity, particularly in the context of Person Re-identification (Re-ID), a widely explored topic in computer vision. Researchers have proposed various approaches to tackle the challenges associated with Re-ID, encompassing deep learning-based methods and multi-feature learning methods. Deep learning-based methods have demonstrated impressive performance in Re-ID by leveraging their capacity to learn discriminative features from extensive datasets. For instance, Zheng et. al. [14] introduced a deep neural network architecture dedicated to learning a discriminative embedding space for Re-ID. Similarly, Hermans et. al. [15] utilized triplet loss for the deep feature embedding space learning in Re-ID.

Multi-feature learning methods have also garnered attention in the Re-ID domain. These methods focus on extracting complementary features from diverse modalities, such as color, texture, and shape, combining them to enhance Re-ID performance. An example is OSNet [16], which proposes a lightweight network architecture tailored for efficient person Re-ID. OSNet integrates spatial and channel attention mechanisms with the ResNet backbone for multi-scale feature extraction. However, it's noteworthy that OSNet concentrates solely on body features, overlooking considerations for face features.

3. DATASETS

Datasets Market-1501 [17] and DukeMTMC-reID [18] 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. Sample of the dataset images are shown in figure 4 and figure 5.

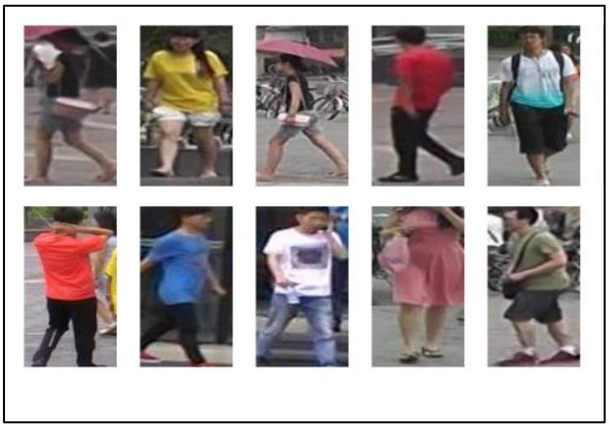

Figure 4: Samples of Market-1501 Dataset

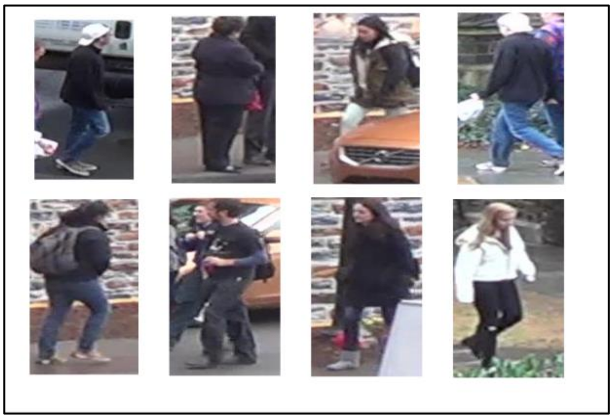

Figure 5: Samples of DukeMTMC-reID dataset

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. The dataset also includes manually labelled bounding boxes and a training/testing split. See Table 1 for the datasets description.

Table 1: Summary of Market-1501 and DukeMTMC-reID Datasets

Dataset	ID	Boxes	Cameras	Labeled
Market-1501	1,501	32,217	6	DPM+Handcrafted
DukeMTMC-reID	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.

4. PROPOSED APPROACH

In this section we have proposed a novel body feature extraction model and a multi-model person re-identification (MPRe-ID) Framework which are explained as below.

4.1 Proposed Body Feature Extraction Model

The propose model uses the following components to extract the body features –

a. **Backbone architecture-ResNeXt Model [19]:** ResNeXt, introduced by researchers at Facebook AI Research (FAIR) in 2016, is a deep learning model. It serves as a variation of the ResNet (Residual Network) model, initially proposed by Microsoft Research in 2015. The core concept behind ResNeXt is to enhance the representational capability of deep neural networks by consolidating the output from multiple parallel pathways, referred to as "cardinality," allowing for more diverse and expressive feature representations. This enhancement is accomplished by substituting the single convolutions in the original ResNet model with groups of parallel convolutions, each group employing a distinct set of filters.

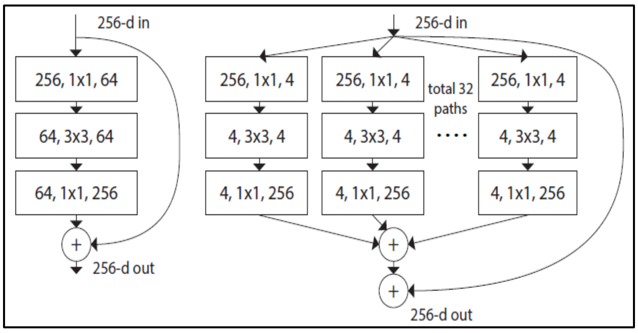

Figure 6: ResNeXt architecture [19]

By aggregating the output from multiple convolution groups, ResNeXt excels in capturing diverse and intricate features, leading to enhanced performance across various computer vision tasks. Not only does ResNeXt exhibit robust performance, but it also boasts scalability, allowing easy adaptation to different network architectures and datasets. Since its inception, ResNeXt has gained significant popularity in the field of computer vision and has been instrumental in achieving state-of-the-art results in tasks such as image classification, object detection, and semantic segmentation. In the proposed model, we utilized two loss functions: cross-entropy loss and triplet loss.

1. **Cross-Entropy Loss [20][21]:** This loss measures the dissimilarity between the predicted probability distribution and the true distribution of classes. The Softmax function is applied to convert the logits (unnormalized scores produced by the model) into a normalized probability distribution across classes. The formula for cross-entropy loss is –

$$Loss = -\frac{1}{N}\sum_{i=1}^{N} \log \left(\frac{(e^{W_{y_{i}*f_i}^{T}})}{\sum_{j=1}^{C} e^{(W_{y_{i}*f_i}^{T})}}\right)$$

   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 ith person image and W_j represents the weight vector for the jth class.

2. **Triplet Loss [22]:** Triplet loss is a common loss function for deep learning applications like face recognition and image retrieval. It aims to learn a feature embedding space where the gaps between samples of the same class are minimized, while the gaps between samples of different classes are maximized.

   The fundamental concept involves grouping three samples together: an anchor (A), a positive sample (P) (from the same class as the anchor), and a negative sample (N) (from a different class). The goal is to ensure that the distance between the anchor and the positive sample is smaller, by a certain margin (α), than the distance between the anchor and the negative sample.

$$L_{triplet} = \lfloor \| f(A) - f(P) \|_2^2 - \| f(A) - f(N) \|_2^2 + \alpha \rfloor_+ \quad \ldots (1)$$

   Where $[z]_+ = max\,(z, 0)$, α is a margin parameter and f is the embedding function learned during the stage of training.

b. Residual Block Architecture for Person Re-Identification [16]

The proposed 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 [16] are the fundamental component of our architecture, containing the Lite 3 × 3 layer as Shown in Figure. 4.1(a). Given an input x, the goal of this bottleneck is to discover a residual x˜ with a mapping function F, such that y = x + x˜, where 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). Because 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 [23]. In this context, y represents the output of the residual block.

c. Depthwise Separable Convolutions [24][16]

Depthwise Separable Convolutions are used to reduce the number of parameters [24]. The main idea is to split the convolutional layer. ReLU(w ∗ x) as in Figure 7 with kernel w ∈ $R^{k\times k\times c\times c_0}$ into two separate layers ReLU((v ◦ u) ∗ x) with depthwise kernel u ∈ $R^{k\times k\times 1\times c}0$ and pointwise kernel v ∈ $R^{1\times 1\times c\times c_0}$.

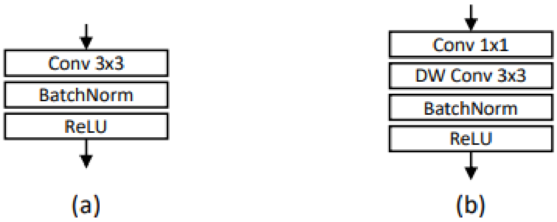

Figure 7: (a) Standard 3 × 3 convolution (b) Lite 3 × 3 convolution

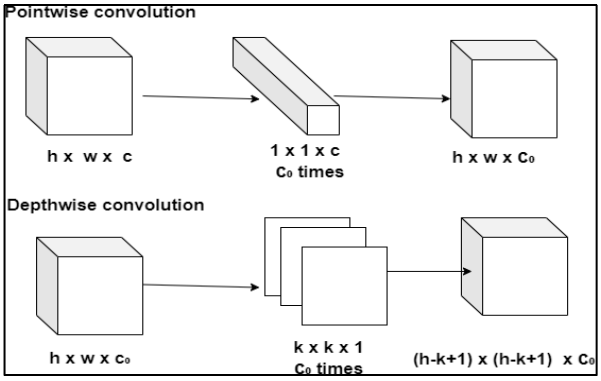

Figure 8: Point wise and Depth wise separable convolution.

Where ∗ 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 ∈ $R^{h\times w\times c}$ of height h and width w, y computational cost (number of multiplications) without DSC (Depth wise Separable Convolution). y′ is the 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′.

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

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

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

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

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 the proposed model, ReLU((u ◦ v) ∗ x) pointwise → depthwise is used instead of depthwise → pointwise , which performed better for omni-scale feature learning [16] and such layer is called as Lite 3 × 3. The Depth wise Separable Convolution is shown in Figure 8.

d. Aggregation Gate [16]

In the context of person re-identification, the Unified Aggregation Gate plays a significant role in aggregating features to capture a wide range of scales, allowing for dynamic multi-scale feature fusion through a fine-grained fusion of input features [25]. This gate is designed to dynamically fuse multi-scale features with input-dependent channel-wise weights, enhancing the model's performance in capturing different visual concepts.

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 channel wise weights and efficient model training due to shared parameters. The aggregation gate is a crucial component of our proposed model 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, ..., F_n, where n is the number of branches. The aggregation gate G_i for the i^{th} branch is computed as follows:

$$G_i = \sigma \left(W \cdot g\left(F_i\right) + b\right) \qquad \dots \ (4.5)$$

Where σ represents the activation function, such as sigmoid or softmax, W is the learnable weight matrix, g(·) 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 as:

$$F_{\text{aggregated}} = G_1 \cdot F_1 + G_2 \cdot F_2 + \cdots + G_n \cdot F \qquad \dots \ (4.6)$$

The aggregation gate is explained in figure 4.4. 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.

In the proposed model double aggregation channels are used after 2 Lite blocks to extract body features for Improved Feature Discrimination, Enhanced Model Capacity, Better Robustness to Variations and Reduced Over fitting for 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.

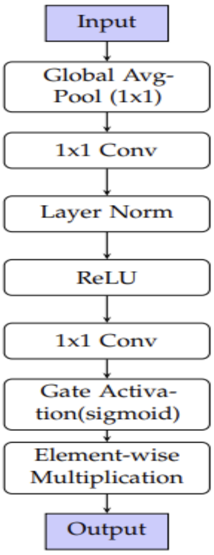

Figure 9: Flow diagram of Aggregation Channel Gate

The architecture of the proposed body feature extraction model, including the above components is illustrated in Figure 10.

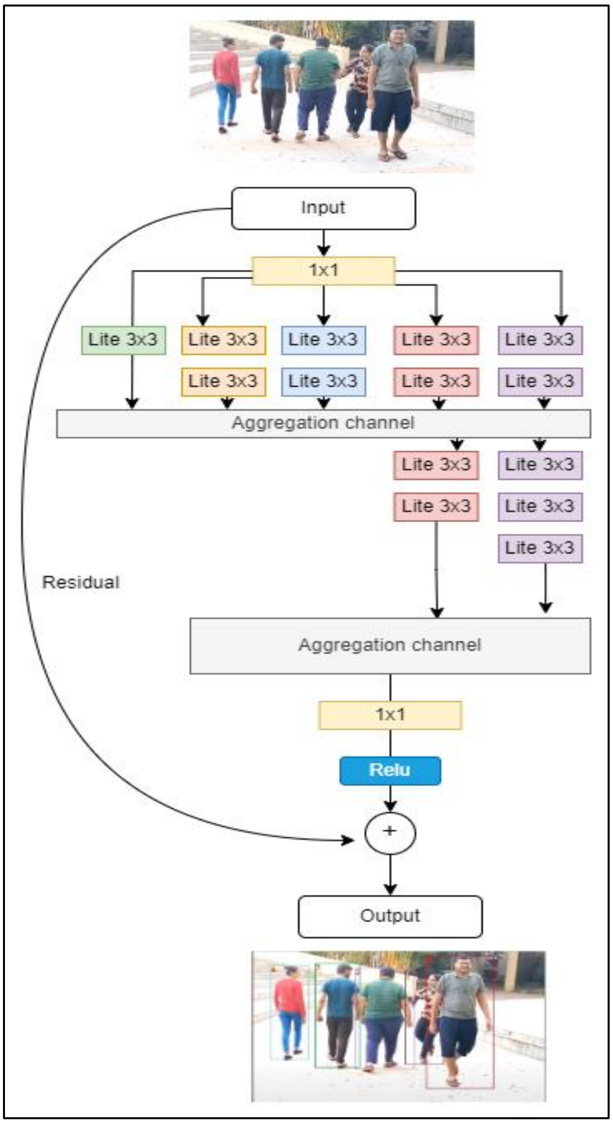

Figure 10: Proposed Body Feature Extraction Model

4.2 Proposed Multi Model Person Re-Identification Framework (MPRe-ID)

In the proposed multi-scale feature learning framework for Person Re-identification (MPRe-ID) combines body and face features to improve the accuracy of the task. MPRe-ID utilizes two aggregation channels to extract body features and FaceNet model [26] to extract face features.

For the extraction of face features, we use the MTCNN (Multi-task Cascaded Convolutional Networks) [27] 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 [26]. FaceNet 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 11. For tracking the person, YOLOv4 (You Only Look Once version 4) [2] object detection model and DeepSORT [3] is used. In DeepSORT, the "Deep" refers to the integration of deep learning-based features for improved object tracking. It utilizes a deep neural network to extract appearance features from object detections, allowing it to handle appearance variations and occlusions more effectively compared to traditional methods. DeepSORT works by first detecting objects in each frame using YOLOv4 detection algorithm. Then, it associates these detections with existing tracks using a combination of motion prediction and appearance matching. The appearance matching is facilitated by the deep feature embedding’s extracted from the detected objects. By combining motion prediction, appearance matching, and deep learning-based features, DeepSORT is able to achieve robust and accurate object tracking in video streams. 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. Flow diagram of proposed MPRe-ID framework is shown below.

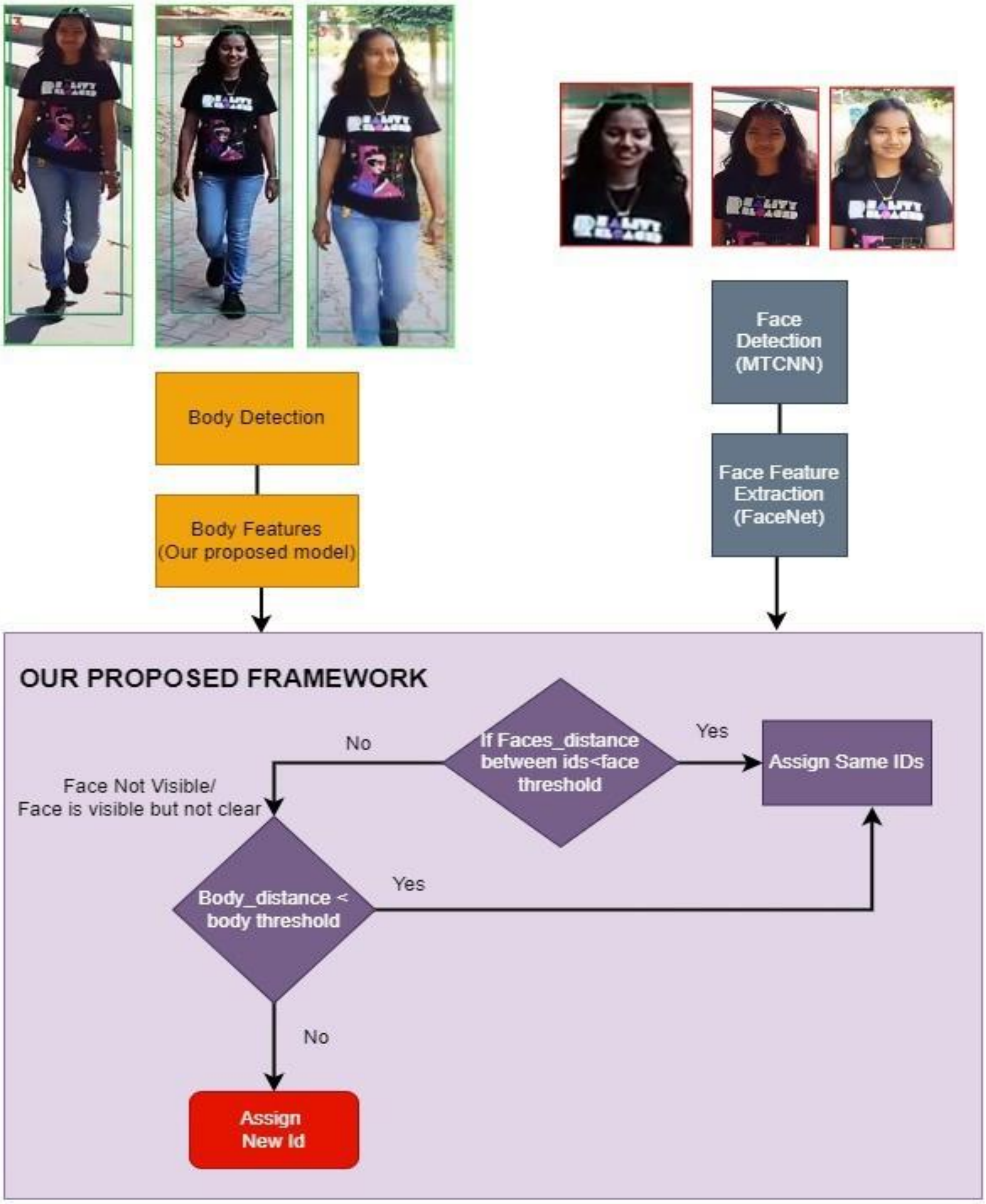
Body Detection
Body Features (Our proposed model)
Face Detection (MTCNN)
Face Feature Extraction (FaceNet)
OUR PROPOSED FRAMEWORK
If Faces_distance between ids<face threshold
Yes
No
Assign Same IDs
Face Not Visible/ Face is visible but not clear
Body_distance < body threshold
Yes
No
Assign New Id

Figure 11: Flow diagram of proposed MPRe-ID framework

5. RESULTS AND ANALYSIS

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 flipping. For training our model has used SoftMax loss.

The Experimental Results provides an overview of our research findings. We present the results obtained by using Market-1501, DukeMTMC-reID datasets on MPRe-ID framework, showcasing the performance on metrics including mAP, rank-1, rank-5, rank-10, and rank-20. We compared the performance of MPRe-ID with other existing state-of-the-art methods. The result and analysis of the experimental outcomes providing insights into the effectiveness and capabilities of our model is as follows.

5.1 Result

The results that are obtained from the proposed MPRe-ID on the Market-1501 and DukeMTMC-reID datasets are presented in Table 2. MPRe-ID 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, MPRe-ID achieved 89.1% and 79.8% on Market-1501 and DukeMTMC-ReID, respectively.

Table 2: Result of proposed MPRe-ID

	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%

5.2 Analysis

Table 3: Comparison of proposed MPRe-ID with state of the art Approaches

Method	**Market-1501**		**DukeMTMC-ReID**	
	Rank-1	**mAP**	**Rank-1**	**mAP**
Verifi-Identifi [14]	79.5	59.9	68.9	49.3
DCF [28]	80.3	57.5	-	-
SVDNet [29]	82.3	62.1	76.7	56.8
PAN [30]	82.8	51.5	71.6	51.5
DeformGAN [31]	80.6	61.3	-	-
LSRO [32]	84.0	66.1	67.7	47.1
PT [33]	87.7	68.9	78.5	56.9
Multi-pseudo [34]	85.8	67.5	76.8	58.6
ShuffleNet [35]	84.8	65.0	71.6	49.9
MobileNetV2 [36]	87.0	69.5	76.2	55.8
MPRe-ID	89.1	72.1	**79.8**	**59.5**
PN-GAN [37]	89.4	72.6	73.6	53.2

As shown in Table 3, the MPRe-ID method outperforms other techniques, including PN-GAN, when applied to the DukeMTMC-reID dataset. This dataset is known for its complexity, featuring eight camera views and numerous unique identities. The proposed model achieves a higher Rank-1 accuracy of 79.8% and a mean Average Precision (mAP) of 59.5% on the DukeMTMC-ReID dataset, surpassing the performance of PN-GAN.

In contrast, PN-GAN exhibits lower performance scores on the DukeMTMC-reID dataset. This indicates that the PN-GAN model lacks sufficient generalization capabilities, which are crucial for adapting across different datasets and real-world scenarios. Our model, however, demonstrates superior results in terms of both Rank-1 accuracy and mAP, outperforming the other methods evaluated.

6. CONCLUSION AND FUTURE SCOPE

The proposed MPRe-ID Model is a lightweight and robust solution for person re-identification (Re-ID) that leverages multi-scale feature learning and the fusion of body and face features. For pedestrian detection, MPRe-ID utilizes the YOLOv4 object detection model, and for tracking, it employs the DeepSORT algorithm. The model's backbone is built on the ResNeXt architecture, which is designed to learn discriminative features at multiple semantic levels in a novel body feature extraction process. In addition to body features, MPRe-ID incorporates face features using the FaceNet model, which significantly enhances the model's accuracy. By combining body and face features, the experimental results of the proposed MPRe-ID model demonstrate its effectiveness in person re-identification tasks. Future work may involve exploring advanced versions of YOLO that are more suitable for person re-identification, as well as investigating other face recognition models or alternative methods for fusing body and face features.

7. REFERENCES

[1] Singh, Nikhil Kumar, Manish Khare, and Harikrishna B. Jethva. "A comprehensive survey on person re-identification approaches: various aspects." Multimedia Tools and Applications 81, no. 11 (2022): 15747-15791.

[2] Mukherjee Arnab, “YOLO: Algorithm for Object Detection Explained,” *LinkediIn*, 2023. Available: https://www.linkedin.com/pulse/yolo-algorithm-object-detection-explained-arnab-mukherjee/

[3] N. Wojke, A. Bewley, and D. Paulus, “Simple online and real-time tracking with a deep association metric,” in *Proceedings - International Conference on Image Processing, ICIP*, 2017, vol. 2017-September, doi: 10.1109/ICIP.2017.8296962.

[4] Y. C. Chen, W. S. Zheng, J.-H. Lai, and P. C. Yuen. An asymmetric distance model for cross-view feature mapping in person re-identification. IEEE transactions on circuits and systems for video technology, 27(8):1661–1675, 2016.

[5] R. De Maesschalck, D. Jouan-Rimbaud, and D. L. Massart. The mahalanobis distance. Chemometrics and intelligent laboratory systems, 50(1):1–18, 2000.

[6] 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.

[7] L. An, X. Chen, and S. Yang. Person re-identification via hypergraph-based matching. Neuro computing, 182:247–254, 2016.

[8] 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.

[9] N. Pu, W. Chen, Y. Liu, E. M. Bakker, and M. S. Lew. Lifelong person reidentification via adaptive knowledge accumulation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 7901– 7910, 2021.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] Z. Zheng, L. Zheng, and Y. Yang. A discriminatively learned CNN embedding for person re-identification. ACM transactions on multimedia computing, communications, and applications (TOMM), 14(1):1–20, 2017.

[15] A. Hermans, L. Beyer, and B. Leibe. In defense of the triplet loss for person re-identification. arXiv preprint arXiv:1703.07737, 2017.

[16] 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.

[17] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian. Scalable person reidentification: A benchmark. In Proceedings of the IEEE international conference on computer vision, pages 1116–1124, 2015.

[18] 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.

[19] G. Pant, D. Yadav, and A. Gaur. Resnext convolution neural network topologybased deep learning model for identification and classification of pediastrum. Algal research, 48:101932, 2020.

[20] V. Yathish, “Loss Functions and Their Use In Neural Networks,” *Towards Data Science*, Aug. 04, 2022. https://towardsdatascience.com/loss-functions-and-their-use-in-neural-networks-a470e703f1e9 (accessed Apr. 12, 2023).

[21] K. Mahendru, “Understanding Loss Functions to Maximize Machine Learning Model Performance (Updated 2023),” Aug. 14, 2019. https://www.analyticsvidhya.com/blog/2019/08/detailed-guide-7-loss-functions-machine-learning-python-code/ (accessed May 25, 2023).

[22] A. Hermans, L. Beyer, and B. Leibe, “In Defense of the Triplet Loss for Person Re-Identification,” 2017, [Online]. Available: http://arxiv.org/abs/1703.07737.

[23] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. cvpr. 2016. arXiv preprint arXiv:1512.03385, 2016.

[24] 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.

[25] Zhou, Yalei, Peng Liu, Yue Cui, Chunguang Liu, and Wenli Duan. "Integration of multi-head self-attention and convolution for person re-identification." *Sensors* 22, no. 16 (2022): 6293.

[26] 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.

[27] 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.

[28] Li, Dangwei, Chen, X., Zhang, Z. and Huang, K., 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 (pp. 384-393), 2017.