UNIVERSITY
OF UDINE

Polytechnic Department
of Engineering and Architecture (DPIA)

PH.D. THESIS IN
INDUSTRIAL AND INFORMATION ENGINEERING

Deep Neural Networks Approaches for Person and Vehicle Re-Identification

CANDIDATE
Asad Munir

SUPERVISOR
Prof. Christian Micheloni

Cycle XXXIII — A.Y. 2020-2021

Dedicated to my parents, teachers, siblings and friends.

This work was supported by EU H2020 MSCA through Project ACHIEVE-ITN (Grant No 765866).

Acknowledgements

In The Name of Allah, The Most Beneficent, The Most Merciful.

All praise be to Him, who has given me this opportunity and got me through in every tough situation. This thesis got completed due to the guidance and consolation of several kind and caring people around. I would love to submit my sincere thanks to all of them.

First of all, I would like to give my regards to my supervisor Prof. Christian Micheloni, thank you for providing me this opportunity and your continuous support. It has been a very healthy experience working with you. Your friendly and kind behaviour always gave me confidence every time I stuck in my research. I would also thank Prof. Niki Martinel as he helped a lot in research as well as other many issues. I would like to thank the EU H2020 MSCA Project ACHIEVE-ITN (Grant No 765866) for supporting me and providing the training activities.

Thanks to Machine Learning and Perception (MLP) lab which provided a very comfortable environment and the required hardware. I would thank to my lab mates Rao Muhammad Umer, Matteo Dunnhofer and all others for their support and help. One of the great pleasures of my stay in Italy is to have the company of my fellows M. Tanseef Shahid, Abdullah Bhatti, Umair Zeb, Bilal Ahmed, M. Ghawas, Ammar ul Hassan, Harshit Bhatia, Sajid Hussain, M. Shoaib, Yousaf Hemani and Faraz Malik Awan for their undying moral support through thick and thin. I thank them for their valuable support, discussions and precious time they spared for me. To my parents and siblings, I will always love you irrespective to that where I live. Thank you for making me able to achieve this degree. Thank you, Mama, Papa and Atif Munir.

Abstract

The task of Re-Identification (re-id) is to retrieve any entity's images from a gallery set of multiple non-overlapping cameras for a given probe image. Active re-id research is composed of two entities named as person re-identification and vehicle re-identification. Person re-id is a very challenging task due to the presence of illumination, appearance, background, viewpoint, domain and pose variations, lighting and occlusions in the persons images. Domain variations occur due to camera's field of view environment (indoor and outdoor cameras) and light intensity as there are many different cameras present in the surveillance networks. Pose variations take place due to the person's movement and position when captured in different cameras. These variations make it difficult to match images of the same persons. To overcome this issue, we adopt Generative Adversarial Network (GAN) based approach to generate images in multiple domains and poses. Another solution for such problem is proposed by generating images from one camera domain to all other camera domains present in the environment and then merge these generated samples with the original data to enhance the method's performance. The proposed mechanisms magnify the matching between two images of the same person. The cameras used in video surveillance systems usually generate low resolution degraded images. The neural networks fail to learn various salient features in the existence of noise and other degradations in the data. The general architectures of neural networks learn features through neighbourhood similarities and hence forgot the similar patches which are discriminative for a specific person. The ignorance of such long-range similarities (dependencies) in the learned features halts

the performance of neural networks. We introduce attention mechanisms in the existing neural networks to get rid of such limitations. Attention is used to include non local and distant computations in the local receptive field of the convolution operations in networks. We propose multiple designs and solutions by the addition of a couple of attention mechanism which improve the performance of the network in different aspects. Another active research line is cross resolution person re-id in which query and gallery images are of different resolutions and hence reduce the performance of the existing person re-id models. We propose a distillation process based on resolution features in addition with channel attention to tackle this problem. Vehicle re-id is different from person re-id in the aspects of orientations present and the existence of multiple vehicles of a same model in the data. Vehicles have limited number of colors and models so require more discriminative features to represent a specific vehicle. We propose an oriented splitting of the features to learn local features along with global features to create a strong descriptor for each vehicle.

Contents

List of Tables

List of Figures

1

Introduction

1.1 Re-Identification

Re-Identification (re-id) is a challenging task that requires to recover the entity of interest's (probe) images from the gallery sets obtained across multiple disjoint cameras. Due to its importance in video surveillance applications, the problem of re-id is gaining more and more attention and is an active research topic in the field of computer vision having a long range of applications like video surveillance, person tracking, person searching , vehicle tracking, vehicle searching, environmental monitoring, disaster response, elder monitoring and computational forensics [79, 10, 11]. Since the problem of re-id is more related to video surveillance systems, so the two most important entities are persons and vehicles. Re-Identification has also a significant role in long-term tracking because unlike the general tracking problem re-id has many blind regions which are not captured by the field of view of any camera present in the surveillance system. Therefore, for a specific time, the system has no information about the person or vehicle when they are in any blind region as shown in Fig 1.1.

In this dissertation, most of the algorithms proposed are handling the persons so we are

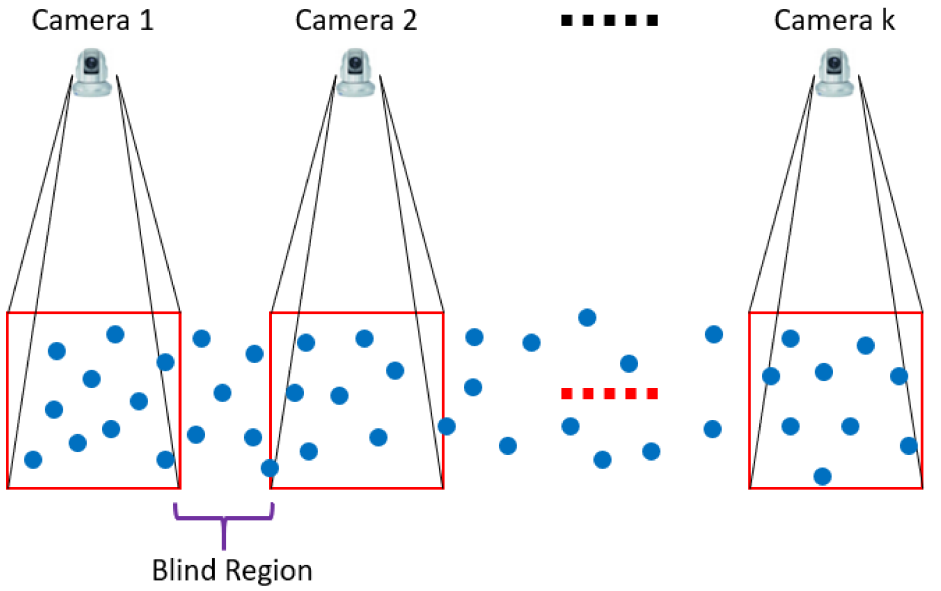

Figure 1.1: Video surveillance system with k cameras, blue circles are representing the entities observed by the system. Red rectangles are the field of view of corresponding cameras.

using persons as entities. To perform the person re-id task, a probe image, that is the image of the person of interest in a specific camera is searched in the images present in the gallery sets of the other cameras. The gallery set of each camera contains the images of persons captured in the field of view of that camera. The data collected from all cameras in the system undergoes many variations and effects which make it difficult to match two images of the same person. Light intensity and illumination variations occur due to lighting sources (e.g. indoor and outdoor cameras) and capturing images in different times of the day and night. Each camera has its own and different field of view due to which it creates its own background for every image in its gallery set. Appearance changes occur due to the clothing of the persons if they are wearing something (e.g. a hoodie) in one camera and remove it while enter in the field of view of the other camera. Pose and viewpoint variations happen due to the movement of the body parts of the person and the capturing angles (e.g. front, back and side views) of the camera. The cameras

in surveillance system usually produce multiple resolutions which is due to the distance of the persons from the camera and hence data generated by the cameras is degraded and contains low resolution images. Occlusions occur when the camera captures some part of the person instead of the whole person and hence it is difficult to identify as the image does not have the required information. Occlusions usually occur when the persons are entering or leaving the field of view of any camera as shown in the Fig 1.1 (the blue circles at the border of the field of view of camera 2).

In the early stage, traditional methods were used to compute the defined features, which are then used to match two images. These traditional methods compute color histograms to estimate the appearance and handcrafted features to resolve the pose and viewpoint changes. Texture descriptor were determined to detect the textural variations. Some of the methods also incorporated the spatio-temporal information which include topology of the network, time travel between adjacent cameras and distance between adjacent cameras to achieve better performance. With the growing trends and increasing development in deep learning, deep learning methods overshadow these traditional methods in performance and speed. Most of the deep learning methods adopt image classification networks to learn discriminative features and create strong representations for persons to improve re-id. A research trend in deep learning has also introduced the use of Generative Adversarial networks (GANs) to generate new data or better learn from the original data. The details of all such methods are shown in Fig 1.2 and discussed in the next chapter.

The task of vehicle re-id is similar to person re-id but having more complexities in data. The appearances of the vehicles are very limited compared to clothing in person re-id due to which they are more difficult to distinguish. Each model of vehicle has a large number of similar (e.g. lights, bumper etc) vehicles which force the networks to yield similar features for different vehicles. Another issue present in the vehicle re-id data is the orientations of the vehicles. Vehicles are present in every orientation unlike persons which almost are in standing position.

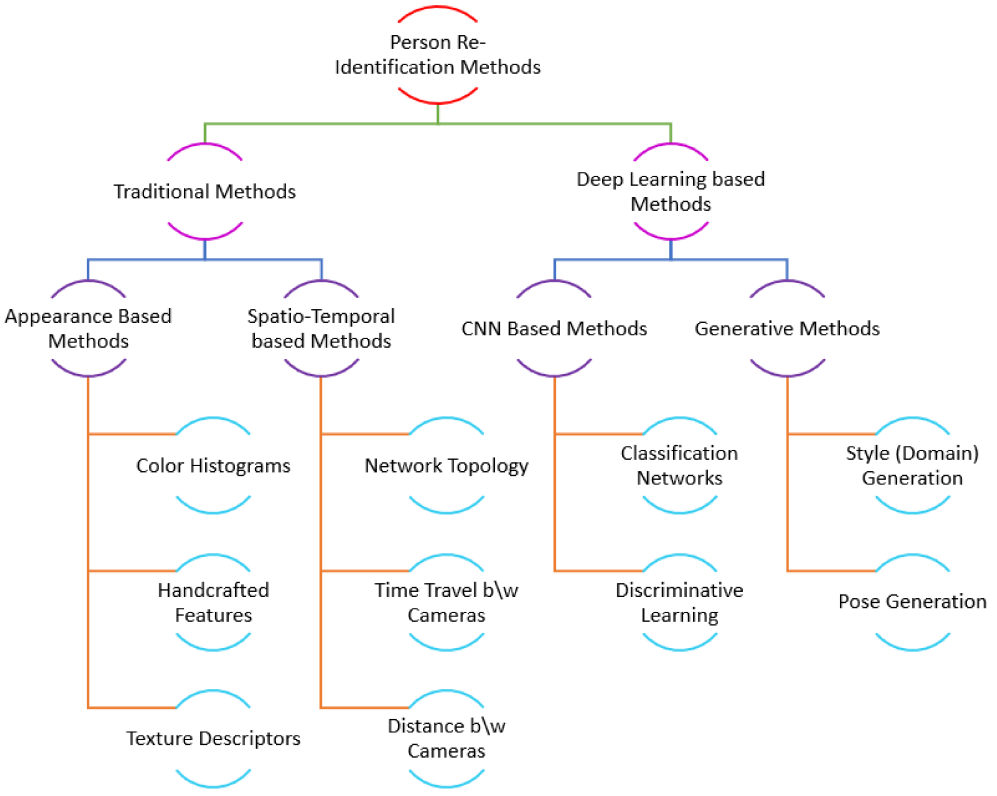

Figure 1.2: The Hierarchy of the methods proposed for person re-id.

Due to such complexities, person re-id models do not perform well and there is a need to design new models which can capture all variations by learning the most discriminative features and provide a strong descriptor for each vehicle.

1.2 Motivations

Person re-identification is usually solved as an image classification task. For this purpose, image classification networks are used to train with person re-id data. To perform the re-id, last classifier layers are removed and the features from the previous layers are used as the final rep-

resentation for each person. The similarity (euclidean distance, cosine similarity) is calculated between the two feature vectors obtained from the query and gallery image. The distance matrix computed by these feature representations is used to rank the gallery images with respect to each query image to perform re-id. Unlike image classification, in person re-id the person identities (classes) in training data are not similar to the identities in testing data, which makes re-id more challenging than a simple classification task. Therefore, general image classification networks perform within specific range. To learn more discriminative representations for persons, some enhancements are needed in the training mechanism or architectures of image classification network. We proposed multi-classifier training scheme to overcome such issues in our algorithms.

The general convolutional neural networks rely on the convolution operations within them which have a local receptive field. Since these convolution operations only depend on the local neighbourhood (size of the kernel) so they misses the long range dependencies and relationship between the pixels. For re-id task, these long range dependencies have a strong impact in calculating the distance matrix (e.g. occlusions, person's backpack, carrying some item). Part based approaches are proposed to overcome such an issue but the alignment of body parts is very difficult to achieve. Due to which, the part based stripes obtained from images have different parts to match. Attention mechanisms solve such kind of problems and several methods based on attention are also proposed to find long range similarities. They perform well but fail to learn unique representations from the noisy and blurry data. Unlike the previous approaches, we designed three different networks based on the mixture of self and channel attention mechanisms to overcome the above issues. The proposed algorithms eliminate the need of part-based solution and enhance the learning capability of the network in the presence of degraded data.

Person re-id datasets are collected from multiple non-overlapping cameras, which causes domain variations in the data as shown in Fig 1.3. Each camera possesses its own domain which can be due to light intensity or other environment variables. These domain variations

Figure 1.3: Domain (style) difference between camera 1 and camera 6 of Market1501 dataset. The appearance changes occur due to different camera field of view environment

cause changes in appearances of the same person and hence make it difficult for the networks to match them. Pose variations occur frequently due to the movement of the body parts and viewpoint of the cameras. It is also very challenging to match two different views of the poses (front and back sides) of a person. Several methods based on GANs have been introduced to generate augmented data to capture such variations. The generated images by such methods either focus on domain or pose and are unable to merge both of them. The generated augmented data have a lot of artifacts and degradations in them, which can not be treated as real samples in training. We propose a mechanism to generate both pose and domain variations within a single network and propose a technique to learn from the domain (style) transferred generated degraded data.

The cameras used in video surveillance systems usually produce low resolution images to avoid the high usage of bandwidth during the transfer of data. For this reason, the images used for person re-id are degraded, blurry and noisy. The persons present in the captured images undergo multiple resolutions due to their distances from the camera. This scenario creates a multi-resolution person re-id problem in which we have to match two images of different resolutions. Match Low Resolution (LR) query image with High Resolution (HR) gallery images is referred as Low resolution or cross resolution person re-id problem. Some methods were proposed to solve this problem, they adopt image super resolution techniques to make HR images from the LR query images and then perform the matching with similar resolutions. These

methods have high computation costs as image super resolution networks or other networks (e.g discriminators in GANs) are also needed and they did not take into count the degradation (noise and blur) present in the images. To solve such problems, We propose a technique to learn resolution invariant features from a single network, which eliminates the need of extra networks. We introduce channel attention mechanisms at multiple positions in the network to force the learned features to be sharp (removal of noise) [133] which improve the learning capability of the network in the presence of noisy and blurry data present for person re-id.

Vehicles are other important entities in the surveillance system and vehicle re-identification is a similar task to person re-id with more complexities. Vehicles are limited in appearance (color) and require strong descriptors to distinguish between them. Persons are usually in single orientation (standing) while the vehicles exist in multiple orientations. Due to which, matching between vehicles is a challenging task when using person re-id models. Many approaches are proposed with the enhancement in the person re-id networks to discriminate similar appearances, but few approaches tackle the orientation problem. We propose a novel oriented splitting of the feature maps within the network to obtain the local features for correlated parts and hence improve matching and re-id scores.

1.3 Problem Statements

With the above mentioned motivations, we come up with the problem statements for each algorithm we propose in this dissertation. We generalize the issue along with the proposed solution for it.

- Several recent person re-identification methods are focusing on learning discriminative representations by designing efficient metric learning loss functions. Other approaches design part-based architectures to compute an informative descriptor based on local features from semantically coherent parts. Few efforts learn the relationship between distant

similar regions and parts by adjusting them to their most feasible positions with the help of soft attention. However, they focus on calibrating distant similar parts features and ignore to learn the noise (blur) free and distinct feature representations as the person re-identification datasets contain degraded images. To tackle these issues

  - We propose a self attention based multi branch (classifier) network to directly model long-range dependencies in the learned features. Multi classifiers assist the model to learn discriminative features while self attention module encourages the learning to be independent of the feature map locations. Spectral normalization is applied in the whole network to improve the training dynamics and for the better convergence of the model.
  - We propose a Self and Channel Attention Network (SCAN) to model long-range dependencies between channels and feature maps. We add multiple classifiers to learn discriminative global features by using classification loss. Self Attention (SA) module and Channel Attention (CA) module are introduced to model non-local and channel-wise dependencies in the learned features. Spectral normalization is applied to the whole network to stabilize the training process.
  - We propose a Consistent Attention Dual Branch Network (CadNet) that has ability to model long-range dependencies (correlations) between channels as well as feature maps. We adopt multiple classifiers trained to learn the most discriminative global features for a unique representation of a person. Correlations between channels are consistently computed by using the channel attention mechanism to make the learned feature noise free and distinct from noisy and blurry data. Feature correlations interpret the relationship between distant similarities in the images computed by the self attention mechanism.

- The lack of cross-view (pose variations) training data and significant intra-class (domain) variations across different cameras make re-id more challenging. To solve these issues,

this work proposes a Domain and Pose Invariant Generative Adversarial Network (DPI-GAN) to generate images for both domain and pose variations capture. It is based on a CycleGAN structure in which the generator networks are conditioned on a new pose. Identity and pose discriminator networks are used to monitor the image generation process. These generated images are used for learning the domain and pose invariant features to improve the performance of person re-identification.

- Style or domain variations occur in person re-id data due to of the field of view of each camera. Most deep Re-ID models learn single scale feature representations which are unable to grasp compact and style invariant representations. We present a multi branch Siamese Deep Neural Network with multiple classifiers to overcome the above issues. The multi-branch learning of the network creates a stronger descriptor with fine-grained information from global features of a person. Camera to camera image translation is performed with a generative adversarial network to generate diverse data and add style invariance in learned features.

- Resolution mismatch occurs due to varying distances between person of interest and cameras, this significantly degrades the performance of re-id in real world scenarios. Most of the existing approaches resolve the re-id task as a low resolution problem in which a low resolution query image is searched in a high resolution images gallery. Several approaches apply image super resolution techniques to produce high resolution images but ignore the multiple resolutions of gallery images which is a better realistic scenario. We introduce channel correlations to improve the learning of features from the degraded data. In addition, to overcome the problem of multiple resolutions, we propose a Resolution based Feature Distillation (RFD) approach. Such an approach learns resolution invariant features by filtering the resolution related features from the final feature vectors that are used to compute the distance matrix. We tested the proposed approach on two synthetically created datasets and on one original multi resolution dataset with real

degradation.

- The orientations of the vehicles in the images make the learned models unable to learn multiple parts of the vehicle and relationship between them. The combination of global and partial features is one of the solutions to improve the discrminative learning of deep learning models. Leveraging on such solutions, we propose an Oriented Splits Network (OSN) for an end to end learning of multiple features along with global features to form a strong descriptor for vehicle re-identification. To capture the orientation variability of the vehicles, the proposed network introduces a partition of the images into several oriented stripes to obtain local descriptors for each part/region. Such a scheme is therefore exploited by a camera based feature distillation (CBD) training strategy to remove the background features. These are filtered out from oriented vehicle representations which yield to a much stronger unique representation of the vehicles.

1.4 Overview

This dissertation is divided into five chapters and the details of all these chapters are given as follows:

In **Chapter 1**, we introduced the re-identification problem and explained the types of solutions proposed to solve it. We also expressed the motivations for the algorithms we proposed in this work. In the end, we provide problem statements for each task we performed in this dissertation.

In **Chapter 2**, we introduced state of the art methods proposed for person and vehicle re-id. We further split these methods into several categories and explained their contributions. We also described the issues with the current state of the art methods which motivate us to perform the proposed solutions.

In **Chapter 3**, We described the proposed solutions for person re-identification. Firstly, we first introduced attention based solutions and we propose three methods with different network archi-

tectures and design to overcome the problems explained in the motivations section. Secondly, we expressed GAN based network for data generation and algorithm for the use of generated samples. In the first approach, we proposed to generate samples for person re-id with multiple poses and domains. In the second approach, we proposed solution for the use of style (domain) transferred images for re-id. Lastly, we solved the problem of cross-resolution person re-id by designing a network which learns resolution invariant features in the presence of degraded data. Each approach has experimental result section which presented the quantitative results along with their comparison with other state of the art methods. Some approaches have visual results as well to show the performance of the proposed technique.

In **Chapter 4**, we explained the proposed solution for vehicle re-identification. We described the proposed network to capture similarities for same vehicles with multiple orientations with the help of local features created by oriented splits of feature maps. We also show the quantitative and visual results and their comparison with other state of the art methods.

In **Chapter 5**, we conclude our work and write the summary for each proposed approach. The future aspects of re-identification is also discussed in this chapter.

2

State of the Art Methods

2.1 Traditional Methods

2.1.1 Hand Crafted Person Re-ID

Before the arise of deep learning and convolutional neural networks, a lot of research efforts [6, 26, 88] were made to design robust handcrafted features such as color histograms, local binary patterns, Gabor features, etc. These methods work well to eliminate the variations in lighting, poses and viewpoints. Their partition schemes consist of horizontal stripes, body parts and patches which are all performed in the spatial dimension.

2.1.2 Hand Crafted Vehicle Re-ID

Early works design hand-crafted feature descriptors to perform vehicle re-id. Liu et al. [70] extracted licence plate numbers with Gaussian mixture model, Haar like features and AdaBoost for vehicle recognition. Zapletal et al. [128] proposed a vehicle re-id model by using color histograms, histograms of oriented gradients and a linear regressor. However, these approaches

are unable to perform for large scale datasets.

2.2 Deep Learning Methods

2.2.1 CNN Based Person Re-ID

A variety of person re-id methods have been proposed to address various challenges in re-id task. Several approaches [72, 112] are based on the ResNet-50 structure along with modifications suitable for person-re-id. Many methods [129] have been developed based on different loss functions to learn the most discriminative features for re-id. Another [83, 140, 153, 49] trend focuses on learning pose and domain variations by using generative adversarial networks (GANs). Several efforts [13, 84] have been made to learn long range dependencies in features by introducing attention mechanism in the networks. Recently, many works introduce unsupervised person re-identification [111, 124].However, these methods typically assume that the query and gallery images have similar resolution (HR), which is not practically true in real world applications.

2.2.1.1 Metric and classification loss

Compared to hand-crafted feature design, deep neural networks learn the required features and metrics from the datasets by training with suitable loss functions. Ding et al [28]. adopts a triplet loss to calculate the relative distance between images. Chen et al. [16] enlarge inter-class variations and reduce intra-class variations by introducing quadruplet loss. Yu et al. [127] introduces a soft hard sample mining technique by assigning weights to hard samples.

Different research groups focused on addressing the person re-id problem as a classification problem. Some [2, 57] take pair wise images as input and compute the cross entropy loss for these images pairs. Some [108] propose margin based losses and others [104] adopt simple classification networks for multiple parts of a single image. In several proposed techniques, we

compute multiple cross entropy losses from each added classifier for the same image, which shows much higher performance than a standard single classifier.

2.2.1.2 Part based deep neural networks

Recently, many methods [136, 137, 104, 131] introduce learning local part-based feature representations to enhance the re-id performance. Some works directly split images into local stripes and create inaccurate part localization. Thus, some research deals with aligning the local parts by pose estimation and region proposal generation. Zhao et al. [137] proposes a part-aligned network for better body part partition. Zhang et al.[131] computes local loss by partitioning the body parts into horizontal stripes along with the global loss. Sun et al. [104] partitions the person image into horizontal stripes by introducing a uniform partition strategy.

2.2.1.3 Attention based models

Attention-based methods are used to handle localization and misalignment issues. Liu et al. [71] proposed HydraPlus network to learn low-level attentive features for better representation of the images. Li et al. [58] proposed a scheme to simultaneously learn hard region-level and soft pixel-level attentive features for a multi-scale feature representation. Other approaches [123, 139] proposed attentive learning by focusing the required attention maps for the better matching of the features.

2.2.2 GAN Based Person Re-ID

GAN based re-id methods utilize GANs to augment the training data. Zheng et al. [142] first generate images from random vectors by using unconditional GAN. Huang et al. [3] proceed in this direction by assigning pseudo labels to the generated images [44] with WGAN. Li et al. [59] propose to share weights between re-id model and the discriminator of GAN. Some recent methods utilize pose estimation to perform pose-conditioned image generation. In [73], A two-

stage generation pipeline is proposed based on pose to refine the generated images. Similarly, some approaches [29, 65, 89] used pose to generate images of a pedestrian in different poses to make the learned features more robust to pose variances. Siarohin et al. [99] improved pose-conditioned image generation by replacing the traditional $L1$ or $L2$ losses with nearest neighbor loss. Meanwhile, some recent methods also tried to compensate for the disparity between the source and target by exploiting synthetic data for style transfer of pedestrian images. CycleGAN [151] is applied in [27, 146] to transfer pedestrian image style from one domain to another. StarGAN [21] is used in [145] to generate pedestrian images with different camera styles to avoid computational complexity. Wei et al. [118] introduce semantic segmentation to extract the foreground mask in assisting style transfer.

2.2.3 Cross Resolution Person Re-ID

Several methods have been proposed to solve the resolution mismatch problem in person re-id. Li et al. [60] carry out cross scale image domain alignment and multi-scale distance metric learning jointly. Jing et al. [51] perform a mapping between HR and LR images with the help of a semi-coupled low-rank dictionary. Wang et al. [115] develop a framework to learn a discriminative scale-distance function by varying the image scale of LR images when matching with HR images. All these approaches adopt handcrafted descriptors which cannot enhance the person re-id performance compared to CNNs. Recently, several CNN-based models have been proposed to perform cross resolution person re-id. SING [47] proposed a network composed of several SR sub-networks and a person re-id module for low resolution person re-id. RIPR [74] jointly trained the foreground focus SR module and resolution invariant feature extractor to learn high resolution features for LR person re-id. A number of GAN based methods have been introduced and they perform significantly better than the CNN-based methods. CSR-GAN [117] cascades multiple GANs to progressively recover the LR image details. In all the above methods, SR models are a necessary part of the training and some of them use pretrained SR

models. RAIN [19] and CRGAN [62] align the feature distributions of LR and HR images to improve the LR person re-id performance. MSA [1] performs SR, denoising and re-id separately to achieve the final performance. [35, 20] use image super resolution techniques and predict the scaling factors to make the query image high resolution when matching with high resolution gallery images. All these methods adopt a single resolution (HR) gallery set while the proposed approach considers multiple resolutions gallery sets. Our approach uses a resnet-50 [38] based MGN [112] like baseline for extracting the features and then filters out the resolution dependent features to create a strong descriptor to match LR and HR images.

2.2.4 CNN Based Vehicle Re-ID

Deep learning methods perform efficiently as compared to the traditional methods. Liu et al. [67] reduced the gap between hand-crafted and deep learning features. An other work of Liu et al. [69] proposed a progressive multi-model search framework by using siamese neural networks trained with contrastive loss. Bai et al. [5] introduced triplet embeddings used in person re-id into vehicle re-id. Sochor et al. [101] mixed orientation information into the learned features to form viewpoint invariant descriptor. Wang et al. [115] implicitly learned local features for vehicle sides by giving credit to the one that is visible. Chu et al. [22] merged two matrices for similar and dissimilar viewpoints to train viewpoint aware network. Tang et al. [106] adopted reasoning about pose and shape via keypoints, heatmaps and segments obtained by predicting vehicle pose. Khorramshahi et al. [53] proposed a dual path attention network to capture global and local features. He et al. [36] integrated the vehicle part constraints with global re-id modules through a detection network. Khorramshahi et al. [55] present Self-supervised Attention for effectively learn vehicle-specific discriminative features to avoid large scale annotations. Sebastian et al. [93] utilizes horizontal and vertical splits to generate the local descriptor but these splits fail in several orientations of the vehicles. He et al. [40] used camera ID for the hard discrimination by assuming two similar vehicles as different in same camera.

3

Person Re-Identification

The presence of highly variable factors like illumination, resolution, clothing, view angle, human pose, occlusions and background in the images make re-id a very challenging task. Person re-id is usually tackled as an image classification task but the complication with respect to classical classification tasks is that in re-id the identities set (classes) mismatch between the training and testing stages. Precisely, the image classification task has similar classes (i.e. person identities) in training and testing sets while re-id has different identities in both sets. Therefore, the task of re-id requires a strong and discriminative feature descriptor to distinguish unseen, in the testing set, similar images belonging to new identities. With the development of neural networks and deep learning algorithms, the ConvNets [56, 100, 38], originally well designed for image classification tasks, perform impressively by providing discriminative feature representations for person images. Such a representation capability outperforms the traditional handcrafted low-level features by a large margin. To exploit ConvNets in Re-Id solutions, a research trend aims to design better metric learning loss functions [76, 41, 16] such as triplet loss, triplet hard loss, quadruplet loss, etc. for a better description of the person's image. These loss functions enlarge and reduce the inter-class and intra-class variations respectively, thus improve

the generalization capability of the model. The performance of such metric learning based loss functions is highly controlled by the sampling method and by hard sample mining techniques. On the other hand, many approaches [129, 143, 83] address the person re-identification task as a general image classification task. The basic idea of these studies is computing the cross entropy softmax classification loss for the person's images. While testing, these classification based approaches compute the distance matrix from the output features of the images to distinguish the person identities. Due to the mismatch between training targets and testing targets, the performance of metric learning loss functions becomes inferior in re-id task. To overcome this issue, we propose multi classifiers training instead of a general single classifier to learn the most discriminative features from person images. The effectiveness of the proposed multi classifiers learning is presented in the ablation studies of the works. Recently, part-based models [136, 137, 104, 131, 97, 135, 144] have represent the state of the art performance in person re-id by learning part-based local feature representations from the person's image. Some of these methods [136, 104] compute a strong discriminative representation of the image by splitting it into several body parts and then evolve the local features from all parts into a single representation. Other approaches [137, 131] horizontally partition the deep neural networks feature maps into several parts to learn more informative and fine-grained salient features in individual local parts. They distinguish one identity from an other by using discriminative cues from these parts. To learn salient part features, such methods require well aligned body parts for the same person. This is one of the main drawbacks due to lack of part consistency. Lately, many attention-based approaches [71, 123, 58, 139, 86, 12, 144, 122] have been proposed to overcome such partitioning and misalignment issues occurring in part based techniques. Attention is a powerful tool to perform spatial localization in the neural networks to interpret their decisions. AACN [123] tackle the misalignment and occlusions issues that occur in re-identification tasks by masking out the undesirable background with pose-guided attention mechanism. Others [71, 139] focus on better matching features and essential attention regions by learning superior attention maps.

However, the drawbacks of these approaches are the lack of learning the key-part features due to random part selection and in considering the noise (e.g. blur) effect in the learned features since most of the re-identification images are blurry and noisy. The general convolutional neural networks rely on the convolution operations within them which have a local receptive field. Since these convolution operations computation only depends on the local neighbourhood (size of the kernel) so they misses the long range dependencies and relationship between the pixels. Self-attention helps to compute the features correlations [130] by providing more weight to similar parts in the image and modeling long-range dependencies in a statistically efficient way. Channel attention mechanisms at multiple positions in the network force the learned features to be sharp (removal of noise) [133] which improve the learning capability of the network in the presence of noisy and blurry data present for person re-id.

Unlike the above mentioned approaches, we designed three different network designs based on the mixture of self and channel attention mechanisms to overcome the above issues. The proposed algorithms eliminate the need of part-based solution and enhance the learning capability of the network in the presence of degraded data.

3.1 Problem Definition and Notations

Let a set of n training images $\left\{I_i\right\}_{i=1}^{n}$ with corresponding identities labels $\left\{y_i\right\}_{i=1}^{n}$ be acquired by a camera network. The task of person re-identification is to, given a probe, retrieve the person's images from the galleries of different cameras. The problem of person re-id is usually treated as an image classification task when using cross entropy loss. The difference between these two tasks is that the training and testing classes (person identities) are identical in image classification while different in re-id. With the help of classifiers, the most discriminative features for each person are learned from the dataset. During testing, these features are used to compute the distance matrix between the probe and the persons identities to achieve the person re-identification.

3.2 Datasets

We performed our experiments and evaluated the proposed network on two person re-id benchmark datasets, market-1501 [138] and DukeMTMC-reID [91]. We adopted rank-1 accuracy, rank-5 accuracy, rank-10 accuracy and mean average precision (mAP) as our evaluation metrics. We used the standards splits for training and testing identities. The details about the two datsets are:

Market-1501 dataset has 32668 images of 1501 person identities automatically detected from six disjoint cameras. The training set consists of 12936 images of 751 identities. The query set has 3368 probe images of 750 identities and the gallery set has 19732 images with 750 identities.

DukeMTMC-reID dataset contains manually annotated boxes generated by eight cameras. It is composed by 36411 images of 1404 identities. There are 16522 images of 702 identities in the training set. The query and testing sets have 2228 and 17661 images of 702 identities, respectively.

The images from both datasets are shown in Fig 3.1

Figure 3.1: Images from the two datasets are shown in this figure. The images on the right and left side of the dashed line are taken from Market1501 and DukeMTMC-reID datsets respectively.

3.3 Self Attention based multi branch Network for Person Re Identification

To process long range dependencies, several convolutional layers are required because the convolution operator has a local receptive field. Due to which small models are unable to get these long range dependencies while large models have high computational costs. On the other hand, self attention [130] has a better trade off between capturing long range dependencies and having a reasonable computational ability. The self attention calculates the weighted sum of all features at all positions to build the response at the current position and prevents high computational cost.

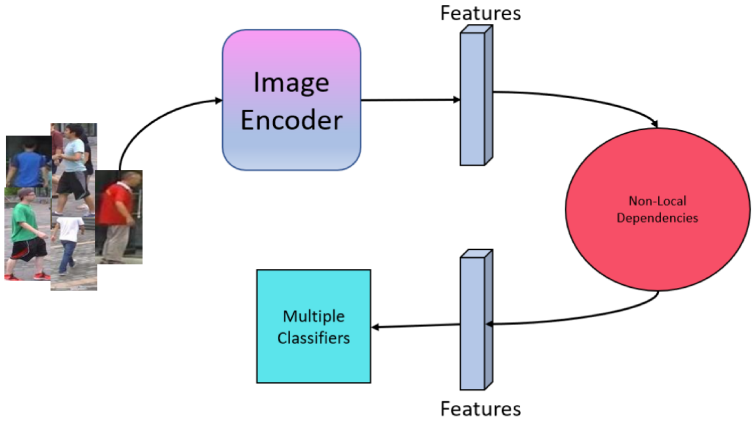

Figure 3.2: Overall framework of the proposed approach. ResNet-50 is used as image encoder to compute the features for the input images. Non local dependencies are calculated within the features to capture long range similarities. Multiple classifiers scheme is proposed to predict person identities.

In this work, we introduce self attention a module to model long range dependencies, which helps to emphasize the similarities at different positions in the backbone network for person re-id. Along with multi classifier training, the addition of a self attention module encourages the network to capture discriminative and robust person feature representations. The proposed

framework significantly enhances the person re-id performance and is shown in Fig 3.2.
The main contributions in this work are given as follows:

- A multi branch (classifier) network to learn the most discriminative features representations for person images to overcome the issue of mismatching identities in the training and testing stage.
- The addition of a self attention module in the backbone network to model long range dependencies for finding similarities between different locations in the learned features.

With the addition of the above mentioned contributions, we perform experiments on two benchmarks for person re-id. Results on these datasets show the performance and robustness of the proposed technique.

3.3.1 Self attention based multi branch Network

3.3.1.1 Proposed Architecture

Recent works have shown that Convolutional Neural Networks (CNNs) are efficient for learning deeper and robust feature representations from images and are accurate to train if they have shorter connections between layers. Relying on such outcomes, we define ResNet-50 [38] as our backbone network with several adjustments. We modify the stride (stride=1) of the last downsampling block to make the spatial size of convolutional features larger before global average pooling by following the work of R-FCN [24]. We apply global max pooling instead of global average pooling on the features from the last downsampling block. A 1×1 convolution layer is added after the pooling to reduce the size of features channels from 2048 to 1024. This added convolution layer learns the most discriminative features from a person image. These features are then sent through batch normalization, Rectified Linear Unit (ReLU) and dropout layers and finally passed to multiple fully connected layers (classifiers) to predict the identity of the person in the input image. Since the gradients from all classifiers are gathered at the previ-

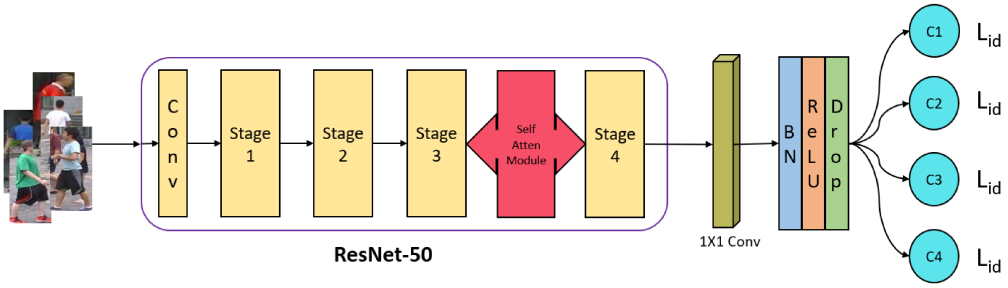

Figure 3.3: Overview of the proposed self attention based multi branch network. $C1$, $C2$, $C3$ and $C4$ are four fully connected layers for the predictions of person identity and their output losses are added to obtain the final loss. BN, Drop represent Batch-Normalization and Dropout Layers respectively.

ous convolution layer, thus they force that layer to learn the most discriminative global features for computing distance matrix in the testing stage. The learned global features are depending on the local neighbourhood as the convolution layers have a local receptive field.

To learn the long-range dependencies, we add a self attention module in the backbone network to model these dependencies and to capture the similar parts at different regions in the image. We append self attention block at the end of stage 3 as ResNet-50 consists of four stages after the first convolution block. The details of the self attention block is discussed in the next section. The proposed network is shown in Fig 3.3 and it is trained by using cross entropy loss which is given as:

$$L_{id} = -\sum_{c=1}^{C} \log(p(c))q(c) \tag{3.1}$$

where $p(c)$ is the output probability of the input belonging to class c. C, $q(c)$ are the total number of classes (person identities) in the dataset and ground truth distribution respectively.

In the testing stage, we remove the classifier layers of the network and the features from the last added 1×1 convolution layer donate the final representation of each person. These representations are used to compute the distance matrix between query and gallery images.

3.3.1.2 Self-Attention module

Most of the person re-id models are built using convolutional layers. Due to the fact that the convolution processes the information in the local neighbourhood, convolutional layers are computationally unable to grasp long-range decencies in images. In the proposed method, we adapt a non-local model [113] to introduce self-attention in a convoultional framework for the association of widely separated spatial regions.

From the previous hidden layers, the image features $x \in R^{C \times N}$ ($N = W \times H$) are first modified into two feature spaces f, g such that $f(x) = W_f x$ and $g(x) = W_g x$ to compute the attention. $s_{ij} = f(x_i)^T g(x_j)$ and attention map $\beta_{j,i}$ is calculated as:

$$\beta_{j,i} = \frac{exp(s_{ij})}{\sum_{i=1}^{N} exp(s_{ij})} \tag{3.2}$$

where $\beta_{j,i}$ represents to which extant the model takes part in i^{th} location when synthesizing j^{th} region. Here, C, N are the number of channels and the number of feature locations of the previous layer's features. The outputs of attention layer are $a = (a_1, a_2, ..., a_j, ..., a_N) \in R^{C \times N}$, and,

$$a_j = v\left(\sum_{i=1}^{N} \beta_{j,i} h(x_i)\right) \tag{3.3}$$

where $h(x_i) = W_h x_i$ and $v(x_i) = W_v x_i$. The formulation in eq 4.3 has learned weight matrices $W_f \in R^{\bar{C} \times C}$, $W_g \in R^{\bar{C} \times C}$, $W_h \in R^{\bar{C} \times C}$, and $W_v \in R^{\bar{C} \times C}$ which are implemented using 1×1 convolutions. $\bar{C}$ are the number of channels after reduction C/k, where $k = 1, 2, 4, 8$ and we are using $k = 8 (i.e., \bar{C} = C/8)$ in our experiments for memory efficiency. For scaling, we multiply the output of the attention layer by a scale parameter and add back to the input feature map. The final output is given by

$$y_i = \gamma a_i + x_i \tag{3.4}$$

where γ is learnable scalar parameter and is initialized as 0. γ is encouraging the network to rely first on the cues in the local neighbourhood and then progressively assign more weight to the non-local evidence. We apply self attention module after the second last stage of the backbone network as shown in Fig 3.4.

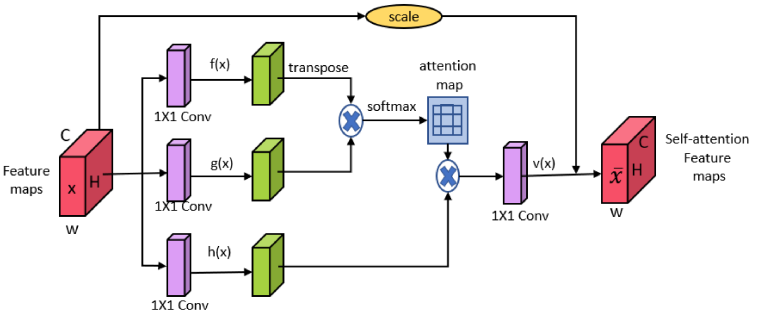

Figure 3.4: Self attention module which is added after stage 3 in ResNet-50. The dimension of the output (self-attention) features is same as input because they are the input to stage 4 of the ResNet-50 Network.

3.3.1.3 Spectral Normalization

Miyato et al. [80] originally proposed spectral normalization in the discriminator of Generative Adversarial Networks (GANs) [31] to stabilize the training which bounds the Lipschitz constant of the network by restricting the spectral norm of each layer. Spectral normalization does not need extra hyper parameter tuning like the other normalization techniques and have relatively low computational cost. We also apply spectral normalization at every convolutional layer in the network to stabilize the training process without which the network starts diverging after a few epochs.

3.3.2 Experimental Results

The details of the proposed network training parameters and implementation details are described in this section. The results and comparison with other state of the art methods is also

discussed.

3.3.2.1 Implementation Detail

We implemented the proposed network using Pytorch. The backbone network contains ResNet-50 network pretrained on ImageNet dataset with the modifications mentioned in section $II - B$. The network is optimized by using stochastic gradient Descent (SGD) with momentum 0.9 and the batch size is set to 64. The initial learning rates for backbone network and layers added at the end are 0.001 and 0.01 respectively. We trained our model for 260 epochs in total with learning rate divided by 10 after 160 epochs. All the images are resized to 256×128 with random horizontal flipping and random erasing data augmentations. The dropout probability is 0.5 and weight decay is set to $1e - 5$.

Table 3.1: Results and their comparison with the state-of-the-art person re-id methods of the proposed self attention based multi branch network on Market-1501. Rank1, rank5 accuracy and mAP are recorded and the top 1 and 2 results are in red and blue.

Methods	**Reference**	**Market-1501**		
		Rank-1(%)	***Rank-5(%)***	***mAP(%)***
SpindleNet [136]	CVPR17	76.9	91.5	-
Part-Aligned [137]	ICCV17	81.0	92.0	63.4
HydraPlus-Net [71]	ICCV17	76.9	91.3	-
LSRO [143]	ICCV17	84.0	-	66.1
SVDNet [102]	ICCV17	82.3	92.3	62.1
DPFL [17]	ICCV17	88.9	92.3	73.1
PSE [92]	CVPR18	87.7	94.5	69.0
HA-CNN [58]	CVPR18	91.2	-	75.5
MLFN [8]	CVPR18	90.0	-	74.3
DuATM [98]	CVPR18	91.4	97.1	76.6
DKP [96]	CVPR18	90.1	96.7	75.3
GCSL [9]	CVPR18	93.5	-	81.6
PCB [104]	ECCV18	92.3	97.2	77.4
IDCL [129]	CVPRW19	93.1	-	78.9
CASN(IDE) [139]	CVPR19	92.0	-	78.0
SFT [72]	ICCV19	93.4	97.4	82.7
Proposed	-	93.8	97.4	80.8

3.3.2.2 Comparison with state-of-art methods

Table 3.1 shows the result of the proposed method and its comparison with other state of the art methods on market-1501 dataset. Highest results are shown in red while the second highest in blue. The proposed method has highest rank-1 accuracy while in terms of mean average precision (mAP) our methods produce comparable results. The results of the proposed method are comparable with other state of the art methods. The objective of this work is to improve the ResNet-50 baseline with the proposed modification and we achieve this result by marginally enhance the rank1 score which is 89.0% for baseline.

Table 3.2: Comparisons to the state-of-the-art person re-id methods of the proposed self attention based multi branch network on DukeMTMC-reID dataset. The top 1 and 2 results are shown in red and blue.

Methods	**Reference**	**DukeMTMC-reID**		
		Rank-1(%)	***Rank-5(%)***	***mAP(%)***
Verif-Identif [141]	TOMM18	68.9	-	49.3
LSRO [143]	ICCV17	67.7	-	47.1
SVDNet [102]	ICCV17	76.7	86.4	56.8
DPFL [17]	ICCV17	73.2	-	60.6
PSE [92]	CVPR18	79.8	89.7	62.0
HA-CNN [58]	CVPR18	80.5	-	63.8
AACN [123]	CVPR18	76.8	-	59.2
MLFN [8]	CVPR18	81.0	-	62.8
DuATM [98]	CVPR18	81.8	90.2	68.6
DKP [96]	CVPR18	80.3	89.5	63.2
GCSL [9]	CVPR18	84.9	-	69.5
PCB [104]	ECCV18	81.8	-	66.1
IDCL [129]	CVPRW19	83.9	-	68.2
CASN(IDE) [139]	CVPR19	84.5	-	67.0
Proposed	-	84.6	91.6	68.6

In table 3.2, results on DukeMTMC-reID dataset are presented with its comparison with other state of art methods. The Rank-1 accuracy and mean average precision of the proposed method is the second highest with a very small difference and we obtain the highest results in the case of rank-5 accuracy. In all experiments, we reported our results while those of other

methods are taken directly from the papers. The results of the proposed method are considered without any re-ranking.

Table 3.3: Ablation study of the proposed self attention based multi branch network. The improvements in terms of rank1 accuracy and mean average precision (mAP) with respect to each proposed component and their placement in the baseline network.

Network Components	**Market-1501**	
	Rank-1(%)	***mAP(%)***
backbone (single classifier)	89.0	70.2
backbone (multiple classifiers)	92.4	78.6
backbone (attention after stage-1)	92.8	79.6
backbone (attention after stage-2)	93.5	80.4
backbone (attention after stage-4)	92.2	79.0
Proposed (attention after stage-3)	93.8	80.8

3.3.2.3 Ablation Study

In the ablation study, we perform component analysis on the proposed network using market-1501. Firstly, we have shown the effect of multi classifiers training as compared to the single classifier. Only one loss is calculated when using single classifier while multiple losses are added to get the final loss in case of multi classifiers. We use 4 classifiers at the end of the proposed network. Multi classifiers outperform the single classifier marginally as shown in table 3.3 (first and second row). In the second phase, we placed our self attention block at several positions in the backbone network and record its performance in table 3.3. Backbone (ResNet-50) network is composed of 4 stages and initial convolution block. All the four stages have 3, 4, 6 and 3 resisual blocks and we put the self attention module at the output of every stage. Among all other places, self attention block performance is higher when we place it at the end of stage 3 of ResNet-50. The self attention performs better when the spatial size of the feature maps is smaller thats why produce better results in the later stages of the network since each residual block reduces the dimension of the feature maps.

We proposed a novel self attention based multi branch network for person re-identification in section 3.3. Multi classifier training learns most discriminative features of the given person image to overcome the identities mismatching in training and testing stage. We added self attention module at the end of stage three in backbone (ResNet-50) network which models long-range dependencies to capture similarities at different feature locations instead of local neighbourhoods. Spectral normalization is applied to stabilize training and to avoid divergence of the model. The proposed network has the capability to learn robust feature representations and performs better than other state of art methods on two person re-id benchmarks.

3.4 Self and Channel Attention Network for Person Re Identification

Self-attention [130] models long-range dependencies by giving more weight to the similar parts in the image, thus it is statistically efficient. Self-attention computes the response at one position as a weighted sum of the features at all other positions. We propose a self attention (SA) module in the network to make it learn essential region features to enhance the final matching of the features. To address the issue of noise and make the learned features sharp, we propose a channel attention (CA) module applied consistently in the network for better feature correlation learning. With the above modifications, we propose a self and channel attention network (SCAN) to better learn the discriminative features thus to improve the matching score.

Multiple classifiers are used at the end of the network to classify different identities by learning the most discriminative features. The overall framework of the proposed network is shown in Fig 3.5 and our novel contributions are :

- We add a self attention (SA) module in the backbone network to model long-range dependencies for finding similarities between different locations in the learned features to enhance the matching score.

- We propose a channel attention (CA) module in every residual block of the backbone network to model inter dependencies and achieve better feature correlation learning to make the features sharp and robust to noise.

- Multiple classifiers are proposed for better prediction of a person's identity by learning the most discriminative features representations for person images to overcome the issue of mismatching identities in the training and testing stage.

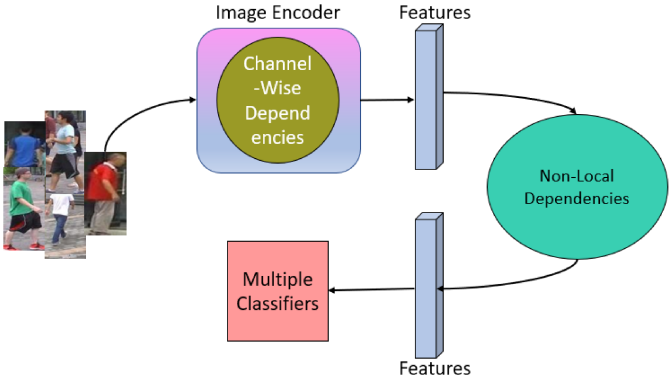

Figure 3.5: Overall framework of the proposed SCAN method. Channel wise dependencies are calculated with the encoder network while the non local dependencies are computed with the features obtained at the end of the encoder network. Finally multiple classification layers are used to make predicitions.

Unlike previous works, we introduce the self and channel attention mechanism to assign more weight to the necessary features and make them sharp (independent of noise) by constantly calculating channel correlations to avoid noise effect, which amplifies the matching of features even with the similarity at different parts in the image. With the above mentioned contributions, we performed experiments on two person re-id benchmarks. The proposed components enhance the performance of the baseline model and produce competitive results with other state of the art methods.

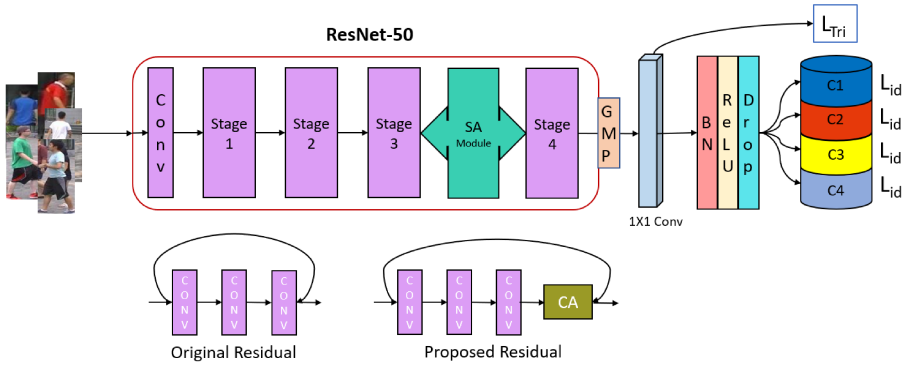

Figure 3.6: Overview of the proposed Network (SCAN). $C1$, $C2$, $C3$ and $C4$ are four fully connected layers for the predictions of person identity and their output losses are added to obtain the global loss. Original residual connection are modified with a CA module and a SA module is inserted in the network. GMP, BN, Drop represent global max pooling, Batch-Normalization and Dropout Layers respectively.

3.4.1 Self and Channel Attention Network

We explain the proposed modules and their insertion in the baseline network to form the proposed SCAN in this section.

3.4.1.1 Proposed Network Architecture

Recent works have shown that Convolutional Neural Networks (CNNs) perform efficiently to learn deeper and robust feature representations from images and are accurate to train if they have shorter connections between layers. Leveraging on such outcomes, we adopt ResNet-50 [38] as our backbone network with the addition of several layers and modifications. To make the spatial size of convolution features larger, we change the stride (stride=1) of the last downsampling block before the global average pooling following the work of R-FCN [24]. We perform global max pooling on these features instead of global average pooling. After max pooling, a 1×1 convolution layer is added to reduce the size of features from 2048 to 1024. Batch normalization, Rectified Linear Unit (ReLU) and dropout layers are appended

before multiple fully connected layers (classifiers) to predict the identity of the person in the input image. The gradients from all classifiers are gathered at the previous 1×1 convolution layer. They force that layer to learn the most discriminative global features for computing the distance matrix to overcome the issue of identity mismatching in the testing stage. However, these learned global features depend on the local neighbourhood since the convolution layers have a local receptive field. To learn the long-range dependencies and channel-wise correlation between the features, we add a self attention (SA) and channel attention (CA) modules in the backbone network to capture the similar parts at different regions in the image. Since ResNet-50 [38] consists of four stages after the first convolution block, we append the SA module at the end of stage 3 because self attention produces better results where the spatial size of features is smaller [130] and CA module before every residual connection inside the network. The details of SA and CA modules are explained in the next sections. The resulting proposed network (SCAN) is shown in Fig 3.6 and is trained by adding cross entropy loss (L_{id}) from every classifier along with triplet loss. The proposed SCAN learned better in the presence of noise and produce better representations to enhance person re-id performance.

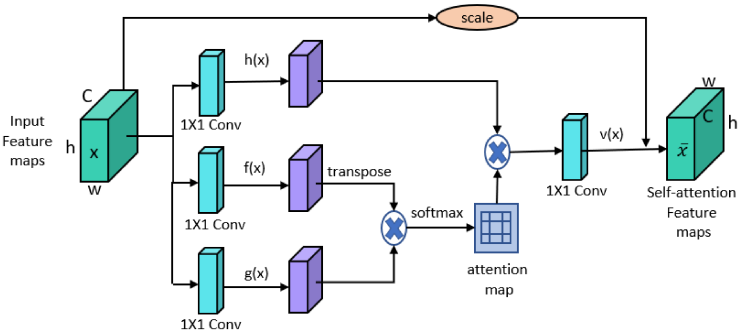

Figure 3.7: Self attention (SA) module which is added after stage 3 in ResNet-50. The dimension of the output (self-attention) features is the same of the input because they are the input to stage 4 of the ResNet-50 Network.

3.4.1.2 Self Attention Module

Self attention module is inserted in the network as discussed above. Self attention benefits from long range similarities in the images and improve match scores. To learn from the degraded data, we modified the previous discussed network to make a new design with the insertion of channel attention modules. The detailed mechanism of self attention module is discussed in section 3.3.1.2 and we showed in Fig 3.7 to match the color scheme with the proposed network design.

3.4.1.3 Channel Attention Module

Since person re-identification is applied to surveillance cameras, commonly real scenarios and used datasets consist of blurry and noisy images. Most of the existing methods are unable to grasp deep salient features from them. To build a stronger descriptor against such a degradation, noise free and distinct feature learning is required. To fulfill this objective, we introduce several channel attention modules to compute channels correlation consistently during the feature learning process.

Let $K = [k_1, k_2, ..., k_C]$ be the learned set of filter kernels for C output channels with k_l being the parameters of the l^{th} filter in a general convolution operation. The output from this convolution operation can be written as $U = [u_1, u_2, ..., u_C]$, where

$$u_l = k_l * X = \sum_{n=1}^{C'} k_l^n * x^n \tag{3.5}$$

In the above equation, $k_l = [k_l^1, k_l^2, ..., k_l^{C'}]$, $X = [x^1, x^2, ..., x^{C'}]$ (X being the input feature maps and C' is the number of input channels). The convolution operation is denoted by * and 2D spatial kernel k_c^n represents a single channel of k_l which interacts with the corresponding channel of X. The output of the convolutional layers is obtained through a channel-wise sum of the computed feature values. Therefore, the channel dependencies are introduced along with

the spatial correlation captured by the convolutional filters in the learned weights. We follow the work in [43] for computing these channel dependencies (correlations) but apply them at compacted features (convolutional blocks) instead of residual connections (used in [43]).

Each unit of the output U is unable to exploit contextual information outside of its region because the convolution operation has a local receptive field. To resolve this issue, global spatial information is squeezed into a channel descriptor. This operation generates channel-wise statistics and is achieved by using global average pooling. A statistic $z \in \mathbb{R}^C$ can be generated by shrinking U through the spatial dimension $H \times W$. The l^{th} element of z, computed by global average pooling, can be written as:

$$z_l = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} u_l(i, j) \tag{3.6}$$

For better modeling channel-wise dependencies, the learned function must have the ability to capture the nonlinear interaction between channels and permit multiple channels to oppose one-hot activation. The sigmoid activation fulfills these requirements and can be written as:

$$n = \sigma(g(z, W)) = \sigma(W_2 \delta(W_1 z)) \tag{3.7}$$

where $W_1 \in \mathbb{R}^{\frac{C}{r} \times C}$ are the parameters of the dimensionality reduction layer and $W_2 \in \mathbb{R}^{C \times \frac{C}{r}}$ are the parameters of dimensionality-increasing layer while δ denotes the ReLU function and r is the reduction ratio (Please refer to the experiments section for the evaluation of r). Two 1×1 convolution layers implement W_1 and W_2 around the non-linearity. The final output of the channel attention is obtained by rescaling the output U by means of the activations:

$$\bar{x}_l = n_l \cdot u_l \tag{3.8}$$

where $\bar{X} = [\bar{x}_1, \bar{x}_2, ..., \bar{x}_C]$. The dot product refers to channel-wise multiplication of feature maps $u_l \in \mathbb{R}^{H \times W}$ and the scalar n_l. The overall operation of the channel attention for computing channels correlation is shown in Fig. 3.12 and it helps to boost feature discrimination.

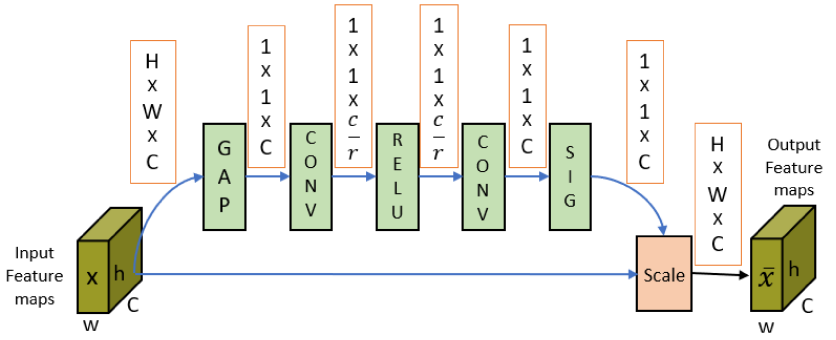

Figure 3.8: Channel attention (CA) module consists of 5 global average pooling layers , 2 convolution layer (1×1), ReLU and Sigmoid. Dimensions are indicated at the output of every layer where r is the reduction ratio.

3.4.1.4 Spectral Normalization

For stable training of the proposed SCAN architecture, we applied spectral normalization at each convolution layer in the network. The details of spectral normalization are already discussed in section 3.3.1.3.

3.4.2 Experimental Results

In this section, we discuss the implementation details and parameters of the proposed network. The results along with their comparison with other state of the art methods are also explained. Effect of each proposed component and technique is also expressed quantitatively and visually to show the improvements occurred due to them. We also perform the component analysis to show the effectiveness and contribution of each proposed component in the performance of person re-id.

3.4.2.1 Implementation Details

We implemented the proposed network using Pytorch. The backbone network consists of a ResNet-50 network. We modified the network with the addition of the proposed SA and CA modules along with all the adjustments mentioned in section 3.4.1.1. We optimized the network by using Adam optimizer with momentum 0.9. The initial learning rate is $3e-4$ and is divided by 10 after 60 epochs. We trained our model with a batch size of 64 for 140 epochs. All the images are resized to 256×128 with random horizontal flipping and random erasing data augmentations. The dropout probability is set to 0.5 and the weight decay is $5e-4$.

Table 3.4: Comparisons of the proposed SCAN to the state-of-the-art re-id methods on Market-1501. The top 1 and 2 results are mentioned in red and blue.

Methods	**Reference**	**Market-1501**		
		Rank-1(%)	***Rank-5(%)***	***mAP(%)***
SpindleNet [136]	CVPR17	76.9	91.5	-
Part-Aligned [137]	ICCV17	81.0	92.0	63.4
HydraPlus-Net [71]	ICCV17	76.9	91.3	-
LSRO [143]	ICCV17	84.0	-	66.1
SVDNet [102]	ICCV17	82.3	92.3	62.1
DPFL [17]	ICCV17	88.9	92.3	73.1
PSE [92]	CVPR18	87.7	94.5	69.0
HA-CNN [58]	CVPR18	91.2	-	75.5
AACN [123]	CVPR18	85.9	-	66.9
MLFN [8]	CVPR18	90.0	-	74.3
DuATM [98]	CVPR18	91.4	97.1	76.6
DKP [96]	CVPR18	90.1	96.7	75.3
GCSL [9]	CVPR18	93.5	-	81.6
PCB [104]	ECCV18	92.3	97.2	77.4
OGSL [46]	ICPR18	87.1	-	70.2
PRFF [121]	ICPR18	86.3	94.8	69.4
IDCL [129]	CVPRW19	93.1	-	78.9
PyrNet [76]	CVPRW19	93.6	98.2	81.7
CASN(IDE) [139]	CVPR19	92.0	-	78.0
SFT [72]	ICCV19	93.4	97.4	82.7
SCAN(ID)	-	94.1	97.7	82.1
SCAN(ID+Tri)	-	94.2	97.8	83.6

3.4.2.2 Comparison with state-of-art methods

Table 3.4 shows the results of the proposed (SCAN) method along with the comparison with other state of the art methods on market-1501 dataset. The highest and second highest scores are shown in red and blue respectively. The SCAN (ID) and SCAN (ID+Tri) represent the training of the model with single loss (ID loss) and two losses (ID and triplet), respectively. The proposed SCAN has the highest scores for rank-1 accuracy, rank-5 accuracy and the second highest in terms of mean average precision (mAP), when the model is trained with the two losses combined. PyrNet [76] has higher rank-5 accuracy because it computes multi-level features at several positions in the network, which is more computationally expensive than the proposed method. The results of SFT [72] and PyrNet [76] are shown without their proposed re-ranking since we are reporting our results without any re-ranking.

Table 3.5: Comparisons to the state-of-the-art re-id methods on DukeMTMC-reid dataset. The highest and second highest results are shown in red and blue.

Methods	**Reference**	**DukeMTMC-reid**		
		Rank-1(%)	***Rank-5(%)***	***mAP(%)***
Verif-Identif [141]	TOMM18	68.9	-	49.3
LSRO [143]	ICCV17	67.7	-	47.1
SVDNet [102]	ICCV17	76.7	86.4	56.8
DPFL [17]	ICCV17	73.2	-	60.6
PSE [92]	CVPR18	79.8	89.7	62.0
HA-CNN [58]	CVPR18	80.5	-	63.8
AACN [123]	CVPR18	76.8	-	59.2
MLFN [8]	CVPR18	81.0	-	62.8
DuATM [98]	CVPR18	81.8	90.2	68.6
DKP [96]	CVPR18	80.3	89.5	63.2
GCSL [9]	CVPR18	84.9	-	69.5
PCB [104]	ECCV18	81.8	-	66.1
OGSL [46]	ICPR18	76.3	-	63.7
PRFF [121]	ICPR18	72.1	83.8	53.4
IDCL [129]	CVPRW19	83.9	-	68.2
CASN(IDE) [139]	CVPR19	84.5	-	67.0
SCAN(ID)	-	84.9	92.0	69.2
SCAN(ID+Tri)	-	85.3	92.7	71.0

The performance of the proposed method on DukeMTMC-reid dataset is shown in the table 3.5 with the comparisons with other state of the art methods. With SCAN (ID), we achieved the highest rank-1 accuracy along with GCSL [9] and the second highest mean average precision very comparable with GCSL. SCAN (ID+Tri) outperforms all other methods in all scores. In all experiments, we reported the results of the other methods directly from their papers. The results in table 3.4 and table 3.5 show the superior performance of the proposed method as compared to other state of the art methods. The proposed SCAN is also compared with other attention methods (presented in the table 3.5) as well but the proposed channels correlations improve the training of the network and hence enhance the performance over other attention based methods.

3.4.2.3 Ablation Study

3.4.2.3.1 Effect of classifiers The improvement in the performance of the proposed method with the help of multiple classifiers compared to a single classifier is shown in Fig 3.9. Both the measurements top1 accuracy and mAP increased along with the number of classifiers before decreased again. We got the peak at 4 classifiers, so we use such a number in the proposed SCAN.

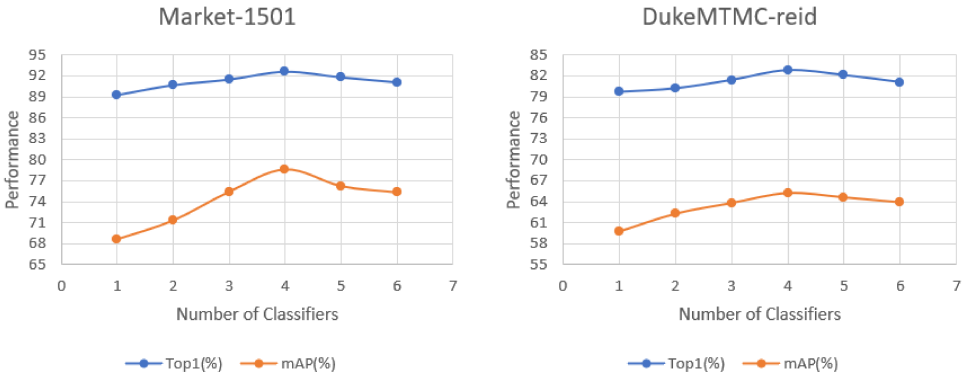

Figure 3.9: Effect of number of classifiers on the performance of the network on two benchmarks. Both the measurements represent peak values when 4 classifiers are selected for each datasets.

3.4.2.3.2 Component Analysis We performed a component analysis to demonstrate the effectiveness of all the components in the proposed network. We separated all the three proposed components and reported the scores of a simple baseline in the first row of the table 3.6. We replaced the single classifier with 4 classifiers to make Muti-C and described their effect on the performance of the network in the second row. Then, we appended the CA modules to create CA-baseline and verified the performance of the network. By replacing CA module with SA module, CA-baseline became SA-baseline and is mentioned in the second last row of the table 3.6. Finally, all the modules are added to the network to build the proposed SCAN and the results are recorded in the last row. The proposed combination of all the components produces much higher performance than each individually.

Table 3.6: Component Analysis of the proposed SCAN on Market-1501 and DukeMTMC-reid datasets in terms of mAP(%) and top-1 accuracy(%). CA and SA donate the channel and self attention modules. ID and Tri are the cross entropy and triplet losses respectively.

Networks	**Components**			**Market**		**Duke**	
	CA	**SA**	**Multi-C**	**mAP**	**R1**	**mAP**	**R1**
baseline	×	×	×	68.6	89.3	59.8	79.7
multi-C	×	×	✓	78.6	92.6	65.2	82.8
CA-baseline	✓	×	✓	81.6	93.5	68.9	83.9
SA-baseline	×	✓	✓	80.8	93.8	68.2	84.5
SCAN (ID)	✓	✓	✓	82.1	94.1	69.2	84.9
SCAN (ID+Tri)	✓	✓	✓	**83.6**	**94.2**	**71.0**	**85.3**

3.4.2.3.3 Parameters Selection All parameters settings mentioned in the section $IV - B$ are used from the previous standard person re-identification networks. The parameter k in self attention module can be chosen from $1, 2, 4, 8$. We did not notice any significant performance decrease when reducing the channel number of $\bar{C}$ to be C/k. For memory efficiency, we chose$k = 8$ (i.e., $\bar{C}$ = C/8) in all our experiments. The increment in the value of k reduces the computational cost and reduces the memory usage.

We proposed a novel Self and Channel Attention Network (SCAN) with multiple classifiers for person re-identification in section 3.4. Since we address re-id problem as a classification task, the multi classifiers training learns the most discriminative features from the person's image such that identities mismatching is handled. The addition of SA module encourages the network to capture similarities from distant patches and provides better matching scores. To learn the salient and sharp features from degraded person re-identification data, the CA module is introduced in the network. This learns the channel correlation (channel dependencies) of the features. The proposed SCAN model learns the most discriminative, sharp (invariant to noise) and salient features for feature matching. We apply spectral normalization to stabilize the training dynamics for the convergence of the model. The SCAN has the ability to learn robust feature representations and performs significantly better than other state of the art methods on two person re-id benchmarks.

3.5 Consistent Attentive Dual Branch Network for Person Re Identification

To overcome aforementioned issues, thus to enhance the final matching relevant region features should be computed as well as the feature correlation. For such purposes, we propose the introduction of a self attention module. In addition, to address the issue of noise, e.g. blur, thus to learn noise free features, e.g. features learned from sharp patches, we propose a consistent computation of channels correlation of multi scale features by exploiting the channel attention module. The proposed Consistent Attention Dual Branch Network (CadNet) is a modified version of the self and channel attention network (SCAN) [85] work such that noise free, salient and discriminative features are learned. The SA module is complimentary to convolution and enables the model to capture long-range, multi-level dependencies across image regions. To model inter dependencies and for better feature correlation learning, we propose a channel at-

tention (CA) module in the network. Similar to SCAN we train CadNet with multiple classifiers to classify different identities by learning the most discriminative features. The overall scheme of the proposed approach is shown in Fig. 3.10 and our contributions with respect to the SCAN [85] are :

- The channels correlation are used at the end of each stage instead for every residual connection, thus reducing their number. This is motivated by the hypothesis that the features computed at the end of each stage are a better representation of the image. Such an hypothesis is experimentally supported by the fact that computing the channels correlation at this stage enhances the performance and reduces the computational cost. In addition, these inter dependencies make the learned features robust to noise (e.g. blur), thus contributing to improve the matching score.

- A dual branch mechanism composed by a residual and an attentive branch is introduced. The former aims to provide *noise free* features while the latter provides similarities between patches at different location in the matching images (e.g. a backpack carried by hand in the probe and on shoulders on the gallery images). The final representation is the concatenation of both branches.

With the above mentioned changes, the performance of CadNet significantly improves on person re-id benchmarks compared to SCAN [85] as well as to other state of the art methods.

3.5.1 Consistent Attentive Dual Branch Network

3.5.1.1 Overview

To make a better matching in person re-id, strong and discriminative feature representations of person's images are required by the neural networks. To learn such representations, neural networks are trained in a supervised fashion by using the data of persons with known identities. In the testing stage, the features of unseen persons are extracted to match with other unknown

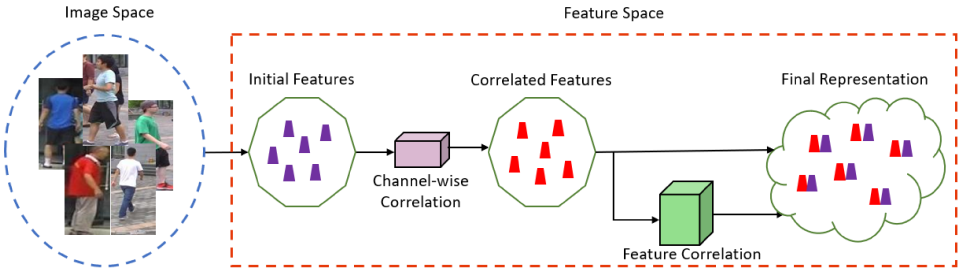

Figure 3.10: The explanation of the mechanism of the proposed approach. Images are first converted into feature space with the help of neural network and then channels correlation are applied within the network. The network is split into two branches, one having the original feature while the other calculating feature correlations to generate the final representation.

persons. The presence of unknown identities significantly reduces the matching performance. The training mechanism of neural networks plays an important role to learn from the person's image specific things (cloths, handbags, etc) that are important features to disambiguate between different people. We propose a training mechanism that exploits 4 classifiers. The predictions from all the classifiers are merged to make the final decision. We name it multi classifier (Multi-C) training. During training, several convolution operations take place across multiple channels of the features produced by the network's layers. The final output of network's layers are the sum through all the channels. This induces channel dependencies in the learned features. Such channel dependencies cause to miss the tiny effective details in the output features especially in case of person re-id since the images are blurry and noisy. We compute channels correlation to enhance the convolution features at every stage of the network so that the network is able to increase its sensitivity to missing informative features due to degraded data. Another aspect is that the convolution operations have just local information, hence they miss the long range similarities present in the images. These long range similarities has an essential impact on re-id when matching images. We capture these similarities with feature correlations which can be exploited by self attention mechanism. The details of computing channel and feature correlations are explained in the next sections.

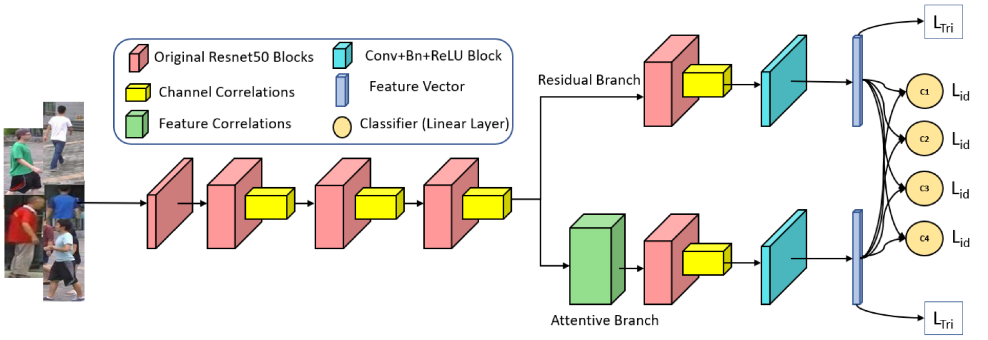

Figure 3.11: Overview of the proposed Network. $C1$, $C2$, $C3$ and $C4$ are four fully connected layers for the predictions of person identity and their output losses are added to obtain the identity loss. Features from both residual and attentive branches are fed to multiple classifiers having shared weights.

3.5.1.2 Proposed Network Architecture

Recent research works have shown that Convolutional Neural Networks (CNNs) efficiently learn deeper and robust feature representations from images and are precise to train if they have shorter connections between layers. Leveraging on such outcomes, we adopt the ResNet-50 [38] as our backbone network with several additions and modifications. We adopt multi classifiers training with multiple fully connected layers which are shown as classifiers in Fig. 3.11 to predict the identity of the person in the input probe image. The gradients from all added classifiers are gathered at the previous 1×1 convolution layer and force that layer to learn the most discriminative global features. Such features are used to compute the distance matrix to overcome the issue of identity mismatching in the testing stage. Since the convolution layers have local receptive fields then the learned global features depend on the local neighbourhood similarities and ignore the long-range dependencies. To capture the similar parts at different regions in the image and to work with re-id degraded data, we compute feature and channels correlation with the help of self and channel attention mechanisms to learn noise-free and salient features. These two modifications are expressed as channels correlation and feature correlations in Fig.

3.11. After the third stage of the backbone network [38] we designed two branches named residual and attentive branches. We added the self attention module at the start of the attentive branch because self attention produces better results when the spatial size of features is small [130]. Channels correlation are computed after every block of the ResNet-50. The details for computing the channel and feature correlations are explained in the next sections. The resulting proposed network (CadNet) is shown in Fig. 3.11 and is trained with cross entropy losses (L_{id}) from all classifiers and the triplet loss. Due to the modifications in the placements of channel attention (channels correlation) modules and the introduction of dual branch mechanism which preserves the orignal features along with attentive features, the proposed CadNet architecture performs better than the previously discussed methods.

3.5.1.3 Channels Correlation

Channels correlations are embedded in the proposed CadNet to learn sharp and noise free features from degraded data. These channels correlations compute the dependencies between all the channels of output feature maps at each residual block. The relationship between all channels generates more informative representations and hence improves the learning process. Channels correlations are calculated with the channel attention module which is discussed in details in section 3.4.1.3. The channel attention module is shown in Fig 3.12 to maintain the color scheme with respect to proposed CadNet.

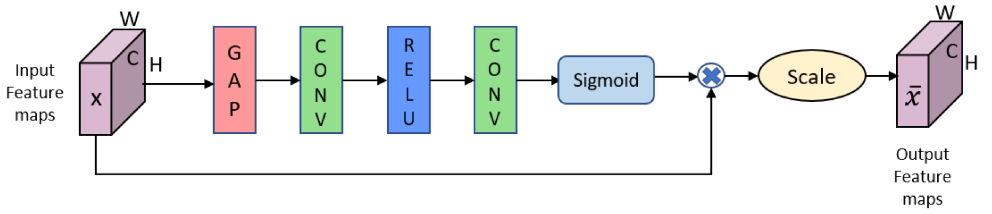

Figure 3.12: Computation of channels correlation via channel attention module which consists of global average pooling, 1×1 convolution, ReLU and sigmoid layers.

3.5.1.4 Feature Correlations

Feature correlations are computed through self attention mechanism. Unlike the previous designs of the proposed networks, In CadNet we also designed a residual branch to keep the original features and their combination with attentive branch produce a much stronger descriptor. The details for calculating feature correlations are explained in section 3.3.1.2 and shown again in Fig 3.13 to match the color coding with CadNet architecture.

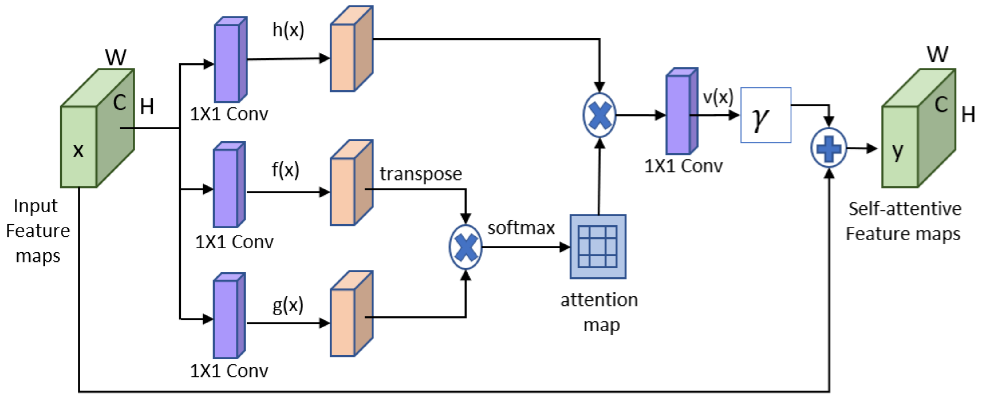

Figure 3.13: Computation of feature correlations via self attention module. The dimension of the output (self-attention) features is the same of the input because they are the input to the next residual block of the ResNet-50 Network.

3.5.2 Experimental Results

In this section, we discuss the results obtained from the proposed CadNet along with their comparison with the other state of the art methods and the proposed SCAN. We also show the visual effects of the attention mechanism through the computation of class activation maps.

3.5.2.1 Implementation Details

The backbone of the proposed CadNet network consists of a ResNet-50 network and is implemented using Pytorch. We trained CadNet on a nvidia RTX2080Ti gpu. Following the work of

R-FCN [24], we modified the stride (stride=1) of the last downsampling block before the global average pooling to make the spatial size of convolution features larger. We used global max pooling on these features instead of global average pooling. A 1×1 convolution layer followed by Batch normalization, Rectified Linear Unit (ReLU) and dropout layers are appended after the max pooling to reduce the size of features from 2048 to 1024. We add several modifications in the network as specified in section 3.2. The channels correlation of features at each stage are computed by the channel attention modules embedded throughout the network. The two branches of the proposed CadNet provide two feature vectors of length 1024 and are trained separately (without concatenation) by using shared multiple classifiers. The concatenation of the features from two branches defines the final representation vector of length 2048 which is used for feature matching.

We optimized the network by using Adam optimizer with momentum 0.9. The initial learning rate is set to $3e-4$ and is divided by 10 after 80 epochs. We trained our model for 140 epochs with a batch size of 64. Photometric distortions [42] and the AlexNet-style color augmentation [39] are applied to 256×128 sized images followed by random horizontal flipping and random erasing data augmentations. The dropout probability is set to 0.5 and the weight decay is $5e-4$. The ResNet-50 baseline training time on the exploited testing configuration is 2.5 hours for Market-1501 and 3 hours for DukeMTMC-reID. The proposed CadNet converges in 3.5 hours for DukeMTMC-reID and takes 3 hours for Market-1501 dataset to train. The training time for the CadNet is comparable with respect to the baseline while the performance is significantly higher than the baseline.The inference time is identical for both baseline and CadNet (0.175 sec per batch).

3.5.2.2 Comparison with state-of-art methods

The results of the proposed CadNet along with the comparison with other state of the art methods on market-1501 and DukeMTMC-reID datasets are presented in Table. 3.7 and Table. 3.8.

Table 3.7: Comparisons of the proposed CadNet to the state-of-the-art person re-id methods on Market-1501. The dashed line splitting the state of the art methods the proposed methods. and The top 1 and 2 results are mentioned in red and blue.

Methods	**Reference**	**Rank-1(%)**	**Rank-5(%)**	**mAP**
SpindleNet [136]	CVPR17	76.9	91.5	-
Part-Aligned [137]	ICCV17	81.0	92.0	63.4
HydraPlus-Net [71]	ICCV17	76.9	91.3	-
LSRO [143]	ICCV17	84.0	-	66.1
SVDNet [102]	ICCV17	82.3	92.3	62.1
DPFL [17]	ICCV17	88.9	92.3	73.1
PSE [92]	CVPR18	87.7	94.5	69.0
HA-CNN [58]	CVPR18	91.2	-	75.5
AACN [123]	CVPR18	85.9	-	66.9
MLFN [8]	CVPR18	90.0	-	74.3
DuATM [98]	CVPR18	91.4	97.1	76.6
DKP [96]	CVPR18	90.1	96.7	75.3
GCSL [9]	CVPR18	93.5	-	81.6
PCB [104]	ECCV18	92.3	97.2	77.4
Part-aligned [137]	ECCV18	91.7	96.9	79.6
SGGNN [94]	ECCV18	92.3	96.1	82.8
Mancs [110]	ECCV18	93.1	-	82.3
IDCL [129]	CVPRW19	93.9	97.8	80.5
PyrNet [76]	CVPRW19	93.6	98.2	81.7
AWPCN [97]	MMTA20	94.0	-	82.1
MMHPN [135]	MMTA20	94.6	-	83.4
APA [144]	MMTA20	93.6	-	81.7
MSMP [109]	NC20	93.7	-	81.2
CASN(IDE) [139]	CVPR19	92.0	-	78.0
SFT [72]	ICCV19	93.4	97.4	82.7
Baseline(ResNet-50)	-	92.6	-	78.6
SCAN [85]	ICPR20	94.2	97.8	83.6
CadNet(Proposed)	-	94.6	98.0	85.2

Unlike the other state of the art methods, the proposed CadNet introduce multi classifiers training mechanism which enhance the performance. The gradients from the each added classifiers are gathered at the previous convolution layer and make that layer to learn more and more refined and discriminative features with each addition. With 4 classifiers we got the highest performance and further addition of classifiers starts reducing the scores because of the classifiers errors which are also getting added for each classifier. To handle the blurriness and noise in the data, the proposed network computes channels correlation continuously at various spatial sizes.

Table 3.8: Results and their comparison with the state-of-the-art person re-id methods on DukeMTMC-reID dataset. The results of the state of the art methods are recorded above the dashed line and the results of the baseline used, the proposed SCAN and the proposed CadNet are reported below the dashed line. The highest and second highest results are shown in red and blue.

Methods	**Reference**	**Rank-1(%)**	**Rank-5(%)**	**mAP**
Verif-Identif [141]	TOMM18	68.9	-	49.3
LSRO [143]	ICCV17	67.7	-	47.1
SVDNet [102]	ICCV17	76.7	86.4	56.8
DPFL [17]	ICCV17	73.2	-	60.6
PSE [92]	CVPR18	79.8	89.7	62.0
HA-CNN [58]	CVPR18	80.5	-	63.8
AACN [123]	CVPR18	76.8	-	59.2
MLFN [8]	CVPR18	81.0	-	62.8
DuATM [98]	CVPR18	81.8	90.2	68.6
DKP [96]	CVPR18	80.3	89.5	63.2
GCSL [9]	CVPR18	84.9	-	69.5
PCB+RPP [104]	ECCV18	83.3	-	69.2
Part-aligned [137]	ECCV18	84.4	92.2	69.3
SGGNN [94]	ECCV18	81.1	88.4	68.2
Mancs [110]	ECCV18	84.9	-	71.8
IDCL [129]	CVPRW19	84.7	-	69.4
CASN(IDE) [139]	CVPR19	84.5	-	67.0
AWPCN [97]	MMTA20	85.7	-	74.1
APA [144]	MMTA20	84.7	-	69.4
MSMP [109]	NC20	84.4	-	70.4
Baseline(ResNet-50)	-	82.8	-	65.2
SCAN [85]	ICPR20	85.3	92.7	71.0
CadNet(Proposed)	-	86.3	92.8	72.7

These correlations produce noise free feature maps from degraded data [25, 134] and proceed them towards the classifiers. The sharpness in features makes them easy to distinguish from each other and hence improve the matching scores. For further refinement of features, the proposed network adopts a dual branch mechanism. The contribution of the residual branch in the performance of the CadNet is to provide the noise-free and discriminative features of the image which enlarge the difference between two different identities. The attentive branch merge the information from the distant image location which has higher contributions in the prediction

of person's identity. The concatenation of the attentive features with residual features amplifies the information in the learned feature vectors. The final representation of the proposed method produce better matching between to person and accelerate the performance. The effects of each of the components on the results of the proposed method are explained in section 4.5. In all experiments, we reported the results of the other methods directly from their papers. The results in Table. 3.7 and Table. 3.8 show the superior performance of the proposed method as compared to other state of the art methods. Unlike other methods, the learned features with the proposed CadNet consist of distant similarities and hence provide better and unique representation of the person. Therefore, the performance of the proposed method is higher than the others.

3.5.2.3 Visual Results

The visual/qualitative results from the proposed network are illustrated in Fig. 3.1. We used the trained CadNet model to obtain the feature representations of all images and then followed [149] to compute the visual results. We computed the class activation maps [149] for both datasets and present them visually. The proposed network is unable to compute the class activation maps at the last convolution block because we reduce the size of the features to 1024. The last block returns 2048 feature maps and the input to the classifiers is 1024. Since they have different sizes, thus we calculated class activation maps at the third convolution block where the network is split into two branches.

In Fig. 3.14 first row shows the original images from Market1501 dataset and the second row represents the corresponding class activation maps. Similarly, the third and fourth rows demonstrate the class activation maps for DukeMTMC-reID dataset. Class activation maps represents how much each image region contributes in the prediction of classes (person identities) probabilities. The highest contribution in the predictions is carried out by the red regions while the blue regions represent the lowest contribution (or no contribution). The images clearly show that the proposed solution takes into higher consideration regions belonging to persons while is

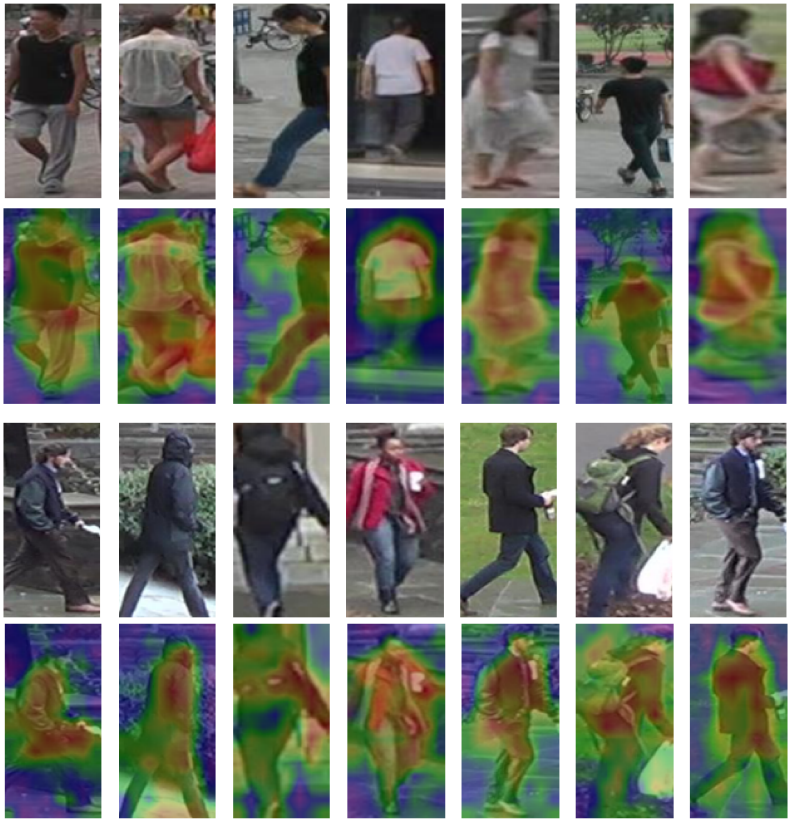

Figure 3.14: Class activation maps obtained with the proposed method. First and third rows are the original images and second and fourth rows consists of corresponding class activation maps for Market1501 and DukeMTMC-reID datasets respectively.

discarding the background. This behaviour contributes to improve the quantitative performance.

3.5.2.4 Ablation Study

3.5.2.4.1 Effect of classifiers To evaluate the contribution of a single classifier versus multiple classifiers, we modified the ResNet-50 baseline to get unbiased results. In this view, we trained the ResNet-50 [38] baseline network with different number of classifiers and reported the rank1 accuracy and mAP for both the datasets. In particular, we used a ResNet-50 pretrained on ImageNet dataset and removed its last linear layers. New linear layers according to the num-

ber of classes present in the datasets have been appended before training on the re-id datasets. Fig. 3.9 shows the contribution of different classifier layer on the ResNet-50 baseline. Both the measurements rank1 accuracy and mAP linearly increase until 4 classifiers. Then the slopes reaches a plateau or decreases gently. Since the highest performance has been reached with 4 classifiers we exploited such a number of layers in the CadNet solution and in the aforementioned/comparison results.

3.5.2.4.2 Parameters Selection Most of the parameters and specifications expressed in section 3.2 are used throughout our experiments and are gathered from the previous standard person re-identification techniques. Instead, the parameter r, used in the self attention module, can be set to $2, 4, 8, 16$. The parameter r is the division factor to generate patches from the input features. We reported the results by selecting multiple sizes of the generated patches in self attention module. The performance is slightly effected with different values of r because the number of channels $\bar{C}$ are reduced to C/r. The impact of different values for r is shown in Fig. 3.15. Evaluating such information, we chose $r = 16$ (i.e., $\bar{C}$ = C/16) in all our experiments. Such a selection not only improves qualitative performance but also reduces the computational costs and improves memory efficiency.

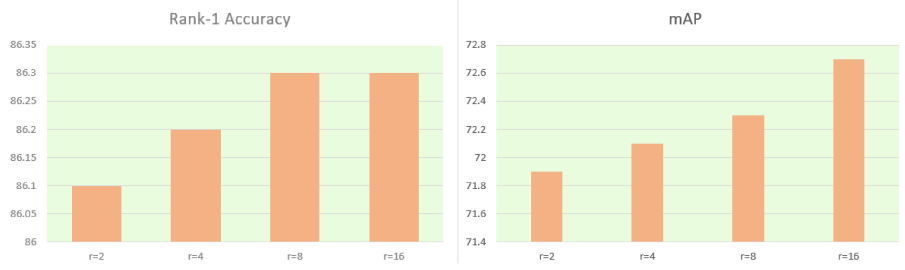

Figure 3.15: The impact of choosing different values for the reduction ratio r on DukeMTMC-reID dataset. Left side represents the rank-1 accuracy scores and right side shows the mean average precision (mAP)

3.5.2.4.3 Component Analysis To illustrate the effectiveness of the proposed contributions, we provide a component analysis for the proposed network. First, we performed the separation of the three components (Channels Correlation CC, Feature Correlation FC and multi classifiers Multi-C) according to SCAN [85] and reported the results in Table. 3.9. The ResNet-50 baseline proposed in section $4.5.1$ (one classifier) has been exploited as performance reference (the first row). Second row shows the impact of the exploitation of the four classification layers. Such first two rows show the numerical values of points 1 and 4 of Fig.3.15.
Third and forth rows in Table. 3.9 demonstrate the contributions of the channel attention (CA) and self attention (SA) modules. SCAN [85] is the ResNet-50 4-C with both modules. Third and forth row of Table 3.9 show performance of SCAN without SA and CA respectively. The performance improvement of the new exploitation of the channel attention with respect to SCAN [85] show the superiority of CadNet(residual branch) in row 6 with respect to SCAN row 5. To evaluate the CadNet(attentive branch), we trained the proposed network shown in Fig. 3.11 with only such a branch (e.g. the residual branch has been removed). The results are in row 7 of Table. 3.9. Finally, both branches have been used (row 8) to show the results of the complete proposed solution. Each of the residual and attentive branches improve the performance over SCAN [85] separately but the combination of both branches has a greater effect on the performance. This implies that the mixture of all proposed contributions together in the form of CadNet provides a stronger descriptor.

With the proposed placement of channels correlation (CC), $R1$ and mAP are increased by 0.9% and 2.5% for Market1501 and 1.6% and 2.8% for DukeMTMC-reID compared to SCAN. Similarly, the dual branch design with the proposed CC and feature correlation (FC) improves $R1$ and mAP by 0.4% and 1.6% for Market1501 and 1.0% and 1.3% for DukeMTMC-reID.

The proposed CadNet is the modified form of the proposed SCAN and shows a significant improvement in all measurements recorded in table 3.9. The component design is similar in both networks but the placement of these components in an appropriate place makes the net-

Table 3.9: Component Analysis of the proposed CadNet on Market-1501 and DukeMTMC-reID datasets in terms of mAP(%) and top-1 accuracy(%). CC and FC represents the channel and feature correlations respectively.

Networks	Components			Market1501		DukeMTMC-reID	
	CC	FC	Multi-C	mAP	R1	mAP	R1
ResNet-50 1-C baseline	×	×	×	68.6	89.3	59.8	79.7
ResNet-50 4-C	×	×	✓	78.6	92.6	65.2	82.8
SCAN [85] - FC	✓	×	✓	81.6	93.5	68.9	83.9
SCAN [85] - CC	×	✓	✓	80.8	93.8	68.2	84.5
SCAN [85]	✓	✓	✓	83.6	94.2	71.0	85.3
CadNet(residual branch)	✓	×	✓	84.1	94.4	71.7	85.5
CadNet (attentive branch)	✓	✓	✓	84.4	94.4	71.9	85.7
CadNet (proposed)	✓	✓	✓	**85.2**	**94.6**	**72.7**	**86.3**

work to learn more discriminative features.

We proposed a novel Consistent Attentive Dual Branch Network with multiple classifiers for Person Re-Identification (CadNet) in section 3.5. We exploited a multi classifiers training strategy in which each classifier contributes in distinguishing between identities and helps the model to learn the most discriminative and unique features for each person. Due to blurry and noisy person re-id data, general re-id models miss small and tiny details. The introduction of channels correlation makes the learned features noise free and highlights these small details to build a stronger descriptor. Channels correlation are computed through channel attention module consistently at multiple positions in the network to flow the tiny information towards the final representations. Local and non-local similarities in the person images are computed by the two branches, respectively residual and attentive ones, and merged to create a strong, unique and discriminative feature representation of each person. This has been shown to improve the matching score between two persons. Spectral normalization is applied while computing channel and self correlations to stabilize the training dynamics for the convergence of the model. The visual results show the participation of each tiny component of person in predicting the identity. The proposed CadNet learns small details that help to significantly en-

hance the person re-id performance with respect to other state of the art methods as shown on two widely adopted benchmarks datasets. The proposed network only focused on learning similarities/dissimilarities present for a single person. Cross correlations can be introduced in the future work to learn distinguishing features between two persons.

3.6 Generating Domain and Pose Variations between Pair of Cameras for Person Re Identification

In person re-id, person image encounters many changes independent of the person's identity. These changes include appearance, background (domain variations), viewpoint, pose variations, lightning and occlusions. Wide range of deep learning methods have been proposed to improve the performance of re-id. Generative adversarial network (GAN) [32] is gaining popularity in image generation to increase the re-id performance. Existing GAN based methods consider either domain variations [147] or pose variations [90] to generate new images.

In this work, we propose Domain and Pose Invariant Generative Adversarial Network (DPI-GAN) to generate images by changing both domain and pose in a pair of cameras. The proposed DPI-GAN uses CycleGAN [152] approach to translate images from one domain to another. The generators are conditioned on a new pose to generate an image in new domain with given pose. Identity and pose discriminators are used with the each generator to preserve the identity and conversion to new poses in the generated images. For each of the two training cycles, the proposed framework trains the two generators and two discriminators. The images are generated from one camera's domain to other camera's domain with a new pose and returning back to the original domain with a new pose. The working mechanism of the proposed method is shown in Fig 3.17 and the next section explains the proposed DPI-GAN framework. Section 3 describes the experimental results and parameters.

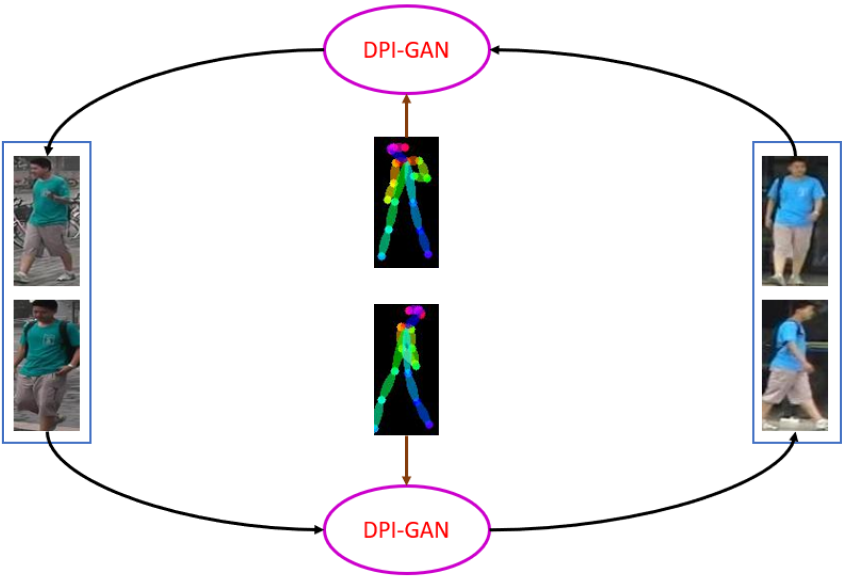

Figure 3.16: The proposed technique to translate between different domains and poses. Input is conditioned with a pose map to generate the new pose and then return back to same domain with new pose to complete a cycle.

3.6.1 Proposed DPI-GAN

Our proposed DPI-GAN aims to generate an image with a new pose and domain from an input image and a skeleton pose image. Skeleton pose images are calculated using the human pose estimator [7]. Input image and skeleton pose image are concatenated and fed into the generator network to generate the image with the given pose. We used CycleGAN [152] approach to train the network for capturing the domain changes between a pair of cameras. First Cycle of our framework is shown in Figure 3.17 which is transferring an image from camera A to camera B and then reconstruct that image back to camera A with a different pose. The second cycle is the same with starting from camera B to A and reconstruct to Camera B. With this, we are training the two generators which are generating images from domain A to B and vice versa. Identity and pose discriminators are used to identify real and fake generated images.

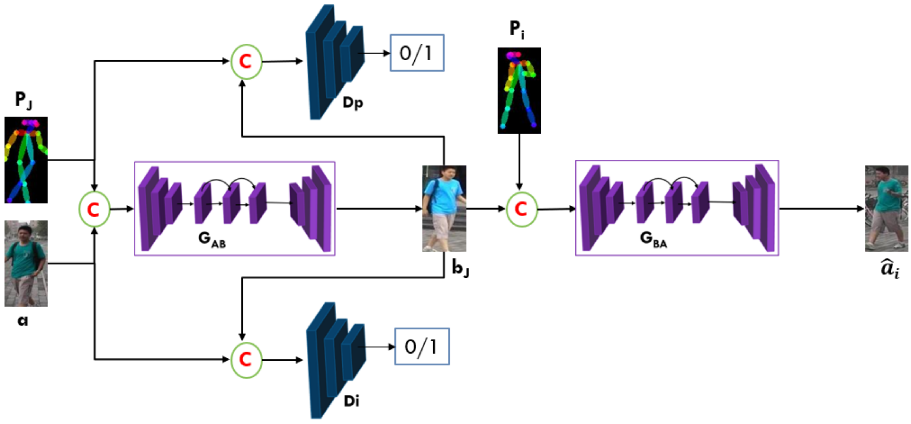

Figure 3.17: Overview of our framework. Gen A-B and Gen B-A are the generators to transfer from domain A to B and from B to A respectively. Dp and Di are the pose and identity discriminators. C is symbol for concatenation

3.6.1.1 Training

Assume we have $\{a_i^m, x_m\}_{i=1...Mi}^{m=1...M}$ and $\{b_j^n, y_n\}_{i=1...Nj}^{n=1...N}$ persons images in domain A and B respectively, where M and N are the number of images in both domains. i and j are the pose indexes from the total poses M_i and N_j of a person. The skeleton pose images are denoted as P_i and P_j for ith and jth pose respectively. x_m and y_n are the persons identities and for every training sample $x_m = y_n$. The full loss function is denoted as:

$$L = \arg\min_G \max_D \lambda_1 L_{GAN} + \lambda_2 L_{cycle} + \lambda_3 L_{identity} \tag{3.9}$$

where

$$\begin{aligned} L_{GAN} &= \mathbb{E}_{a,a_i \in \rho, p_i \in \rho_p} \log(D_p(P_i, a_i).D_i(a, a_i)) \\ &+ \mathbb{E}_{a \in \rho, p_i \in \rho_P, \hat{a}_i \in \hat{\rho}} \log[(1 - D_p(P_i, \hat{a}_i)).(1 - D_i(a, \hat{a}_i))] \end{aligned} \tag{3.10}$$

$$L_{cycle} = ||\hat{a}_i - a_i||_1 + ||\hat{b}_j - b_j||_1 \tag{3.11}$$

$$L_{identity} = ||G_{AB}(b, P_j) - b_j||_1 + ||G_{BA}(a, P_i) - a_i||_1 \tag{3.12}$$

L_{GAN} is the adversarial loss for first cycle and is calculated in the same way for the other cycle. As we are using two discriminators with each generator so the final outputs of these discriminators are multiplied to get the final score. ρ,$\hat{\rho}$ and ρ_P denote the distributions for real,fake and skeleton pose images. We use least square loss which is more stable [152]. In L_{cycle}, $\hat{a}_i$ and $\hat{b}_j$ are the reconstructed images as shown in figure 3.17. a_i and b_j are the ground truth images for skeleton poses P_i and P_j. Identity mapping loss is used to preserve the color composition between input and output [152].

Figure 3.18: Results generated by the two generators. (a) and (c) are the ground truths from camera **1** ($\boldsymbol{A}$ domain) and camera **6** ($\boldsymbol{B}$ domain) respectively. (b) shows the output of generator $\boldsymbol{B - A}$. The output of generator $\boldsymbol{A - B}$ is shown in (d).

3.6.2 Experimental Results

We select camera 1 and camera 6 images from Market-1501 [138] dataset and all the images are resized into 256 x 128. Adam optimizer is used with $\beta_1 = 0.5$ and $\beta_2 = 0.999$. Learning rates for generator and discriminator are 0.0002 and 0.0001 respectively. Generator consists of encoder decoder network with 9 ResNet basic blocks and PatchGAN [45] structure is used for all discriminators. The qualitative results of the proposed method are shown in figure 3.18. Generated images between two camera domains and their ground truths are shown. The inputs are from opposite domain and having different poses in each case.

We have proposed an image generation method which captures both domain and pose changes for re-id in section 3.6. In contrast with the previous approaches the proposed method merges both these variations in a single network. Generated images with the proposed approach provide domain and pose invariant features for person re-identification. Experimental results prove the image generation with the above mentioned variations.

3.7 Multi Branch Siamese Network for Person Re Identification

Variations like changing in view point, background clutters, different camera domains and occlusions make re-id a very difficult task. To resolve these issues, existing techniques focus on either robust feature representations [64, 120] or learn optimal matching metrics [78, 63]. Currently, deep learning based methods [18, 34] with the combination of the above mentioned solutions provide superior results outperforming traditional handcrafted low level feature representations for re-id. With the rapid growth of deep learning and convolutional neural networks designed for image classification, retrieving robust and impressive feature representations of person images in person re-id is more reliable. On the contrary to classification, the learned

descriptor discriminates between unseen similar images as the training and testing classes (identities) are different in re-id. Different researches [77, 16] aim to design better metric learning loss functions apart from feature learning including triplet loss, triplet hard loss, quadruplet loss etc for improving the generalization of the model. These metric learning losses have higher performance than classification losses because of dissimilar identities in the testing stage. Classification based approaches need to calculate the distance matrix of features for unseen person images during the inference time that creates mismatching due to different categories (i.e. person identities) in training and testing. To avoid this mismatching and learn more robust global features for person re-id, we propose a siamese network (metric learning) based on classification loss. For the better use of classification loss and to overcome the mismatch of features during testing, we add multiple classifiers to learn more disrciminative features from person images. The problem of person re-id undergoes many variations in images such as pose variations (different views of a single person) and domain variations (different camera domains i.e. camera environment and illuminations) as shown in Fig.1.3. To learn these type of variations many generative adversarial network (GAN) [31] based approaches [90, 66, 29, 148, 82] has been proposed. In these methods, new data is generated with the help of GAN and is added to the original training data to make it more robust to these variations. Generative models with pose variations and style (camera domains) variations have significantly improved the performance of person re-id.

In the proposed approach, the learned features are more discriminative and robust to overcome the pose variations. Different camera domains have different environments and illuminations (i.e. indoor and outdoor cameras) and produce images of their own style. To learn these variations, we generate augmented data for every camera style. Since Cycle-GAN translates images between two domains, we trained CycleGAN [151] models for each pair of cameras. Generated samples are added to the original data with a soft labeling [143] which are improving the performance of the proposed approach. We propose a multi classifier siamese network inte-

grated with CycleGAN [151] to learn discriminative features for person re-identification. Our contributions are as follows

- A Multi branch (classifier) siamese network with classification loss to learn the most discriminative features for person images.
- CycleGAN is integrated with the proposed siamese network to capture the style variations and to enhance the network performance.

By introducing the above contributions, the proposed approach produces better results than the existing methods on benchmark datasets for person re-id as mentioned in the experimental results section.

3.7.1 Multi-branch Siamese Network

We introduce the proposed network and its training mechanism in this section. The generated images have artifacts in them so can not be treated as original samples. We propose the loss function to use this augmented data.

3.7.1.1 Proposed Architecture

Recent works have shown that Convolutional Neural Networks (CNNs) are deeper and efficient in learning feature representations and accurate to train if they consist of shorter connections between layers. Leveraging on such outcomes, image encoders are defined upon ResNet-50 architecture in the proposed network. We modify the last downsampling block to make the spatial size of convolutional feature maps larger before global average pooling layer. We set the stride to 1 by following the work of R-FCN [24]. At the end of image encoder, we add a 1×1 convolutional layer to reduce the feature size from 2048 to 1024. This added convolutional layer learns the most discriminative global features from the entire person image. Two images features from encoders and convolution layers are feeded into element-wise subtraction, element-wise square,

batch normalization and fully connected (fc) layer to calculate the similarity score. To predict the person identity from features at convolutional layer, multiple classifiers (fully connected layers) are added along with batch normalization and Rectified Linear Unit (ReLU) layers. Since in back-propagation the gradients from the classifiers gather into the previous convolutional layers, thus the classifiers are responsible to focus the learned model on global features for computing distance matrix [129]. The overall architecture of the network is shown in Fig 3.20. Both the image encoders and added convolutional layers are sharing weights since they are performing the same task.

Figure 3.19: Camera style transferred images by Cycle-GAN from one camera to all other cameras in Market-1501. Two images in column one are taken from camera 3 and camera 1 and translate to all the remaining cameras in the dataset.

3.7.1.2 Camera to Camera Style Translation

We employ CycleGAN [151] to generate new samples in each camera style. The goal of the CycleGAN is to learn two mapping functions $G : X \rightarrow Y$ and $F : Y \rightarrow X$ such that the distribution X is indistinguishable from distribution of images from $F(Y)$ using adversarial loss and two discriminators D_X and D_Y. The overall loss function of CycleGAN is:

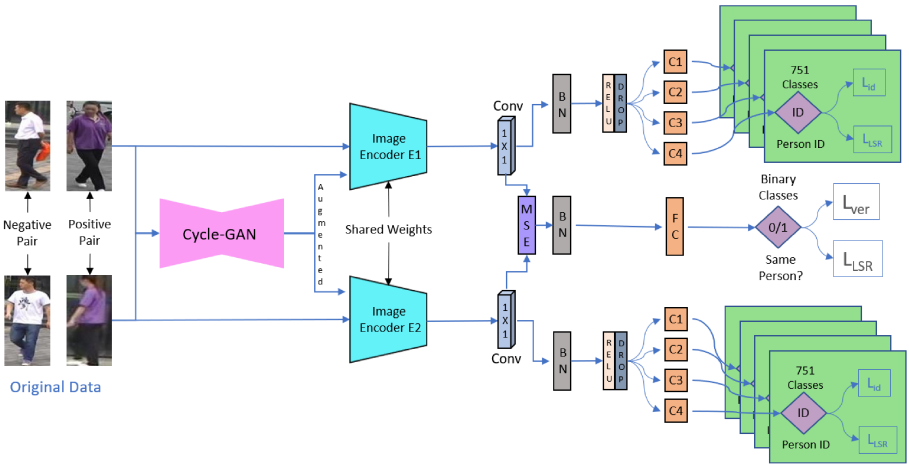

Figure 3.20: Overview of our framework. $C1$, $C2$, $C3$ and $C4$ are four fully connected layers for the predictions of person identity and their output losses are added to obtain the final loss. We show 751 classes from Market-1501 [138] dataset.

$$\begin{aligned} L(G, F, D_X, D_Y) &= L_{GAN}(D_Y, G, X, Y) \\ &+ L_{GAN}(D_X, F, Y, X) \\ &+ \lambda L_{cyc}(G, F) \end{aligned} \tag{3.13}$$

where $L_{GAN}(D_Y, G, X, Y)$ and $L_{GAN}(D_X, F, Y, X)$ are the adversarial loss functions for both mapping functions. $L_{cyc}(G, F)$ is the cycle consistency loss to reconstruct the image after cycle mapping and is given by:

$$L_{cyc}(G, F) = ||F(G(X)) - X||_1 + ||G(F(Y)) - X||_1 \tag{3.14}$$

λ is a weight parameter in eq 3.22 while X and Y are images from two different camera domains. To preserve the color consistency between input and output images, an identity mapping loss is added, which is expressed as:

$$L_{idt}(G, F) = ||F(X) - X||_1 + ||G(Y) - Y||_1 \tag{3.15}$$

We train CycleGAN [151] models for every pair of cameras in the datasets and follow the settings and networks architectures used in CamStyle [148]. Given a re-id dataset consisting of images collected from M cameras, we generate $M - 1$ new images for every image in the training set and refer them as style augmented images as shown in Fig 3.19. Since the contents of the original images are preserved in augmented images so we assign the same identity labels to newly generated samples. Along with the original training images, we use style augmented images in training to make the network robust to style variations.

3.7.1.3 Training

We use cross entropy classification losses for the training of original images. Two types of losses named as verification loss L_{ver} and identity loss L_{id} are used for similarity and identity learning respectively in the network. L_{ver} is the binary cross entropy loss and is given as:

$$L_{ver} = -C \log d(x_1, x_2) - (1 - C)(1 - \log d(x_1, x_2)) \tag{3.16}$$

where x_1, x_2 represent the two input person images and $d(x_1, x_2)$ is the output score of the network. C is the ground-truth label i.e if x_1, x_2 belongs to same person then $C = 1$ and $C = 0$ otherwise. To predict the identity of the person image, we use the cross entropy loss L_{id} which is written as:

$$L_{id} = -\sum_{c=1}^{C} \log(p(c))q(c) \tag{3.17}$$

where $p(c)$ is the output probability of the input belonging to class c and C is the total number of classes (person identities) in the dataset. $q(c)$ is the ground truth distribution and it is expressed as:

$$q(c) = \begin{cases} 1 & c = y \\ 0 & c \neq y \end{cases} \tag{3.18}$$

The generated augmented samples contain noise so they cannot be treated as real samples. To address this issue, we apply the label smoothing regularization (LSR) [143] for the augmented samples to assign soft labels to them. The redefinition of eq 3.18 is

$$q_{LSR}(c) = \begin{cases} 1 - \epsilon + \frac{\epsilon}{C} & c = y \\ \frac{\epsilon}{C} & c \neq y \end{cases} \tag{3.19}$$

where $\epsilon \in [0, 1]$ and we use $\epsilon = 0.1$ in the proposed method. With eq 3.24 cross entropy loss becomes L_{LSR} loss written as

$$L_{LSR} = -(1 - \epsilon) \log p(y) - \frac{\epsilon}{C} \sum_{c=1}^{C} \log p(c) \tag{3.20}$$

The overall loss function is the addition of all the losses from every classifier to train the whole network on the original and augmented samples.

$$L = L_{ver} + L_{id} + L_{LSR} \tag{3.21}$$

L_{id} and L_{LSR} losses are calculated at the output of each classifier which is represented in Fig 3.20. All these losses are added to make the final loss. In the testing stage, only image encoder is used along with the added convolutional layer to calculate image features and Euclidean distances between probe and gallery images. The proposed training scheme uses the generated images more efficiently with the original samples and enhances the performance of the proposed method compare to other methods having generated data.

3.7.2 Experimental Results

Experimental results with the original data and the mixture of original and augmented data are presented in this section.

3.7.2.1 Datasets

We conduct our experiments on two benchmark datasets Market-1501 [138] and DukeMTMC-reID [91]. The statistics of these two datasets are shown in table 3.10. We adopt standard data split setting and single query test.

Table 3.10: Statistics of two person re-id benchmark datasets Market1501 and DukeMTMC-reID. The details about the splitting of the persons identities into training and testing sets are shown.

Benchmark	Item	Total	Train	Test
Market-1501	ID	1501	751	750
	Image	32668	12936	19281
DukeMTMC-reID	ID	1404	702	702
	Image	36411	16522	17661

3.7.2.2 Implementation Details

We implemented the proposed model using Pytorch. Resnet50 [38] is used as image encoder E1 and E2 pretrained on ImageNet with the settings mentioned in section 2.2. The network is optimized by Stochastic Gradient Descent (SGD) with momentum 0.9. The initial learning rates are 0.001 and 0.1 for image encoders and all the other layers respectively, and they are divided by 10 after 80 epochs as we train 100 epochs in total. The batch size is set to 64 with positive-negative ratio and generated-original ratio are $1:3$. All the images are resized to 256×128 with random cropping and random horizontal flipping data augmentations. The dropout probability is 0.5. For the training of Cycle-GAN [151] we followed the setting used in Cam-style [148].

3.7.2.3 Comparison with the State-of-Arts Methods

Table 3.11 shows the performance comparison with the previous methods. The dashed line in table 3.11 is splitting two types of methods based on training data. The methods below the dashed line add extra augmented data generated by Generative Adversarial Networks (GANs) [31] for training compared to the above methods which are trained on the original data only. Range-s [119] has better Mean Average Precision (mAP) because it is based on re-ranking algorithm while we present our results without any type of re-ranking. The posted results of IDCL [129] are based only on multi branch strategy as we are using that type of multi classifier technique (table 3.12, 3^{rd} row). We perform better than all other methods in terms of Rank 1 Accuracy (R1). Compared to GAN methods (below the dashed line), our performance with the augmented data is much higher than them in the case of both measurements. The limitation of the proposed method is the number of cameras because large number of cameras have very high computational cost when calculating style transfer between each pair of cameras.

Table 3.11: Comparisons to the state-of-the-art re-id methods on Market-1501 and DukeMTMC-ReID. The top 1 and 2 results are in red and blue. Methods between the two dashed lines are using generated data along with the original data.

Methods	**Reference**	**Market**		**DukeMTMC-ReID**	
		mAP	**R1**	**mAP**	**R1**
SVDNet [103]	ICCV17	62.1	82.3	56.8	76.7
DPFL [17]	ICCV17	73.1	88.9	60.6	73.2
BraidNet [114]	CVPR18	69.5	83.7	59.5	76.4
PSE [92]	CVPR18	69.0	87.7	62.0	79.8
MLFN [8]	CVPR18	74.3	90.0	62.8	81.0
Range-s [119]	ICIP19	81.0	90.7	70.1	81.9
IDCL [129]	CVPRW19	73.3	89.2	59.1	79.4
LSRO [143]	ICCV17	66.1	84.0	47.1	67.7
PT [66]	CVPR18	68.9	87.7	56.9	78.5
PN-GAN [90]	ECCV18	72.6	89.4	53.2	73.6
CAM-Style [148]	CVPR18	71.5	89.4	57.6	78.2
OURS	-	75.9	91.1	62.9	82.2

3.7.2.4 Component Analysis

We divide the proposed network into four components to make an ablation study and verify the effectiveness of each component. The Muti-C in table 3.12 represents four classifiers and removing it means we are using a single classifier. The existence of L_{ver} makes the network a siamese structure, without which the network is only a single line structure. We carry out experiments on Market dataset with different combinations. The detailed results are described in table 3.12.

Table 3.12: Component Analysis of the proposed Multi Classifier Siamese Network on Market-1501 dataset in terms of mAP(%) and top-1 accuracy(%). The combinations of the selected components is in the middle of the table.

Networks	**Components**				**Market1501**	
	L_{ver}	L_{id}	**Multi-C**	**Cycle-GAN**	**mAP**	**R1**
baseline single	×	✓	×	×	65.8	85.6
baseline siamese	✓	✓	×	×	71.5	88.4
proposed(no ver)	×	✓	✓	×	73.3	89.2
proposed(no cgan)	✓	✓	✓	×	75.0	90.2
proposed	✓	✓	✓	✓	**75.9**	**91.1**

We propose a Multi Branch (classifier) Siamese Network along with Cycle-GAN for person re-identification in section 3.7. With multiple classifiers and losses, proposed network learns robust global features at the added convolutional layers. Multiple identity losses are merged with verification loss to build a stronger and disciminative descriptor . To overcome the camera style variations, we generate augmented data with the help of cycle-GAN. During training, augmented data is utilized by providing soft labeling loss function along with original data. Experimental results demonstrate the benefits of the proposed method in enhancing the performance of person re-id on two benchmark datasets.

3.8 Resolution based Feature Distillation for Cross Resolution Person Re Identification

The presence of illumination changes, occlusions, background clutter and viewpoint changes makes re-id a challenging task for practical applications. With a rapid advancement and success of deep learning and convolutional neural networks (CNNs), many learning based approaches [83, 112, 76, 72, 13, 129, 84] have been proposed for re-id. These methods achieve promising results but assume that query and gallery images have similar resolutions (e.g. high resolution). However, this assumption is not true for realistic scenarios since image resolutions would vary drastically because query images captured by the surveillance cameras are often of low resolution (LR). Usually, the gallery images are chosen beforehand and are of high resolution (HR). This creates a non-trivial resolution mismatching problem due to the direct matching of LR query images with HR gallery images. To address low resolution person re-id, most existing methods develop image super resolution (SR) based solutions [48, 117, 20, 35, 4] by converting the LR images into HR images and perform re-id. However, the limitation of such methods is that they operate on known upscaling factors to generate the HR images. A specific SR model is required for each scaling factor which limits generalization (e.g. both source and target resolutions must be known). Several generative adversarial network (GAN) based methods [62, 19, 117] are also proposed to resolve the above limitations. These methods learn all the degradation in a single network by using adversarial training mechanism and make the learned features independent of the resolution. Although these methods produce promising results, they are computationally expensive and unable to capture highly degraded images which limits exploitation in re-id. Some image super resolution works [133, 25] used attention mechanisms to improve the learning ability of the network from degraded data and hence provide noise-free sharp features to generate a super resolved image. Benefiting from such mechanism, we compute multiple channels correlations in the baseline network [112] to learn enhanced features in

the presence of degraded data.

Along with the discussed limitations, low resolution person re-id methods also work on the assumption that all the gallery images are HR images. In a better realistic scenario, the gallery sets are also collected in the form of LR images (e.g. persons at different distances from the camera) so there should be multiple resolutions even in the gallery images instead of a single resolution (HR).

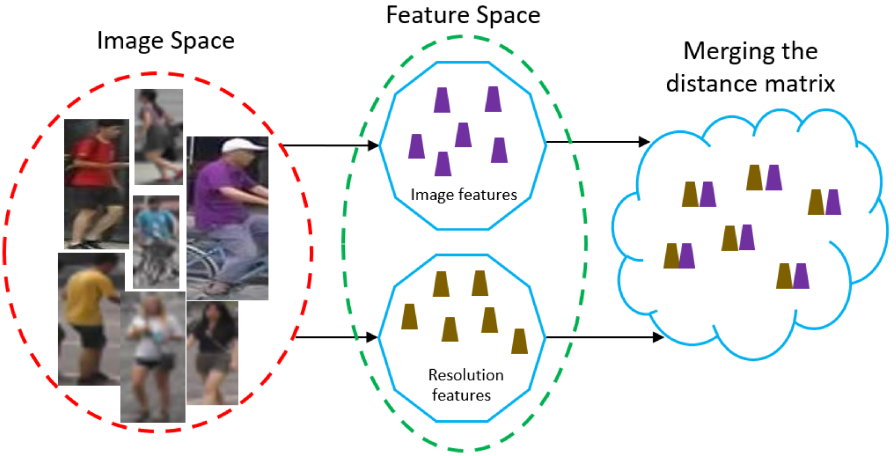

Figure 3.21: From the input degraded data image and resolution based features are learned by the proposed baseline and their distance matrices are merged to get the final distance matrix.

In this work, we propose a Resolution based Feature Distillation (RFD) approach to compute resolution invariant features to solve the multi resolution gallery set problem. First, we train a feature baseline (B-F) based on MGN [112] approach with several modifications and train a resolution baseline (B-R) to distinguish each resolution present in the synthetic dataset. For the real multi resolution dataset we propose pseudo labeling of the resolutions and then train the B-R. Finally, we filter out the resolution variant features by combining the distance matrix from both the baselines. The overview of the proposed approach is shown in Fig. 3.21. This work has the following contributions:

- Improve the baseline network with the addition of multiple channels correlation modules to learn better representations from degraded data.
- Propose a resolution based feature distillation with feature and resolution baselines to match the features of different resolutions to perform the cross resolution re-id.
- Adopt a new and more realistic scenario assuming that the gallery images are also collected in LR form. Thus, the LR query image is matched with multiple resolutions (HR and LR) instead of a single resolution (HR).

With the above mentioned contributions, we performed the experiments on two synthetically created re-id datasets and one real dataset with multiple resolutions. Our feature baseline produces competitive results on the low resolution re-id (single resolution in gallery set) with other state of the art methods. The proposed RFD approach improves the results in the case of multiple resolutions in the gallery set when compared to a single baseline (B-F).

3.8.1 Proposed Method

In this section, we provide the approach overview, network architecture and resolution based feature distillation (RFD) technique.

3.8.1.1 Notations and Overview

Let $X_H = \{x_i^H\}_{i=1}^N$ be a set of N HR training images, with corresponding identities labels $Y_H\{y_i^H\}_{i=1}^N$, acquired by a camera network where $x_i^H \in \mathbb{R}^{H\times W\times 3}$ and $y_i^H \in \mathbb{R}$ are the i^{th} HR image and its label respectively. To enable the model to learn different resolutions, we generate the synthetic LR image set $X_L = \{x_i^L\}_{i=1}^N$ by downsampling and upsampling (bilinear) back to the original image size each image in X_H i.e. $x_i^L \in \mathbb{R}^{H\times W\times 3}$. The Y_L labels for the corresponding X_L images are the same of Y_H.

Low resolution person re-id requires to retrieve images of the same identity from a HR gallery

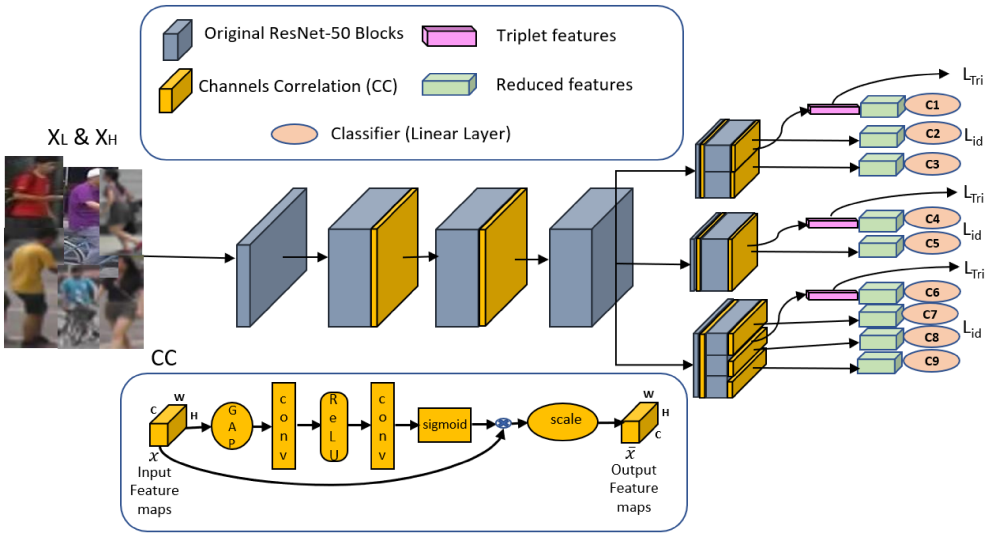

Figure 3.22: Architecture of the baseline and overview of the proposed Resolution based Feature Distillation (RFD) approach.

images of different cameras by giving a LR probe image. While in this work we make a cross resolution person re-id by building a gallery set of multiple resolutions instead of single resolution (HR). During the training stage, a feature baseline B-F is used to extract the features from the images of all resolutions and persons identities are predicted by exploiting classification and metric learning losses. The networks details and the loss functions are explained in the next section. A resolution baseline B-R is trained to classify the resolutions present in the dataset. In the testing stage, we compute the distance matrix from each baseline and then merge them to get the final distance matrix. This final distance matrix is used to perform cross resolution person re-id.

3.8.1.2 Network Architecture

Recent works have shown that Convolutional Neural Networks (CNNs) are efficient for learning deeper and robust feature representations from images and are accurate to train if they have

shorter connections between layers. Relying on such outcomes, we define ResNet-50 [38] as our baseline network with several adjustments. We follow the MGN [112] architecture and introduce the final convolutional block three times to make global and local modules. Concerning the global module, we modify the stride (stride=1) of the last downsampling block to make the spatial size of the convolutional features larger before global average pooling. Channel correlations are computed at each stage of the network as in [133] to make the learned feature sharp and discriminative. Unlike [133], we introduce the channel attention modules after each stage of the network instead of using them only before residual connections as shown in Fig. 3.22. Cross entropy loss is used to predict the identity of each person and all the cross entropy losses from each branch are added to get the final loss which is given by:

$$L_{id} = -\sum_{b=1}^{B_i}\sum_{c=1}^{C}\log(p_b(c))q_b(c) \tag{3.22}$$

where B_i is the number of classification branches. Hard mining triplet loss from each branch is added to achieve the final loss to make the features discriminative and it is computed as:

$$L_{tri} = \sum_{b=1}^{B_t} max(\|x_b^a - x_b^p\| - \|x_b^a - x_b^n\| + \alpha, 0) \tag{3.23}$$

where B_t is the number of branches to compute triplet loss. This loss ensures that the representation of the positive sample is closer, by at least a margin α, to the anchor sample than to the negative one. For both the baselines, we use the same structure and modifications discussed above. Now we review channels correlations which are computed with the help of channel attention mechanism. We have discussed the details of channels correlations in section 3.4.1.3.

3.8.1.3 Resolution based Feature Distillation

In this section, first we perform the learning of resolution similarity as a resolution re-id problem. This task is performed by using the baseline B-F to compute the resolution features and

provide a classification between the available resolutions. However, instead of outputting a regression result or a classified direction resolution re-id outputs an embedding that can be used to compute resolutions similarity. To train the baseline B-R we need the resolution labels which are explained in the next section. In the testing stage, for an x image we retain the resolution features $f_r(x)$ from the baseline B-R. Then calculate the distance between the two images x_i and x_j is computed as:

$$D_r(x_i, x_j) = \frac{f_r(x_i) \cdot f_r(x_j)}{\|f_r(x_i)\| \, \|f_r(x_j)\|} \tag{3.24}$$

The distance matrix from the baseline B-F is computed as the euclidean distance between the query and gallery images and then merged with the distance in eq.3.24 to compute the final distance matrix for person re-id. The final distance matrix provides resolution invariant matching of features and it is given by:

$$D(x_i, x_j) = D_f(x_i, x_j) - \lambda D_r(x_i, x_j) \tag{3.25}$$

where D_f is the distance matrix calculated from the baseline B-F similar to eq.3.24 and λ is the scaling parameter (we use $\lambda = 0.1$ for all our experiments). This distillation process enhances the performance when we have multiple resolutions in the gallery set as well as in the query set and is shown in Fig. 3.23.

3.8.1.4 Resolution pseudo labeling

The training of the baseline B-R for multiple resolutions requires labels for each resolution in the dataset. For synthetic datasets we have known downsampling scaling factors and we used them as training labels for baseline B-R. For synthetic datasets, we have 4 different classes (labels) which include 3 downscaling factors i.e.$2, 3, 4$ and one original HR resolution. This type of downsampling is not available for the real dataset having multiple different resolutions. Therefore, for them, we propose a pseudo labeling process that computes the total number of

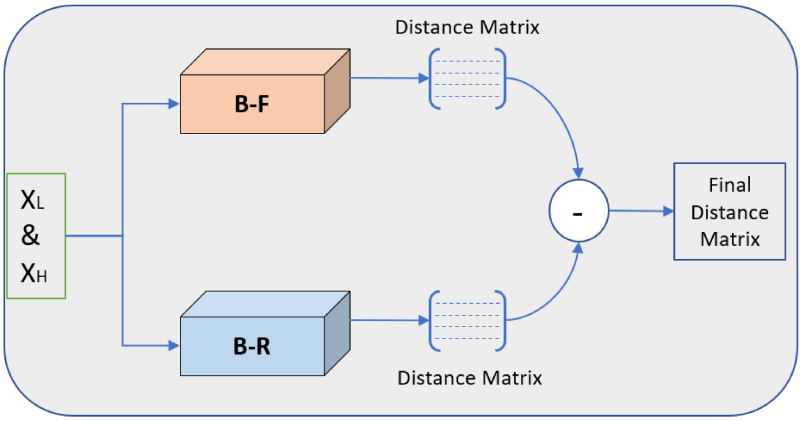

Figure 3.23: Proposed Resolution based Feature Distillation (RFD) training mechanism. x_L and x_H are the multiple low resolution and high resolution images. B-F and B-R are the baselines trained with person ids and resolution ids respectively.

pixels for each image. Then we divide these number of pixels into 5 patches of equal length. Each image within the same patch is represented by single resolution pseudo label. With this technique, we build the training set with 5 resolutions for the real dataset.

3.8.2 Experiments

3.8.2.1 Implementation Details

We implemented the proposed network using Pytorch. The baselines are built upon a ResNet-50 network pretrained on ImageNet. We modified the network with all the adjustments mentioned in section 3.2. We optimized the network by using Adam optimizer with momentum 0.9. We trained the models for $500 \times (\textit{number of identities})$ iteration with the initial learning rate is $2e-4$ and is divided by 10 after half iterations. We used a batch size of 32 (having 8 identities with 4 images each) and $(\textit{number of resolutions}) \times 16$ for baseline B-F and baseline B-R respectively. All the images are resized to 384×128 with random horizontal flipping and random erasing data augmentations. The dropout probability is set to 0.5 and the weight decay is $5e-4$.

3.8.2.2 Datasets

We evaluated the proposed method on three datasets which are described as follows:

MLR-market dataset consists of 32668 images of 1501 identities captured from 6 different cameras. Training and testing set consists of 751 and 750 identities respectively. We follow SING [47] to create the MLR-market1501 dataset.

MLR-Duke dataset contains 36411 images of 1404 identities with 8 camera views. Following SING [47] and 702/702 splits for training and testing, we create MLR-Duke dataset.

CAVIAR is a real multi resolution dataset and is composed of 1220 images of 72 persons taken from 2 different cameras. We discard 22 identities those appear in one camera and split the dataset into two non-overlapping halves by following SING [47].

Table 3.13: Results of the proposed (B-F + RFD) method and their comparison with the state-of-the-art person re-id methods on MLR-Market dataset. The best and second best rank1, rank5, rank10 accuracies are highlighted in red and blue respectively.

Methods	**Reference**	**MLR-Market**		
		R1	***R5***	***R10***
SING [47]	AAAI18	74.4	87.8	91.6
CSR-GAN [117]	IJCAI18	76.4	88.5	91.9
CamStyle [149]	CVPR18	74.5	88.6	93.0
FD-GAN [30]	NeurIPS18	79.6	91.6	93.5
RIPR [74]	IJCAI19	66.9	84.7	-
CRGAN [62]	ICCV19	83.7	92.7	95.8
MSA [1]	IEEE Access20	68.3	85.7	-
INTACT [20]	CVPR20	88.1	95.0	96.9
PRI [35]	ECCV20	88.1	94.2	96.5
baseline	-	84.1	92.3	95.9
baseline B-F	-	85.5	94.1	96.0
(B-F+RFD) Proposed	-	86.9	95.6	97.4

3.8.2.3 Comparison with state-of-the-art

Table 3.13, 3.14 and 3.15 show the results of the proposed method and its comparison with other state of the art methods on two synthetic datasets and one real dataset. We refer MLR-Market

and MLR-Duke as synthetic datasets because the degraded process is known in datasets while the CAVIAR dataset has unknown degradation. We computed rank-1 (R1), rank-5 (R5) and rank-10 (R10) accuracy by using our baseline B-F. We used bicubic downsampling to generate the low resolution images with a randomly selected scaling factor of $2, 3, 4$. The results in Table 3.13, 3.14 and 3.15 are computed for low resolution person re-id in which the query image is randomly down sampled and the gallery set has all HR images. Like for the existing methods, we make random splits of the data for 10 times and then take the average to calculate the final scores. The performance of the proposed method is significant when compared to the other state of the art methods.

Table 3.14: Results and comparisons of the proposed (B-F + RFD) method with the state-of-the-art perosn re-id methods on MLR-Duke dataset. The best and second best results are highlighted in red and blue respectively.

Methods	**Reference**	**MLR-Duke**		
		R1	***R5***	***R10***
SING [47]	AAAI18	65.2	80.1	84.8
CSR-GAN [117]	IJCAI18	67.6	81.4	85.1
CamStyle [149]	CVPR18	64.0	78.1	84.4
FD-GAN [30]	NeurIPS18	67.5	82.0	85.3
CRGAN [62]	ICCV19	75.6	86.7	89.6
MSA [1]	IEEE Access20	79.1	90.0	-
INTACT [20]	CVPR20	81.2	90.1	92.8
PRI [35]	ECCV20	82.1	91.1	92.8
baseline	-	81.2	90.1	91.9
baseline B-F	-	82.0	90.8	92.7
(B-F+RFD) Proposed	-	82.9	92.0	94.0

Table 3.13 shows the results and comparison of the proposed method with state of the art methods on MLR-Market dataset. The proposed contributions improved the R1 accuracy by 2.8% when compared to the baseline. The R1 accuracy of the proposed method is the second highest in the table when compared to the state of the art methods but the improvement in the baseline is quite significant. The results for MLR-Duke dataset are presented in Table 3.14 and the performance of the proposed approach is superior to the state of the art methods in terms

of R1, R5 and R10 accuracy. The proposed approach performs better and uses training data only which contains low and high resolution images and no extra data generated or any super resolution models are introduced.

The results and comparison of the proposed approach on real world degraded dataset CAVIAR are shown in Table 3.15. We obtained the highest results among all other methods in the presence of real world degradation. The proposed RFD enhanced the R1 accuracy by 3.3% when compared with baseline B-F and the reason for such an improvement is the proposed pseudo labeling technique which separate resolutions through baseline B-R. The CAVIAR dataset is a very small dataset compared to MLR-Market and MLR-Duke so improving the rank 1's of few queries has a greater impact on the performance.

Table 3.15: Comparisons of the proposed (B-F+RFD) method with the state-of-the-art person re-id methods on the real world multi resolution CAVIAR dataset. We compute rank1, rank5, rank10 accuracy and present best and second best results with red and blue colors respectively.

Methods	**Reference**	**CAVIAR**		
		R1	***R5***	***R10***
SING [47]	AAAI18	33.5	72.7	89.0
CSR-GAN [117]	IJCAI18	34.7	72.5	87.4
CamStyle [149]	CVPR18	32.1	72.3	85.9
FD-GAN [30]	NeurIPS18	33.5	71.4	86.5
RIPR [74]	IJCAI19	36.4	72.0	-
CRGAN [62]	ICCV19	42.8	76.2	91.5
INTACT [20]	CVPR20	44.0	81.8	93.9
PRI [35]	ECCV20	45.2	84.1	94.6
baseline	-	42.1	87.6	92.9
baseline B-F	-	44.3	88.4	94.2
(B-F+RFD) Proposed	-	47.6	89.2	96.0

3.8.2.4 Ablation Study

We performed an ablation study for the proposed approach in Table 3.16 and 3.17. We created the synthetic data (low resolution images) by using two types of interpolations, bicubic and

Table 3.16: Ablation study of the proposed B-F baseline with RFD method on two synthetic datasets MLR-Market and MLR-Duke. We create datasets with two traditional methods (first two rows) and then create different splits for query and (single and multi resolution) gallery

Components	**MLR-Market**		**MLR-Duke**	
	R1	***R5***	***R1***	***R5***
Bicubic	83.1	96.0	82.4	92.0
Bilinear	82.9	95.9	82.2	92.1
Single-Reso	85.5	94.1	82.0	90.8
Multi-Reso	84.2	93.8	81.5	90.0
Multi-Reso+RFD	85.6	94.4	83.0	91.8

bilinear for MLR-Market and MLR-Duke dataset. We performed the experiments and recorded the results in the first two rows of Table 3.16. The performance remains almost similar with both interpolations. Third and fourth rows show the results on single and multi resolution gallery set respectively, and results are without using RFD method.

Table 3.17: Results in terms of rank1, rank5 and rank10 accuracy on CAVIAR dataset of the proposed B-F baseline and RFD method method for single and multi resolution gallery set.

Components	**CAVIAR**		
	R1	***R5***	***R10***
Single-Reso	44.3	88.4	94.2
Multi-Reso	55.5	90.6	91.5
multi-Reso+RFD	56.4	90.6	92.2

We generated the multi resolution gallery set by randomly downsample images from synthetic datasets and by randomly pick one image for each identity from all resolutions for CAVIAR dataset. The performance for CAVIAR dataset is recorded in the first and second rows of the Table 3.17 for single and multiple resolutions in the gallery set. This multi resolution causes performance reduction for synthetic datasets while performs better in the case of real dataset. The reason for this better performance is the multiple HR query images are also present unlike single resolution. The improvement with the proposed RFD for multiple resolutions in the gallery is shown in the last row of Table 3.14 and 3.17 for synthetic and real data respectively.

The pseudo labeling significantly enhance the performance for CAVAIR dataset. Rank-5 scores remain almost similar because of the presence of only single image for each query in the gallery set.

We presented the effects of the proposed pseudo labeling for CAVIAR dataset in Fig. 3.24. We generated ten random splits of the dataset for testing and recorded their performance (rank-1 accuracy) along with RFD approach in Fig. 3.24. Orange and blue lines represent the baseline B-F and RFD performances respectively. Some splits are significantly better while the others have similar performance. We noticed a reduction in the performance of $3rd$ and $4th$ split which is due to the fact that the gallery and query set in that split do not have much resolution variance. The average scores are higher than the baseline when using RFD for 10 splits of the data.

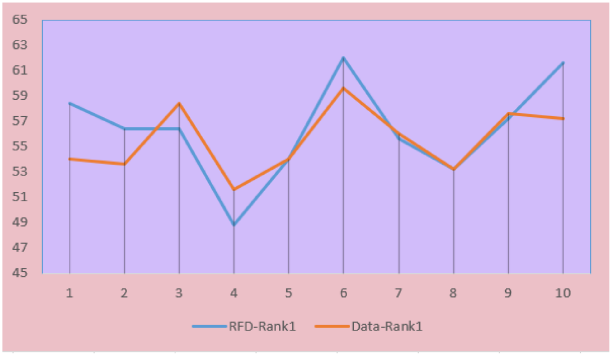

Figure 3.24: The effect of proposed pseudo labeling technique with the B- baseline on CAVIAR dataset on each split (quries and gallery) generated from data. Random splits of the dataset are shown on horizontal axis with their performance on vertical axis.

The visual results of the proposed approach are shown in Fig. 3.25. Five query images from CAVIAR dataset are presented along with their first ten matches in the gallery set. Query images are on the left side of the red dashed line while the right side of the red dashed line are the top 10 rankings from the gallery images. There is only one true match since the gallery set only has single images for each query to match. We did not resize the images so the original resolutions have been used and are shown. The green rectangle is used to show the true match

in the gallery set.

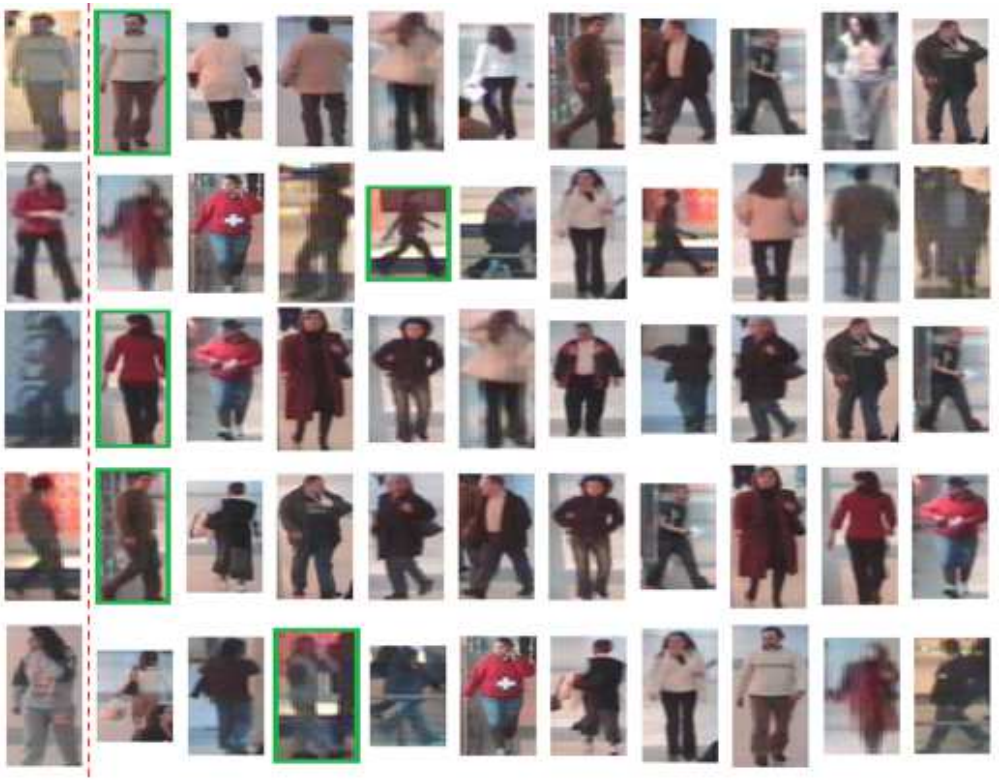

Figure 3.25: Visual results extracted by the proposed B-F baseline and RFD method. Query image is on the right side (first column) of the red dashed line and its first ten matches in the gallery on the right side for CAVIAR dataset.

We proposed a resolution based feature distillation (RFD) approach for cross resolution person re-identification in section 3.8. Firstly, we improved a baseline by means of channels correlation to solve the general low resolution person re-id problem (only HR images in the gallery set). We achieved competitive results on three datasets. Secondly, we proposed a resolution based feature distillation technique which filters out the resolution dependent features to compute the final distance matrix for matching. A pseudo labelling technique for computing the

resolution label is also introduced to train the RFD. The proposed approach considers the better realistic scenario in which gallery sets contain multiple resolutions instead of single resolution (HR). This is a novel scenario that we think should be further investigated in future studies within the person re-id community. Experimental results have shown how the RFD approach improves with respect to both the baseline and other state of the art approaches.

4

Vehicle Re-Identification

4.1 Oriented Splits Network to Distill Background for Vehicle Re-Identification

A variety of intra-class variations like the diversity of car shapes from different viewpoints and inter-class variations (models produced by various manufacturers have limited colors and shapes) make vehicle re-id a very challenging task. Preliminary approaches adopted hand-crafted visual features [70, 128] designs and used them for matching vehicles. Recent approaches have utilized the benefits of person re-identification methods [83, 76, 112] and their network architectures, loss functions and data augmentations. These approaches exploit deep neural networks to learn the discriminative feature representations. Other methods introduce attention mechanism [55, 54, 53] and part-based solutions [36, 132, 15] and significantly improve the performance of vehicle re-id with these deeply-learned part-based representations. Vehicle parts provide stable semantic cues but additional label data and human efforts are required to obtain such parts. In addition, these part-based solutions have also increased the computational cost which is not feasible when the solution is deployed in a real world environment. Vehicle

re-id has a large number of applications in video surveillance and traffic analysis so requires fast real time and low computational power inference. To handle these challenges, horizontal and vertical splits [93] of the feature maps are introduced, which overcome the issue of generating parts. Vehicle datasets have images of vehicles in several varying orientations. These splits perform well with the few orientations (horizontal and vertical) and fail when other orientations are present in the image.

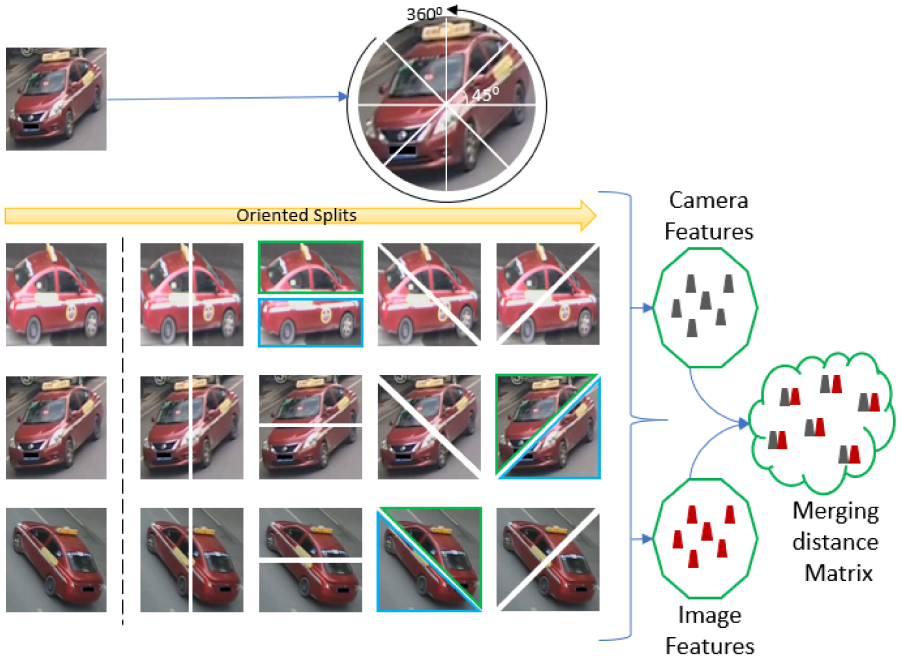

Figure 4.1: Overall framework of the proposed method. The input image is divided into four splits i.e. vertical, horizontal and two diagonal. Correlated parts are shown with the same color contour. Features from these splits along with the camera features are used to compute the final distance matrix.

To cover all the orientations and learn the local descriptor through all regions (angles), we proposed an Oriented Splits Network (OSN) which provides an alternative and efficient technique compared to the part-based solutions. The proposed OSN learns the global and local features in an end to end manner without providing parts and alignments. The global features

focus on representing the most salient clues of appearances of different vehicles. However, some non-salient or infrequent detailed information is ignored and has no contribution for better discrimination due to the limited scale and weak diversity of vehicle re-id datasets. For covering these missing details, the proposed OSN also learns local features by using several types of oriented stripes and splits from the input image as shown in Fig 4.1. The upper and lower parts of the vehicles are shown in the same colors contours for different orientations present within the same vehicle. These oriented splits make each part/region to contribute in the prediction of the vehicle and effective for occlusion cases. The proposed OSN is composed of one global and four local branches in which the globally pooled feature maps are divided into multiple oriented splits and stripes as shown in fig 4.1.

Vehicle images from a single camera undergo similar backgrounds. The presence of background in the input images contributes to the learned features and affects the matching performance in re-id process. To filter out the dependency of the background, we learn the similarity of the background by interpreting it as a camera re-identification problem. We use the proposed OSN for camera based feature learning by taking cameras as labels and named the learned features from this network as camera features. We introduce a distillation process based on these camera features to remove the effect of the background. The proposed oriented splits enhance the learning capability of OSN and learns discrminative representations for each camera. The image features as well as camera features are merged to obtain the final distance matrix in the testing stage. The final distances are computed with the features independent of background and hence improve the performance of vehicle re-id. The overall framework of the proposed approach is shown in fig 4.1 and we made the following contributions in this work:

- Proposed an Oriented Splits Network (OSN) to learn global and local representations in an end to end manner. The local stripes and splits are obtained from the pooled feature maps instead of image parts which are computationally expensive to obtain.
- Introduction of a oriented distillation training process to filter out the features having

background information. Similar backgrounds are classified with the proposed OSN and then camera features from this network are used to compute the final distance matrix.

Unlike the existing methods, we propose a novel splitting strategy which covers all the orientations present in the vehicle images instead of adopting part-based and horizontal and vertical splits mechanism. To remove the effect of background features in vehicle predictions, a distillation training technique for oriented splits is proposed to filter the background similarity from the learned features. By contributing the above mentioned modifications, We perform experiments on two benchmarks of vehicle re-id. Results on these datasets show the improvement in the performance and robustness of the proposed technique when compared with other state of the art methods.

4.1.1 Proposed Oriented Splits Network

4.1.1.1 Problem Definition and Notations

Let a set of n training images of vehicles $\left\{I_i\right\}_{i=1}^{n}$ with corresponding identities labels $\left\{y_i\right\}_{i=1}^{n}$ be acquired by a camera network. The task of vehicle re-identification is to retrieve similar vehicles to the given probe image from the gallery sets of different cameras. Unlike image classification task, re-identification tasks required the most discriminative and unique representations of the objects since the training and testing classes (vehicle identities) are not identical. With the help of classifiers, the discriminative features for each vehicle are learned from the dataset. During testing, these features from all classifiers are merged to compute the distance matrix between the probe and the gallery images to achieve vehicle re-identification.

4.1.1.2 Network Architecture

Recent works have shown that Convolutional Neural Networks (CNNs) are efficient for learning deeper and robust feature representations from images and are accurate to train if they have shorter connections between layers. Relying on such outcomes and to benefit from batch and

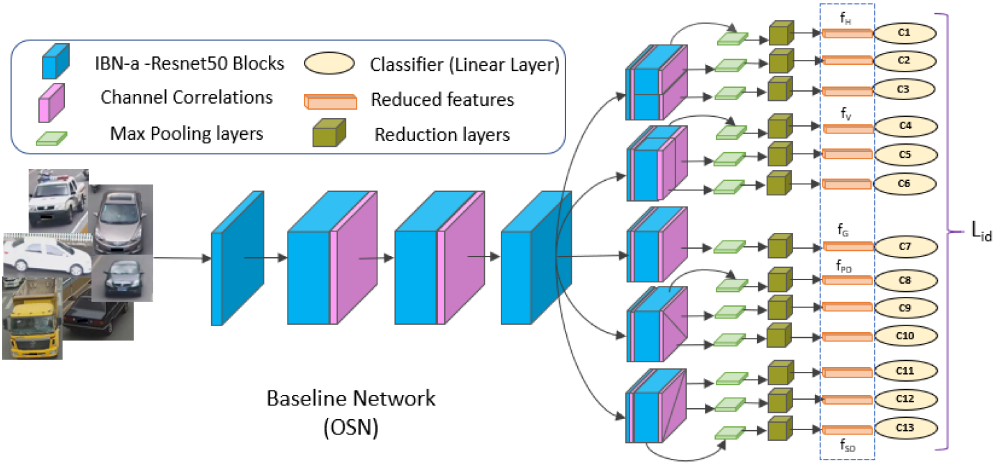

Figure 4.2: The proposed OSN Network. IBN-a-Resnet50 is used with channel attention modules at the end of each convolution block. Oriented splits produce a global feature vector along with its local feature vectors. Global features are used to compute the triplet loss before reduction. $c1...c13$ are the linear layers to predict the identity of the vehicle.

instance normalization, we define IBN-a-ResNet-50 [87] as our baseline network with several adjustments. We follow the MGN [112] architecture and scheme with multiple oriented novel splits of image feature maps. First two convolutional blocks are used with the addition of a channel attention (channels correlations) block at the end of each of them. Third block is split and its last convolution layer is embedded to fourth block with a channel attention module in the middle. Five duplicates are generated from this fourth module to represent the horizontal, vertical, global and two diagonal splits. The feature vectors from all splits are then fed into 1×1 convolution layers to obtain the reduced feature vectors. These reduced feature vectors are then fed into linear layers (classifiers) to predict the identities of the vehicle based on the local splits and stripes.

Concerning the global module, we modify the stride (stride=1) of the last downsampling block to make the spatial size of the convolutional features larger before global average pooling. Global average pooling layers are replaced with global max pooling layers. The proposed

network (OSN) architecture is shown in fig 4.2. Channel correlations are computed at each stage of the network as in [84, 43] to make the learned feature sharp and discriminative. Unlike [84, 43], we introduce the channel attention modules after each stage of the network instead of using them only before residual connection. In the testing stage, the feature vectors from all reduction layers (reduced features) are concatenated to create the final representation of the vehicle which is used to compute the distance matrix between probe and gallery.

To enhance the the discrimination ability of the OSN, we employ softmax and triplet losses for classification and metric learning respectively. To learn the discriminative features, we address the re-id task as a multi-class problem like the other existing re-id models. All the classification losses from each branch are added to get the final loss, which is given as:

$$L_{id} = -\sum_{i=1}^{N} \log \frac{e^{W_{yi}^{T} f_i}}{\sum_{k=1}^{C} e^{W_{yi}^{T} f_i}} \tag{4.1}$$

where W_k is the weight matrix for class k with total C classes in the dataset. N is the size of a mini-batch. Hard mining triplet loss from each branch is added to achieve the final loss to make the features discriminative and it is computed as:

$$L_{tri} = \sum_{b=1}^{B_t} max(\|f_b^a - f_b^p\| - \|f_b^a - f_b^n\| + \alpha, 0) \tag{4.2}$$

where B_t is the number of branches which are $B_t = \{f_H, f_V, f_G, f_{PD}, f_{SD}\}$ to compute triplet loss and f_b^a, f_b^p and f_b^n are the features from branch b for anchor, positive and negative respectively. This loss ensures that the representation of the positive sample is closer, by at least a margin α, to the anchor sample than to the negative one.

4.1.1.3 Background Distillation

In this section, we introduce a camera based feature distillation technique to learn the background similarity and filter out its effect from the final representations. Each camera in a surveil-

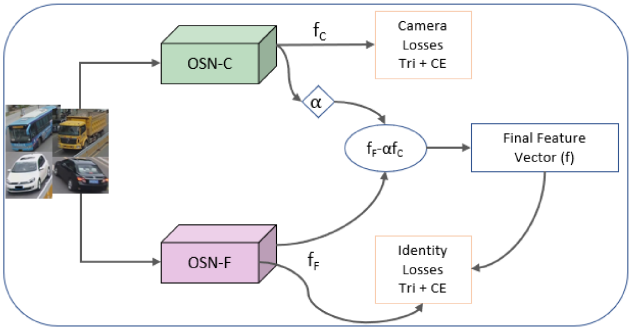

Figure 4.3: Camera base distillation process to learn the background similarity. OSN-C computes the camera features and OSN-F provides image features. The features from both baselines are merged in final feature representation and are used to obtain the final distance matrix.

lance system provides a unique background which takes part in the prediction of vehicles. We learn the similar backgrounds and address the learning of background similarity as a camera re-id problem. The proposed OSN is trained with cameras as their labels to classify between multiple cameras (backgrounds) and we call it OSN-C baseline. The baseline trained with the original vehicle labels is referred as OSN-F. With the OSN-C, we compute the camera features and provide a classification between the available cameras (backgrounds). However, instead of outputting a regression result or a classified direction, the camera re-id outputs an embedding that can be used to compute the distance matrix. In the proposed CBD training stage, for an x image, we compute the image features $f_F(x)$ by OSN-F and camera features (background features) $f_C(x)$ from OSN-C. And these features are merged linearly to get the final feature vector $f(x)$.

$$f(x) = f_F(x) - \alpha f_C(x) \tag{4.3}$$

where $f(x)$ is the final representations of image x with the distilled background which are used in the classifiers to predict the identity of the vehicle. This proposed CBD training mechanism significantly enhance the performance and is shown in Fig. 4.3.

In the testing stage, the distance matrix between the two images x_i and x_j from camera i and j is computed as:

$$D(x_i, x_j) = \frac{\{f_F(x_i) - \alpha f_C(x_i)\} \cdot \{f_F(x_j) - \alpha f_C(x_j)\}}{\|f_F(x_i) - \alpha f_C(x_i)\| \, \|f_C(x_i) \cdot f_F(x_j\|} \tag{4.4}$$

where $D(x_i, x_j)$ is the distance matrix between any query image x_i and and a gallery image x_j. The distance matrix is the difference or similarity between two vehicles without any background contribution. The proposed oriented splits tackle noise as background as well hence improve the performance as shown in the experiments section.

4.1.2 Experimental Results

In this section we provide the implementation details, followed by the datasets descriptions, then by quantitative and qualitative results and ablation studies.

4.1.2.1 Datasets

We perform our experiments on two vehicle re-id benchmark datasets the VeRi-776 [68] and the VRIC [52].

VeRi-776 dataset is a comprehensive vehicle re-id carrying footages captured from 20 surveillance cameras installed along several roads in $1.0km^2$ area. The dataset represents real-world scenarios like two-lane roads, four-lane roads and cross roads. Approximately 50000 images of 776 different vehicles are obtained after annotaion. The training set consists of 37781 images of 576 vehicles and testing set has 11579 iamges of 200 vehicles.

VRIC is a recent dataset composed of 60430 images of 5656 identities collected from by 60 cameras. VRIC have been collected in unconstrained settings thus have significant appearance variations. There are 54808 images of 2811 identities in training set. Remaining 5622 images of 2811 identities are in test set.

4.1.2.2 Implementation Details

We implemented the proposed network using Pytorch. The baselines are built upon a IBN-a-ResNet50 network pretrained on ImageNet. We modified the network with all the adjustments mentioned in section 4.1.1.2. We optimized the network by using Adam optimizer with momentum 0.9. We trained the models for $500 \times$ (*number of identities*) iteration with the initial learning rate is $2e-4$ and is divided by 10 after half iterations. We used a batch size of 32 (having 8 identities with 4 images each). All the images are resized to 320×320 with random horizontal flipping and random erasing data augmentations. The dropout probability is set to 0.5 and the weight decay is $5e-4$.

Table 4.1: Results and Comparison of the proposed OSN and CBD approaches with the state-of-the-art vehicle re-id methods on VeRi-776 dataset. The dashed line is used to split the proposed results with the results of other methods. The top 1 and 2 results are in red and blue.

Methods	**Reference**	**VeRi-776**		
		Rank-1(%)	***Rank-5(%)***	***mAP(%)***
NuFACT [69]	TMM18	76.7	91.4	48.4
VAMI [150]	CVPR18	77.0	90.8	50.1
PROVID [69]	TMM18	81.5	95.1	53.4
AAVER [54]	ICCV19	88.9	94.7	61.1
VANet [23]	ICCV19	89.7	95.9	66.3
PAMTRI [107]	ICCV19	92.8	96.9	71.8
He et al. [37]	CVPR19	94.3	98.7	74.3
VCAM [14]	CVPR20	94.4	96.9	68.6
SPAN [15]	ECCV20	94.0	97.6	68.9
SAVER [55]	ECCV20	96.4	98.6	79.6
GLAMOR [105]	arXiv20	96.5	98.6	80.3
RiDiNet [75]	TII20	94.6	99.0	78.2
PGAN [132]	TIT20	96.5	-	79.3
VARID [61]	TIT20	96.0	99.2	79.2
KAE-Net RN50 [81]	WACV21	93.6	96.8	70.1
WCVL [125]	arXiv21	95.3	99.1	80.4
CPL [126]	arXiv21	96.0	-	80.6
AT [33]	PAA21	78.1	98.2	85.4
MCRL [50]	TIT21	96.1	99.4	81.1
OSN	-	96.7	98.4	83.8
OSN + CBD	-	97.9	99.0	84.4

4.1.2.3 Comparison with state-of-art methods

Table 4.1 shows the result of proposed method and its comparison with other state of the art methods on VeRi-776 dataset. Highest results are shown in red while the second highest in blue. Our proposed OSN with multiple oriented splits outperforms all other methods in terms of rank-1 and rank-5 accuracy. The reason for such an improvement is the splitting of feature maps across the whole image in 360 degrees. Each split is computed within 45 degrees and creates its local descriptor. All these local descriptors along with the global descriptors generate strong feature representations which significantly enhance the matching scores. The proposed method has highest rank-1 and rank-5 accuracy, while in terms of mean average precision (mAP) our methods produce the second highest results.

Table 4.2: Results and comparisons to the state-of-the-art vehicle re-id methods for the proposed OSN and CBD training strategy on VRIC dataset. The proposed results and other state of methods are separated by the dashed line. The top 1 and 2 results are shown in red and blue.

Methods	**Reference**	**VRIC**		
		Rank-1(%)	***Rank-5(%)***	***mAP(%)***
Siamese-Visual [95]	ICCV17	30.5	57.4	-
OIFE [116]	ICCV17	24.6	50.9	-
MSVF [52]	GCPR18	46.6	65.6	-
RiDiNet [75]	TII20	70.2	88.9	-
VARID [61]	TIT20	65.3	89.0	-
GLAMOR [105]	arXiv20	78.6	93.6	76.5
PGAN [132]	TIT20	78.0	93.2	84.8
WCVL [125]	arXiv21	76.2	92.6	-
CPL [126]	arXiv21	76.6	91.8	-
AT [33]	PAA21	78.8	93.4	85.4
MCRL [50]	TIT21	79.1	93.0	85.1
OSN	-	78.9	94.3	85.6
OSN + CBD	-	79.7	95.2	86.0

In table 4.2, we present the results on VRIC dataset with its comparison with other state of art methods. VRIC is a new and difficult dataset due to degradations present in the images along with multiple appearance variations. The OSN baseline without CBD strategy performs compareably with other state of the art methods. The Rank-5 accuracy and mean average precision

Figure 4.4: Visual results of the proposed method. We showed first 10 ranks and true matches are shown in green rectangle. There is only one true match because gallery set has only one image image of every vehicle in query. Right side of the dashed line consists of query images while left side has the rankings from gallery.

of the OSN baseline is highest while rank 1 is second highest with a very small difference. The proposed CBD training technique further improves the performance of the OSN baseline and produce state of the art performance. In all experiments, we reported our results while those of other methods are taken directly from the papers and the results of the proposed method are reported without any re-ranking.

4.1.2.4 Visual Results

The visual results of the proposed method are shown in Fig. 4.4 in which the top 10 ranks for each query image are presented in each row. The proposed method retrieves the true match at the first rank even if the query image is noisy due to the addition of channel attention modules in the OSN. Fourth query image has shown the significance of the diagonal splits as the secondary diagonal split captures the vehicle in its local descriptor and hence find the true match at rank 1 even in the presence of large portion of background in the query image. The last query image is a

part of any vehicle and the oriented splits of the proposed OSN has its true match at rank 2 in the retrieved rankings. The proposed splitting strategy learns the key and salient parts/regions of the vehicles in its local descriptors and helps to match images with those parts/regions. The visual results have shown that the proposed approach learns strong representations of the vehicles, which performs accurately in the presence of background and occlusions.

4.1.2.5 Ablation Study

In this section, we discuss the quantitative and visual effect of multiple components of the proposed OSN and CBD.

Table 4.3: Component analysis of the proposed network (OSN). We create add and remove several components to generate different designs. With the insertion of all components, the final OSN is built and its results are shown in **bold** text.

Network	**VeRi-776**		
Components	***R1(%)***	***R5(%)***	***mAP(%)***
baseline (IBN-a-ResNet)	82.9	90.6	61.3
OSN (HV)	93.1	95.8	77.8
OSN (HV) attentive	93.3	96.2	78.0
OSN (HVD)	96.3	-	83.2
OSN (HVD) attentive	96.7	98.4	83.8
OSN + CBD proposed	**97.9**	**99.0**	**84.4**

4.1.2.5.1 Component Analysis Table 4.3 shows the outcomes of various components and combinations of the proposed method on VeRi-776 dataset. In the first row, we recorded the results of the baseline (IBN-a-ResNet50) used in the proposed work. From the baseline, we create horizontal and vertical splits to divide the vehicles into two parts along each axis as shown in the upper two branches in Fig. 4.2 and named it OSN (HV). We placed the results achieved by OSN (HV) in the second row of the Table 4.3 which showed a large improvement as compared to the baseline. We appended a channel attention module to learn better representations from noisy and degraded data by following [133]. Unlike [133], we used channel attentions at the end of each residual block instead of each residual connection which reduces the computational cost

and with OSN (HV) attentive, we slightly improved the performance. OSN (HVD) is created with the addition of diagonal splits across both the primary and secondary diagonals. These diagonal splits capture multiple parts and regions across several orientations and introduce better local features in the final representations to create and strong descriptors for each vehicle. With OSN (HVD), we got 3.2% improvement in Rank-1 accuracy and 5.4% increment in mAP when compared to horizontal and vertical splits. Then we used channel attentions for all proposed splits to obtain the final OSN and recorded its performance in the second last row. Finally, we trained the proposed OSN with the proposed CBD training strategy and got state of the art performance on VeRi-776 dataset which is shown in bold text in Table 4.3.

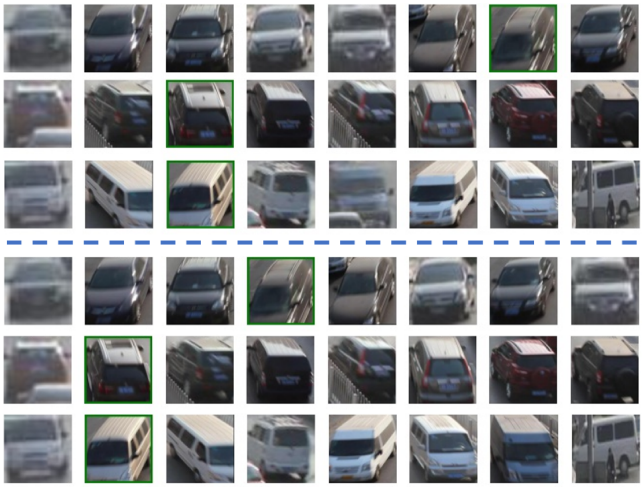

Figure 4.5: Visual results for the background distillation training strategy. The upper rows until dashed line shows the rankings from the proposed OSN and the lower rows represent the rankings when OSN is trained with CBD.

4.1.2.5.2 Effect of Background distillation We presented the visual results for the effect of camera based feature distillation (CBD) in Fig. 4.5. We used OSN baseline to find rankings of three query images from gallery images. First column in Fig 4.5 are the query images and columns $2-8$ are the 7 ranks computed by OSN with respect to each query. First three rows are

the results from OSN and their is no true match at rank 1. When we trained the OSN with the proposed CBD approach, we improved the ranks for the true match by 3 for the first query and by 1 for other two queries. All query images are noisy which shows that the proposed method counts noise as background.

In the section 4.1, we proposed a novel oriented splits network which splits the features into multiple splits to obtain the local features for each part and region of the vehicles. Key and salient parts/regions of the vehicles with many different orientations are captured through horizontal, vertical and diagonal splitting of the feature maps which enhance the discriminative ablility of the proposed OSN. A channel attention mechanism has been introduced to improve the learning from degraded data at the end of each residual block. To remove the effect of background features in the prediction of vehicles, we proposed camera based feature distillation training strategy which filter out the backgrounds which are induced by each camera in the surveillance area. The proposed network has the capability to learn robust feature representations for the vehicles and showed state of art performance on two vehicle re-id benchmarks.

5

Conclusion and Future Work

5.1 Conclusion

In this dissertation, we proposed techniques to improve the performance of person and vehicle re-identification tasks. We modified the general neural networks with the insertion of channel and self attention modules to capture long range similarities. We proposed three different architectures based on general convolutional networks combined with an attention mechanism. The introduction of attention modules brings the network to build the descriptor by computing non-local and long range relationship between the pixels unlike the general convolution operation which has a local receptive field. For better learning from degraded person re-id data, we adopted the channel attention mechanism at various positions in the network, which provide sharp feature maps in the presence of noise and blur. The proposed designs significantly improve the performance when compared to the general convolutional baseline.

To overcome the domain (style) and pose variations, we proposed two types of solutions. Firstly, we proposed a network based on conditional and cycle GANs to generate images in specified domains and pose in a single network. Usually, all the previous works can only gen-

erate one type of variations (pose or domain). We proposed a novel training mechanism in which a new images can be generated based on the conditional pose and domain. Secondly, we used cycle-GAN to translate images from one camera domain to all the other camera domains present in the dataset. Then, we proposed a network to utilize those generated images along with the original data to improve the performance. The proposed approach efficiently utilized the generated images having a number of artifacts in them and can not be treated as original samples.

To solve the problem occurred due to the existence of multiple resolutions within probe and gallery images, we designed a network which computes channels correlations in the feature maps to learn from low resolution degraded data. We also proposed a distillation based training technique which makes the network to learn resolution invariant features for better match between two different resolutions. We also proposed a novel more realistic scenario which has multiple resolutions in both probe and gallery images unlike the low resolution person re-id in which the probe image has LR probe images and HR images in gallery sets. With the proposed approach, we have significantly enhanced the performance on real degraded data set and performed comparably on synthetic datasets.

For vehicle re-identification, we proposed a network which captures the local and global part and merge them to produce the final representations. We designed the splitting of feature maps within the network to find the local features for the correlated parts. We merged these local features with the global features to produce a strong representation for a vehicle to reduce the effect of orientations, which make the features unable to match between different orientations. The proposed oriented split network marginally improves the performance of vehicle re-id when compared to previous methods. In addition, we proposed a camera-based distillation training strategy to remove the effect of background in the prediction of the vehicle. Such training scheme also considers noise as back ground and improves the final ranking as shown in the visual results for vehicle re-identification.

5.2 Limitations

The current methods for person re-identification discussed in this dissertation have several limitations as well. One of the drawbacks of such methods is that they perform well for a single dataset on which they are trained. They perform very poorly when trained on one dataset and used to test on another dataset. Such methods have a lack of generability. Another limitation is the person recognition as these methods made most of their predictions based on the appearance of the person. The general person re-id datasets have a large number of images of a person with the same appearances, which makes the trained networks prediction biased toward appearances. These methods usually fail to match the images of a person if both images have different clothing. The use of similar modality cameras is also one of the limitations of these methods. The cameras used in person re-id works well in the presence of light and unable to capture images in the darkness. For such purposes, thermal or infrared cameras are used to detect the persons and produce different modality images instead of RGB images. This cross-modality person re-id needs some special mechanism to merge different modalities images where general person re-id models fail.

5.3 Future Work

The proposed work can be improved in several ways to resolve the above limitations. Some research works are actively working on domain adaptation and semi supervised learning schemes to improve the performance of the networks when performing transfer learning. They used labelled data from the source dataset and data without labels from the target data set to train models. To improve training, few research efforts are assigning pseudo labels to the unlabelled data through trained networks and some are adopting unsupervised training on this unlabelled data. These works start to overcome some of the limitations but their results have a lot of room to improve. The future person re-id directions are improvements in the performance of such

methods. Recently, cloth changes data set for person re-id is also proposed and the existing methods perform very poorly on it. There is a need of modifying the current methods to adapt to such variations. Datasets with different modality camera are also proposed and one research line is currently working on that but they have a lot of room for improvements.

The generated images with the proposed method having domain and pose variations need improvements so that the match between query and gallery images can be done by converting them in same domain and pose. We can also create a virtual domain and pose to transfer all images into that domain and pose to have a unique image for each person. In case of cross resolution person re-id, dataset with real degradations can be used to train a model which translate images between different resolutions. These degradations can also be learned through a network to transfer the synthetic datasets into real degraded dataset which is more realistic case for cross resolution person re-identification.

Publications

1. **A Munir**, C Lyu, B Goossens, W Philips, C Micheloni, "Resolution based Feature Distillation for Cross Resolution PersonRe-Identification" Proceedings of The IEEE/CVF International Conference on Computer Vision Workshops (ICCV Workshop IWDSC 2021).

2. **A Munir**, N Martinel, C Micheloni, "Multi branch siamese network for person re identification." IEEE International Conference on Image Processing (ICIP 2020).

3. **A Munir**, N Martinel, C Micheloni, "Self and Channel Attention Network for Person Re-Identification." 25th International Conference on Pattern Recognition (ICPR 2020).

4. **A Munir**, C Micheloni, "Self attention based multi branch network for person re identification." 5th International Conference on Smart and Sustainable Technologies (SpliTech 2020).

5. **A Munir**, GL Foresti, C Micheloni, "Generating domain and pose variations between pair of cameras for person re-identification." Proceedings of the 13th International Conference on Distributed Smart Cameras (ICDSC 2019).

6. RM Umer, **A Munir**, C Micheloni, "A Deep Residual Star Generative Adversarial Network for multi-domain Image Super-Resolution." 6th International Conference on Smart and Sustainable Technologies (SpliTech 2021).

7. C Lyu, P Heyer Wollenberg, **A Munir**, L Platisa, C Micheloni, B Goossens, "Visible-thermal pedestrian detection via unsupervised transfer learning." International Conference

on Innovation in Artificial Intelligence 2021.

8. **A Munir**, N Martinel, C Micheloni, ”Oriented Splits Network to Distill Background for Vehicle Re-Identification.” IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS 2021) (Submitted).

9. **A Munir**, N Martinel, C Micheloni, ”Consistent Attentive Dual Branch Network for Person Re-Identification.” Multimedia Tools and Applications (Revision Submitted).

Bibliography

[1] M. Adil, S. Mamoon, A. Zakir, M. A. Manzoor, and Z. Lian. Multi scale-adaptive super-resolution person re-identification using gan. *IEEE Access*, 8, 2020.

[2] E. Ahmed, M. Jones, and T. K. Marks. An improved deep learning architecture for person re-identification. In *CVPR*, 2015.

[3] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In *ICML*, 2017.

[4] D. Avola, M. Cascio, L. Cinque, A. Fagioli, G. L. Foresti, and C. Massaroni. Master and rookie networks for person re-identification. In *CAIP*, 2019.

[5] Y. Bai, Y. Lou, F. Gao, S. Wang, Y. Wu, and L.-Y. Duan. Group-sensitive triplet embedding for vehicle reidentification. *IEEE Transactions on Multimedia*, 2018.

[6] L. Bazzani, M. Cristani, A. Perina, M. Farenzena, and V. Murino. Multiple-shot person re-identification by hpe signature. In *2010 20th International Conference on Pattern Recognition*, pages 1413–1416, 2010.

[7] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7291–7299, 2017.

[8] X. Chang, T. M. Hospedales, and T. Xiang. Multi-level factorisation net for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2109–2118, 2018.

[9] D. Chen, D. Xu, H. Li, N. Sebe, and X. Wang. Group consistent similarity learning via deep crf for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern*

recognition, pages 8649–8658, 2018.

[10] J. Chen, K. Li, Q. Deng, K. Li, and S. Y. Philip. Distributed deep learning model for intelligent video surveillance systems with edge computing. *IEEE Transactions on Industrial Informatics*, 2019.

[11] J. Chen, K. Li, Q. Deng, K. Li, and S. Y. Philip. Distributed deep learning model for intelligent video surveillance systems with edge computing. *IEEE Transactions on Industrial Informatics*, 2019.

[12] L. Chen, H. Yang, Q. Xu, and Z. Gao. Harmonious attention network for person re-identification via complementarity between groups and individuals. *Neurocomputing*, 2020.

[13] T. Chen, S. Ding, J. Xie, Y. Yuan, W. Chen, Y. Yang, Z. Ren, and Z. Wang. Abd-net: Attentive but diverse person re-identification. In *ICCV*, 2019.

[14] T.-S. Chen, M.-Y. Lee, C.-T. Liu, and S.-Y. Chien. Viewpoint-aware channel-wise attentive network for vehicle re-identification. In *CVPRW*, 2020.

[15] T.-S. Chen, C.-T. Liu, C.-W. Wu, and S.-Y. Chien. Orientation-aware vehicle re-identification with semantics-guided part attention network. In *ECCV*, 2020.

[16] W. Chen, X. Chen, J. Zhang, and K. Huang. Beyond triplet loss: a deep quadruplet network for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 403–412, 2017.

[17] Y. Chen, X. Zhu, and S. Gong. Person re-identification by deep learning multi-scale representations. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 2590–2600, 2017.

[18] Y. Chen, X. Zhu, and S. Gong. Person re-identification by deep learning multi-scale representations. In *ICCVW*, 2017.

[19] Y.-C. Chen, Y.-J. Li, X. Du, and Y.-C. F. Wang. Learning resolution-invariant deep representations for person re-identification. In *AAAI*, 2019.

[20] Z. Cheng, Q. Dong, S. Gong, and X. Zhu. Inter-task association critic for cross-resolution person re-identification. In *CVPR*, 2020.

[21] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *CVPR*, 2018.

[22] R. Chu, Y. Sun, Y. Li, Z. Liu, C. Zhang, and Y. Wei. Vehicle re-identification with viewpoint-aware metric learning. In *ICCV*, 2019.

[23] R. Chu, Y. Sun, Y. Li, Z. Liu, C. Zhang, and Y. Wei. Vehicle re-identification with viewpoint-aware metric learning. In *ICCV*, 2019.

[24] J. Dai, Y. Li, K. He, and J. Sun. R-fcn: Object detection via region-based fully convolutional networks. In *Advances in neural information processing systems*, pages 379–387, 2016.

[25] T. Dai, J. Cai, Y. Zhang, S.-T. Xia, and L. Zhang. Second-order attention network for single image super-resolution. In *CVPR*, 2019.

[26] A. Das, A. Chakraborty, and A. K. Roy-Chowdhury. Consistent re-identification in a camera network. In *European conference on computer vision*, pages 330–345, 2014.

[27] W. Deng, L. Zheng, Q. Ye, G. Kang, Y. Yang, and J. Jiao. Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. In *CVPR*, 2018.

[28] S. Ding, L. Lin, G. Wang, and H. Chao. Deep feature learning with relative distance comparison for person re-identification. *Pattern Recognition*, 48(10):2993–3003, 2015.

[29] Y. Ge, Z. Li, H. Zhao, G. Yin, S. Yi, X. Wang, et al. Fd-gan: Pose-guided feature distilling gan for robust person re-identification. In *Advances in Neural Information Processing Systems*, pages 1222–1233, 2018.

[30] Y. Ge, Z. Li, H. Zhao, G. Yin, S. Yi, X. Wang, and H. Li. Fd-gan: Pose-guided feature distilling gan for robust person re-identification. In *NeurIPS*, 2018.

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

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

[33] J. Gu, W. Jiang, H. Luo, and H. Yu. An efficient global representation constrained by angular triplet loss for vehicle re-identification. *Pattern Analysis and Applications*, 2021.

[34] Y. Guo and N.-M. Cheung. Efficient and deep person re-identification using multi-level similarity. In *CVPR*, 2018.

[35] K. Han, Y. Huang, Z. Chen, L. Wang, and T. Tan. Prediction and recovery for adaptive low-resolution person re-identification. In *ECCV*, 2020.

[36] B. He, J. Li, Y. Zhao, and Y. Tian. Part-regularized near-duplicate vehicle re-identification. In *CVPR*, 2019.

[37] B. He, J. Li, Y. Zhao, and Y. Tian. Part-regularized near-duplicate vehicle re-identification. In *CVPR*, 2019.

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

[39] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[40] Z. He, Y. Lei, S. Bai, and W. Wu. Multi-camera vehicle tracking with powerful visual features and spatial-temporal cue. In *CVPRW*, 2019.

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

[42] A. G. Howard. Some improvements on deep convolutional neural network based image classification. *arXiv preprint arXiv:1312.5402*, 2013.

[43] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. In *CVPR*, 2018.

[44] Y. Huang, J. Xu, Q. Wu, Z. Zheng, Z. Zhang, and J. Zhang. Multi-pseudo regularized label for generated data in person re-identification. *TIP*, 2018.

[45] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[46] N. Jiang, J. Liu, C. Sun, Y. Wang, Z. Zhou, and W. Wu. Orientation-guided similarity learning for person re-identification. In *ICPR*, 2018.

[47] J. Jiao, W.-S. Zheng, A. Wu, X. Zhu, and S. Gong. Deep low-resolution person re-identification. In *AAAI*, 2018.

[48] J. Jiao, W.-S. Zheng, A. Wu, X. Zhu, and S. Gong. Deep low-resolution person re-identification. In *AAAI*, 2018.

[49] X. Jin, C. Lan, W. Zeng, Z. Chen, and L. Zhang. Style normalization and restitution for generalizable person re-identification. In *CVPR*, 2020.

[50] Y. Jin, C. Li, Y. Li, P. Peng, and G. A. Giannopoulos. Model latent views with multi-center metric learning for vehicle re-identification. *IEEE Transactions on Intelligent Transportation Systems*, 2021.

[51] X.-Y. Jing, X. Zhu, F. Wu, X. You, Q. Liu, D. Yue, R. Hu, and B. Xu. Super-resolution person re-identification with semi-coupled low-rank discriminant dictionary learning. In *CVPR*, 2015.

[52] A. Kanacı, X. Zhu, and S. Gong. Vehicle re-identification in context. In *GCPR*, 2018.

[53] P. Khorramshahi, A. Kumar, N. Peri, S. S. Rambhatla, J.-C. Chen, and R. Chellappa. A dual-path model with adaptive attention for vehicle re-identification. In *ICCV*, 2019.

[54] P. Khorramshahi, A. Kumar, N. Peri, S. S. Rambhatla, J.-C. Chen, and R. Chellappa. A dual-path model with adaptive attention for vehicle re-identification. In *ICCV*, 2019.

[55] P. Khorramshahi, N. Peri, J.-c. Chen, and R. Chellappa. The devil is in the details: Self-supervised attention for vehicle re-identification. In *ECCV*, 2020.

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

[57] W. Li, R. Zhao, T. Xiao, and X. Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *CVPR*, 2014.

[58] W. Li, X. Zhu, and S. Gong. Harmonious attention network for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2285–2294, 2018.

[59] X. Li, A. Wu, and W.-S. Zheng. Adversarial open-world person re-identification. In *ECCV*, 2018.

[60] X. Li, W.-S. Zheng, X. Wang, T. Xiang, and S. Gong. Multi-scale learning for low-resolution person re-identification. In *ICCV*, 2015.

[61] Y. Li, K. Liu, Y. Jin, T. Wang, and W. Lin. Varid: Viewpoint-aware re-identification of vehicle based on triplet loss. *IEEE Transactions on Intelligent Transportation Systems*, 2020.

[62] Y.-J. Li, Y.-C. Chen, Y.-Y. Lin, X. Du, and Y.-C. F. Wang. Recover and identify: A generative dual model for cross-resolution person re-identification. In *ICCV*, 2019.

[63] S. Liao, Y. Hu, X. Zhu, and S. Z. Li. Person re-identification by local maximal occurrence representation and metric learning. In *CVPR*, 2015.

[64] G. Lisanti, I. Masi, A. D. Bagdanov, and A. Del Bimbo. Person re-identification by iterative re-weighted sparse ranking. *IEEE transactions on pattern analysis and machine intelligence*, 2014.

[65] J. Liu, B. Ni, Y. Yan, P. Zhou, S. Cheng, and J. Hu. Pose transferrable person re-identification. In *CVPR*, 2018.

[66] J. Liu, B. Ni, Y. Yan, P. Zhou, S. Cheng, and J. Hu. Pose transferrable person re-identification. In *CVPR*, 2018.

[67] X. Liu, W. Liu, H. Ma, and H. Fu. Large-scale vehicle re-identification in urban surveillance videos. In *ICME*, 2016.

[68] X. Liu, W. Liu, T. Mei, and H. Ma. A deep learning-based approach to progressive vehicle re-identification for urban surveillance. In *ECCV*, 2016.

[69] X. Liu, W. Liu, T. Mei, and H. Ma. Provid: Progressive and multimodal vehicle reidentification for large-scale urban surveillance. *IEEE Transactions on Multimedia*, 2017.

[70] X. Liu, H. Ma, H. Fu, and M. Zhou. Vehicle retrieval and trajectory inference in urban traffic surveillance scene. In *ICDSC*, 2014.

[71] X. Liu, H. Zhao, M. Tian, L. Sheng, J. Shao, S. Yi, J. Yan, and X. Wang. Hydraplus-net: Attentive deep features for pedestrian analysis. In *Proceedings of the IEEE international conference on computer vision*, pages 350–359, 2017.

[72] C. Luo, Y. Chen, N. Wang, and Z. Zhang. Spectral feature transformation for person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4976–4985, 2019.

[73] L. Ma, X. Jia, Q. Sun, B. Schiele, T. Tuytelaars, and L. Van Gool. Pose guided person image generation. In *NeurIPS*, 2017.

[74] S. Mao, S. Zhang, and M. Yang. Resolution-invariant person re-identification. In *IJCAI*, 2019.

[75] N. Martinel, M. Dunnhofer, R. Pucci, G. L. Foresti, and C. Micheloni. Lord of the rings: Hanoi pooling and self-knowledge distillation for fast and accurate vehicle re-identification. *IEEE Transactions on Industrial Informatics*, 2021.

[76] N. Martinel, G. Luca Foresti, and C. Micheloni. Aggregating deep pyramidal representations for person re-identification. In *CVPRW*, 2019.

[77] N. Martinel, G. Luca Foresti, and C. Micheloni. Aggregating deep pyramidal representations for person re-identification. In *CVPRW*, 2019.

[78] N. Martinel, C. Micheloni, and G. L. Foresti. Kernelized saliency-based person re-identification through multiple metric learning. *IEEE Transactions on Image Processing*, 2015.

[79] C. Micheloni, P. Remagnino, H.-L. Eng, and J. Geng. Intelligent monitoring of complex environments. *IEEE Intelligent Systems*, 2010.

[80] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018.

[81] O. Moskvyak, F. Maire, F. Dayoub, and M. Baktashmotlagh. Keypoint-aligned embeddings for image retrieval and re-identification. In *WACV*, 2021.

[82] A. Munir, G. L. Foresti, and C. Micheloni. Generating domain and pose variations between pair of cameras for person re-identification. In *ICDSC*, 2019.

[83] A. Munir, N. Martinel, and C. Micheloni. Multi branch siamese network for person re-identification. In *ICIP*, 2020.

[84] A. Munir, N. Martinel, and C. Micheloni. Self and channel attention network for person re-identification. In *ICPR*, 2020.

[85] A. Munir, N. Martinel, and C. Micheloni. Self and channel attention network for person re-identification. In *ICPR*, 2020.

[86] A. Munir and C. Micheloni. Self attention based multi branch network for person re-identification. In *SpliTech*, 2020.

[87] X. Pan, P. Luo, J. Shi, and X. Tang. Two at once: Enhancing learning and generalization capacities via ibn-net. In *ECCV*, 2018.

[88] S. Pedagadi, J. Orwell, S. Velastin, and B. Boghossian. Local fisher discriminant analysis for pedestrian re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3318–3325, 2013.

[89] 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 *ECCV*, 2018.

[90] 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.

[91] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *European Conference on Computer Vision*, pages 17–35, 2016.

[92] M. Saquib Sarfraz, A. Schumann, A. Eberle, and R. Stiefelhagen. A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 420–429, 2018.

[93] C. Sebastian, R. Imbriaco, E. Bondarev, et al. Dual embedding expansion for vehicle re-identification. In *CVPRW*, 2020.

[94] Y. Shen, H. Li, S. Yi, D. Chen, and X. Wang. Person re-identification with deep similarity-guided graph neural network. In *ECCV*, 2018.

[95] Y. Shen, T. Xiao, H. Li, S. Yi, and X. Wang. Learning deep neural networks for vehicle re-id with visual-spatio-temporal path proposals. In *ICCV*, 2017.

[96] Y. Shen, T. Xiao, H. Li, S. Yi, and X. Wang. End-to-end deep kronecker-product matching for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6886–6895, 2018.

[97] X. Shu, D. Yuan, Q. Liu, and J. Liu. Adaptive weight part-based convolutional network for person re-identification. *Multimedia Tools and Applications*, 79(31):23617–23632, 2020.

[98] J. Si, H. Zhang, C.-G. Li, J. Kuen, X. Kong, A. C. Kot, and G. Wang. Dual attention matching network for context-aware feature sequence based person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5363–5372, 2018.

[99] A. Siarohin, E. Sangineto, S. Lathuiliere, and N. Sebe. Deformable gans for pose-based human image generation. In *CVPR*, 2018.

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

[101] J. Sochor, A. Herout, and J. Havel. Boxcars: 3d boxes as cnn input for improved fine-grained vehicle recognition. In *CVPR*, 2016.

[102] 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.

[103] Y. Sun, L. Zheng, W. Deng, and S. Wang. Svdnet for pedestrian retrieval. In *ICCV*, 2017.

[104] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 480–496, 2018.

[105] A. Suprem and C. Pu. Looking glamorous: Vehicle re-id in heterogeneous cameras networks with global and local attention. *arXiv preprint arXiv:2002.02256*, 2020.

[106] Z. Tang, M. Naphade, S. Birchfield, J. Tremblay, W. Hodge, R. Kumar, S. Wang, and X. Yang. Pamtri: Pose-aware multi-task learning for vehicle re-identification using highly randomized synthetic data. In *ICCV*, 2019.

[107] Z. Tang, M. Naphade, S. Birchfield, J. Tremblay, W. Hodge, R. Kumar, S. Wang, and X. Yang. Pamtri: Pose-aware multi-task learning for vehicle re-identification using highly randomized synthetic data. In *ICCV*, 2019.

[108] R. R. Varior, B. Shuai, J. Lu, D. Xu, and G. Wang. A siamese long short-term memory architecture for human re-identification. In *ECCV*, 2016.

[109] C. Wang, L. Song, G. Wang, Q. Zhang, and X. Wang. Multi-scale multi-patch person re-identification with exclusivity regularized softmax. *Neurocomputing*, 382:64–70, 2020.

[110] C. Wang, Q. Zhang, C. Huang, W. Liu, and X. Wang. Mancs: A multi-task attentional network with curriculum sampling for person re-identification. In *ECCV*, 2018.

[111] D. Wang and S. Zhang. Unsupervised person re-identification via multi-label classification. In *CVPR*, 2020.

[112] G. Wang, Y. Yuan, X. Chen, J. Li, and X. Zhou. Learning discriminative features with multiple granularities for person re-identification. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 274–282, 2018.

[113] X. Wang, R. Girshick, A. Gupta, and K. He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018.

[114] Y. Wang, Z. Chen, F. Wu, and G. Wang. Person re-identification with cascaded pairwise convolutions. In *CVPR*, 2018.

[115] Z. Wang, R. Hu, Y. Yu, J. Jiang, C. Liang, and J. Wang. Scale-adaptive low-resolution person re-id via learning a discriminating surface. In *IJCAI*, 2016.

[116] Z. Wang, L. Tang, X. Liu, Z. Yao, S. Yi, J. Shao, J. Yan, S. Wang, H. Li, and X. Wang. Orientation invariant feature embedding and spatial temporal regularization for vehicle re-identification. In *ICCV*, 2017.

[117] Z. Wang, M. Ye, F. Yang, X. Bai, and S. Satoh. Cascaded sr-gan for scale-adaptive low resolution person re-identification. In *IJCAI*, 2018.

[118] L. Wei, S. Zhang, W. Gao, and Q. Tian. Person transfer gan to bridge domain gap for person re-identification. In *CVPR*, 2018.

[119] G. Wu, X. Zhu, and S. Gong. Person re-identification by ranking ensemble representations. In *ICIP*, 2019.

[120] Z. Wu, Y. Li, and R. J. Radke. Viewpoint invariant human re-identification in camera networks using pose priors and subject-discriminative features. *IEEE transactions on pattern analysis and machine intelligence*, 2014.

[121] J. Xiang, R. Lin, J. Hou, and W. Huang. Person re-identification based on feature fusion and triplet loss function. In *ICPR*, 2018.

[122] S. Xiang, Y. Fu, H. Chen, W. Ran, and T. Liu. Multi-level feature learning with attention for person re-identification. *Multimedia Tools and Applications*, 79(43):32079–32093, 2020.

[123] J. Xu, R. Zhao, F. Zhu, H. Wang, and W. Ouyang. Attention-aware compositional network for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2119–2128, 2018.

[124] S. Xuan and S. Zhang. Intra-inter camera similarity for unsupervised person re-identification. In *CVPR*, 2021.

[125] L. Yang, H. Liu, J. Zhou, L. Liu, L. Zhang, P. Wang, and Y. Zhang. Pluggable weakly-supervised cross-view learning for accurate vehicle re-identification. *arXiv preprint arXiv:2103.05376*, 2021.

[126] L. Yang, Y. Wang, L. Liu, P. Wang, L. Chi, Z. Yuan, C. Wang, and Y. Zhang. Center prediction loss for re-identification. *arXiv preprint arXiv:2104.14746*, 2021.

[127] R. Yu, Z. Dou, S. Bai, Z. Zhang, Y. Xu, and X. Bai. Hard-aware point-to-set deep metric for person re-identification. In *ECCV*, 2018.

[128] D. Zapletal and A. Herout. Vehicle re-identification for automatic video traffic surveillance. In *CVPRW*, 2016.

[129] Y. Zhai, X. Guo, Y. Lu, and H. Li. In defense of the classification loss for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

[130] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena. Self-attention generative adversarial networks. *arXiv preprint arXiv:1805.08318*, 2018.

[131] X. Zhang, H. Luo, X. Fan, W. Xiang, Y. Sun, Q. Xiao, W. Jiang, C. Zhang, and J. Sun. Alignedreid: Surpassing human-level performance in person re-identification. *arXiv preprint arXiv:1711.08184*, 2017.

[132] X. Zhang, R. Zhang, J. Cao, D. Gong, M. You, and C. Shen. Part-guided attention learning for vehicle instance retrieval. *IEEE Transactions on Intelligent Transportation Systems*, 2020.

[133] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu. Image super-resolution using very deep residual channel attention networks. In *ECCV*, 2018.

[134] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu. Image super-resolution using very deep residual channel attention networks. In *ECCV*, 2018.

[135] Y. Zhang, S. Liu, L. Qi, S. Coleman, D. Kerr, and W. Shi. Multi-level and multi-scale horizontal pooling network for person re-identification. *Multimedia Tools and Applications*, 79(39):28603–28619, 2020.

[136] H. Zhao, M. Tian, S. Sun, J. Shao, J. Yan, S. Yi, X. Wang, and X. Tang. Spindle net: Person re-identification with human body region guided feature decomposition and fusion. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1077–1085, 2017.

[137] L. Zhao, X. Li, Y. Zhuang, and J. Wang. Deeply-learned part-aligned representations for person re-identification. In *Proceedings of the IEEE international conference on computer vision*, pages 3219–3228, 2017.

[138] 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.

[139] M. Zheng, S. Karanam, Z. Wu, and R. J. Radke. Re-identification with consistent attentive siamese networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5735–5744, 2019.

[140] Z. Zheng, X. Yang, Z. Yu, L. Zheng, Y. Yang, and J. Kautz. Joint discriminative and generative learning for person re-identification. In *CVPR*, 2019.

[141] 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.

[142] Z. Zheng, L. Zheng, and Y. Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *ICCV*, 2017.

[143] 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.

[144] W. Zhong, L. Jiang, T. Zhang, J. Ji, and H. Xiong. A part-based attention network for person re-identification. *Multimedia Tools and Applications*, 79:22525–22549, 2020.

[145] Z. Zhong, L. Zheng, S. Li, and Y. Yang. Generalizing a person retrieval model hetero-and homogeneously. In *ECCV*, 2018.

[146] Z. Zhong, L. Zheng, Z. Luo, S. Li, and Y. Yang. Invariance matters: Exemplar memory for domain adaptive person re-identification. In *CVPR*, 2019.

[147] Z. Zhong, L. Zheng, Z. Zheng, S. Li, and Y. Yang. Camera style adaptation for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5157–5166, 2018.

[148] Z. Zhong, L. Zheng, Z. Zheng, S. Li, and Y. Yang. Camera style adaptation for person re-identification. In *CVPR*, 2018.

[149] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning deep features for discriminative localization. In *CVPR*, 2016.

[150] Y. Zhou and L. Shao. Aware attentive multi-view inference for vehicle re-identification. In *CVPR*, 2018.

[151] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017.

[152] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.

[153] Y. Zou, X. Yang, Z. Yu, B. V. Kumar, and J. Kautz. Joint disentangling and adaptation for cross-domain person re-identification. In *ECCV*, 2020.