Multi Stage Quasi-labeled Vehicle Re-Identification Using Branched Convolutional Neural Networks

By

Ali Haider

FALL-2018-MSCS 00000276806 SEECS

Supervisor

Dr. Muhammad Shahzad

Department of Computing
School of Electrical Engineering and Computer Science (SEECS)
National University of Sciences and Technology (NUST)
Islamabad, Pakistan

(June 2022)

THESIS ACCEPTANCE CERTIFICATE

Certified that final copy of MS/MPhil thesis entitled "Multi Stage Quasi-labeled Vehicle Re-Identification Using Branched Convolutional Neural Networks" written by ALI HAIDER, (Registration No 00000276806), of SEECS has been vetted by the undersigned, found complete in all respects as per NUST Statutes/Regulations, is free of plagiarism, errors and mistakes and is accepted as partial fulfillment for award of MS/M Phil degree. It is further certified that necessary amendments as pointed out by GEC members of the scholar have also been incorporated in the said thesis.

Signature: M. Shahzad

Name of Advisor: Dr. Muhammad Shahzad

Date: 25-Jul-2022

HoD/Associate Dean: ______________________

Date: ______________________

Signature (Dean/Principal): ______________________

Date: ______________________

Approval

It is certified that the contents and form of the thesis entitled "Multi Stage Quasi-labeled Vehicle Re-Identification Using Branched Convolutional Neural Networks" submitted by ALI HAIDER have been found satisfactory for the requirement of the degree

Advisor : Dr. Muhammad Shahzad

Signature: M. SHAHZAD

Date: 25-Jul-2022

Committee Member 1:Dr. Muhammad Imran Malik

Signature: Imran Malik

Date: 26-Jul-2022

Committee Member 2:Dr. Arsalan Ahmad

Signature:

Date: 26-Jul-2022

Committee Member 3:Dr. Muhammad Moazam Fraz

Signature: M Moazam Fraz

Date: 25-Jul-2022

Dedication

Dedicated to my father and mother for their never-ending support. And also I would like to dedicate my thesis to people all around the world who are unable to access educational institutes.

Certificate of Originality

I hereby declare that this submission titled "Multi Stage Quasi-labeled Vehicle Re-Identification Using Branched Convolutional Neural Networks" is my own work. To the best of my knowledge it contains no materials previously published or written by another person, nor material which to a substantial extent has been accepted for the award of any degree or diploma at NUST SEECS or at any other educational institute, except where due acknowledgement has been made in the thesis. Any contribution made to the research by others, with whom I have worked at NUST SEECS or elsewhere, is explicitly acknowledged in the thesis. I also declare that the intellectual content of this thesis is the product of my own work, except for the assistance from others in the project's design and conception or in style, presentation and linguistics, which has been acknowledged. I also verified the originality of contents through plagiarism software.

Student Name:ALI HAIDER

Student Signature: _______________

Acknowledgments

I am thankful to my Creator Allah Subhana-Watala to have guided me throughout this work at every step and for every new thought which You set up in my mind to improve it. Indeed, I could have done nothing without Your priceless help and guidance.

Whosoever helped me throughout my thesis, whether my parents or any other individual was Your will, so indeed none be worthy of praise but You. I am utterly thankful to my parents for their support, I am nothing but a piece of everything you have wanted me to be.

I would also like to express my utmost gratitude to my supervisor DR. MUHAMMAD SHAHZAD for his help throughout my thesis and his guidance, tremendous support, and cooperation. I can safely say that I wouldn't have been able to complete my thesis without his in-depth knowledge of this field. Each time I stumbled because of my personal and professional issues; he was there with a helping hand that got me where I am right now in my professional education. There are not enough words to appreciate his patience and guidance throughout my entire thesis. I am heartily thankful to him. His encouragement, guidance, and support from the initial to the final level enabled me to develop an understanding of the subject.

I would also like to pay special thanks to, Dr. Arsalan Ahmad, Dr. Muhammad Imran Malik, and Dr. Moazam Fraz for being on my thesis guidance and evaluation committee and for being kind enough to help me whenever I needed their guidance and support.

I offer my regards and blessings to all my friends who supported me in any respect during the completion of the project. Finally, I would like to express my gratitude to all the individuals who have rendered valuable assistance to my study.

Ali Haider

Table of Contents

List of Figures

List of Tables:

List of Equations

Abstract

Vehicle re-identification has become an important field of interest in computer vision, especially after the rise of importance of smart cities, and also with the evolution of Convolutional Neural Networks it has seemed possible to take advantage of technology to perform vehicle re-identification, Due to vehicles being an important object in the formation of Smart cities many scholars have applied different techniques to perform vehicle re-identification as it has many applications such as unmanned vehicle parking, vehicle tracking, vehicle identification, and vehicle counting and also arial traffic flow management. Many scholars have tried to solve vehicle re-identification challenges using computer vision conventional methods such as feature extraction and histogram of oriented gradients but with the evolution of CNN which itself tries to learn the representation of the vehicle from images with the labels provided and this method has proved to be very successful instead of its predecessors and has achieved great results in the field of Re-identification. Person re-identification as well as vehicle re-identification. Although supervised vehicle re-identification has achieved great results in vehicle re-identification but due to the complication of implementing supervised re-identification methods which are huge data annotation costs and due to their classification nature supervised methods are always in need of finetuning on new data to produce good results, so we have tried to solve this problem with unsupervised learning combining it with the training of CNN's using the technique of pseudo label generation which are used to learn representation for CNN's and they are gathered from data using some conventional clustering techniques used for unsupervised learning. Vehicle reidentification has different challenges present in real-world scenarios and one of the major problems is vehicle appearance or different views of vehicles in different cameras which can be described as domain discrepancy in different cameras we have tried to solve that problem for vehicle re-identification using unsupervised learning by training a CNN in for intra camera to adjust model feature

generalization across the camera and then fine-tuning backbone for inter camera training to further make the model backbone robust to view discrepancy present in the real world.

We have also infused vehicle contextual information (color, type) information in the model for the model to better learn the representation and learn class discrimination for a single vehicle in different cameras and to converge faster.

We First trained the color model in the same manner as we are training the re-id model and then after training, using an unsupervised hierarchical technique for clustering we produced pseudo labels and fed those pseudo labels for the color label to Re-id Network after converting it into the multi-class multilabel network. Idea was to increase feature similarity between same class boundary learning and decrease feature similarity between different class samples.

Chapter:1

Introduction

1.1. Introduction and Motivation

In 2017, Garther[1] predicted that the world will be using around 20.4 billion devices by the end of 2021 and all of these are producing a huge amount of data. These connected devices are everywhere like in airports, schools, traffic, universities, train stations, and shopping centers. and they are already solving a lot of problems for humans. As urban areas are populating most of the world's population, they have bigger problems as well like, traffic congestion, security surveillance, etc. These problems can be solved by using the visual data being generated by the connected devices. For example, traffic can be made more effective, safer, and smarter by getting useful insights from the existing traffic data. Moreover, this data can help to solve problems including visual surveillance, crowd behavior analysis, player tracking, anomaly detection, and suspicious activity detection.

The evolution of deep neural networks and the internet of things paved the way for smart city concept. And vehicle being the important factor in smart city transportation system has gotten high attraction and many research articles related to vehicles are being carried out using deep learning. Vehicle tracking by H. Wang at el [2], vehicle detection, vehicle type, and vehicle color classification are some of them but vehicle reidentification being the hottest and main technology for the formation of intelligent traffic/transport systems for the concept of smart cities caused more attention in the research area.

However, many successful deep neural networks are trained in a supervised manner, in other words, it needs a large amount of labeled data which are then annotated using expensive man force and also strictly limits the deployment of deep models proposed by Z. Wang at el[3].

Nonetheless, there exist many issues that can hinder the solution of these problems like poor quality data, poor illumination conditions, occlusion, and lack of labels because cameras are not properly positioned and their field of view is limited to a small area; not collecting enough information that can be feed to deep learning models.

And to enable deep learning models to learn from unlabeled data, the use of the unsupervised method has also been the center of interest in the field of computer vision. Techniques in unsupervised learning can particularly cope with these kinds of issues by inferencing from the unlabeled data and understanding underlined representation and these models have shown promise in the field of object re-ID by Z. Zheng at el[4], [5]. And also there is a very major upper hand to unsupervised learning that it does not need continuous training of the model with data that the model is seeing for the first time. Unsupervised models perform very better than supervised models with unseen data which is why they are gaining popularity and are more feasible for practical use in the context of Re-identification.

1.2. Background:

Object re-identification solves the problem of identifying the same object present in non-overlapping cameras if it has been observed by the network before. And this object re-identification has many vital applications in the smart city such as tracking and identifying objects in urban surveillance systems. And with increasing traffic on road, vehicles being one the main object in surveillance systems captured a lot of attraction in object re-identification and many other applications such as vehicle type recognition and vehicle detection in which vehicle re-identification is the front runner. And the purpose of vehicle reidentification is to identify the same vehicle in a network of non-overlapping cameras J. Wang at el[6]. The reason behind vehicles being a very important object in urban surveillance in the formation of smart cities; are with vehicle re-identification we can easily identify and detect vehicles in a city and obtain their target location which can also help us in numerous other application such as intelligent parking, suspicious vehicle tracker, automatic car charging and live monitoring of vehicular objects in urban surveillance.

Although persons and vehicles are both very important objects in smart city formation person re-identification has been given a lot more attention as compared to vehicle reidentification in the past. Although vehicle re-identification is considered more difficult in terms of single-class object appearance in the multicamera environment as compared to person re-identification. In other words vehicle, reidentification has less inter-class similarity and more intra-camera discrepancy which makes it more challenging to identify the same make and model car to identify between different cameras because it can have a different appearance in different cameras J. Peng at el[7].

There are many benchmark datasets available for vehicle reidentification to mimic the real-world problems for vehicle re-identification and some of the most famous are VeRi(776), VehicleID, CityFlowReID, and VERI-Wild.

VeRi(776): “VeRi-776 is a vehicle re-identification dataset which contains 49,357 images of 776 vehicles from 20 cameras. The dataset is collected in the real traffic scenario, which is close to the setting of CityFlow. The dataset contains bounding boxes, types, colors, and brands.”[8] **VERI-Wild**: “Veri Wild is the largest available vehicle-reidentification dataset and it was published in CVPR 2019. It was captured by a large CCTV surveillance system comprising 174 cameras and it is captured within 1 month. It has data available for almost 40,671 vehicles. And it is also divided into three subsets. Veri Wild (small, medium and large)”.[9], **VehicleID:** Vehicle ID contains images of multiple vehicles in the daytime from multiple surveillance cameras distributed in a small city of china. It has 26,267 unique vehicle data which is and the vehicle has information available for its vehicle id, camera, color, and type.

These Datasets are captured in real-world traffic scenarios, And these are some of the properties these datasets have in common to address the challenges of vehicle re-identification in real-world scenarios.

1. All datasets contain a minimum of 50,000 or more images taken by different cameras placed perfectly and arranged in a way so that they can capture a single vehicle from different viewpoints which are also non-overlapping.
2. All these images are captured in a real-world environment so that they have real-world present vehicle re-identification challenges present in them such as occlusion, different viewpoints, Illuminations, and different resolutions.

3. They are all labeled with vehicle id’s, spatio temporal, color and BBoxes, and license plate information available or visible in them.

Vehicle re-identification traditionally used to be solved with a combination of conventional machine learning methods and different hardware information such as passing time information of the vehicle and electromagnetic hardware information and sensory data which adds extra hardware cost; also those methods can not produce good results considering the challenges of re-identification. But there are other methods that use Optical character recognition and try and read license plate information on basis of that and try to perform vehicle re-identification but it is very expensive to install high-resolution cameras across the network and also not feasible for many scenarios which are visible in vehicle re-identification datasets such as occlusion, different camera views, decorated license plate, and different formation of license plates in different cities which makes it very difficult to read license plate information hence optical character recognition field being mature enough still it is not feasible to solve the challenges of vehicle re-identification. Therefore, vehicle re-identification done with representation learning and appearance learning combined with different vehicle attributes has received more popularity in the field of vehicle re-identification. There are methods and pipelines available for vehicle and person reidentification in both supervised and unsupervised manner. But most successful vision applications are trained in a perfectly supervised manner because deep neural networks work best when they are provided with suitable labels to learn the representation underlining visible pictures.

1.3. Problem Statement:

Vehicle re-identification is mainly identifying one vehicle in a network of connected non-overlapping cameras. Despite supervised learning being the major technique to tackle vehicle re-identification, it is still not feasible for practical implementation due to expensive data annotation and the dynamic growth of data. In the rescue, unsupervised feature learning/re-identification can cope with these challenges because it uses direct inference from the unlabeled input image.

1.4. Solution Statement:

Techniques using unsupervised learning methods can potentially reduce the data annotation time. Representation learning and inferring directly from the unlabeled input data have been effectively employed in the context of person re-Identification.[5],[10]. We are developing an unsupervised technique for vehicle re-identification by training a base deep convolutional neural network in two stages; In the first instance a convolutional neural network is used to learn vehicle appearance features and to transfer the learned representations in the first stage; a clustering technique is used to cluster the similar looking vehicles on the basis that appearance representation to generate pseudo labels rather than using original labels to incorporate the learned representation in the second stage. The proposed approach is based on learning contextual information from input data and it is an extension of the proposed idea presented in [10]; that representation is then used to cluster input images with similar-looking representations according to contextual features. which are then used to fine-tune another convolutional neural network, which also works by training the base convolutional neural network by adding the appearance information into the framework learned in the first stage. This base convolutional neural network is also trained in two stages. 1. intra-camera stage, 2. Inter camera stage.

1.5. Thesis Contribution:

We are proposing a full fledge Unsupervised Vehicle reidentification pipeline in which in the first stage the base Convolutional neural network learns the contextual or appearance information of vehicles and then these learned features are used in the second stage as a classifier and clusters all similar feature vehicles into one class and generates pseudo labels to feed into baseline[10] extending it in the second stage which already uses pseudo labels generated with feature vectors of Imagenet and then used as input to the convolutional neural network as labels, now that baseline also uses pseudo labels generated by contextual representation features as labels and uses that information to further better the classification. These appearance-based pseudo labels embed local class information and helps the model refine the boundaries between different classes which results in better clustering and better reidentification in the end.

1.6. Thesis Outline:

This thesis Document is further organized in the following way.

1.6.1. Chapter 2: Literature Review

This section will contain a brief description of papers of the past/literature review on the topic of vehicle re-identification and a brief discussion of their methodology and results. All the methods that were and are being used to perform re-identification and their pros and cons, the discussion includes major work done in the field of and state-of-the-art performances for re-identification.

1.6.2. Chapter 3: Design and Methodology

In this chapter the proposed methodology is explained in detail with the help of figures and tables and the design of our methodology is also explained with each module and the proposed network and the overall design of the model have also been explained in detail. The design includes the previous paper model configurations, architecture, and our proposed model with the structure of the model. Moreover, settings of different hyper-parameters have also been discussed which elaborates the optimal setting as well.

1.6.3. Chapter 4: Implementation

In this chapter Detailed discussion is on how implementation of the project is done and what are the important technologies and tools used to implement the project. It is a multi-stage model; this chapter also contains the detailed explanation of implementation of each stage. First stage is to implement the color-classification model, second; How to generate pseudo labels from color classification and then lastly, to train Re-id model with those pseudo labels which is also done in a multistage setup 1. intra-camera 2. inter-camera.

1.6.4. Chapter 5: Experiments and Results

The results of all tests performed using our model have been discussed in this chapter. Datasets used for the model testing and evaluation are discussed in the first section of the chapter. Hyper-parameters, evaluation metrics, and model settings are also part of this chapter. To future explain the complete results; different tables and figures are included that show the comparison of our model with other models on different datasets.

1.6.5. Chapter 6: Conclusion

Conclusions include the summary of work, conclusion, and future work. In Conclusion and for future work explain the contributions of our work and future improvements in the code that can further improve the work done. On the other hand, a summary of work briefly describes the work done with the scope of vehicle re-identification.

1.6.6. Chapter 7: Bibliography/References

The final chapter includes the references (in IEE-tran style) to the related work that has been quoted in the thesis.

Chapter:2

Literature Review

2.1.Literature Review:

Re-identification is technically defined as a task to identify objects in a network of surveillance cameras without overlapping vision. Re-identification has multiple applications like object tracking, intelligent unmanned car parking, object identification object counting, and many else. Recently re-identification has gained a lot of attention; after deep learning models are mature enough in the field of computer vision. Although it has applications in many other fields such as Natural language processing, which is used to reidentify a person's voice in different streams of audio. They are all very important for the concept of smart cities. And by looking at it in detail we can finalize that person and vehicle are the two most important objects in smart cities, They need 24-hour surveillance which is impossible to be done by humans due to the limitations of the human body and that’s where artificial surveillance systems come handy but at the moment they lack the abilities to perform re-identification accurately because of the challenges that exist in real-world scenarios. But with the evolution of deep neural networks and computer vision, we can safely say that time is not far away when we would be able to fully rely on the artificial system and their ability to do tasks with more accuracy as compared to humans. And these technologies are very crucial for the formation of the intelligent transportation system and smart cities as with each passing day we are moving forward towards the connected physically (internet of things).

Coming back towards Re-ID we can formulate re-identification in vision by classifying it as a matching task, in which we have a query image that needs to be matched with the gallery set of images that are collected from a network of surveillance cameras and try to identify by matching the same vehicle. Thus matching this target image in a pool of images already captured can be formulated as:

$$T = arg_{T_i} minD(T_i, Q), T_i \in \tau \qquad (1)$$

Where T $= \{Ti, \dots., Tn\}$ is a gallery set of images with N number of images to be matched with the target image. And the metric to be used for matching is the Distance metric which is described with D. Furthermore, the rest of the thesis document investigates the topic of object re-identification.

Before deep learning became a norm, conventional machine learning methods were used to perform re-identification and hand-crafted features were used for that task. And there were only some parameters at hand which could be changed, But after the rise of Deep learning methods there were thousands of parameters for the model to adjust automatically and the model also learned different features from images automatically with different settings of parameters which abolished the need for handcrafted features. So we will discuss some methods which used hand-crafted features and then some methods based upon deep learning respectively which are categorized into two categories Supervised methods and Unsupervised methods. And some other re-identification methods are also mentioned.

2.2. Conventional Re-Identification Methods:

Conventional Re-identification used to be done by feature engineering which was the combination of three main techniques, feature extraction, feature encoding, and feature classification. “Some famous methods of feature extraction were SIFT[11], Histogram of oriented gradients HOG[12] and local binary patterns (LBP)[13]”[2].SIFT works on local features of the image and tries to maintain the local feature invariant to rotation, brightness, and scaling of viewing angles. which is particularly beneficial when matching images for identification. These features can easily be matched with the image base containing a lot of other images because of the uniqueness of features preserved by SIFT. Matching algorithms using SIFT as a feature extractor produce good results whose dataset contains the challenges of scaled images with different viewing angles with different color brightness. And these features joined with other features can make a good pipeline for re-identification.

Then came the Histogram of Oriented Gradients which did not have scale/rotation invariant features like SIFT but HOG can handle noise very efficiently as compared to SIFT. HOG Unlike SIFT describes the whole image instead of focusing on local features one by one. Zapletal and Herout[14] used a Histogram of color gradients and also HOG features and applied regression for classification to perform vehicle re-identification. While these traditional hand-crafted features have their own characteristics, they also have many disadvantages as some of them are very poor for generalization as they are effective only for specified tasks in which they can be used like the histogram of oriented gradients is only beneficial in the classification of the image but it can not be used for image semantic segmentation. SIFT only focuses upon local features of the image and HOG is used mainly for edges information and LBP focuses on the texture of the image also they mainly focus upon low-level features; hence not feasible for generalization.

2.3. Deep Learning-Based Re-Identification Methods:

Deep learning methods, unlike conventional computer vision, introduce many hidden layers in the network which tries to learn low-level features and also high-level features which improve the generalization of the model and extend its uses in the field of computer vision. And it has also achieved good results in the field of object re-identification. Other applications of deep learning-based methods in computer vision are Video tracking, Object identification, and Semantic segmentation. Which gave rise to the interest of researchers in the field of Vehicle re-identification. Vehicle reidentification is being pursued by many researchers using different measures such as supervised, unsupervised, and semi-supervised, and different approaches are being implemented for each field to produce better and more competitive results. Some of the approaches used are explained below:

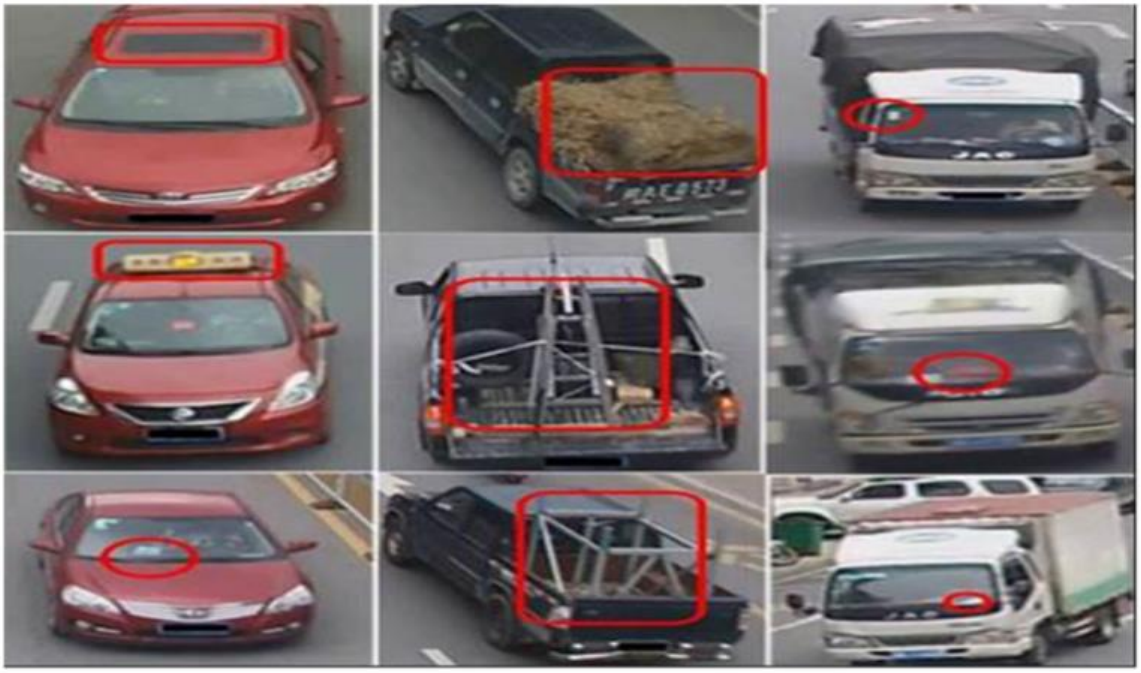


Figure 1: Same global appearance and different local region: From top-left to top-right each box contain images of different vehicles and each column displays the similar looking vehicles visually but they have some local appearance difference as can be seen in image.

2.4. Vehicle Re-Identification Using Local Features:

The reason behind results not being enough to implement techniques in real-world scenarios was; that traditional computer vision methods used global features for vehicle re-identification although after critically analyzing vehicle re-identification data sets we can safely say that major changes in objects are in local features representing images as shown in [Fig.1] the figure describes the objects that look similar in different views but have different local region features. And the red shape is highlighting local areas that are different in the same vehicle id's. we can clearly see that feature representation has a very critical role in vehicle re-identification. And there are a lot of features in hand which is why deep learning-based feature representation is popular in-vehicle re-identification produced by GoogLeNet[15], AlexNet[15], and ResNet[add reference]. And various types of loss functions are used to improve these features and the triplet loss function is used in re-id tasks for discriminative feature extraction to train the deep learning base model. Commonly used methods based upon local features are applied to vehicle re-identification. Wang et al[15] used key points and segmentation of local regions and marking of regions based upon key points and marked exactly twenty key points and formulated results by mixing regions of segmentation of target vehicle for vehicle re-identification. They used a convolutional neural network for local region feature extraction and then after fusing them with a global feature vector to obtain the appearance-based contextual feature vector of the target vehicle. And they compare directly with the feature vector of the different vehicles which solves the problem of comparison of different regions between different vehicles. Because it was hard to distinguish between vehicle images from different cameras so many scholars then studied the local region feature vector methods.

2.5. Similarity Matric For Vehicle Re-Identification:

We know that learning a good feature vector for similarity matching is very important for vehicle reidentification but selecting a good matrix for similarity matching is also very important for vehicle re-identification tasks. “Several distance matrics and learning approaches are studied in image retrieval and recognition tasks”[16]. The described metrics are defined in such a way that the same class features are kept close to each other and features from different classes are kept far away from each other so that they can be classified as different classes. The descriptor is what defines the image. And it should be able to maximize the distance between different class features and minimize the distance between features belonging to the same class. There are different types of distance metrics used for the feature distance calculation such as Euclidean distance and cosine distance for similarity calculation in different face recognition[17] algorithms. “Furthermore, Deep Relative Distance Learning (DRDL)[18] studied a two-branch convolutional neural network to convert the raw vehicle images into Euclidean space, so that distance can be used directly to measure the similarity of two individual vehicles.”

In a supervised manner, matrix learning is done by using pairwise constraints, when in the training phase the image appearance descriptors are labeled as positive and negative in pairs. And they decide whether they belong to the same class or different classes. Here they are described as x_1,x_2,…..,x_n, and n is the total amount of images in a dataset or training images. And dimensionality is m and metric is to learn D Ɛ R_{mxm} represent; so the distance between pair of descriptors can be described as: x_i and x_j :

$$d(xi, xj) = (x_i - x_j)^T D(x_i - x_j) \tag{2}$$

"d(x_i , x_j) is a true metric only possible when matrix D is symmetric positive semi-definite. This issue is resolved by adopting convex programming as follows:*"* [2]

$$min_D \sum_{(x_i - x_j \in Pos)} ||(x_i - x_j)||_D^2 \; std \geqq 0 \;, \; and \sum_{(x_i - x_j \in Neg)} ||(x_i - x_j)||_D^2 \geqq 1 \tag{3}$$

In this equation, Pos defines positive label data and Neg represents negative labeled data that belongs to other vehicle classes in training samples, both describe appearance descriptors of images.

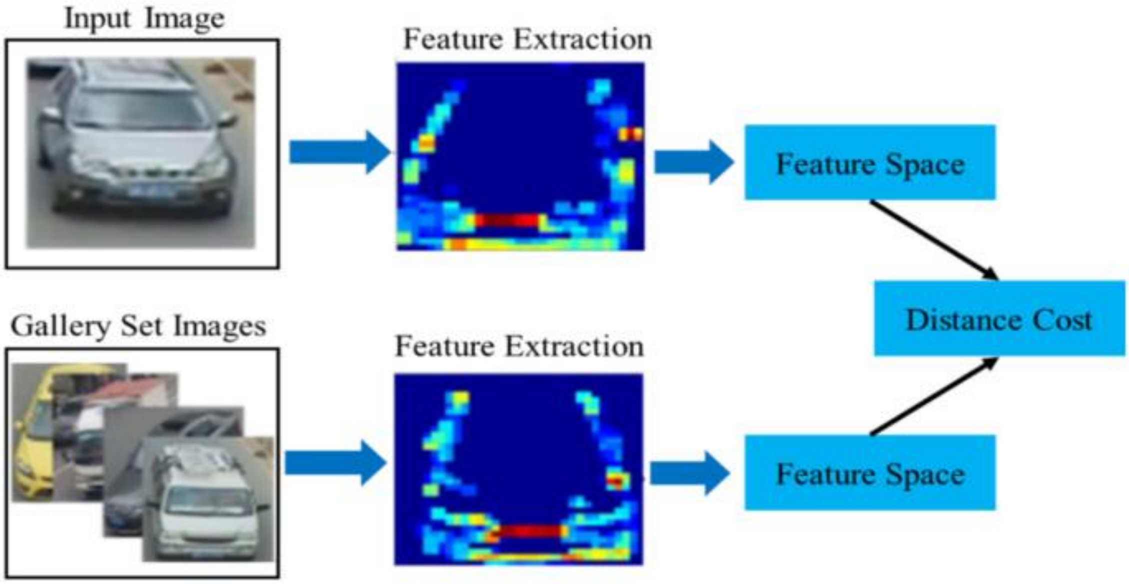


Figure 2: Classification of similar looking images using distance metric: Distance is calculated of input vehicle feature vector by comparing it with Gallery set in gallery set feature space to put similar looking vehicles together in Gallery set feature space.

2.6. Vehicle Re-Identification Using Representation Learning:

In the real world, scenarios vehicle re-identification can not only be done on the basis of local features because the camera angles can produce significant differences in vehicle local feature areas. So to increase accuracy in vehicle re-identification we need to move forward from only local feature differences. Representation learning[19] is gaining popularity with the rapid evolution of CNNs because we can form different useful representations using different non-linear transformations on input data. Representation learning is a very important task in re-identification as by applying CNN to a huge amount of data, it is already formulating different representations according to the job of classification or prediction. Due to its natural stable training and robustness, it is already applied to different jobs of person re-identification[18]. So that is why these are being applied to vehicle re-identification also.

According to Zheng *et al.*[20] learning different representations from vehicle images is very important which are also discriminative. He proposed a unified deep convolutional neural network that was guided by these discriminative attributes of vehicles. Attributes are color, type, and camera view of the vehicle for the job of vehicle re-identification. Because of the unification of model components, the learned representation was discriminative. Another representation-based vehicle re-identification model is VS-GAN which uses the baseline of generative adversarial network and also uses the method of multiple granularities, it is a generative model for vehicle re-identification that uses both global and part feature fusion. And partitioned vehicle images use two directions such as horizontal and vertical. And also included multiple other discriminative information. It includes two branches for horizontal and two for vertical local feature representation and one branch is specified for global feature representation[21]. There are other methods that are based upon representation learning with unique ideas like Hou *et al* proposed a method that uses the idea of including images into training data set which has augmented occlusion it also increased the training dataset and prevented the model to overfit on the dataset and also added the intended attribute to mimic the real world occlusion. Then they used the joint identification and verification methods to train the model with occlusion and training images.

An Deep learning based framework was introduced which could lead to an accurate representation of vehicles was published [22], in which variational feature learning was used to learn feature relationships and learn those relationships an LSTM network was used which proved that variational feature learning can improve the performance of vehicle re-identification. There are many other methods that focus on data labeling, domain mismatch, and variance of vehicle appearance in different viewpoints that uses visual representation. Later a multiattribute driven vehicle re-identification method used multibranch training and a re-ranking strategy to learn the representations of variate vehicle images, which also took advantage of contextual and appearance-based attributes of vehicle *i.e* color, type, and model to make it more generalized and variance prone. And re-ranking introduced the spatiotemporal information of vehicles across different cameras to construct a similar appearance of the vehicle and used Jaccard distance for similar vehicle clustering.

Recently vehicle re-identification is solved by two famous re-identification methods, one is to take it as a classification problem and provide fine-grained vehicle labeled images as train sets to already existing classification pipelines in a supervised manner and calculate loss by forward and backward propagation and then classifying vehicles. But to the increasing traffic day by day, this method would cost more to label and with the increasing number of training samples, it would be difficult to effectively classify vehicles. Which would lead to accuracy Bottlenecks. And the second one is to use the process of verification which works by inputting two vehicle images containing the vehicle ID information and determining in the process whether they belong to the same vehicle class or not. It is done by using verification loss for learning. The loss is reduced by taking into account the measures to increase the discrimination between two separate classes. Despite it being more generalizable it can not only be done by simple vehicle id information rather one would need to include vehicle attribute information such as color, and other attributes to the better representation learning ability of the neural network.

2.7. Vehicle Re-Identification Based On Unsupervised Learning:

Most of the work that produced good results in vehicle re-identification is carried out in a supervised manner in which you directly take labeled data and input it into the network and it learns representation according to that, but it has its own downsides such as generalization is difficult to achieve with supervised labels with increasing data day by day and also the cost of labeling data is very high. To solve these issues unsupervised learning came in place to feed the network directly inferencing from the unlabeled data. Some techniques of unsupervised learning have already been employed in the field of person re-identification[23] - [24] and Deng *et al* [25] which did image-to-image cross domain adaption by changing the source domain images into target domain images in an unsupervised manner. Then they trained re-identification models on this target domain with supervised labels, They used GANS for the similarity preservation in both source and target domain and contrastive loss for the re-identification task.

Some scholars have also applied unsupervised techniques to vehicle re-identification. In which a two-step progressive cascaded framework was presented for unsupervised vehicle re-identification. “Which used the combination of CNN for feature extraction and an unsupervised technique to enable self-paced progressive learning”[26]. And they also incorporated the contextual information into the network which increased the convergence, In [27] Marin-Reyes *et al* applied the method of unsupervised to the vehicle re-identification task and generated labels in an unsupervised manner and along with used visual tracking to generate a weakly labeled training set. Bashir *et al* [28] proposed an unsupervised approach that trained an unlabeled vehicle re-identification dataset using the self-paced progressive network, This technique transferred the learned representation from a deep neural network to an unlabeled dataset. Also, a Generative adversarial network [29] is an emerging technique for unsupervised learning which includes a generator and a discriminator the task of the generator is to generate an image that looks like the source domain and the discriminator tries to discriminate between the label and generated image and the idea is that both generator and discriminator reach a balance, to perfect both sides.

2.8. GAN-Based Vehicle Re-Identification:

Generative adversarial network[29] is one of the hot techniques that are being used for the semi-supervised and unsupervised manner in the field of computer vision, it was proposed by Sir, Ian Goodfellow and it has been applied and tested in many fields such as image synthesis, image super-resolution, classification of person and vehicle re-id.

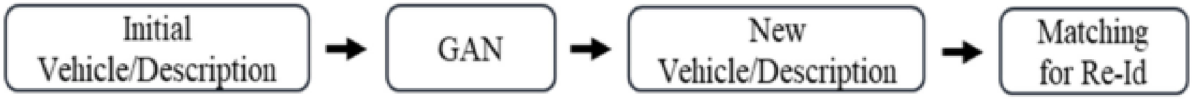


Figure 3: Vehicle re-identification system based on GAN: This image shows the visual layer of GAN usage for Vehicle re-identification , Visual description of vehicle is learned by Generator after frequent matching with discriminator and then new learned description is used for matching of vehicle for vehicle re-id [29].

There are many papers that use GAN to solve the problems of vehicle reidentification. And existing vehicle re-identification datasets have a small scale and low level of diversities which can lead to low-level generalization. To solve these kinds of problems latest trend for the re-identification task is GAN. GAN has proved its worth recently in other main fields of computer vision such as image synthesis [25], and translation[30], now they are utilized for vehicle re-identification[31]. In their paper Zheng *et al* [32] proposed a method that was implemented using DCGAN[30], combining with gaussian noises to generate an unlabeled person image for training. PT-GAN was studied by Wei *et al* [33] to minimize the domain gap for style transfer of person images. Zhou *et al* [34] solved the problem of multiview points of vehicle images by using GAN by generating different views of vehicle images. Lou *et al* [35] generated multi-view vehicle images from the source domain to facilitate model training. And Zhou *et al* proposed a GAN to generate cross/ Multiview vehicle images from desired vehicle domain.

Aihua *et al* formulated an idea that used view transformation for the vehicle Reid model. And these different views were generated for the vehicle Reid model by the generated adversarial network. “The vehicle re-id model consists of one backbone, three subnetworks, and one embedding network.

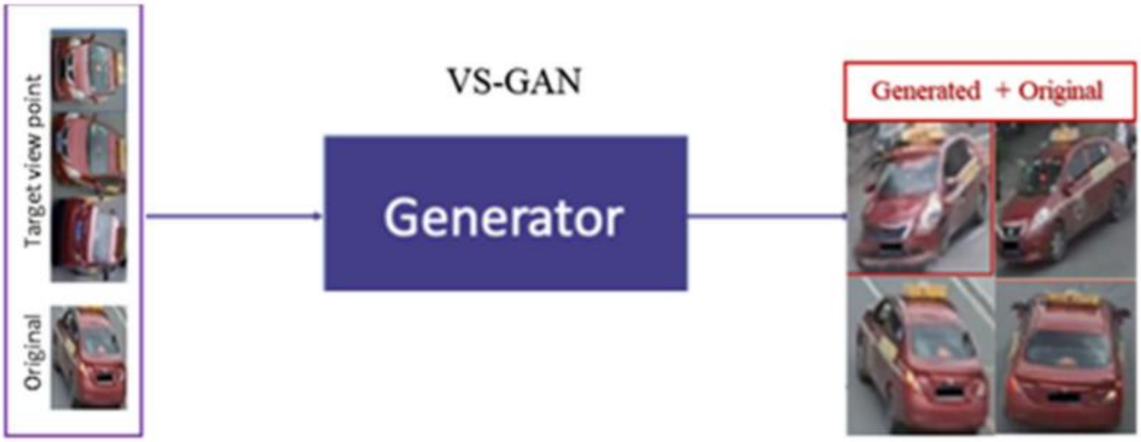


Figure 4: GAN-based vehicle image generation: This technique describes another usage of GAN for Vehicle re-id in which GAN is used to generate different looking copies of same vehicle. With orientation changes such as inverted angles. And then those generated vehicle images along with original images are used in the process of vehicle re-identification.

2.9. Comparison Of Vehicle Re-Identification Datasets:

Datasets are the benchmarks that are developed by the research community to validate the performance and accuracy of vehicle re-identification methods, they should depict the real-world scenarios and different real-world challenges present in these datasets, such as occlusion, background, clutter, and illumination problem in different views of different cameras. Multiple benchmark datasets are available such as VeRi776, VehicleID, and VeRi-Wild.

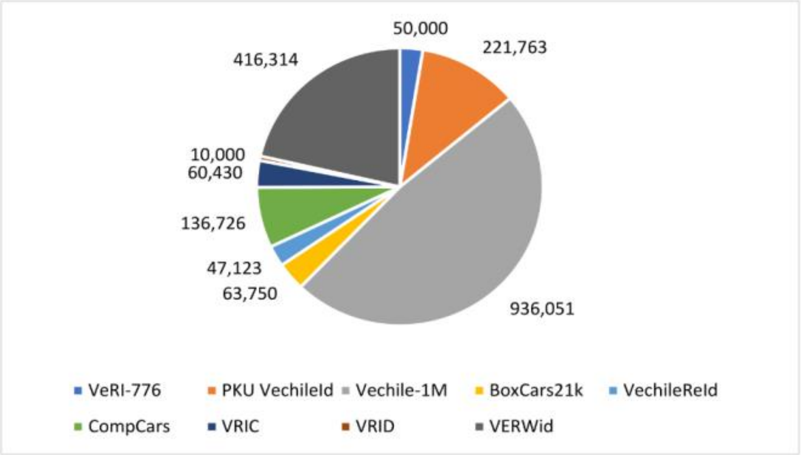


Figure5: Berief comparison and description of differently available datasets: Pie-chart compares the number of images available for training and testing in mainstream datasets for vehicle re-identification.

Table 1: 1 Comparison of different benchmark datasets:

S.no	***Datasets***	***year***	***Total Images***	***No. of Models***	***No. of Vehicles***	***No. of viewpoints***	***No. of Cameras***
1	VeRi-776	2016	50,000	10	776	6	18
2	PKU Vehicle ID	2016	221,763	250	26,267	2	12
3	Vehicle-1M	2018	936,051	400	55,527	……	……
4	BoxCars21K	2016	63,750	148	21,250	4	……
5	VehicleReId	2016	47,123	……	1232	……	……
6	CompCars	2015	136,726	1716	……	5	……
7	VRIC	2018	60,430	……	5622	……	60
8	VRID	2017	10,000	10	1000	……	326
9	VERIWild	2019	416,314	……	40,671	Unconstrained	174

Chapter:3

Design and Methodology

3.1. Problem Definition:

To address the problems and challenges of vehicle re-identification we have used a novel procedure[10] for vehicle re-identification which uses the method of producing pseudo labels by taking into consideration different camera views and their vehicle distribution discrepancy according to different camera views and generates sample pseudo label to train the CNN baseline in two stages, In the first stage; pseudo labels are generated of source domain directly using simple CNN features and their feature similarity to train a multibranch network for each camera. The second stage; also includes the classification scores of each class in the first stage for the generation of pseudo labels with this new feature vector which solves the problem of this intra-camera distribution difference and produces better pseudo labels for the second stage.

In addition to that, we have also trained a contextual information classifier to input pseudo labels of source domain generated by this classifier result in the first stage by modifying the base model into a multilabel, multiclassifier to help the base-model learn better representation in the first stage, and produce better feature vector to be used in the second stage. which will also help in the faster convergence.

3.2. Architecture & Design:

When each sample of the dataset is captured from different cameras, it is then captured with different variations in images with varied parameters and environmental factors which make each sample appear different in different cameras, This intra-inter camera similarity method[10] focuses on solving the problem of intra camera discrepancy by training a multibranch classifier, which consists of a classifier for each camera, because it would be easy to recognize each sample within a camera than to classify inter camera without having the knowledge of intra camera appearance. In this process, the backbone learns the discriminative features of each sample so backbone learned features are used to generate pseudo labels that do not have the problem of intra-camera discrepancy.

3.2.1. Formulation:

Given an unlabeled vehicle image with its view information in cameras $\boldsymbol{X = \{X^c\}}$, where $\mathbf{X^c}$ represents the whole collection of vehicle images with respective camera **c = 1: C**, and the goal is for every query image **q**, the base model should be able to produce feature vector to narrow down the same vehicle's image from gallery set **G** . And the baseline should guarantee that the classified image $\mathbf{I_g}$ has more similar features with **q** as compared to other images available in gallery set.

$$g * = \ arg\ max_{g \in G}\ sim(f_g\ , f_q) \tag{4}$$

Where $\boldsymbol{f}$ belongs to R^d which is the d-dimensional vector which is computed for every query image by baseline and similarity comparison is done with f feature vector available in gallery set and image with the maximum similarity in features is returned. In other words it should maximize the distance of feature map in different classes or images from different classes and also minimize the distance between vehicle images from same class. Suppose we have a single vehicle v that is captured by many cameras then the collection can be represented as ${X_v}^c$ where c is number of cameras and v belongs to number of vehicles. So the label generation idea can be formulated as.

$$\tau * \ = \ arg\ min_{\tau} D(\tau\ , \{Xv\}v = 1{:}\, V\ ) \tag{5}$$

Where T is the clustering result of the source domain and D (.) is to compute the difference between pseudo label T and $\{X_v\}$*v=1:V.*

According to equation 5 by performing the steps for each camera and predicting the labels we can see this specifically solves the issues of intra-camera discrepancy and effects of camera setting $\boldsymbol{S_c}$ such as environment and viewpoint as well as other stochastic factors that were altering the image of the vehicle to variate between different cameras. And then train the model with predicted pseudo labels, training of feature according to clustering result suppose the cth camera produces the clustering result as $\boldsymbol{T^c}$, Loss can be defined as:

$$\int_{intra}^{c} \quad = a_0 + \sum_{I_n \in X^c, I_n \in \tau_m^c} loss(f_n + m) \tag{6}$$

Where m denotes the cluster id that is assigned as a label which will be used in loss as the label to compute loss and after computing equation 6 on all cameras we will have separately we will have multibranch CNN in which each classifier is being trained for each camera. By using the learned feature f for clustering we see it still suffers with discrepancy. The idea is to enhance the similarity between single-class vehicles captured in different cameras and increase the difference between different class vehicles in the same camera. And it can be formulated as:

$$SIM_{inter}(I_n\ , \ I_m\ ) \ = \ sim(f_n\ , f_m\ ) + \Delta(s_n\ , \ s_m) \tag{7}$$

In which S_n is used for denoting the classifier result for the image $I_{n,}$ and S_m denotes that same for image I_m and $\Delta(S_m\ , S_n)$ is the probability between two images, we can enhance feature camera discrimination by applying loss in the second stage.

$$\int inter \ = \sum_{I_n \in \tau_m} loss_c(f_n + m) \tag{8}$$

3.2.2. Architecture:

The architecture of the whole is described below in detail:

3.2.3. Intra Camera Training For Vehicle Re-Identification:

In the first stage subsets of whole datasets ***X*** are created according to each camera which can be written as ***{Xc}*** with c being the index for the camera. Then each subset is clustered using the CNN directly and with the similarity of ***f***, each subset image is assigned a label according to cluster id. Which then is fed to the network as a labeled dataset. And the loss can be written as

$$loss^{c}(f_n, m) = l\,(F(w^{c}, f_n), m) \tag{9}$$

This formula describes learnable parameters as ***F(w^c, .)*** in which the learnable parameters are ***w^c*** and loss is computed with cross-entropy softmax loss and label are described as ***m***. As described above the intra-camera training stage trains the baseline in such a manner that with it each classifier for each camera is a new classifier thus explaining the multi-branch network architecture. And overall training loss can be described as below.

$$\int intra = \sum_{c=1}^{C} \int_{intra}^{c} \tag{10}$$

Where we can see that C is for each camera which increases the discrimination ability for feature ***f*** for each camera. And it also alleviates the effect of variate appearance in different cameras and improves the generalization of the model.

3.2.4. Inter Camera Training For Vehicle Re-Identification:

In the inter-camera training phase the most important part is to identify that two images of the same identity from different cameras belong to the same class for that we need a domain-independent feature vector and the images from the class should produce the same probability of distribution classification even so in a different camera. They used Jaccard similarity for the probabilistic classification similarity to check the probability that two images I_m and I_n are from the same class. For that we need to computer similarity between I_m and I_n as $\boldsymbol{\Delta(s_m , s_n)}$ which can be described as:

$$\Delta(s_n , s_m) = \frac{s_n \cap s_m}{s_n \cup s_m} \tag{11}$$

To compute the similarity we need to take in the input from the previous stage where we trained multi-classifiers of cameras and we need to take voting for each image or source image that from which class it belongs in each cameras classifier for that we take element-wise min and max of vectors, and s_m was gathered by concatenating the result of classification from each camera classifier for 1 image. Now after calculating the similarity between cameras we can easily generate pseudo labels that will have the generalization ability to discriminate between different classes. And the loss of inter camera can be calculated as follows:

$$\ell_{inter} = \frac{1}{|B|} \sum_{I_n \in B} l\left(\mathrm{F}\left(w^c , f_n\right), m\right) + \lambda L_{triplet} \tag{12}$$

In which training mini-batch can be seen as B and, ℓ describes the loss of pseudo labels generated by clustering result which is mentioned as m, $L_{triplet}$ is the triplet loss used for re-identification process validation and K number of clusters are selected randomly.

Also inter stage loss can also be described in simple words as a loss to minimize the loss between single class samples and increase loss between different classes, and to reiterate the idea behind this intra-inter camera similarity for vehicle reidentification that is to take these classification of cluster results in the end and to perfectly classify which image of vehicle belong to what cluster of images. In simple words this loss is computed for the better cluster results that are learned by this re-id model.

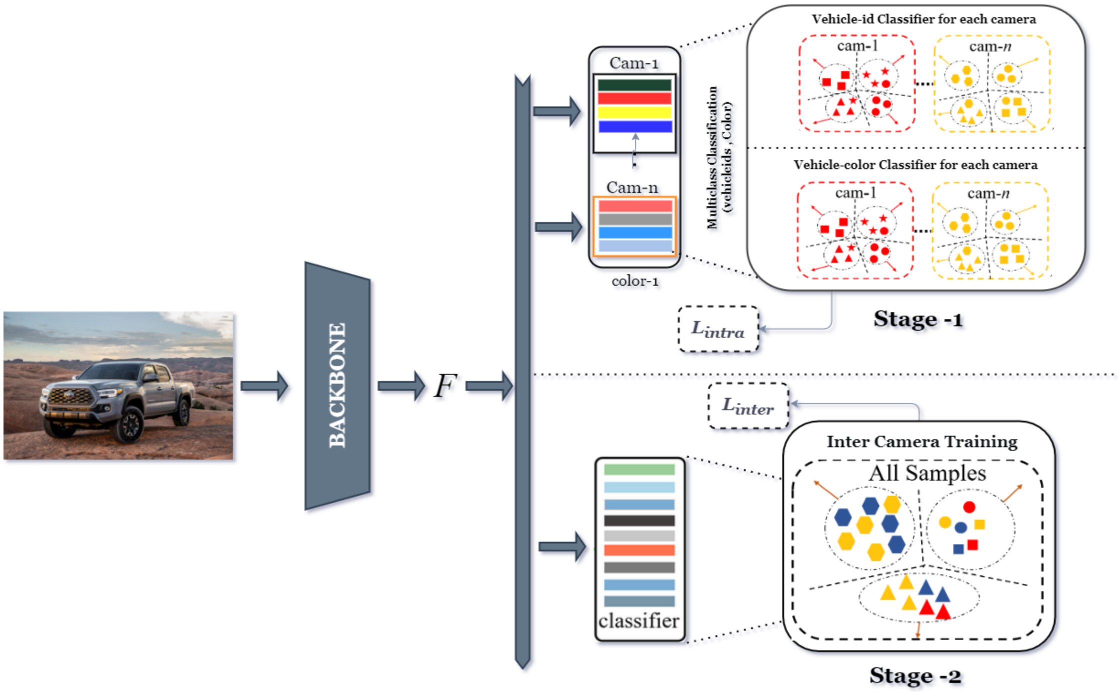


Figure 6: Intra-inter Camera training with multiclass multilabel classification: Stage-1 includes the intra-camera multibranch classifier training and its loss is computed with pseudo labels for color and vehicle id's , Stage-2 is backbone finetuning for all samples and its loss is computed with pseudo labels generated from stage one feature vector results.

3.2.5. Pseudo Label Generation For Training Of Color Classifier :

First, we used a simple contextual information classifier that would recognize the information present in the dataset and we trained the model in a supervised manner because pseudo label generation in the next step would require better accuracy, the better the model learns this information the better in the next step it would be able to generate pseudo labels for source domain to feed in the intra camera trainer. And the loss of this model can be formulated as below:

$$\int color = \sum_{I_{n \in X}} loss(n, m) \tag{13}$$

In this equation the L is loss for every image when it is forwarded and backproped in the network for learning and ***f*** is the predicted label for color and loss is computed between the original label and predicted label for each image.

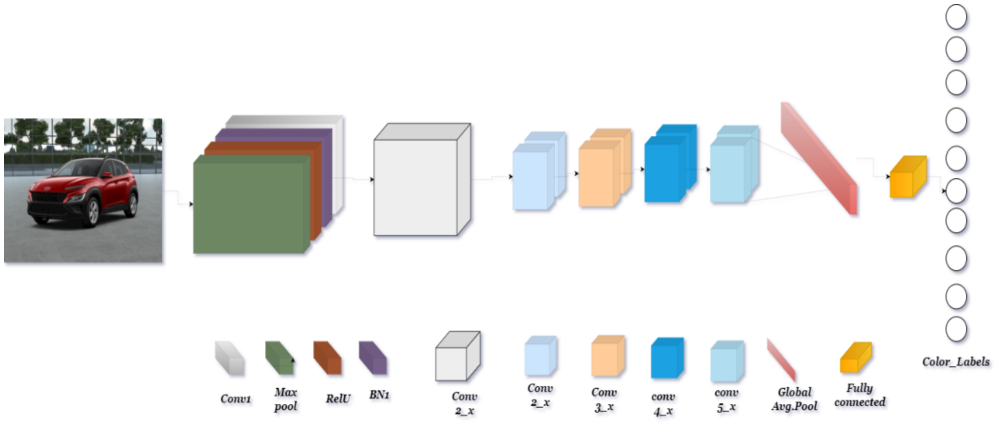


Figure 7: Color model based on ResNet: Said Architecture contains the full layer display of ResNet 18 which was used for color classification of VeRi 776 dataset whose learned feature vector would be used to produce color pseudo labels for re-id model.

After training this color classifier for the vehicle images model we clustered images according to the vehicle re-identification dataset every sample from the same class of source domain should have the same cluster label but the results of this model suffer the same domain discrepancy and camera view variation which was the challenge for this intra-inter camera similarity for re-identification that different same class images have different views in different cameras, and after generating clusters ids through its feature vector we observed the same problem exists with color variation and we solved the problem for color variation in different cameras as the same vehicle in different camera seems like it has different color due to illumination problem. So we tried to solve this problem by using the same structure to train an intra-camera model for color clusters.

In this model, we followed the same structure for the intra-camera trainer but this model was trained in a supervised manner and contained labels for each image. In which we first categorized images and put together an image that was from the same camera and trained each classifier for each camera so it somehow solved the problem of illumination and color variation of images in different cameras. And the loss of this model can be formulated as :

$$\int_{intra}^{c} = \sum_{I_n \in X_m} loss_c(+m) \tag{14}$$

In which the I_n is the image of n sample or class and the whole dataset can be represented as X^c and c is the index/number of cameras and the loss of each camera can be calculated as calculating cross-entropy loss between label a predicted label by the camera and each classifier is finetuned for each camera images. which seems to have removed this intra-camera discrepancy for color variation.

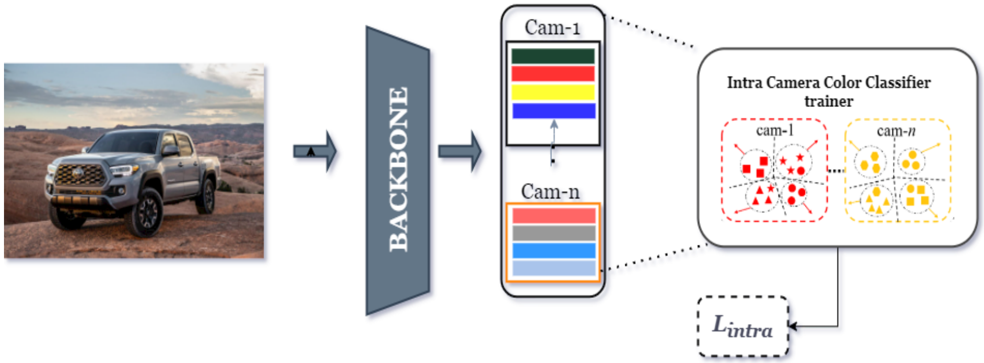


Figure 8: Intra camera trainer for color classifier: This model is used the multibranch structure of CNN to learn color variation of vehicle for each camera to produce better pseudo labels to be used in vehicle re-id model.

After generating clusters for color pseudo labels from feature vector that was finetuned on color labels, and by observing the cluster results we can easily say that it removed the intra-camera color variation problem up to some extent and generated clusters better than the previous clusters. Then we used these features to perform clustering to produce pseudo labels to be used as labels in the first stage of intra-inter camera similarity for vehicle re-identification.

Chapter:4

Implementation

4.1. Details:

Combining the contextual information model with the whole vehicle reidentification model we have our whole model and we have explained in detail what the model does and how, so now we explain the whole implementation and settings on which we have trained our model and in what sequence. Diagrams describing the color feature extractor model and then the vehicle re-identification model exploit those learned feature domains as extra information for vehicle re-identification and generate pseudo labels for color information and take them as labels to train its backbone for better feature extraction.

4.2. Re-id Model:

“We use ResNet-50 pre-trained on ImageNet as the backbone to extract the feature. The layers after pooling 5 layers are removed and a BN-Neck is added behind it. During testing and clustering, we extract the pooling-5 feature to calculate the similarity. All models are trained with PyTorch. During training, the input image is resized to **256 × 256**. Image augmentation strategies such as random flipping and random erasing are performed. At each round, we perform the intra-camera stage and inter-camera stage in order. The number of training rounds is set as **50**. At the intra-camera training stage, the batch size is **4** for each camera. The **SGD** is used to optimize the model. The learning rate for **ResNet-50** base layers is **0.0005**, and the one for other layers is **0.005**

At the inter-camera training stage, a mini-batch of **16** is sampled with $\boldsymbol{P} = \mathbf{16}$ randomly selected clusters and $\boldsymbol{K} = \mathbf{4}$ randomly sampled images per cluster. The SGD is also used to optimize the model. The learning rate for ResNet-50 base layers is **0.001**, and the one for other layers is **0.01**. The margin in triplet loss is fixed to **0.3**. The training progressively uniforms the distribution of features from different cameras. we train the model for 2 epochs at both stages. We use the standard Agglomerative Hierarchical method for clustering. The number of clusters is 600 for each camera at intra-camera stage and 800 at the inter-camera stage.”[10]

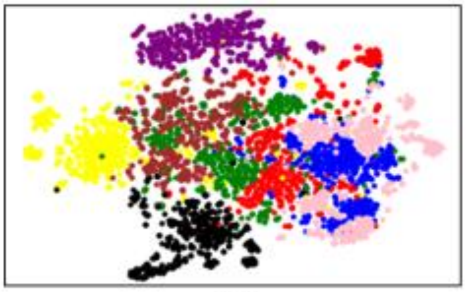
4.3. Color-Model:

We use the standard Agglomerative Hierarchical method for clustering. The number of clusters is 10 for each camera at the intra-camera stage and 10 at the inter-camera stage.

Feature Representation: The architecture of the color feature extractor is described in the figure **(Figure 8: Intra camera trainer for color classifier)** that module takes the whole 3d image as an input and also the labels provided in the dataset and tries to learn single sample representation in different cameras without color variation tries to discriminate between different samples altogether at the same time tries to learn single class boundary to label all images present in that class as same. and particularly focus on feature representation of single sample in different cameras which can be categorized as local features.

Which(learned representations of color information) are then used to cluster the whole training set in different color labels. And then those color labels are assigned as pseudo labels combined, with other pseudo labels that were generated from the feature network of vehicle re-id model backbone, and both are fed to the model to learn better representation for vehicle re-identification. Which is described in figure **(Figure 6: Intra-inter Camera training with multiclass multilabel classification:)**

And then that model scans through the whole training set and learns different representation and tries to learn how one vehicle or sample look through different cameras and tries to learn the feature representation of one sample boundary and discriminate between different samples. And the whole idea for intra-inter camera training of features was to make the model invariant to different camera views of a single sample and learn the existing boundary for a single class and also to distinguish the boundaries of different classes altogether. Which can be seen in the image below.



(a)

Figure 9: (a) is the simple distribution learned from a different model and it can be seen that it suffers from feature discrepancy as different color describes the sample from a different camera.

And the idea is to distinguish that boundary that the same sample from different cameras can have the same distribution and can easily be categorized as the same sample and that is what we have achieved from this model.

4.4. Network Architecture:

In this section, we are explaining the fusion of two models **1. Color classification model** and **2. Vehicle re-identification model**.

Architecture: The color model has an architecture of a single label single classifier multibranch network which has RESNET 50 as backbone which outputs feature dimension of 2048 and then it has a classification layer, A fully connected layer at the output from the backbone module predicts the confidence score for all classes for each camera in a multibranch module. And loss is computed by comparing the corresponding label with model prediction to fine-tune weights. And the overall model loss is Cross Entropy loss which can be seen in the equation**(Equation 14)** And the model is fine-tuned on IMAGENET pre-trained weights.

And the second network that performs the whole vehicle re-identification has the same model structure but it is a multilabel multiclassifition model which at the same time can have multiple labels to compute the loss in a multibranch network architecture for one classifier of each camera. Each camera will have its color classifier and vehicle classifier in its first stage. which can be seen in figure **(Figure 6: Intra-inter Camera training with multiclass multilabel classification:)** which also has a backbone of RESNET50 and after that a fully connected layer at the output from backbone module predicts the confidence score for all classes for each camera in a multibranch module. And this stage also has a loss of Cross Entropy loss which is computed for each class and used to finetune weights for the next iteration.

After the first stage of intra-camera training we use another stage in which clustering is done on the whole training dataset and for that clustering to be able to generate better cluster results classifier trained in the last stage takes part in the second stage similarity matching result which is called Jaccard similarity, and after that results of all classifiers are concatenated and in a vector and then from their distance result cluster results are produced and pseudo labels for the second stage are generated,

In this sequence whole procedure is repeated until it produces desired results. Another module to alter the bottleneck module in the backbone is used which is to increase the generalization of the model which is explained below:

4.5. AIBN Module:

Another module in the back backbone is used to increase the generalization of the model, instance normalization is good for detecting changes to appearance but it reduces the variance in different classes on the other hand Batch normalization reduces the internal covariate shift but increases the internal class variance so by looking at both of them we can see that by combining these two both of them are complementary to each other.

In order to gain the advantage of both of them, they proposed the AIBN, Adaptive instance and batch normalization,

“In order to gain the advantages of both IN and BN, we propose the AIBN. It is computed by linearly fusing the statistics (mean and var) obtained by IN and BN, *i.e.*,”[10]

$$\hat{\mathbf{x}}[i,j,n] = \gamma \frac{\mathbf{x}[i,j,n] - (\alpha\mu_{bn} + (1-\alpha)\mu_{in})}{\sqrt{\alpha\sigma_{bn}^2 + (1-\alpha)\sigma_{in}^2 + \epsilon}} + \beta \tag{15}$$

By using this module in the backbone as a bottleneck we can increase the generalization of the model and reduce its overfitting and make it robust to variance.

Chapter:5

Experiments and Results

5.1. Dataset:

we have used **VeRi776** for vehicle re-identification testing and training and the main reason behind the use of this dataset was that it is a challenging dataset for benchmarking, as it has in total of **20 cameras,** so 20 different views for a single vehicle and all of them will have different illumination and occlusion problem. Data consists of **776 vehicles** of data at a timeframe of 24 hours which contains different types, and colors of vehicles. Each vehicle is captured by each camera at one time for about 6 times so we have at least 6 images of 1 vehicle in 1 camera. All of the data is captured in non-overlapping views in cameras. Data lists the labels, e.g., vehicle ID, camera ID, color, and type, of the training images. And for the comparison with the state of the Art table we have used Standard Train, test data split as mentioned in the dataset.

But for our training and testing purposes, we have used other settings in which we have used different splits in data to adjust data according to the resource.

For color_model training, we have split the training and test data into more chunks to fit our usage and resource management.

5.2. Evaluation Matrix:

During the training of the vehicle re-identification model from the dataset, we do not use any other annotation other than camera labels. And camera labels are used just to categorize data according to camera for intra-camera training.

And for performance evaluation, we have used MAP (Mean Avg. Precision).

Other Evaluation metrics that are used for vehicle re-identification are CMC (Cumulative matching characteristic).

For color classification model training we have used color labels/annotations from the dataset and also camera id's as color-classifier is trained in a supervised manner so that is why end-to-end labels are used for that model.

5.3. Results:

We trained and evaluated our model on a single Nvidia Tesla P40 GPU with a batch size of 4 in the first intra-camera stage and 16 in the second inter-camera stage training. The model training is done for 40 epochs with each epoch having separate intra-camera stage 2 epochs and 2 inter-camera stage epochs. We used a cluster size of 300 for the intra-camera stage and 800 for inter camera stage. we have implemented the project in Linux with Python and PyTorch.

The performance of our network on the VeRi776 dataset is analyzed with other state-of-the-art methods. We have quantitatively presented this comparison in Table (Table 2: Comparison with the state of the Art:)

Table 2: Comparison with the state of the Art:

Methods	*Year*	*Dataset*	*Rank-1*	*Rank-5*	*mAP*
SPGAN	2018	VeRi (776)	57.4	70.0	16.4
VR-PROUD	2019	VeRi (776)	55.7	70.0	22.4
ECN	2019	VeRi (776)	60.8	70.9	27.7
UDPA	2020	VeRi (776)	76.9	85.8	35.8
VACP-DA	2020	VeRi (776)	77.4	84.6	40.3
PAL	2020	VeRi (776)	68.2	79.9	42.0
OUR's	2022	VeRi (776)	86.8	91.4	38.6

5.3.1. Results with Changed Parameters:

We trained our model on a single Nvidia GRID V100 GPU with a batch size of **4** in the first intra-camera stage and **16** in the second inter-camera stage training. The model training is done for **50** epochs with each epoch having separate intra-camera stage **2** epochs and **2** inter-camera stage epochs. We used a cluster size of **300** for the intra-camera stage and **800** for inter camera stage. we have implemented the project in Linux with Python and PyTorch.

We trained models with multiple different settings to fit the needs of our resources.

- The Reid model with half dataset VeRi776:
- The Reid model with Supervised color labels with half dataset Veri 776:
- The Reid model with Unsupervised color pseudo Labels with half dataset Veri 776:

These are three models that we trained and checked results in different settings. The first model (***- The Reid model with half*** *dataset VeRi776:*) was trained with half dataset because the whole dataset contained training images of **37k** which could not be loaded with the memory of 16 GB and reduced to 14k and evaluated results. The second model (***- The Reid model with Supervised color labels with half*** *dataset Veri 776:*) with 14k training images was trained and evaluated. And on basis of the results of the second model, we produced pseudo labels for the third model (***- The Reid model with Unsupervised color pseudo Labels with half*** *dataset Veri 776:*) and provided those pseudo labels in the form of labels with 14k images and evaluated results.

Three models that are described above are trained in the same settings and the results are observed and explained in the table below.

Table 3: Evaluated results:

Model	***Performance Measure***		
- The Reid model with half **dataset VeRi776:**	Rank-1	Rank-5	mAP
	71	86.2	27.3

Model	***Performance Measure***		
Re-id - The Reid model with Supervised color labels with **half dataset Veri 776:**	Rank-1	Rank-5	mAP
	81.1	88.7	30.3

Model	***Performance Measure***		
Re-id - The Reid model with Unsupervised color pseudo Labels with **half dataset Veri 776:**	Rank-1	Rank-5	mAP
	62.5	77.2	23.4

Chapter:6

Conclusion

6.1. Conclusion:

“This paper proposes an intra-inter camera similarity method for unsupervised vehicle Reid which iteratively optimizes Intra-Inter Camera similarity through generating intra- and inter-camera pseudo-labels. The intra-camera training stage is proposed to train a multi-branch CNN using generated intra-camera pseudo-labels. Based on the classification score produced by each classifier trained at the intra-camera training stage, a more robust inter-camera similarity can be calculated. Then the network can be trained with the pseudo-label generated by performing clustering across cameras with this inter-camera similarity. Moreover, AIBN is introduced to boost the generalization ability of the network. Extensive experimental results demonstrate the effectiveness of the proposed method in unsupervised vehicle Reid.”[10]

6.2. Future Work:

After analyzing the results of the trained Re-Identification model and comparing it with different state-of-the-art models that are using the same technique and producing good results. We have concluded that combining pseudo labels of vehicles with the different contextual feature information is a good technique to use because not only it can make the model converge faster, but it also contributes to the feature enhancement which in result produces better re-identification results overall.

So to increase or produce better results same as I trained the color model to produce pseudo labels for the Re-Identification model I will train a separate model for car-type and also input car-types as pseudo labels in Re-Id Model for better results and faster convergence.

Chapter:7

References

[1] R. Meulen, “8.4 Billion Connected Things Will be in Use 2017,” *Gartner*, 2017. https://www.gartner.com/en/newsroom/press-releases/2017-02-07-gartner-says-8-billion-connected-things-will-be-in-use-in-2017-up-31-percent-from-2016

[2] H. Wang, J. Hou, and N. Chen, “A Survey of Vehicle Re-Identification Based on Deep Learning,” *IEEE Access*, vol. 7, pp. 172443–172469, 2019, doi: 10.1109/ACCESS.2019.2956172.

[3] Z. Wang, Y. Wang, Z. Wu, J. Lu, and J. Zhou, “Instance Similarity Learning for Unsupervised Feature Representation,” *Proc. IEEE Int. Conf. Comput. Vis.*, pp. 10316–10325, 2021, doi: 10.1109/ICCV48922.2021.01017.

[4] Z. Zheng, L. Zheng, and Y. Yang, “Unlabeled Samples Generated by GAN Improve the Person Re-identification Baseline in Vitro,” *Proc. IEEE Int. Conf. Comput. Vis.*, vol. 2017-Octob, pp. 3774–3782, 2017, doi: 10.1109/ICCV.2017.405.

[5] H. Fan, L. Zheng, C. Yan, and Y. Yang, “Unsupervised person re-identification: Clustering and fine-tuning,” *ACM Trans. Multimed. Comput. Commun. Appl.*, vol. 14, no. 4, 2018, doi: 10.1145/3243316.

[6] H. Guo, K. Zhu, M. Tang, and J. Wang, “Two-Level Attention Network With Multi-Grain Ranking Loss for Vehicle Re-Identification,” *IEEE Trans. Image Process.*, vol. 28, no. 9, pp. 4328–4338, 2019, doi: 10.1109/TIP.2019.2910408.

[7] J. Peng, Y. Wang, H. Wang, Z. Zhang, X. Fu, and M. Wang, “Unsupervised vehicle re-identification with progressive adaptation,” *IJCAI Int. Jt. Conf. Artif. Intell.*, vol. 2021-Janua, pp. 913–919, 2020, doi: 10.24963/ijcai.2020/127.

[8] Zheng, Z., Ruan, T., Wei, Y., Yang, Y. and Mei, T., 2021. VehicleNet: Learning Robust Visual Representation for Vehicle Re-Identification. *IEEE Transactions on Multimedia*, 23, pp.2683-2693.

[9] Hoang, V., 2019. EarVN1.0: A new large-scale ear images dataset in the wild. *Data in Brief*, 27, p.104630.

[10] S. Xuan and S. Zhang, “Intra-inter camera similarity for unsupervised person re-identification,” *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 11921–11930, 2021, doi: 10.1109/CVPR46437.2021.01175.

[11] M. Stommel and O. Herzog, “Binarising SIFT-descriptors to reduce the curse of dimensionality in histogram-based object recognition,” in *Communications in Computer and Information Science*, 2009, vol. 61, no. 1, pp. 320–327. doi: 10.1007/978-3-642-10546-3_38.

[12] T. Watanabe, S. Ito, and K. Yokoi, “Co-occurrence histograms of oriented gradients for human detection,” in *IPSJ Transactions on Computer Vision and Applications*, 2010, vol. 2, pp. 39–47. doi: 10.2197/ipsjtcva.2.39.

[13] T. Ojala, M. Pietikäinen, and T. Mäenpää, “Gray scale and rotation invariant texture classification with local binary patterns,” *Lect. Notes Comput. Sci. (including Subser. Lect.*

Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 1842, pp. 404–420, 2000, doi: 10.1007/3-540-45054-8_27.

[14] X. Zhang *et al.*, “Re-ranking vehicle re-identification with orientation-guide query expansion,” *Int. J. Distrib. Sens. Networks*, vol. 18, no. 3, Mar. 2022, doi: 10.1177/15501477211066305.

[15] G. Zeng, Y. He, Z. Yu, X. Yang, R. Yang, and L. Zhang, “Preparation of novel high copper ions removal membranes by embedding organosilane-functionalized multi-walled carbon nanotube,” *J. Chem. Technol. Biotechnol.*, vol. 91, no. 8, pp. 2322–2330, 2016, doi: 10.1002/jctb.4820.

[16] L. Yang and R. Jin, “Distance metric learning: A comprehensive survey,” *Michigan State Universiy*, pp. 1–51, 2006, doi: 10.1073/pnas.0809777106.

[17] Y. Sun, Y. Chen, X. Wang, and X. Tang, “Deep learning face representation by joint identification-verification,” *Adv. Neural Inf. Process. Syst.*, vol. 3, no. January, pp. 1988–1996, 2014.

[18] Y. Li, L. Zhuo, X. Hu, and J. Zhang, “A combined feature representation of deep feature and hand-crafted features for person re-identification,” *PIC 2016 - Proc. 2016 IEEE Int. Conf. Prog. Informatics Comput.*, no. 2, pp. 224–227, 2017, doi: 10.1109/PIC.2016.7949499.

[19] Y. Bengio, A. Courville, and P. Vincent, “Representation learning: A review and new perspectives,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1798–1828, 2013, doi: 10.1109/TPAMI.2013.50.

[20] H. Li *et al.*, “Attributes Guided Feature Learning for Vehicle Re-Identification,” *IEEE Trans. Emerg. Top. Comput. Intell.*, 2021, doi: 10.1109/TETCI.2021.3127906.

[21] P. Huang *et al.*, “Deep feature fusion with multiple granularity for vehicle re-identification,” in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 2019, vol. 2019-June, pp. 80–88. Accessed: Jul. 03, 2022. [Online]. Available: https://openaccess.thecvf.com/content_CVPRW_2019/papers/AI City/Huang_Deep_Feature_Fusion_with_Multiple_Granularity_for_Vehicle_Re-identification_CVPRW_2019_paper.pdf

[22] S. Ahmed *et al.*, “VARIATIONAL REPRESENTATION LEARNING FOR VEHICLE RE-IDENTIFICATION School of Electronic and Information Engineering , South China University of Technology , GRG Intelligent Security Institute , Guanghzou 510006 , P . R . China,” pp. 3118–3122, 2019.

[23] R. Zhao, W. Ouyang, and X. Wang, “Unsupervised salience learning for person re-identification,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3586–3593. doi: 10.1109/CVPR.2013.460.

[24] J. Wang, X. Zhu, S. Gong, and W. Li, “Transferable Joint Attribute-Identity Deep Learning for Unsupervised Person Re-identification,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2275–2284. doi:

10.1109/CVPR.2018.00242.

[25] W. Deng, L. Zheng, Q. Ye, G. Kang, Y. Yang, and J. Jiao, “Image-Image Domain Adaptation with Preserved Self-Similarity and Domain-Dissimilarity for Person Re-identification,” *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 994–1003, 2018, doi: 10.1109/CVPR.2018.00110.

[26] R. M. S. Bashir, M. Shahzad, and M. M. Fraz, “VR-PROUD: Vehicle Re-identification using PROgressive Unsupervised Deep architecture,” *Pattern Recognit.*, vol. 90, pp. 52–65, Jun. 2019, doi: 10.1016/j.patcog.2019.01.008.

[27] P. A. Marin-Reyes, L. Bergamini, J. Lorenzo-Navarro, A. Palazzi, S. Calderara, and R. Cucchiara, “Unsupervised vehicle re-identification using triplet networks,” *IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Work.*, vol. 2018-June, pp. 166–171, 2018, doi: 10.1109/CVPRW.2018.00030.

[28] R. M. S. Bashir, M. Shahzad, and M. M. Fraz, “DUPL-VR: Deep unsupervised progressive learning for vehicle re-identification,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2018, vol. 11241 LNCS, pp. 286–295. doi: 10.1007/978-3-030-03801-4_26.

[29] I. J. Goodfellow *et al.*, “Generative adversarial nets,” in *Advances in Neural Information Processing Systems*, 2014, vol. 3, no. January, pp. 2672–2680. doi: 10.3156/jsoft.29.5_177_2.

[30] A. Radford, L. Metz, and S. Chintala, “Unsupervised representation learning with deep convolutional generative adversarial networks,” 2016.

[31] Y. Lou, Y. Bai, J. Liu, S. Wang, and L. Y. Duan, “Embedding Adversarial Learning for Vehicle Re-Identification,” *IEEE Trans. Image Process.*, vol. 28, no. 8, pp. 3794–3807, 2019, doi: 10.1109/TIP.2019.2902112.

[32] W. Sun, F. Liu, and W. Xu, “Unlabeled samples generated by GAN improve the person re-identification baseline,” in *ACM International Conference Proceeding Series*, 2019, vol. Part F1482, pp. 117–123. doi: 10.1145/3323933.3324091.

[33] L. Wei, S. Zhang, W. Gao, and Q. Tian, “Person Transfer GAN to Bridge Domain Gap for Person Re-identification,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2018, pp. 79–88. doi: 10.1109/CVPR.2018.00016.

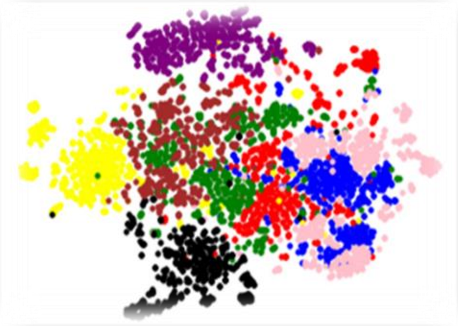
[34] C. Zhang, C. Yang, D. Wu, H. Dong, and B. Deng, “Cross-view vehicle re-identification based on graph matching,” *Appl. Intell.*, 2022, doi: 10.1007/s10489-022-03349-y.

[35] X. Liu, W. Liu, J. Zheng, C. Yan, and T. Mei, “Beyond the Parts: Learning Multi-view Cross-part Correlation for Vehicle Re-identification,” in *MM 2020 - Proceedings of the 28th ACM International Conference on Multimedia*, Oct. 2020, vol. 20, pp. 907–915. doi: 10.1145/3394171.3413578.

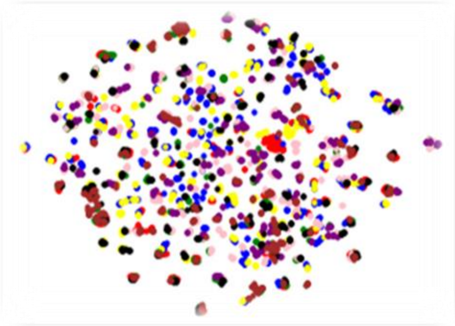
Appendix

Appendix A:

- **Conclusion with comparison:**



(a)



(b)

This visualization is of ***features*** learned from different models and as can be seen from images that both images have different feature representations and different colors in both images represent that images are from different cameras which also means that same color in these images represent that vehicles are from same camera. And by observation we can say that:

- Feature representation from image (a) does have feature distribution discrepancy among cameras as visible from visualization.
- But features from image (b) does not suffer from feature distribution discrepancy.

Image (a) contains the features extracted from ***t-SNE*** model and features in image (b) are from ***IICS*** method. So by observing results we can conclude that ***Intra-inter camera similarity*** alleviates effectively feature distribution discrepancy among cameras and produce better results proposed by Xuan and Zhang [10].