Suspect Tracking through Person Re-ID

By

Muhammad Wasif Ijaz	01-132182-025
Ammara Humayun	01-132182-033

Supervised by
Engr. Ammar Ajmal

Bachelors of Computer Engineering

Bahria University Islamabad Campus
BSEAS H-11, Islamabad

2022

Suspect Tracking through Person Re-ID

By

Muhammad Wasif Ijaz	01-132182-025
Ammara Humayun	01-132182-033

Supervised by

Engr. Ammar Ajmal

Submitted to the Department of Computer Engineering in the partital fulfilment of the requirements for the degree of Bachelors in Computer Engineering.

Bachelors of Computer Engineering

Bahria University Islamabad Campus

BSEAS H-11, Islamabad

2022

Muhammad Wasif Ijaz & Ammara Humayun
01-132182-025 & 01-132182-033

UNDERTAKING

We, Muhammad Wasif Ijaz and Ammara Humayun, guarantee that the project "Suspect Tracking Through Person Re - Identification" is our own original work. There has been no further evaluation of the work. Other sources of information have been properly acknowledged and referred to.

Muhammad Wasif Ijaz
01-132182-025

Ammara Humayun
01-132182-033

DEDICATION

We would like to dedicate this project to our beloved parents and supervisor Engr.Ammar Ajmal who encourage us to choose this project and help us throughout the whole project and dissertation. My dissertation committee members have graciously donated their time and skills to help us improve our work. I like their willingness to help and their pleasant nature. Last but not least, we dedicate this thesis to our best friends who support us and delighted us when we exhausted and tired of the work.

ACKNOWLEDGEMENTS

We, Muhammad Wasif Ijaz and Ammara Humayun, are most thankful to the Almighty Allah. With blessing of him we are able to complete our final year project. We would like to convey our heartfelt gratitude to Mr. Ammar Ajmal, our supervisor, for his unwavering support, encouragement, and vast expertise. His guidance helped us in thesis and throughout the whole project. We could not have imagined such a good mentor who beside the project, encourages us to avail the better opportunities. We would also like to thanks to all the fellows who helped us in doing project. Our sincere appreciation to all the committee members who critically analyze our work and make us prepare to perform more excellently and effectively. Last but not least we are also thankful and grateful to our families who support us through every thick and thin problems. Their encouragement in tough situations is highly appreciated. They are the ultimate role models. Our sincere gratitude to all.

ABSTRACT

This thesis gives a complete framework for re identifying a person in a camera network in order to automate the surveillance system. Attention neural models are used in the current state-of-the-art solutions. We propose combination of different loss functions on top of a temporal attention-based neural network model and apply bag of trick over it to outperform current state of art person re-identification results on PRID2011 dataset. Our combined loss function is combination of Center Loss (CL) and Online Soft Mining Loss (OSM) summation with Attention Loss (AL). As the need for surveillance and camera networks has expanded in past few years, the use of video to re-identify people has become a critical task and has attracted a lot of interest. A typical video-based person Re-Id system consists of an image-level feature extractor (such as CNN), a temporal modelling method for aggregating temporal information, and a loss function. Although several temporal modelling methods were described, direct comparisons are challenging since the used feature extractor and the used loss function have significant effect on the final result. For suspect tracking through person Re-Id, we implement a complete temporal modelling plus bag of tricks strategy. We apply this strategy on PRID2011 as well as custom dataset.

Keywords: Person Re-Identification, Surveillance, Temporal Attention Model, PRID2011, Camera, Deep Learning, Computer Vision, Web Portal

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

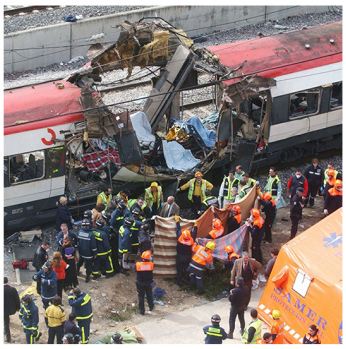
Chapter 1

INTRODUCTION

1.1 Background

Thanks to substantial technological advancements in the previous years and the simple availability to cameras and advanced media capacity, surveillance through cameras are now common and are everywhere in today's societies. An act of terrorism in New York (2001), Madrid (2004), and Pakistan (2012) (Figure 1.1) have probably contributed to this development, in order to ensure the safety of individuals and assets, but not mainly, against terrorism. The need to expand the limit of video surveillance of urban communities and organizations has been exhibited by the anticipation and suppression of wrongdoing and misconduct, the insurance of modern and authoritative structures, security air terminals, train stations, ports, street well being, individuals stream the executives, and different necessities.[1]

(a) New York 2001



(b) Madrid 2004

(c) Pakistan 2012

Figure 1.1: Terrorist attacks over the years

Video monitoring is broadly used and created in all pieces of life in current developments, because of huge technological enhancements over the latest years and the clear accessibility of cameras and automated media limit. The new worldwide security climate, with an act of terrorism in New York (2001), Pakistan(2012), Iraq(2013) and Madrid (2004)[1], and , has without a doubt added to this ascent, to guarantee the insurance of people and resources against psychological warfare activities, however not solely. The recent Boston (2013) incident shows the strength of video surveillance, since the two writers were quickly identified as a result of it. The need to expand the limit of video monitoring of urban areas, organizations, and other concerned partners has been exhibited by the need to forestall and stifle wrongdoing and misconduct, the assurance of modern and managerial structures, getting air terminals, train stations, and ports, street wellbeing, individuals stream the board, and different necessities. Because each video operator is responsible for a large number of video streams, there is a more prominent possibility missing significant occasions on the off chance that the guard who monitors the screen isn't focusing on the perfect camera at the ideal time since the person is observing numerous screens, or on the other hand assuming different camera views are on a screen and introductions the saw scene of all of them occasionally. Moreover, as indicated by current clinical examinations, a camera operator administrator loses 90% of his concentration and vigilance following 20 minutes of zeroing in on screens[1]. More often than not, no particular occasion happens, in this way sooner or later of watching pictures and seeing nothing, watchfulness exhaustion, and diminished concentration much of the time lead to disappointment in distinguishing critical occasions. As of late, re-identification of individuals in camera networks has turned into a critical issue. With the developing number of cameras in both confined and wide districts, it's more imperative than any other time to have a worldwide image of what's happening in a particular site that is being checked by a few cameras. Information from a few cameras ought to never again be seen as discrete snippets of data, yet rather all in all. A few suspicious activities or occasions must be reasoned through long haul following of the person in which we are interested over the different cameras (for example, an individual who abandon their baggage in a station, mall or an air terminal and remains nearby it's anything but a suspicious individual).[1] If, then again, this individual leaves their gear and leaves the air terminal while being watched by various cameras, this activity ought to draw notice). The ability to follow or re identify a specific individual in a camera or to track down person concluded on recorded film is ending up being logically basic in a combi-

nation of usages, including security. Numerous different applications, for example, shopping center showcasing and sports measurements, require following individuals with various cameras to surmise the most regular shopping ways and in this manner perceive stores, or to compute a football player's voyaged distance and the quantity of passes and shots he has made. Without vigorous ways to deal with hold a similar character of a followed individual paying little mind to where the person is found or which camera is seeing the person in question, this worldwide thinking on expansive regions under camera networks can't be led consequently. Re-distinguishing proof is the method involved with keeping an individuals distinguish starting with one camera then onto the next. A developing number of studies have been led as of late, and this issue keeps on provoking scientists' interest[1]. A writing survey of a few group Re-Id techniques is presented in the accompanying segments.
We can separate two essential classifications of approaches in view of the sort of data utilized for Re-Id: biometric approaches and appearance-based approaches.

1.2 Problem Statement

To track a suspect we need to look in hundred or thousand hours of recorded videos taken from large number of different cameras.It requires a large dedicated manpower and takes a lot of time.Also results in high financial cost, unavailability of competent staff and decreased concentration.
So, how an autonomous system defines an individual of interest, monitor a person across multiple cameras, and establish person's location in a massive volume of recorded videos?

Figure 1.2: Monitoring too many screens is impossible for a human

1.3 Proposed System

Person Re-Id tackles the query of locating a specific an individual in many images or videos, sometimes taken with various cameras in various locations. Person re-identification is a technique to re identify an individual of interest from numerous, across various cameras. Deep Neural Networks are advancing, and there is an increasing need for intelligent video monitoring, the computer vision community has become much more interested in this subject. It has gotten more attention in recent years as a result of rising public safety demands and quickly expanding surveillance camera networks. It has an extensive variety of Practical uses; for instance, in an enormous scene, it can save a lot of people and material resources. However because to various uncontrollable difficult environment aspects such as time-varying light conditions, human position changes and partial occlusion, it remain a challenge. Figure 1.3 [2] shows person re-identification system where, (a) Query Images, (b) Gallery Images.

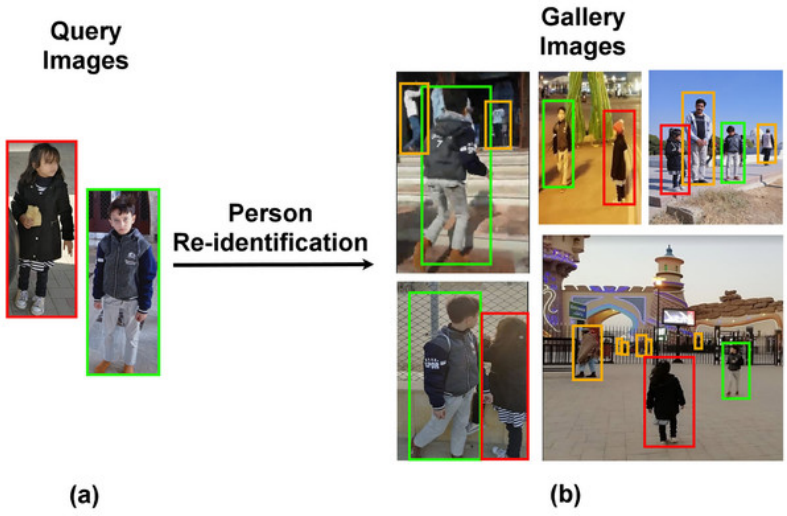


Figure 1.3: A smart Person Re-Identification System

1.4 Aim and Objective

Re – identification means to determine whether an individual of-interest has showed up in one more area at an alternate time caught by an alternate camera, or even a similar camera at an alternate time moment caught by a similar camera. To illustrate a person's inquiry, an image, a video clip, or perhaps a text description could be used. The objective of the project is to use technology to allow assistance for surveillance and bring automation to security systems to efficiently and accurately

identify events.

- Aim to identify a person from multiple camera views.
- To improve the time efficiency.
- To avoid human errors due to tiredness.

1.5 Project Scope

By obtaining sufficient hardware and making algorithms efficient enough to make Person Re-Identification Real Time by overcoming limitations.

- Re identify an individual from numerous camera views.
- A log of a track person will be generated.
- End product is a web page.

1.6 Tools and Techniques

Tools that are being used in our project are discussed below:

1.6.1 Python

A notable and undeniably valuable level programming language for a large number of uses. It was made by Guido van Rossum in 1991, and the Python Software Foundation keeps on creating it. Its sentence structure was made in light of code lucidity, assisting designers with conveying their considerations in less lines of code. Python is a scripting language that permits you to work with structures all the more quickly and successfully.[3] It is employed for the following purposes:

- Programming improvement
- Web improvement
- Framework prearranging
- Mathematics

- It has a key language structure that is like that of English.
- Python's syntax engages engineers to make program in less lines than other programming vernaculars.

- It is an interpreter language, and that suggests that code can be run when it is made. Likewise, prototyping ought to be conceivable reasonably quickly.
- Python can be moved toward in three ways: procedural, object-situated, and utilitarian.

1.6.2 Pycharm

It is an IDE that consolidates many key instruments for Python planners, which are solidly organized to lay out a magnificent environment for valuable web, data science and python improvement. PyCharm provides its customers and developers with some of the top features in the following areas:

- Inspection and completion of the code
- Support for advanced debugging for web programming and frameworks like Django and Flask
- An IDE that is both customizable and cross-platform
- Developers' Integrated Tools
- Refactoring
- Debugging and testing in one package
- Development from afar

1.6.3 Latex

Leslie Lamport, an American PC researcher, planned Latex as an augmentation to the tex typesetting framework in 1985. Plastic was expected to make it simpler for tex clients to create universally useful books and articles. Since Latex is a subsidiary of the tex typesetting framework, it can typeset specialized papers containing

complex numerical estimations. Plastic became well known among researchers and specialists due to this capability.Latex is a report planning framework for top notch typesetting. It is articulated «Lah-tech» or «Lay-tech». It's generally regularly utilized for medium-to-huge specialized or logical works, yet it tends to be utilized for almost any sort of distribution. It's vital to take note of that Latex isn't a word processor. All things being equal, Latex urges creators to zero in on accomplishing the appropriate substance rather from stressing over the presence of their works.

1.6.4 GoogleColab

As far as AI research, Google is genuinely dynamic. Google went through years creating TensorFlow, an AI system, and Colaboratory, an improvement stage. TensorFlow is presently open-source, and Google has made Colaboratory allowed to use starting around 2017new name for colaboratory is Google Colab The utilization of GPU is one more captivating part that Google provides for fashioners. Colab is a free application that maintains GPU. Its product could turn into a norm in scholastics for showing AI and information science in the event that it is made unreservedly accessible to people in general. It could likewise have the drawn out objective of laying out a client base for Google Cloud APIs, which are sold on a for each utilization premise. No matter what the causes, the send off of Colab has made AI application learning and advancement more straightforward.

1.6.5 Django

Django is an undeniable level Python web structure for rapidly making protected and pragmatic regions. Django is a web system made by gifted developers that deals with a ton of the monotonous work so you can zero in on further developing your application as opposed to becoming exhausted. It's free and open source, with a dynamic and dynamic local area, documentation, and a blend of free and business support choices.

- The information you want to display, which is commonly data from a database.
- A request handler that, in response to the user's request, returns the appropriate template and content.
- A text file (similar to an HTML file) that contains the layout of the web page, as well as logic on how to show the data.

1.6.6 GitHub

GitHub is an internet based association point that uses Git, an open source interpretation control system that allows a couple of clients to carry out simultaneous enhancements to site pages. GitHub urges groups to team up to create and change their site content since it takes into account ongoing coordinated effort, as Carpenter brings up. GitHub permits various engineers to chip away at a similar undertaking simultaneously, which brings down the opportunity of copy or clashing work and rates up creation. Designers can utilize GitHub to all the while compose code,

track changes, and think of new answers for issues that might emerge during the site improvement process. Craftsman outlines how non-designers might utilize it to make, change, and update site content in her instructional exercise.

Chapter 2

LITERATURE REVIEW

This segment analyzes the connected work, which incorporates probably the most generally involved loss functions in metric learning, deep learning for re-identification of people, picture based individual re-identification, video-based individual re-identification, and the attention system used in re-identification.

2.1 Deep Learning

Deep learning can be viewed as a subset of machine learning. A field is centered around PC calculations learning and creating all alone. Deep learning utilizes artificial neural networks, which should copy how people think and learn, rather than AI, which utilizes less difficult standards. As of not long ago, the intricacy of brain networks was compelled by computational limit. Bigger, all the more impressive brain networks are currently potential on account of advances in Big Data examination, permitting PCs to screen, learn, and respond to confounded occasions quicker than individuals. Picture arrangement, language interpretation, and discourse acknowledgment have all profited from profound learning. It can handle any example acknowledgment issue without the requirement for human mediation.

2.1.1 Working of Deep Learning

Similar to how the human mind is made up of neurons, brain networks are made up of layers of nodes. Nodes from one layer are connected to nodes from another layer. The more tiers in an organisation, the more complicated it becomes. A single neuron in the human brain receives several driving cues from other neurons. In an artificial neural network, signals pass between nodes and assign burdens to them. The nodes beneath a node with a higher weight will have a big impact. In the last layer, the weighted data sources are pooled to get a result. Deep learning frameworks require powerful hardware since they handle a large quantity of data and perform a variety of difficult numerical calculations. Even with such sophisticated invention, getting the hang of preparing computations can take weeks.

2.2 Re-identification with Deep Learning

Since the initial attempts [4] in 2014, re-id models based on DL have become increasingly popular. With the colossal progress of notable organizations like ResNet [5] and DenseNet [6], the ongoing methodology is to utilize them as base organizations and influence pre-prepared loads for re-id alterations. A few contemporaneous examinations have taken a gander at different re-id misfortunes, the least difficult of which is the grouping (curtailed cls) misfortune [7, 8]. The positioning misfortune, which utilizes trios [9, 10] or even quadruplets [11] to implement constraint between intra-individual and between individual examples, and the Siamese check misfortune, which predicts whether two photographs are from a similar individual [12], are two others.

All of these investigations [13, 14] found that joining at least two preparation misfortunes further developed execution, particularly DarkRank [15]. Person on foot arrangement [16], present assessment [17], quality learning [18], and consideration component [19] are a couple of instances of upgrading strategies. The apparent idea of joining body part models was regularly affirmed in [20, 21]. Information expansion was viewed as worthwhile in preparing with the utilization of GAN [22] and arbitrary deletion [23], while trimming based test-time increase was additionally observed to be valuable in [9]. Multi-stage preparing [24], multi-scale learning [25], and multi-model common learning [26] all delivered essentially prevalent outcomes.

2.3 Re-Identification of a Person Using Images

Most recent work on picture-based individual re-ID upgrades execution in two ways: picture spatial demonstrating and metric learning loss function work. [17] and [27] utilized human joints to parse the picture and wire the spatial information toward spatial element demonstrating. For managing the body part misalignment issue, Zhao et al. [16] introduced a section adjusted portrayal. As far as misfortune works, the character softmax cross entropy misfortune capacity and pivot misfortune in a Siamese organization are usually used. [9] presented a changed trio misfortune work that chooses the hardest positive and negative for each anchor test to gain proficiency with a viable measurement installing, and they got best in class execution.

2.4 Re-Identification of a Person Using video

[28]; [29] [8] have inspected video-based individual re-ID as an expansion of picture based individual re-ID. McLaughlin et al. [28] utilize the Recurrnet Neural Network (RNN) to move the message of each casing recovered from the Convolution Network

(CNN). For a more solid ID, [29] focus on advancing long-range movement setting qualities from successive edges.

2.5 Person Re-Identification Attention Models

Since [30] proposed the attention system, it has been utilized in a few re-ID examinations ([31], [19], [32]). In ([32]), [32] portrays a technique for consequently assessing the quality score of each frame and limiting the effect of noisy sample. The joint Spatial and Temporal Attention Pooling Network, which can remove discriminative casings from examining and display films and ascertain transient consideration loads for one succession in view of the highlights of the others, is introduced by Xu. [**xu2017**]. [33] joins various spatial consideration models with the worldly consideration model to assemble inactive portrayals of unmistakable body portions of every individual.

2.6 Temporal Model

Past work on video-based individual re - id temporal modeling algorithms has been partitioned into two classifications: RNN-based and temporal temporal based. [34] utilized a RNN to encode succession highlights, with the last secret state filling in as video portrayal. [28] pushed utilizing a RNN to demonstrate worldly data among outlines, and [28] supported utilizing a RNN to show fleeting data between outlines. [32] pushed using a Quality Aware Network (QAN), which is basically a consideration weighted normal, to total fleeting elements; [31] suggested encoding the video with fleeting RNN and consideration. [9] likewise utilized a trio misfortune work and an essential transient pooling procedure. In spite of the way that broad trials on the previously mentioned strategies have been accounted for, it is challenging to straightforwardly look at the impact of worldly demonstrating techniques since they utilized different picture level component extractors and misfortune works; these distinctions can altogether affect execution. [28], for instance, utilized a 3-layer CNN to encode the photos; [34] utilized hand-created highlights; and QAN [32] removed VGG [35] highlights as picture portrayals.

2.7 Sample Mining in Deep Metric Learning

Deep metric learning has been broadly explored in an variety of visual works, with promising outcomes [36]. Deep metric learning, in contrast to standard metric learning, which relies on hard-crafted features [37], gains include embeddings straightforwardly from information focuses utilizing deep convolutional neural net-

works (CNN). In spite of the fact that we can make a colossal number of picture matches for deep metric learning, numerous trifling matches will contribute zero to the loss and slope once the model has arrived at a decent degree of execution. During training, it is normal to search for nontrivial matches. To get quicker convergence and better execution, it makes sense to mine nontrivial pairs during training. As a result, sample mining has gotten a ton of consideration, and a various strategies have been created [38]. Mining tough negatives is used in [39] for mining negatives. In this paper, mining semi-difficult negatives is investigated [38]. Mining tough positives is covered in mining positives [36]. In this article, we look at how to find semi-difficult positives [38]. To become familiar with a drawn out complex as opposed to a contracted hypersphere, nearby up-sides mining (i.e., closer up-sides) is offered [38]. They generally mine a subset of matches by relegating a twofold score to each match, indicating whether it should be dropped or kept. Instead, our proposed OSM uses continuous score soft mining, which assigns varying weights to different pairs. When challenging pair mining is used, samples with a higher training loss are selected. As a result, outliers are frequently mined, and their significant losses cause model training to be disrupted.To overcome this, outliers are removed via semi-difficult negatives mining [38].

2.8 Online Soft Mining

For the positive set, Online Soft Positives Mining (OSPM) is used, and for the negative set, Online Soft Negatives Mining (OSNM) is used.

2.8.1 Online Soft Negative Mining

We really want to drive different matches in the negative set N to the side by an edge roused by the standard contrastive misfortune. To dismiss a tremendous piece of piddling matches that don't add to learning, we consign higher OSM scores to negative matches whose distance is inside this edge, similar to past hard irksome negatives mining strategies using twofold mining score ([40], [41]). Negative couples with distances more unmistakable than the pre-described edge have their OSM scores set to 0 since they don't add to the misfortune or incline. The OSM score s_{ij}^{-} of each regrettable pair (x_i, x_j) is determined just by the edge distance for ease:

$$s_i j^{-} = max(0, \alpha - d_{ij}) \tag{2.1}$$

2.8.2 Online Soft Positive Mining

OSPM's motivation is to deliver OSM scores for pairings in the positive set. The interclass distance might be negligible in numerous visual errands, like fine-grained visual categorization, however the intraclass distance might be significant. OSPM grants higher OSM scores to neighborhood up-sides because of complex realizing, where nearby up-sides are chosen out for learning broadened manifolds. This is on the grounds that contracted hyperspheres are learned, and up-sides with huge intraclass distances can't be gotten assuming they are dealt with in basically the same manner. We figure the Euclidean distance d_{ij} between their highlights after L2 standardization for each comparative pair in the positive set, for example $(x_i, x_j)\varepsilon P$. We simply convert the distance d_{ij} to OSM score utilizing a Gaussian capacity with mean=0 in light of the fact that we need to relegate more noteworthy mining scores to additional connected matches. In synopsis, for every positive mix (x_i, x_j), the OSM score s_{ij}^{+} is determined as follows:

$$s_{ij}^{+} = exp(\frac{-d_{ij}^{2}}{\sigma_{ij}^{2}}) \tag{2.2}$$

2.9 Class Aware Attention

According to [42], [43] outliers can easily influence sample mining, resulting to a sub optimal local minimum in the trained model. Outliers are also prone to the proposed OSM; for example, We pay more attention to negatives that are more difficult to interpret and may contain outliers. Faulty labels produce outliers have a negative impact on the performance of model [44]. Outliersalternatively, are often made up of mislabeled images and hence have a lower semantic relationship with their labels. As a result, we propose that noisy images be identified based on the semantic relationship between their labels. We call it Class-Aware Attention since it's guided by the class designation (CAA). Image x_i's CAA score reflects how semantically linked it is to its label y_i.

We figure the similarity between a picture's element vector and its related class setting vector to decide how semantically related an image is to its label. The dot product of two vectors can be used to determine their compatibility. After learning context vectors for all classes, we apply a classification branch. The class context vectors are the fully connected layer's trained parameters, which are ${c_k}_{k=1}^{C}$, where C is the number of training classes in the set and $c_k \varepsilon R^D$ is the class k context vector.

As a result, x_i is CAA score is calculated as:

$$a_i = \frac{exp(f_i^T c_{yi})}{\Sigma_{k-1}^{C} exp(f_i^T c_k)} \tag{2.3}$$

Softmax is used to normalize the semantic relation of an image across all classes. In citegoldberger2016training, the real label is forecasted using the output of softmax(classification confidence/likelihood) . In this paper, we use it to estimate the semantic relationship between a picture and its label, i.e. the label's accuracy degree.

2.10 Intraclass Variance

In fine-grained recognition, intraclass distance may be greater than interclass distance [38] and individual re - id [45], for instance, pictures from various classifications could have comparative tone and shape while pictures in similar Using the information's fundamental design helps various ways. Intraclass distance could be bigger than interclass distance in fine-grained acknowledgment [38] and individual reID [45], for instance, pictures from various classifications could have comparative tone and shape while pictures in similar Using the information's fundamental design helps various ways. To hold nearby component structures, neighborhood closeness mindful implanting was proposed in [46], [38] recommended that as opposed to treating all up-sides similarly, just neighborhood up-sides be utilized to get familiar with a drawn out complex. Along these lines as [38], our OSM gives more prominent load to up-sides that are more equivalent, fully intent on learning consistent manifolds. The primary distinction is that Cui et al [38] just mined a subset of neighborhood up-sides, though our OSM utilizes all up-sides and focuses on nearby up-sides.

2.11 ID cross-entropy loss

Each ID is treated as a class, and the cross-entropy loss is used to achieve multi-class classification:

$$L_{ce} = -\mathbb{E}_x \left[\log \frac{\exp(w_y^\top x + b_y)}{\sum_j^C \exp(w_j^\top x + b_j)} \right] \tag{2.4}$$

2.12 Center Prediction Loss

In center prediction loss paper [47], this paper have used center prediction loss for person ReID. A set of randomly collected same-class samples is used to calculate the suggested CPL. We get a set of feature vectors $x1, x2, ..., x_k$ from those intra-class samples after extracting features via a feature extractor $\phi(.)$. The CPL is described

as follows:

$$\zeta_{CPL} = min_{\theta} E_{x_1,...x_k}[\Sigma_{i=1}^{k} \parallel f(x_i;\theta) - c_i \parallel_2^2] \tag{2.5}$$

where k indicates the amount of intra-class samples, and $f(xi;\theta)$ indicates the centre predictor, which is a multi-layer perceptron (MLP) specified by $\theta, c_i = \frac{1}{k-1}\Sigma_(j \neq i)x_j$. The expectation is denoted by E. CPL is the optimal prediction error from a sample to the center of all other samples in the same class. CPL appears to be quite similar to the center loss, in that it requires x_i to be close to a learned center.

2.13 Center Loss

Center Loss learns a center for each class while penalizing distances between deep features and their linked class centers. [48]

$$L_{center} = \frac{1}{2}\mathbb{E}_x \left[\left\| x - c_y \right\|_2^2 \right] \tag{2.6}$$

2.14 Triplet Loss

Anchor point (x_a), positive point (x_p), and point loss [38] apply to a triplet of samples designated anchor point (x_a), positive point (x_p), and point loss (x_n). Its goal is to move an anchor point by a defined margin m nearer to the positive point (same personality) than to the negative point (different identity).

$$L_{tri} = \mathbb{E}_{(x_a,x_p,x_n)} \left[D(x_a,x_p) - D(x_a,x_n) + m \right]_+ \tag{2.7}$$

2.15 Circle Loss

A randomly picked batch of samples is subjected to Circle Loss [49]. If there are K same-class samples with K similarity scores and L other-class samples with L between similarity of class scores and in the feature space, a single sample x, circle loss is defined as:

$$L_{circle} = \mathbb{E}\left[\log\left(1 + \sum_{j=1}^{L} \exp(\gamma(s_n^j + m)) \sum_{i=1}^{K} \exp(\gamma(-s_p^i))\right)\right] \tag{2.8}$$

2.16 Lifted Structure Loss

Oh Song [50] attempts to bring one positive pair as near together as feasible while pushing all negative samples outside a margin of m.

$$L_{LS} = \mathbb{E}\left[\sum_{y_{ij}=1} \left(D_{ij} + \log\Big(\sum_{y_{ik}=0} \exp(m - D_{ik})\Big) + \log\Big(\sum_{y_{jl}=0} \exp(m - D_{jl})\Big)\right)_+\right] \tag{2.9}$$

2.17 Ranked List Loss (RLL)

Wang et al. [51] recommends using all cases in the gallery to create a set-based similarity structure. Unlike other algorithms, which aim to put positive pairs as close together in the embedding space as possible, the RLL just needs to bring positive examples closer together than a predefined threshold (boundary). The loss can be stated in the form of a pairwise constraint as:

$$L_{RLL} = \mathbb{E}\left[\left(1 - y_{ij}\right) \left[\alpha - d_{ij}\right]_+ + y_{ij} \left[d_{ij} - (\alpha - m)\right]_+ \right] \tag{2.10}$$

2.18 N-pair-mc Loss

[52] learns more discriminative features by using structure information between the data. It pushes a negative point away while bringing one good point closer to the anchor. During each parameter update, N-pair-mc loss considers the relationship between the query sample and negative samples of other classes, pushing the query to maintain a distance from all other classes, potentially speeding up the model's convergence.

2.19 Limitations of the existing losses

We discover the following drawbacks of previous approaches by studying the aforesaid loss functions: Regardless, all of the loss functions recorded above endeavor to decrease the distance between tests of a similar class. The majority of losses cause the same-class samples to condense into a single point. This, however, may make it more difficult to preserve the intrinsic intra-class sample variability that has been found to be favourable for learning transferable traits. Second, to regularise the intra-class sample distribution, some loss functions, such as RLL, use a fixed intra-class margin for all classes. A fixed margin, on the other hand, is unlikely to be appropriate for all classes. CPL, on the other hand, allows for more freedom in intra-class sample distribution and does not require that same-class samples have very tiny distances. It also does not use any pre-defined hyper-parameters and is adaptable to each class.

2.20 Training Tricks

To obtain a solid baseline, we changed the standard baseline using various training methods. Furthermore, we discovered that several research were unfairly compared to other cutting-edge methodologies. Rather than the approaches themselves, the increases were mostly due to training tactics. The study's training procedures, on the other hand, were downplayed, allowing readers to overlook them. The method's

effectiveness would be inflated as a result. We believe that while commenting on academic publications, reviewers should be aware of these techniques. Many local features are always incorporated in academic settings, or semantic information from pose estimation or segmentation approaches is used. Such procedures result in excessive consumption. Large features can slow down the retrieval process significantly. As a result, we intend to increase the ReID model's capabilities by employing some methods and relying solely on global features to attain high performance. The main purpose is:

- We looked at a lot of papers that were presented at prestigious conferences and discovered that the majority of them were built on shaky foundations.
- For researchers in academia, we intend to establish a firm baseline that will allow them to attain higher accuracy in person ReID.
- We intend to provide reviewers with some references as to what tricks will effect the ReID model's performance. Reviewers should consider these strategies while comparing the performance of different methodologies, according to us.
- We intend to share some useful tips for the industry on how to get better models without spending too much money.

Chapter 3

METHODOLOGY

In this segment, we present the general framework pipeline and point by point arrangements of our proposed technique for Person Identification.

3.1 Dataset

We focus on two datasets.

- PRID2011
- Custom

3.1.1 PRID2011

This dataset was created in a joint effort with the Austrian Institute of Technology in order to explore several approaches to human re-identification. Images from several person trajectories collected by two different, static surveillance cameras make up the dataset. The pictures from these cameras show a change in perspective as well as huge contrasts in lighting, foundation, and camera credits. Because images are retrieved from trajectories, each camera view has numerous possible positions per individual. 475 human directions were recorded from one view and 856 from the other, with 245 individuals showing up in the two perspectives. We separate out people who were substantially obstructed, people who had not more than five credible photographs in both camera views, and images that were corrupted due to tracking and annotation issues. As a result, the arrangement is as follows.PRID2011 385 people are visible in camera view A, 749 individuals are noticeable in camera view B. Individual 0001 of view A connects with individual 0001 of view B, etc. People 0201 to 0385 in view An and 0201 to 0749 in view B complete the exhibition set of the significant view. As a result, a typical evaluation entails scanning all humans in the other view for the 200 first persons of one camera view. This demonstrates that either the test set is drawn from view An and the display set is drawn from view B (A to B, as in our paper), or the other way around (B to A).PRID2011

A to B evaluation technique. The first 200 people who saw the video were used as a test group A. All 749 people in view are represented in the gallery B.

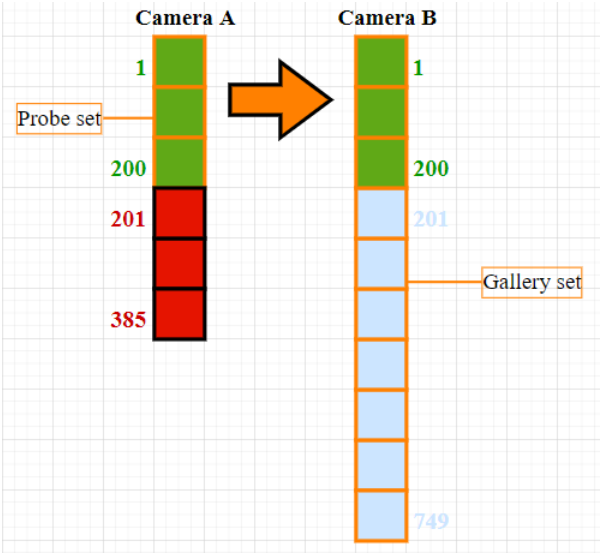


Figure 3.1: Evaluation procedure A to B

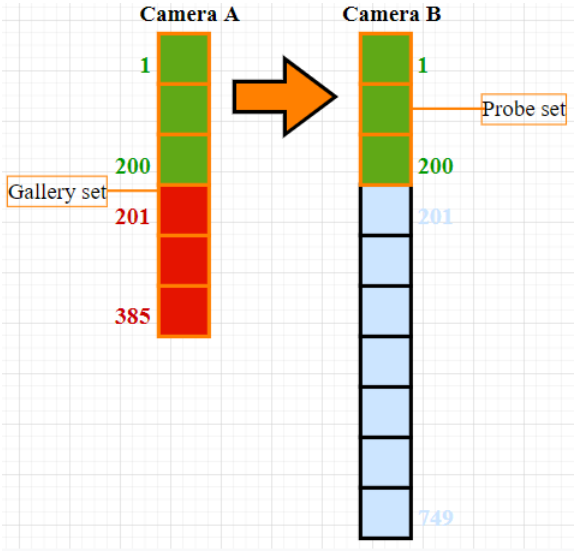


Figure 3.2: Evaluation procedure B to A

B to A evaluation technique. The first 200 people who saw the video were used as a test group B. Gallery set: view A's 385 people.

3.1.2 Single-Shot and Multi-Shot

Two variants of dataset are given, one for single-shot situations and the other for multi-shot situations. Every individual is addressed by numerous photographs in

the multi-shot adaptation (somewhere around five for each camera view). The genuine not entirely settled on the individual's strolling way, speed, and impediments. The single-shot variant incorporates only one (arbitrarily picked) picture per human direction, for example one picture from view An and one picture from view B.PRID2011

Figure 3.3: Image From PRID2011

3.1.3 Custom Dataset

We created a custom dataset by recording a video on an iPhone and extracting frames from the video using ssd. Two iPhones (apple iphone XS apple iphone 12 pro max) were used to collect the videos for this self-generated custom dataset. After collecting the videos we extracted frames from them using python scripts. As videos were shoot at 60 fps (frames per second), so we had to get key frames from

our extracted frames. To get key frames first we subtracted the 2 consecutive frames and then computed their mean and standard deviation. To compute the threshold we multiplied standard deviation (SD) with a constant 'a' and added the result with mean (M). Value of constant was selected a = 4 by hit and trial method. Then using threshold value we collected key frames. Afterwards we performed pre-processing techniques on our dataset to make it clean and usable for model. We resized the size of frames to 128x64 to feed it to our model. Then we used single shot detection (SSD) Model to recognize person from a frame and create bounding box around it. After creating bounding boxes we manually created the folders for all the different identities and placed their images accordingly. Same steps were repeated for both camera views and data structure of our self-generated custom dataset was exact replica of PRID2011, so we can have a fair evaluation between the results of both datasets.

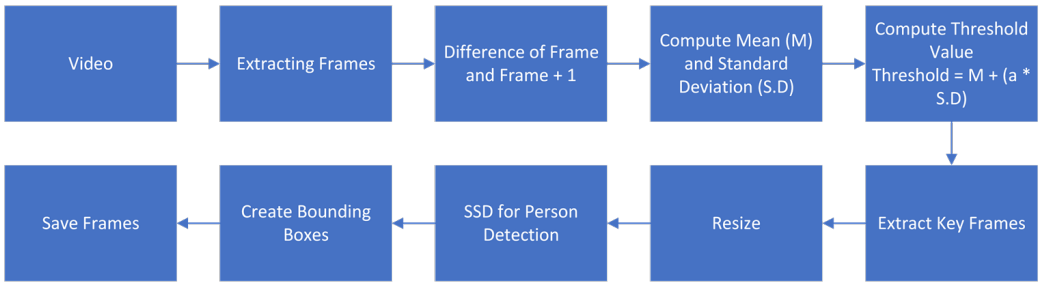


Figure 3.4: Custom Dataset Generation

3.1.4 SSD

SSD is a model that detects objects, but what precisely does that imply? Object detection and picture classification are often confused. In simple terms, image classification identifies the type of image, whereas object detection identifies the various objects in the image and uses bounding boxes to indicate where they are in the image. Let's get into SSD now that we've cleared things up.

Yolo (you just look once) and Single Shot multibox indicators are two famous single shot identifiers. Since it is a more productive and speedier methodology than the YOLO calculation, we will examine the SSD with a solitary shot multibox detector. Single Shot Detector The model's name uncovered most of the model's subtleties. Indeed, dissimilar to different models that navigate the picture at least a time or two to create a result recognition, the SSD model recognizes the item in a solitary ignore the info image.

As recently expressed, the SSD model perceives objects in a solitary pass, which saves a lot of time. Simultaneously, the SSD model seems to have extraordinary identification exactness. The SSD approach makes predictions at various scales from include guides of various scales and expressly isolates expectations by angle proportion to accomplish high recognition precision. Indeed, even on low-goal input photographs, these systems bring about clear start to finish preparing and great exactness.

3.1.5 Structure of SSD

Model consists of 2 parts

- SSD Head
- Backbone Model

3.1.6 SSD Head

The SSD head is comprised of stacked convolutional layers that are added to the highest point of the backbone model. This produces the bounding boxes over the objects as the result. The numerous items in the image are detected by these convolutional layers.

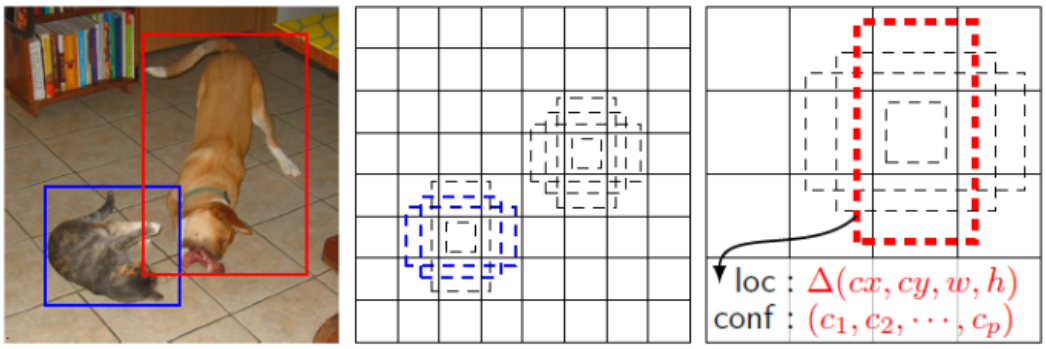
3.1.7 Backbone Model

The feature map extractor is the Backbone model, which is a typical pre-trained picture order network. The model's final picture categorization layers are eliminated here, leaving only the retrieved feature maps.

3.1.8 Working of SSD

The SSD is based on neural networks that generate a variety of bounding boxes of varied fixed sizes and score the presence of the object class instance inside those boxes, followed by a non-greatest suppression phase to obtain the final identifications. The SSD model works like this: each input image is partitioned into networks of varying sizes, with different classes and aspect proportions led at each grid. A score is assigned to each of these networks, indicating how well an object fits into that framework. Non-most extreme suppression is used to obtain the final recognition from the arrangement of covering place. The SSD model is based on this concept. In this case, a variety of matrix sizes are used to recognize objects of various sizes. For example, smaller matrices are used to distinguish a cat in the

image below, whereas larger frameworks are used to distinguish a dog, making the SSD more effective.



(a) Image with GT boxes (b) 8×8 feature map (c) 4×4 feature map

3.1.9 Multi-Scale Feature Maps for Detection

At the very end, the multi-scale feature maps are applied to the truncated backbone model. As the image is shrunk down, the multi-scale feature maps diminish in size, allowing detections at different scales. Each convolutional layer has its own set of feature layers.

3.1.10 Convolutional Predictors for Detection

A certain number of predictions are generated using the convolutional filters in each extra layer. These additional layers are shown at the top of the model in the diagram below. For a feature layer of size m x n with p channels, the lowest prediction parameter that achieves a decent detection is a 3 x 3 x p small kernel. A kernel like this gives the score for a category or a form offset relative to the default box coordinates.

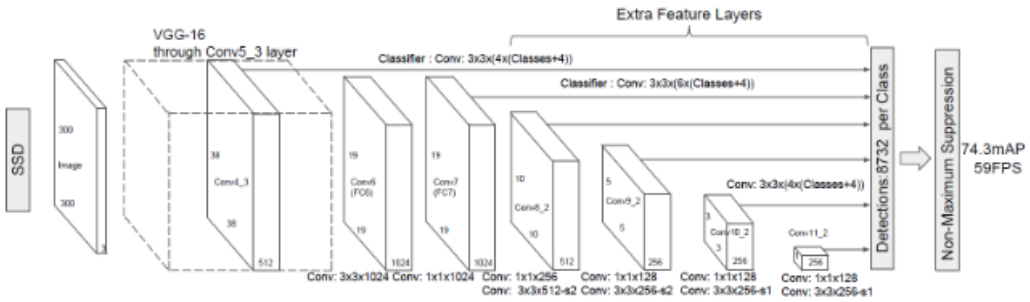


Figure 3.5: Convolutional Predictors for Detection

3.1.10.1 Aspect Ratios and The Use Of Default Boxes

Each feature map cell at the top of the network has default bounding boxes associated with it. The position of each box relative to its corresponding cell is determined by these default boxes, which tile the feature map in a convolutional fashion. C class scores and four offset relatives to the original default box shape are computed for each box out of k at a given location. For a feature map of dimension m x n, this results in a total of (c+4)k filters being applied around each position in the feature map, providing (c+4)kmn outputs. By employing different default box shapes in many feature maps, the model may efficiently discretize the set of possible output box forms.

Table 3.1: SSD Head

Type/Name	Grid Size	Kernel Size
Conv 6	19 x 19	3 x 3 x 1024
Conv 7	19 x 19	1 x 1 x 1024
Conv 8_2	10 x 10	1 x 1 x 256
		3 x 3 x 512 - s2
Conv 9_2	5 x 5	1 x 1 x 128
		3 x 3 x 256 - s2
Conv 10_2	3 x 3	1 x 1 x 128
		3 x 3 x 512 - s1
Conv 11_2	1 x 1	1 x 1 x 128
		3 x 3 x 512 - s1

3.1.11 SSD vs YOLO

YOLO model is a forerunner to the SSD model; it recognises images in a single pass as well, but uses two fully linked layers instead of the SSD's numerous convolutional layers. To anticipate offsets to default boxes of various scales and aspect ratios, as well as their related scores, the SSD model adds many feature layers to a basic network. The SSD averages 8732 detections each class, whereas the YOLO only averages 98 predictions per class. In the VOC2007 test, an SSD with a 300 x 300 inputs size exceeds a 448 x 448 YOLO counterpart in both accuracy and speed.

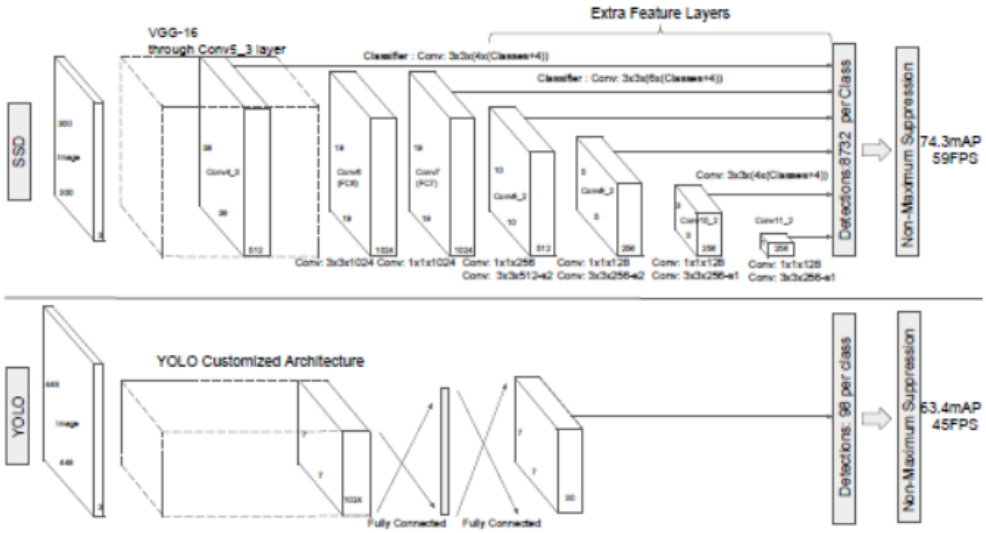


Figure 3.6: SSD vs YOLO

3.1.12 SSD Performance

Prior state-of-the-art detection approaches such as YOLO and Faster R-CNN have been demonstrated to underperform the SSD model. The multi-output layers at various resolutions have had a significant impact on performance; in fact, simply removing a few levels reduced accuracy by 12

System	VOC2007 Test Map	FPS (Titan X)	Number of Boxex	Input Resolution
Faster R-CNN (VGG16)	73.2	7	6000	1000 X 600
YOLO (customized)	63.4	45	98	448 X 448
SSD300* (VGG16)	77.2	46	8732	300 X 300
SSD512* (VGG16)	79.8	19	24564	512 X 512

Table 3.2: Performance Comparison

Dataset	Release Time	No. of Identities	No. of Cameras	No. of Images	Label Method	Crop Size	Multi-shot
PRID2011	2011	934	2	24541	Hand	128X64	✓
Custom		4	2	1059	Hand	128X64	✓

Table 3.3: Details of Datasets

3.1.13 Data Structure

the datasets should have some data structures. So, the data structure for both datasets must be same.

The data structure is given below:

PRID2011	**Custom Dataset**
prid2011/	customdataset/
prid(2011)/	custom(dataset)
singleshot/	singleshot/
Person0001/	Person0001/
.	.
.	.
.	.
Person0200/	Person0004/
multishot/	multishot/
Person0001/	Person0001/
.	.
.	.
.	.
Person0200/	Person0004/

3.2 Base Temporal Attention (Baseline)

We construct our model on Gao and Nevatia's (2018) [53] temporal attention model, which utilizes a pre-trained ResNet-50 on ImageNet to make highlights for each frame of a video clip, and an attention model to calculated a weighted amount of the elements across frames. In temporal attention model, weighted normal on the grouping of picture highlights in applied. The last convolutional layer in Resnet-50 has a tensor size of [w,h,2048], where w and h are subject to the size of the input picture. The attention generation network gets T consideration scores as information sources and results a progression of picture level features [T,w,h,2048]. Two types of attention networks used are:

- Spatial Conv Layer
- Fully Connected Layer

Figure 3.7 Temporal Attention in view of a picture level feature extractor (normally a 2D CNN). Two sorts of attention generation network are shown: "spatial conv +

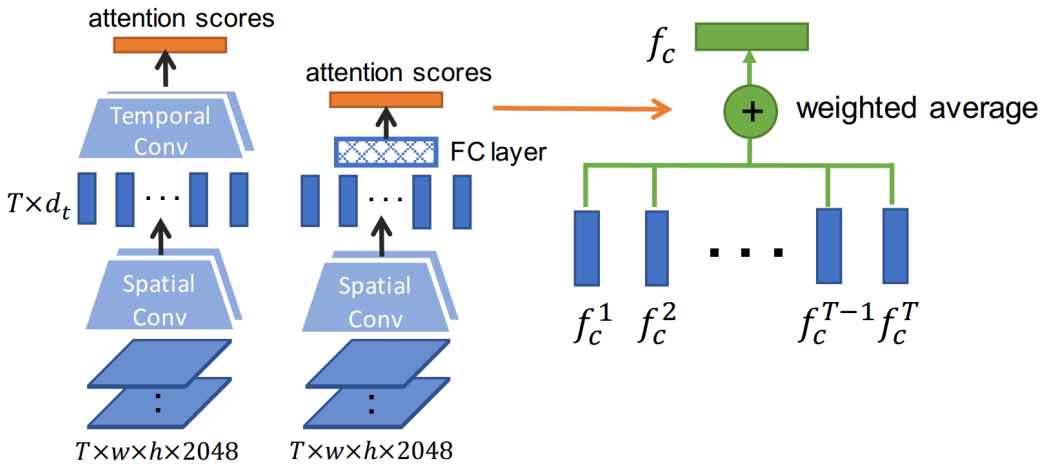


Figure 3.7: Temporal Attention Model

FC [53]" and "spatial conv + temporal conv[33].

3.3 Standard Baseline

As a common baseline, we employ a widely used open-source. ResNet50 is the foundation of the standard baseline. The pipeline includes the following processes during the training stage:

1. On ImageNet, we initialize the ResNet50 with pre-trained boundaries and alter the completely associated layer's aspect to N. The quantity of characters in the preparation dataset is meant by N.

2. We choose P IDs and K photos for each person at random to construct a training batch. Finally, the batch size is equal to B = PK. In this paper, P = 16 and K = 4 were used.

3. Each image is shrunk to 256 128 pixels, then the enlarged image is padded with zero values for the final 10 pixels. Then, at random, crop it into a 256 × 128 rectangular image.

4. With a probability of 0.5, each image is horizontally flipped. With a probability of 0.5, each image is horizontally flipped.

5. In [0, 1], each picture is decoded into raw 32-bit floating point pixel values. The RGB channels are then standardized by taking away 0.485, 0.456, and 0.406 from each and separating by 0.229, 0.224, and 0.225, correspondingly.

6. The model generates re - id features f and ID prediction logits p.

7. Re - id characteristics f [6] are used to determine triplet loss. ID prediction logits p are used to calculate cross entropy loss. The m value for the triplet loss margin is set to 0.3.

8. TThe Adam approach is used to optimize the model. The initial learning rate is 0.00035, and between the 40th and 70th epochs, it decreases by 0.1. There are 120 training epochs in total.

3.4 Bag of Tricks (BOT)

[54] offered a collection of methods to help a ResNet model perform better in image-based person recognition. re-identification, distance), re - id, including decreasing the step of the last layer (more extravagant element space), utilizing warm-up learning rate, irregular eradicating of patches inside frames, label smoothing, center loss in addition to trio loss, cosine-metric based trio loss. In comparison to non-normalized features, results show that batch normalised features produce a more robust model. Table 3.4 shows results of Luo et al BOT [54] using ResNet50 as backbone and achieved 94.5 (85.9) and 86.4 (76.4) on Market1501 and DukeMTMC-reID respectively. Figure 3.8 shows the pipeline of BOT. We are implementing BOT on PRID2011 dataset [55] to improve re-ID results.

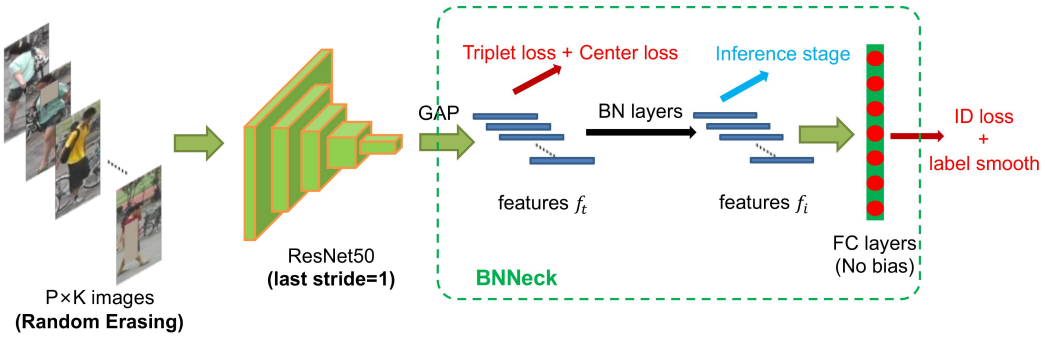


Figure 3.8: Pipeline of Bag of Tricks

3.4.1 Warmup Learning Rate

The learning rate of a ReID model has a significant impact on its performance.The standard baseline is trained using a high and constant learning rate. To bootstrap the network for greater performance, a warmup approach is used. At epoch t the

Table 3.4: Results of BOT on Market1501 and DukeMTMC-reID

Model	Market1501	DukeMTMC-reID
Standard baseline	87.7 (74.0)	79.7 (63.8)
+Warmup	88.7 (75.2)	80.6(65.1)
+Random erasing augmentation	91.3 (79.3)	81.5 (68.3)
+Label smoothing	91.4 (80.3)	82.4 (69.3)
+Last stride=1	92.0 (81.7)	82.6 (70.6)
+BNNeck	94.1 (85.7)	86.2 (75.9)
+Center loss	94.5 (85.9)	86.4 (76.4)
+Reranking	95.4 (94.2)	90.3 (89.1)

learning rate *lr(t)* is calculated in 3.1;

$$lr(t) = \begin{cases} 3.5 \times 10^{-5} \times \frac{t}{10} & \text{if } t \leq 10 \\ 3.5 \times 10^{-4} & \text{if } 10 < t \leq 40 \\ 3.5 \times 10^{-5} & \text{if } 40 < t \leq 70 \\ 3.5 \times 10^{-6} & \text{if } 70 < t \leq 120 \end{cases} \tag{3.1}$$

3.4.2 Random Erasing Augmentation

In person re - id, people in the pictures are in some cases occluded by different objects. [23] presented another method called Random Erasing Augmentation to address the occlusion issue and increment the speculation capacity of Re - id models Sampled example of random erasing augmentation is shown in figure 3.9.

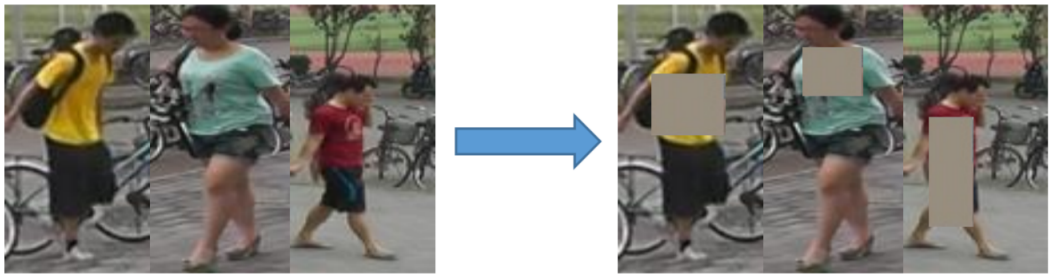


Figure 3.9: Random Erasing Augmentation

3.4.3 Label Smoothing

In person re - id, the ID Embedding (IDE) network [56] is a fundamental premise. The last layer of IDE is a completely associated layer with a hidden size equivalent to the quantity of individuals *N*, which delivers the ID prediction logits of pictures.

Since the testing set's individual IDs didn't present in the training set, Person re - id can be viewed as a single shot learning task. Accordingly, it's basic to keep the re - id model from overfitting training IDs. The technique for mark smoothing (LS) presented in [57] is ordinarily used to stay away from overfitting for grouping errands. It adjusts the *qi's* design to condition 3.2. To work on the exhibition of the model essentially while training set isn't enormous we use LS.

$$q_i = \begin{cases} 1 - \frac{N-1}{N} \in & \text{if } i = y \\ \frac{\in}{N} & \text{otherwise,} \end{cases} \tag{3.2}$$

3.4.4 Last Stride

The granularity of a feature is constantly enhanced by higher spatial resolution.Last stride is the backbone network's final spatial down sampling process. The last stride of ResNet50 is set to two. When fed a 256 x 128 image, the backbone of ResNet50 produces an 8 x 4 spatial dimension feature map. By changing the last stride from 2 to 1, we can get a feature map with a larger spatial size (16 x 8). This modification merely raises the cost of very light computations and does not require any additional training settings. Higher spatial resolution, on the other hand, results in significant improvements.

3.4.5 Center Loss

Center Loss, which all the while learns a middle for deep features of each class and penalize the distances between the deep features and their relating class focuses, compensates for the downsides of the trio loss. The center loss function is calculated as:

$$\mathcal{L}_C = \frac{1}{2} \sum_{j=1}^{B} \| f_{t_j} - c_{y_j} \|_2^2 \tag{3.3}$$

3.5 Attention and CL Loss (Attn-CL loss)

For training Re-ID errands, [58] presented Online Soft Mining (OSM) with Class-Aware Attention (CAA), a changed contrastive loss with attention regarding eliminate noisy frames, as an option in contrast to trio loss. Online Soft Mining contains Online Soft Positives Mining (OSPM) for the positive set and Online Soft Negatives Mining (OSNM) for the negative set. OSPM's motivation is to deliver OSM scores for the positive arrangement of matches and for divergent matches in the negative set N, the objective of OSNM is to drive each pair away by a margin α.

$$s_{ij}^{+} = \mathrm{e}\frac{-\left(\|f_i - f_j\|_2\right)^2}{\sigma_{OSM}^2} \tag{3.4}$$

$$s_{ij}^{-} = \max\left(0, \alpha - \|f_i - f_j\|_2\right) \tag{3.5}$$

Priyank Pathak [59] proposed CL Centers OSM loss for cropping out noisy frames, which employs the center vectors as the class label vector from the center loss representations because they have more variance than the originally proposed classifier weights. They [59] penalized model for assigning high attention scores to frames in which a patch was arbitrarily erased. Randomly erased frames are named as 1 in any case 0. The Attention loss joined with OSM loss and CL Centers is signified as Attn-CL loss..

$$AttentionLoss = \frac{1}{N}\sum_{i=1}^{N} label(i) * Attention_{score}(i) \tag{3.6}$$

3.6 Model

We chose ResNet50 as our base model. We studied multiple techniques and came up with the conclusion that ResNet50 is best option for Person Re-identification on PRID2011. State of art model comparison on PRID2011 is done by paperwithcode website [60], see comparison in figure 3.10. 3 of top 5 models uses ResNet50 as backbone, which clearly shows that ResNet50 is the best option for re-identification on PRID2011. Basic ResNet flow diagram is shown in figure 3.11 and our proposed model with BOT and Attn-Cl Loss is shown in figure 3.12

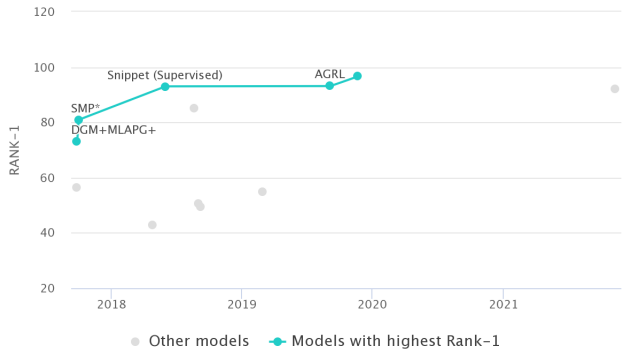


Figure 3.10: Comparison Graph of Models

The graph represents different models used for Person Re-ID on PRID2011 and the top 3 out of 5 higest ranked models uses ResNet50 as base model. Figure 3.11 shows layers of basic ResNet Model. It contains convolutional layer, normalization, activation function, and operation like summation and convolution.

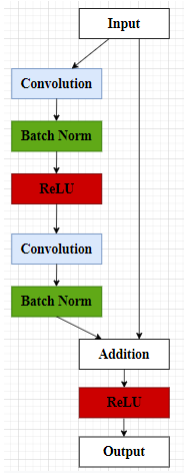


Figure 3.11: Basic ResNet Flowchart

Our proposed model is shown in Figure 3.12 contains ResNet50 as baseline model,

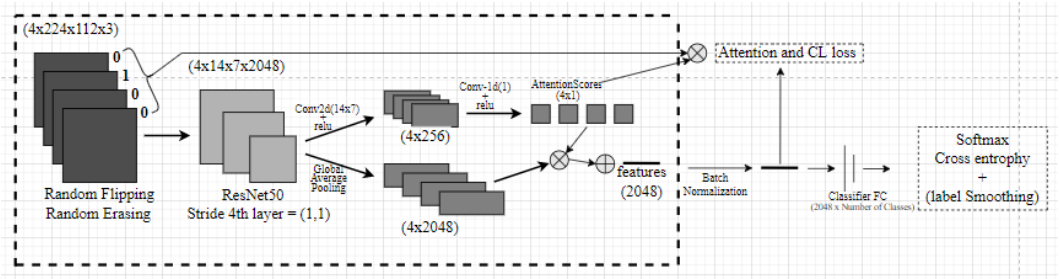


Figure 3.12: The proposed model architecture

uses BOT and Attn-CL Loss to compute results on PRID2011 Dataset.

The proposed system begins by randomly selecting four frames, then randomly flipped them because we want various angles of a person and randomly erased them because the model performs well in occlusion and we don't want our model to train on the same type of images. These frames are fed to ResNet50 to train our temporal attention model. Feature output of our model is 2048. By applying bag of tricks, we set the last layer stride to 1; as a result, we will end up with a single layer of features in the shape of 2048. We will acquire the attention scores of each frame

based on this score, and the scores will be compared and multiplied by attention and cl loss. After that, we'll execute batch normalization and feature classification, and we'll apply loss functions to reduce the loss. Finally, label smoothing will be done.

Chapter 4

EXPERIMENTS

The evaluation metric, and image baseline models are all introduced. Details on the implementation are also presented.

4.1 Metric

At rank-1, rank-5, rank-10, and rank-20, we apply the usual evaluation metrics of mean average accuracy score (mAP) and cumulative matching curve (CMC).

4.2 Image based baseline models

To evaluate the efficacy of temporal modelling, we give an image-based baseline model. This model is identical to [9], however it includes a Softmax cross-entropy loss. There is no temporal modelling approach employed because the clip's sequence length is set to T = 1. The ResNet-50 network is used with the same loss function (triplet loss and cross-entropy loss).

4.3 Implementation

The CNN is standard ResNet-50 pretrained on ImageNet. The video frames have been reduced to 224x112 pixels. The networks are optimised using Adam. The batch size is set to 32; if the total memory usage exceeds the GPU memory limit, the batch size is reduced to the largest size available. For each identity, we choose P = 4 samples in a batch. To attain the optimum result, we test the learning rate with 0.0001 and 0.0003 for different models.

4.4 Temporal Attention

We use an attention weighted average on the succession of picture features in the temporal attention model. The attention for clip c is

$$[a_c^t, t \, \epsilon \, [1, T]] \tag{4.1}$$

then,

$$[f_c = \frac{1}{T} \sum_{t=1}^{n} a_c^t f_c^t] \tag{4.2}$$

The last convolutional layer in Resnet-50 has a tensor size of [w,h,2048], where w and h are dependent on the size of the input image. The attention generating network uses a series of image-level elements to generate attention. [T,w,h,2048] as inputs, and T attention scores as outputs. We create two different types of attention networks. 1) “FC + spatial conv": a conv layer is applied (input channel = 2048, kernel height = h, kernel width = w, output channel = dt; short for {w,h,2048,dt}) and a fully connected (output channel = 1 and input channel = dt) a layer on top of the output tensor, the conv layer’s output is a scalar vector

$$[s_c^t, t \in [1, T] \tag{4.3}$$

This is the score for the clip’s frame t. 2) “temporal con + spatial": After applying a conv layer with the shape {w,h,2048,dt}, we obtain a dt-dimensional feature for each frame of the clip, and finally, we apply a temporal conv layer with the pattern {3,dt,1} on these frame-level features to generate temporal attentions. As soon as we have s_c^t ,there are two methods for determining the final attention scores a_c^t. The first one is Softmax function

$$[a_c^t = \frac{e^{s_c^t}}{\sum_{i=1}^{T} e^{s_c^i}}] \tag{4.4}$$

and the second is L1 normalization + sigmoid function

$$\frac{\sigma s_c^t}{\sum_{i=1}^{T} \sigma s_c^t}] \tag{4.5}$$

4.5 Loss Functions

To train the networks, we employ a triplet loss function and a Softmax cross-entropy loss function. The triplet loss function we utilise is called Batch Hard triplet loss function, and it was first proposed in [9]. To create a batch, we randomly select P identities and K clips for each identity (each clip has T frames); the total number of clips in the batch is PK. When constructing the triplets for computing the loss, the hardest positive and hardest negative samples in the batch are chosen for each sample a.

The Softmax cross-entropy loss function promotes the network to correctly identify the PK clips. We have used combination of following losses.

4.5.1 Center Loss

Center Loss [48] learns a centre for each class at the same time as penalising distances between deep features and their related class centres.

$$L_{center} = \frac{1}{2}\mathbb{E}_x \left[\left\| x - c_y \right\|_2^2 \right] \tag{4.6}$$

4.5.2 Triplet Loss

Anchor point (x_a), positive point (x_p), and point loss [38] apply to a triplet of samples designated anchor point (x_a), positive point (x_p), and point loss (x_n). Its goal is to move an anchor point by a defined margin m nearer to the positive point (same identity) and pushing away the negative point (different identity).

$$L_{tri} = \mathbb{E}_{(x_a, x_p, x_n)} \left[D(x_a, x_p) - D(x_a, x_n) + m \right]_+ \tag{4.7}$$

4.5.3 OSM Loss

We propose CL Centers OSM loss for cropping out noisy frames, which leverages the centre vectors from centre loss as the class label vector representations, because they have more variance than the previously proposed classifier weights. We also punish the model for giving high attention scores to frames where a patch has been randomly erased. These randomly erased frames are called 1; otherwise, they are labelled 0.

4.6 Model Training

To train our model, we have used premium version of Google Colab. We trained multiple models such as:

- Temporal Pooling
- Temporal Pooling + Bag of Tricks
- Temporal Attention
- Temporal Attention + Bag of Tricks
- Temporal Attention + Bag of Tricks + Attention and CL Loss

Our virtual machine specs are show in below figure 4.1. To train our model we divided our datasets into Train, Query, and Gallery set. Details for data loading of PRID2011 dataset (4.2) and Self Generated Custom dataset (4.3) are show below

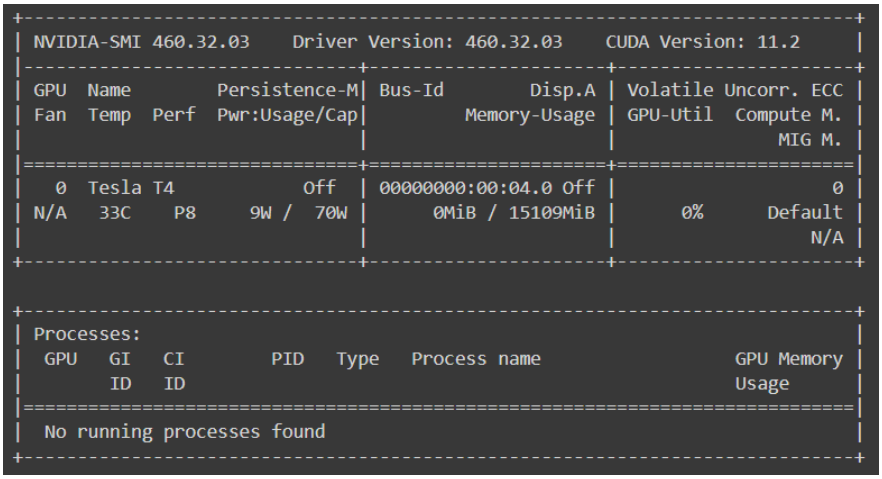


Figure 4.1: Virtual Machine Specifications

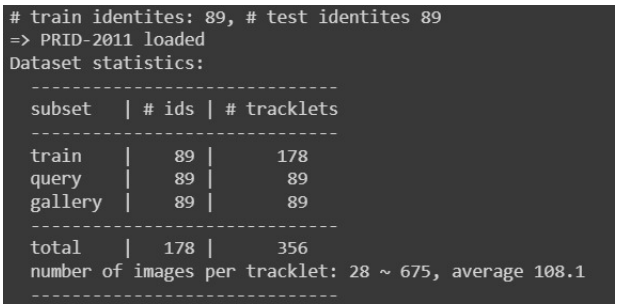


Figure 4.2: PRID2011 Dataset Statistics

respectively. Then we set up our Temporal Attention Model with Bag of Tricks. To verify our model we print summary of the model. Summary of the model is attached below.

After creating a desired model we trained our model our our benchmark dataset and custom dataset. For training purpose we used hyperparameters which were calculated by Fookbooks AX tool. Ax library was used to calculate values of hyperparameters. Arguments were used is script to make it behave anyway want. Different models and Loss functions and many more paramenters were given as key argurment to the script while compiling the script for training. 100 epochs were used to train custom dataset and 800 epochs were used to train PRID2011.

4.7 Web Portal

The end application of this project is a web portal. All the outputs are shown in web portal. The web portal look like:

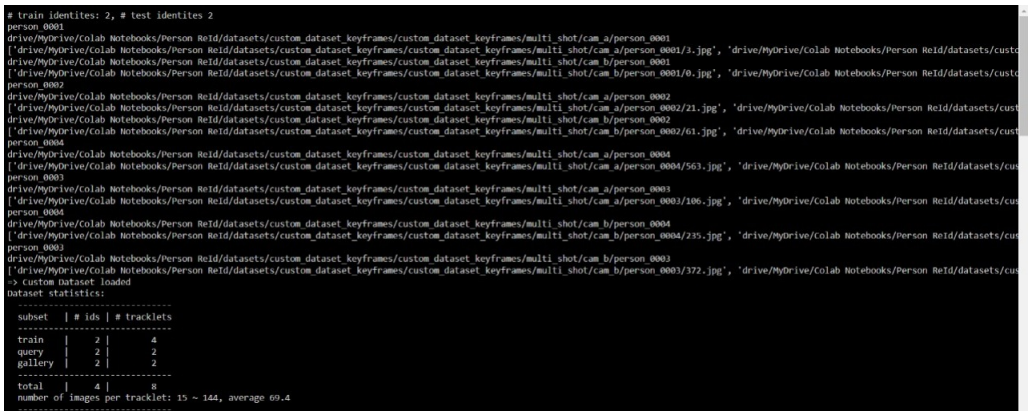


Figure 4.3: Self-generated Custom Dataset Statistics

Figure 4.4: Summary of Model

When we first visit the website, we are presented with a login screen that includes the project title, all group member information, a project overview, and a dataset. As previously stated, we must first visit a login page, which requires us to provide credentials in order to log in. Once we've logged in, we'll see the following: On this page, you'll find the following items:

- The reuslts of benchmark and customized datasets can be obtained using the links provided.

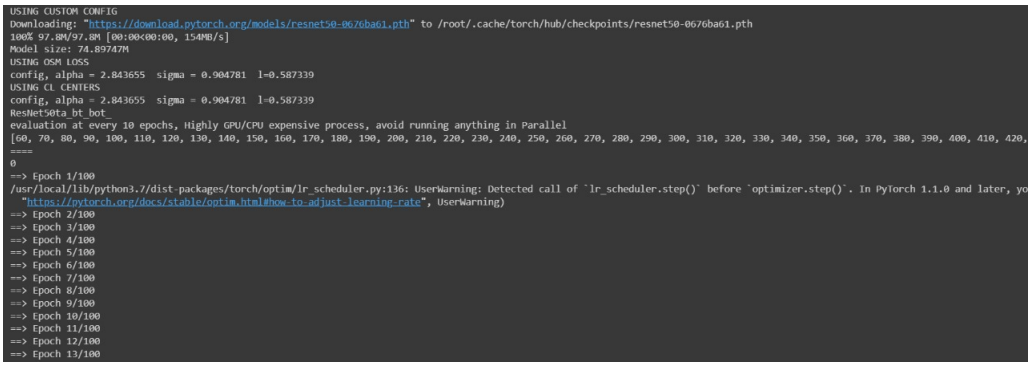


Figure 4.5: Training Model: Custom Dataset

- A comparison link that shows the comparison between the benchmark dataset and the custom dataset.
- There is a link to download the benchmark dataset.
- A link to the project's thesis, where you can view it.
- At the bottom of the page, you'll find all of the information you need about the group members and supervisor. By clicking on the email, you can access it.

The languages that are used for the creation of web portal are:

4.7.1 HTML

HTML (HyperText Markup Language) is the contraction for HyperText Markup Language. Site pages are made utilizing a markup language. HTML is a markup language that consolidates hypertext and markup in one bundle. The expression "hypertext" alludes to the connection between site pages. A markup language is utilized to portray the text record inside the name that depicts the plan of site pages. This language is utilized to explain (add notes to) material with the goal that a PC can comprehend it and control it properly. Individuals can figure out most of markup vernaculars (like HTML). In the language, names are utilized to demonstrate what sort of text dealing with is vital.

4.7.2 CSS

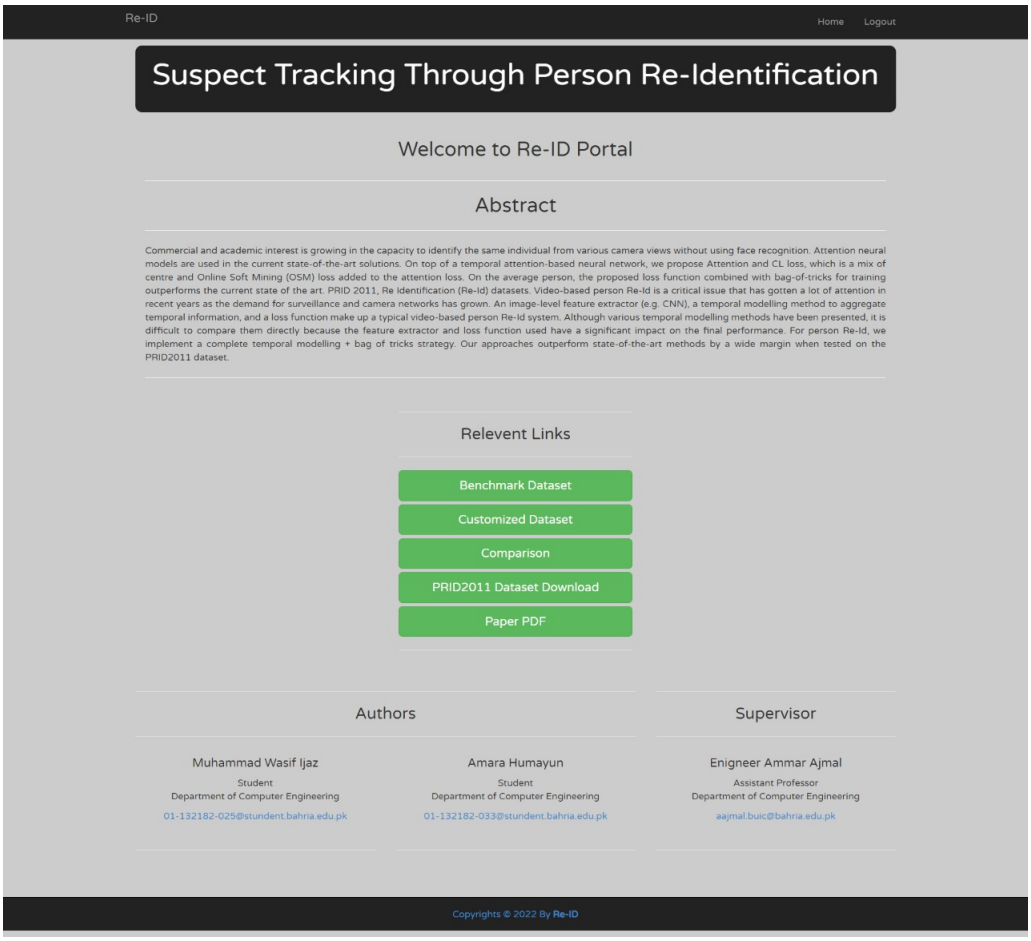
CSS, or Cascading Style Sheets, is a straightforward plan language expected to make the most common way of making website pages satisfactory simpler. Styles can be applied to site pages utilizing CSS. All the more critically, CSS permits you

Figure 4.6: Home Page of Re-ID Web Portal

to do as such without stressing over the HTML code that makes up each website page. It indicates how a website page ought to show up, including colors, text styles, dividing, and then some. As such, you have unlimited authority over the presence of your site. CSS permits engineers and fashioners to control how the program acts, including how things are shown.

4.7.3 JavaScript

JavaScript is a gathered programming language that is lightweight, cross-stage, and deciphered. It is otherwise called the prearranging language for sites. It is notable for page creation, however it is generally utilized in various non-program applications. JavaScript can be utilized to make both client-side and server-side applications. Javascript is a decisive and basic programming language. A standard



library of items, like Array, Date, and Math, as well as an essential arrangement of language parts, for example, administrators, control designs, and proclamations, are remembered for JavaScript.

4.7.4 PHP

PHP represents Hypertext Preprocessor and is an abbreviation for PHP. PHP is a server-side programming language that was made in light of web improvement. It is open-source, and that implies you can download and utilize it free of charge. It's truly simple to get and use. ".php" is the expansion of the records. Rasmus Lerdorf was the main thrust behind the principal rendition of PHP, as well as a supporter of succeeding forms. A deciphered language doesn't require the utilization of a compiler.

4.7.5 Bootstrap

Other than the above dialects, we have additionally utilized the bootstrap subject. Bootstrap is a tool compartment for building responsive sites and web applications

that is free and open-source. It is the most generally utilized HTML, CSS, and JavaScript structure for making versatile first, responsive sites. It fixes various issues that we encountered already, including cross-program similarity. These days, pages are advanced for all programs (Internet Explorer, Firefox, and Chrome) and screen sizes (Desktop, Tablets, Phablets, and Phones). All because of Twitter's Mark Otto and Jacob Thornton, who made Bootstrap, which was ultimately pronounced an open-source project. We used because

- It is now easier and faster to construct a website..
- It generates web pages that are platform agnostic.
- It makes Responsive Websites.
- It's also made to work well on mobile devices.

Chapter 5

RESULTS

5.1 PRID2011 Results

On the benchmark dataset, we trained various models. Trained models are temporal pooling, temporal pooling plus bag of tricks, temporal attention, temporal attention plus bag of tricks, temporal attension plus bag of tricks plus CL loss. From all these trained models, our proposed model is temporal attension plus bag of tricks plus CL loss. The results are given below:

Model	CMC - Rank 1	CMC - Rank 5	CMC - Rank 10	CMC - Rank 20
Temporal Pooling	50.6	85.4	94.4	97.8
Temporal Pooling + Bag of Tricks	78.8	91.0	97.8	97.8
Temporal Attention	65.2	95.8	96.6	98.9
Temporal Attention + Bag of Tricks	83.1	97.8	98.9	100
Temporal Attention + Bag of Tricks + Attention and CL Loss (Proposed)	88.8	98.9	98.9	100

Table 5.1: PRID2011 Dataset Cumulative Match Curve (CMC) Ranking on various models.

Model	mAP
Temporal Pooling	52.3
Temporal Pooling + Bag of Tricks	81.7
Temporal Attention	76.7
Temporal Attention + Bag of Tricks	81.3
Temporal Attention + Bag of Tricks + Attention and CL Loss (Proposed)	92.6

Table 5.2: PRID2011 Dataset Mean Average Precision (mAP) on various models.

5.2 Custom Dataset Results

On the custom dataset, we trained multiple models. Trained models are temporal pooling, temporal pooling plus bag of tricks, temporal attention, temporal attention plus bag of tricks, temporal attension plus bag of tricks plus CL loss. From all these

trained models, our proposed model is temporal attension plus bag of tricks plus CL loss.The results are given below:

Model	CMC - Rank 1	CMC - Rank 5	CMC - Rank 10	CMC - Rank 20
Temporal Pooling	46.7	68.2	75.3	79.5
Temporal Pooling + Bag of Tricks	67.1	71.6	79.3	82.4
Temporal Attention	60.0	75.2	82.6	85.3
Temporal Attention + Bag of Tricks	69.9	78.8	83.2	89.7
Temporal Attention + Bag of Tricks + Attention and CL Loss (Proposed)	73.5	81.6	89.3	94.6

Table 5.3: Custom Dataset Cumulative Match Curve (CMC) Ranking on various models.

Model	mAP
Temporal Pooling	48.5
Temporal Pooling + Bag of Tricks	70.2
Temporal Attention	67.3
Temporal Attention + Bag of Tricks	73.9
Temporal Attention + Bag of Tricks + Attention and CL Loss (Proposed)	84.7

Table 5.4: Custom Dataset Mean Average Precision (mAP) on various models.

5.3 Outputs Through Web Portal

When we click on the benchmark dataset to get the results, we will be presented with three sections, one of which contains all of the input images. In order to re-identify a person of interest, simply click on the person of interest, and all of the images of that individual from camera b will appear on the other side. We'll also have a table of that person's accuracy above this result.

Similarly, if we want customized dataset results, click on the customized dataset, we will be presented with three sections, one of which contains all of the input images. In order to re-identify a person of interest, simply click on the person of interest, and all of the images of that individual from camera b will appear on the other side. We'll also have a table of that person's accuracy above this result.

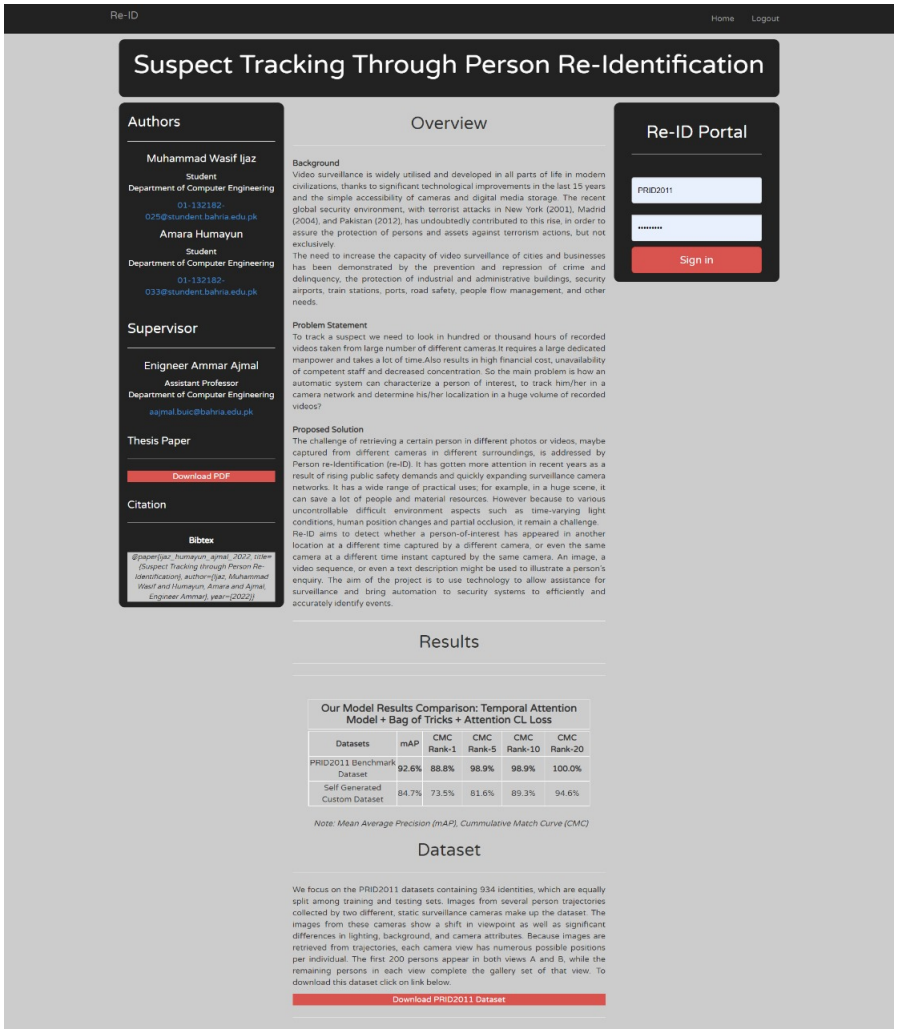


Figure 5.1: Web Portal: Login Page

Figure 5.2: Web Portal: Index Page

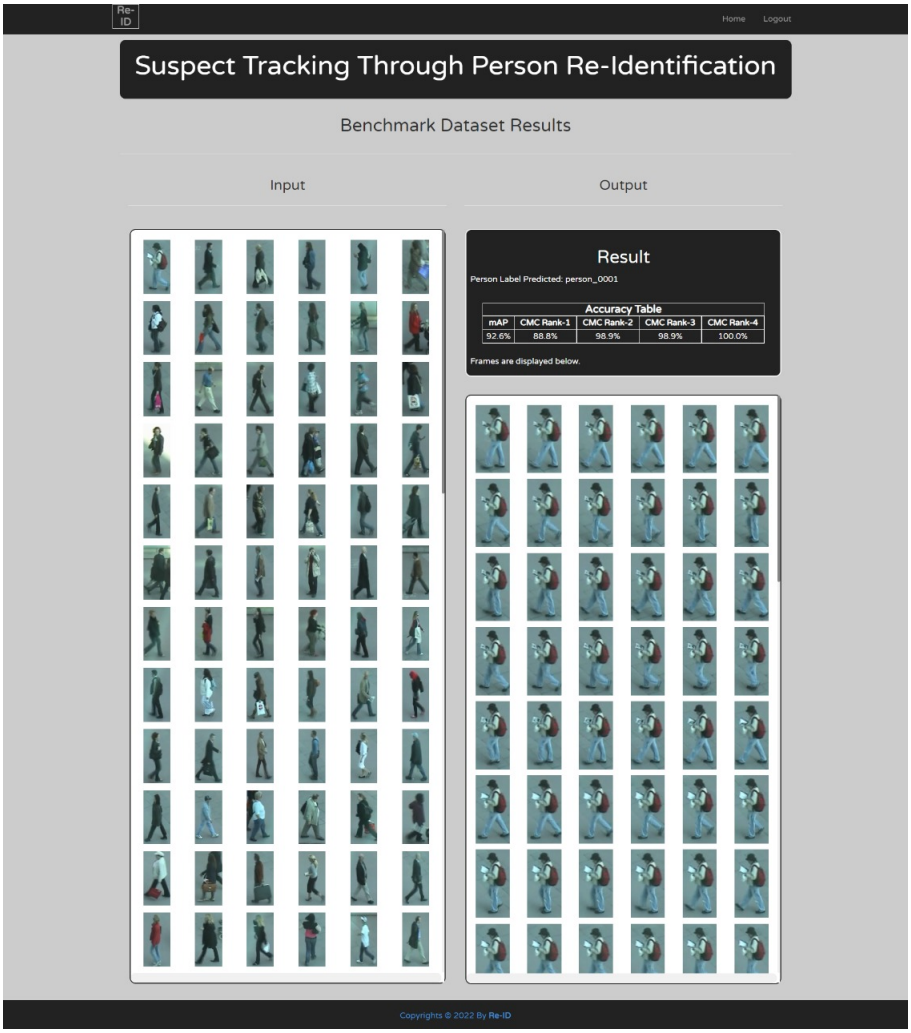


Figure 5.3: Web Portal: Benchmark Results Page

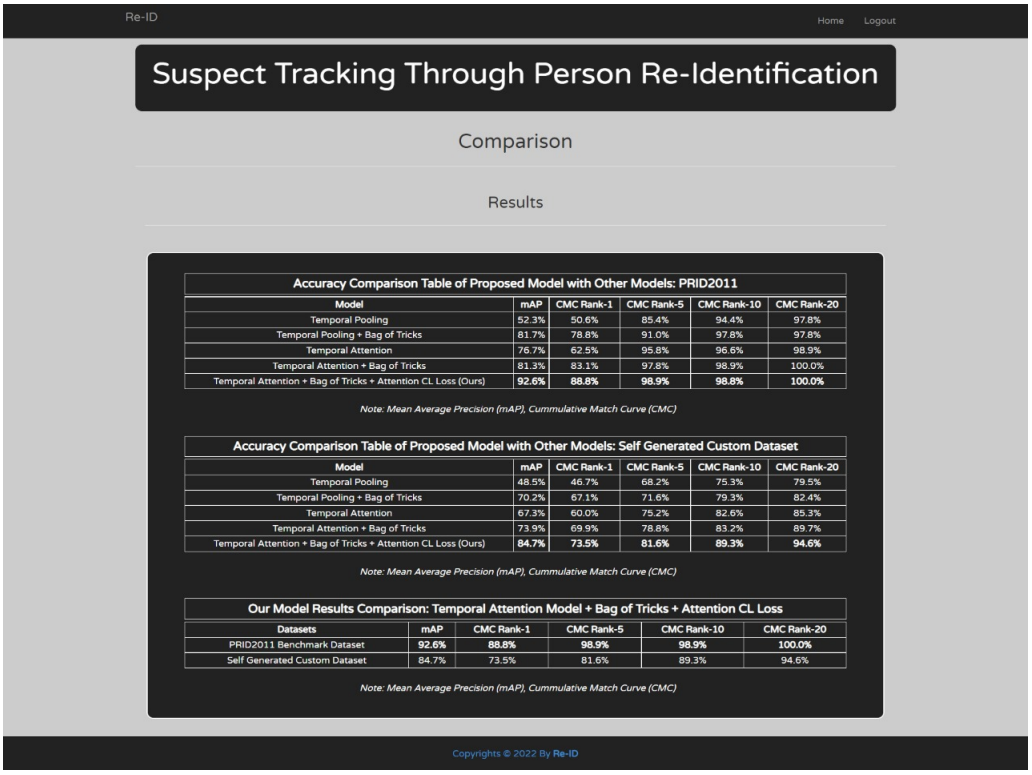


Figure 5.4: Web Portal: Comparison Page

Chapter 6

CONCLUSION

We used deep learning and computer vision for suspect tracking through person re-identification. We work with two datasets: one is a benchmark dataset called PRID2011, and the other is a custom dataset that we built. The model we propose combines a temporal attention model with a bag of tricks and CL loss. As a starting point, we used Resnet 50. To acquire attention scores, a temporal attention model is utilized, and a bag of tricks is used to improve accuracy. We utilized ResNet50 as a baseline. After examining a number of approaches, we determined that ResNet50 is the best option for PRID2011 Person Re-Id. By comparing various models on PRID2011, we can see that three of the top five models employ ResNet50 as a backbone, indicating that ResNet50 is the best option for re-identification on PRID2011. On these two datasets, we trained many models, but our proposed model has the highest accuracy. The limitations and challenges are: similarity in clothing, high cluster of people, blurry images. It is tough for our system to reidentify a person whose clothing is similar to that of others. Because it is natural for two or more people to wear the same outfit. The system will struggle to reidentify a person of interest in a large group of people. Because there is a high possibility of blockage in a high cluster.

Chapter 7

FUTURE SCOPE

Re-identification of people is a crucial task in multimedia, computer vision, and machine learning. Deep learning techniques have made noteworthy progress in a variety of fields. To successfully process varied difficulties, a growing number of developing technologies have been presented in both the academic and industrial domains. As we move towards the digital age and smart cities automation of every key activity becomes a challenge and a necessity. Person re-identification is one of those challenges and it has recently grown in popularity because of its various advantages. Researchers have already achieved absolute results over benchmark datasets, but pure potential of re-identification is still unrevealed. Person re-identification can be used for monitoring, tracking, analytics and much more. Tracking has been a hot domain for person re-identification as every researcher has worked up to their potential to achieve state of art results for tracking. The future of person re-identification is to implement this knowledge of benchmarking datasets to achieve reliable results on real life data. There are various deep learning models which perform person re-identification at run time, but their results are satisfactory compared to models which take time to train and produce output such as our proposed model. We have tried to perform person re-identification on benchmark dataset and a self-generated custom dataset and compare the results. Our next target will be to create a larger dataset from the real world and apply our model over it and try to create a model which can perform person re-identification on run time with state of art results. There has also been progress in the monitoring and analytics sector of person re-identification. Person re-identification can be used to monitor employees and calculate their work efficiency and apply analytics to enhance the efficiency of their employees. This can also be used to monitor the average customer entering your store and the average customer buying your products daily, it is just an example of a use-case. We would love to explore the monitoring side of person re-identification as well in the future.

ABBREVIATIONS

CMC: Cumulative Match Curve
mAP: Mean Average Precision
CNN: Convolutional Neural Network
OSM: Online Soft Mining
CL: Center Loss
RLL: Rank List Lost
HTML: Hyper Text Markup Language
CSS: Cascading Style Sheet
JS: JavaScript
BOT: Bag of Tricks
CAA: Class Aware Attention

BIBLIOGRAPHY

[1] Malik Souded. "People detection, tracking and re-identification through a video camera network". PhD thesis. Université Nice Sophia Antipolis, 2013.

[2] Kaiyang Zhou and Tao Xiang. "Torchreid: A Library for Deep Learning Person Re-Identification in Pytorch". In: *arXiv preprint arXiv:1910.10093* (2019).

[3] Guido Van Rossum et al. "Python Programming language." In: *USENIX annual technical conference*. Vol. 41. 1. 2007, pp. 1–36.

[4] Wei Li et al. "Deepreid: Deep filter pairing neural network for person re-identification". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2014, pp. 152–159.

[5] Kaiming He et al. "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[6] Gao Huang et al. "Densely connected convolutional networks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 4700–4708.

[7] Haibo Jin et al. "Deep person re-identification with improved embedding and efficient training". In: *2017 IEEE International Joint Conference on Biometrics (IJCB)*. IEEE. 2017, pp. 261–267.

[8] Liang Zheng et al. "Mars: A video benchmark for large-scale person re-identification". In: *European Conference on Computer Vision*. Springer. 2016, pp. 868–884.

[9] Alexander Hermans, Lucas Beyer, and Bastian Leibe. "In defense of the triplet loss for person re-identification". In: *arXiv preprint arXiv:1703.07737* (2017).

[10] Jiayun Wang et al. "Deep ranking model by large adaptive margin learning for person re-identification". In: *Pattern Recognition* 74 (2018), pp. 241–252.

[11] Weihua Chen et al. "Beyond triplet loss: a deep quadruplet network for person re-identification". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 403–412.

[12] Dahjung Chung, Khalid Tahboub, and Edward J Delp. "A two stream siamese convolutional neural network for person re-identification". In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 1983–1991.

[13] Mengyue Geng et al. "Deep transfer learning for person re-identification". In: *arXiv preprint arXiv:1611.05244* (2016).

[14] Wei Li, Xiatian Zhu, and Shaogang Gong. “Person re-identification by deep joint learning of multi-loss classification”. In: *arXiv preprint arXiv:1705.04724* (2017).

[15] Yuntao Chen, Naiyan Wang, and Zhaoxiang Zhang. “Darkrank: Accelerating deep metric learning via cross sample similarities transfer”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 32. 1. 2018.

[16] Liming Zhao et al. “Deeply-learned part-aligned representations for person re-identification”. In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 3219–3228.

[17] Chi Su et al. “Pose-driven deep convolutional model for person re-identification”. In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 3960–3969.

[18] Yutian Lin et al. “Improving person re-identification by attribute and identity learning”. In: *Pattern Recognition* 95 (2019), pp. 151–161.

[19] Shuangjie Xu et al. “Jointly attentive spatial-temporal pooling networks for video-based person re-identification”. In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 4733–4742.

[20] Dangwei Li et al. “Learning deep context-aware features over body and latent parts for person re-identification”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 384–393.

[21] Longhui Wei et al. “Glad: Global-local-alignment descriptor for pedestrian retrieval”. In: *Proceedings of the 25th ACM international conference on Multimedia*. 2017, pp. 420–428.

[22] Zhedong Zheng, Liang Zheng, and Yi Yang. “Unlabeled samples generated by gan improve the person re-identification baseline in vitro”. In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 3754–3762.

[23] Zhun Zhong et al. “Random erasing data augmentation”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 07. 2020, pp. 13001–13008.

[24] Yifan Sun et al. “Svdnet for pedestrian retrieval”. In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 3800–3808.

[25] Xuelin Qian et al. “Multi-scale deep learning architectures for person re-identification”. In: *Proceedings of the IEEE International Conference on Computer Vision*. 2017, pp. 5399–5408.

[26] Ying Zhang et al. “Deep mutual learning”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 4320–4328.

[27] Haiyu Zhao et al. “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*. 2017, pp. 1077–1085.

[28] Niall McLaughlin, Jesus Martinez Del Rincon, and Paul Miller. “Recurrent convolutional network for video-based person re-identification”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 1325–1334.

[29] Hao Liu et al. “Video-based person re-identification with accumulative motion context”. In: *IEEE transactions on circuits and systems for video technology* 28.10 (2017), pp. 2788–2802.

[30] Kelvin Xu et al. “Show, attend and tell: Neural image caption generation with visual attention”. In: *International conference on machine learning*. PMLR. 2015, pp. 2048–2057.

[31] Zhen Zhou et al. “See the forest for the trees: Joint spatial and temporal recurrent neural networks for video-based person re-identification”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 4747–4756.

[32] Yu Liu, Junjie Yan, and Wanli Ouyang. “Quality aware network for set to set recognition”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 5790–5799.

[33] Shuang Li et al. “Diversity regularized spatiotemporal attention for video-based person re-identification”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 369–378.

[34] Yichao Yan et al. “Person re-identification via recurrent feature aggregation”. In: *European Conference on Computer Vision*. Springer. 2016, pp. 701–716.

[35] Karen Simonyan and Andrew Zisserman. “Very deep convolutional networks for large-scale image recognition”. In: *arXiv preprint arXiv:1409.1556* (2014).

[36] Md. Zahangir Alom et al. “Inception Recurrent Convolutional Neural Network for Object Recognition”. In: *CoRR* abs/1704.07709 (2017). arXiv: 1704.07709. URL: http://arxiv.org/abs/1704.07709.

[37] Gal Chechik et al. “Large Scale Online Learning of Image Similarity Through Ranking”. In: *Journal of Machine Learning Research* 11.36 (2010), pp. 1109–1135. URL: http://jmlr.org/papers/v11/chechik10a.html.

[38] Florian Schroff, Dmitry Kalenichenko, and James Philbin. “FaceNet: A Unified Embedding for Face Recognition and Clustering”. In: *CoRR* abs/1503.03832 (2015). arXiv: 1503.03832. URL: http://arxiv.org/abs/1503.03832.

[39] Xinshao Wang et al. “Deep Metric Learning by Online Soft Mining and Class-Aware Attention”. In: *CoRR* abs/1811.01459 (2018). arXiv: 1811.01459. URL: http://arxiv.org/abs/1811.01459.

[40] Hyun Oh Song et al. “Deep metric learning via lifted structured feature embedding”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 4004–4012.

[41] Yuhui Yuan, Kuiyuan Yang, and Chao Zhang. “Hard-aware deeply cascaded embedding”. In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 814–823.

[42] Florian Schroff, Dmitry Kalenichenko, and James Philbin. “Facenet: A unified embedding for face recognition and clustering”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 815–823.

[43] Yin Cui et al. “Fine-grained categorization and dataset bootstrapping using deep metric learning with humans in the loop”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 1153–1162.

[44] Chiyuan Zhang et al. “Understanding deep learning (still) requires rethinking generalization”. In: *Communications of the ACM* 64.3 (2021), pp. 107–115.

[45] Liang Zheng et al. “MARS: A Video Benchmark for Large-Scale Person Re-Identification”. In: *Computer Vision – ECCV 2016*. Ed. by Bastian Leibe et al. Cham: Springer International Publishing, 2016, pp. 868–884. ISBN: 978-3-319-46466-4.

[46] Pan He et al. “Reading Scene Text in Deep Convolutional Sequences”. In: *CoRR* abs/1506.04395 (2015). arXiv: 1506.04395. URL: http://arxiv.org/abs/1506.04395.

[47] Lu Yang et al. “Center prediction loss for re-identification”. In: *arXiv preprint arXiv:2104.14746* (2021).

[48] Yandong Wen et al. “A Discriminative Feature Learning Approach for Deep Face Recognition”. In: *Computer Vision – ECCV 2016*. Ed. by Bastian Leibe et al. Cham: Springer International Publishing, 2016, pp. 499–515. ISBN: 978-3-319-46478-7.

[49] Yifan Sun et al. “Circle Loss: A Unified Perspective of Pair Similarity Optimization”. In: *CoRR* abs/2002.10857 (2020). arXiv: 2002.10857. URL: https://arxiv.org/abs/2002.10857.

[50] Hyun Oh Song et al. “Deep Metric Learning via Lifted Structured Feature Embedding”. In: *CoRR* abs/1511.06452 (2015). arXiv: 1511.06452. URL: http://arxiv.org/abs/1511.06452.

[51] Zhizheng Zhang et al. “Densely Semantically Aligned Person Re-Identification”. In: *CoRR* abs/1812.08967 (2018). arXiv: 1812.08967. URL: http://arxiv.org/abs/1812.08967.

[52] Kihyuk Sohn. "Improved Deep Metric Learning with Multi-class N-pair Loss Objective". In: *Advances in Neural Information Processing Systems*. Ed. by D. Lee et al. Vol. 29. Curran Associates, Inc., 2016. URL: https://proceedings.neurips.cc/paper/2016/file/6b180037abbebea991d8b1232f8a8ca9-Paper.pdf.

[53] Jiyang Gao and Ram Nevatia. "Revisiting Temporal Modeling for Video-based Person ReID". In: *arXiv preprint arXiv:1805.02104* (2018).

[54] Hao Luo et al. "Bag of Tricks and a Strong Baseline for Deep Person Re-Identification". In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. June 2019.

[55] Martin Hirzer et al. "Person Re-Identification by Descriptive and Discriminative Classification". In: *Proc. Scandinavian Conference on Image Analysis (SCIA)*. 2011.

[56] Zhedong Zheng, Liang Zheng, and Yi Yang. "A Discriminatively Learned CNN Embedding for Person Re-identification". In: *ACM Transactions on Multimedia Computing Communications and Applications* (2017). doi:10.1145/3159171. DOI: 10.1145/3159171.

[57] Christian Szegedy et al. "Rethinking the Inception Architecture for Computer Vision". In: *CoRR* abs/1512.00567 (2015), pp. 2818–2826. URL: https://ieeexplore.ieee.org/document/7780677.

[58] Xinshao Wang et al. "Deep metric learning by online soft mining and class-aware attention". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33. 2019, pp. 5361–5368.

[59] Priyank Pathak, Amir Erfan Eshratifar, and Michael Gormish. *Video Person Re-ID: Fantastic Techniques and Where to Find Them*. 2019. arXiv: 1912.05295 [cs.CV].

[60] *Person Re-Identification on PRID2011*. URL: https://paperswithcode.com/sota/person-re-identification-on-prid2011.